Full text provided by www.sciencedirect.com

Progress in structure prediction of α-helical membrane proteins

Sarel J Fleishman and Nir Ben-Tal

Transmembrane (TM) proteins comprise 20–30% of the genome but, because of experimental difficulties, they represent less than 1% of the Protein Data Bank. The dearth of membrane protein structures makes computational prediction a potentially important means of obtaining novel structures. Recent advances in computational methods have been combined with experimental data to constrain the modeling of three-dimensional structures. Furthermore, threading and *ab initio* modeling approaches that were effective for soluble proteins have been applied to TM domains. Surprisingly, experimental structures, proteomic analyses and bioinformatics have revealed unexpected architectures that counter long-held views on TM protein structure and stability. Future computational and experimental studies aimed at understanding the thermodynamic and evolutionary bases of these architectural details will greatly enhance predictive capabilities.

Addresses
Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel-Aviv University Ramat Aviv 69978, Israel

Corresponding author: Ben-Tal, Nir (nirb@tauex.tau.ac.il)

Current Opinion in Structural Biology 2006, **16**:496–504

This review comes from a themed issue on
Membranes
Edited by Roderick MacKinnon and Gunnar von Heijne

Available online 5th July 2006

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2006.06.003

Introduction

Transmembrane (TM) proteins comprise ~20-30% of the genome [1,2] and are involved in many crucial cellular processes, such as cell-to-cell signaling, metabolite transport and energy production. Solving the structures of these proteins is therefore imperative for clear mechanistic understanding of central processes in physiology. However, despite recent advances in production of TM protein crystals, membrane protein structures are difficult to obtain and comprise less than 1% of the entries in the Protein Data Bank (PDB) [3].

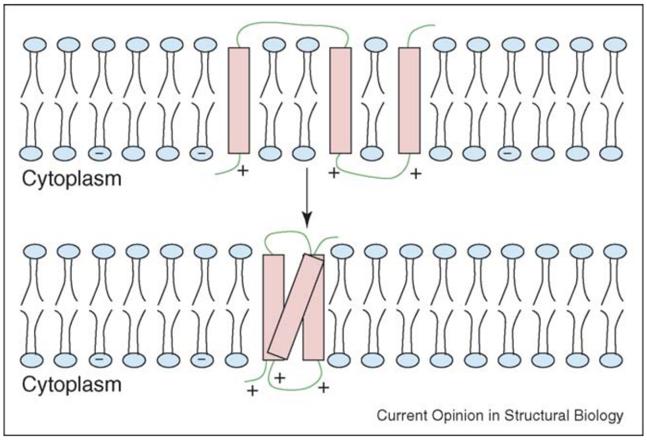
Comparative- or homology-based approaches to structure prediction have been immensely successful with soluble proteins [4]. These methods require a homologous protein, for which a structure has been solved. Because of this requirement, homology modeling has been most useful for the few TM protein families, for which at least one member has been crystallized. A recent analysis of homology-modeling accuracy for membrane proteins has shown that the protocols that are successful in comparative modeling of soluble proteins reach similar achievements for membrane proteins [5•]. However, because at present only few representative atomic-resolution structures of TM protein families are available, homology modeling cannot serve as a general purpose approach for structural modeling. In this review, we will therefore focus on recent advances in structure prediction that do not rely on homology to solve structures (subject covered in [6,7•]).

Membrane protein folding can be conceptually decomposed into two consecutive steps: folding of the individual hydrophobic segments into helices followed by helix association (Figure 1) [8]. Accordingly, the problem of predicting the structure of α-helical TM proteins has been approached by breaking it down into the following steps: (i) delineating the boundaries of the TM segments, each of which will assume a helical conformation; (ii) determining the topology of the protein (i.e. which extra-membrane segments reside inside the cytoplasm and, conversely, which segments reside outside the cell); and (iii) predicting the tertiary conformation of the protein (i.e. the way in which the helices are packed with respect to one another). The past few years have seen considerable advances in all of these steps. In this review, we will describe some of these advances and emphasize the discovery of novel features of TM protein folds that bear on the goal of structure prediction.

Identification of TM α-helices in the protein sequence

Early attempts for predicting the locations in the sequence of membrane-integral segments were based on the notion that a sequence segment would partition into the membrane if it were sufficiently long and hydrophobic. Starting with the method of Kyte and Doolittle [9], various algorithms for detecting membrane-embedded sequence segments were proposed on the basis of experimental and computational data. At the core of these methods lies a hydrophobicity scale that assigns to each amino acid residue a score that can be roughly interpreted as the free energy of its transfer from hydrophilic to hydrophobic media, corresponding to its insertion probability into the membrane. The typical approach would then be to search the sequence for a sufficiently hydrophobic stretch of residues comprising

Figure 1



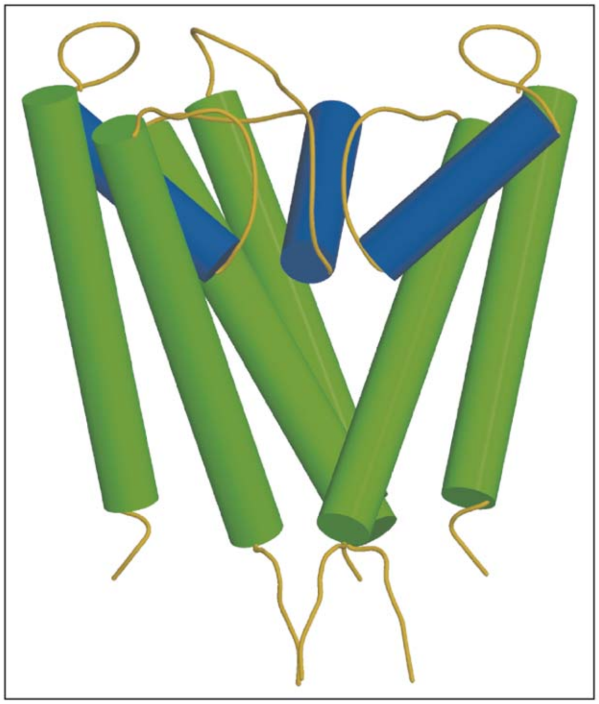
TM protein folding can be thought to proceed in two stages [8]: the folding of individual TM segments into helices (top) followed by helix packing (bottom). The topology of the protein is often determined by the positive-inside rule [17], with the cytoplasmic loops tending to be enriched by positively charged residues in comparison with the extracellular loops.

approximately 20 amino acids, which is the minimal length necessary for an α-helix to traverse the 30 Å hydrophobic core of the membrane [10].

During the 90s, there was a departure from physicochemically based approaches to methods that rely on statistical inference, such as hidden Markov models, support vector machines and neural nets, all of which make use of the existing knowledge on the partitioning of particular sequence segments to the membrane. These methods appeared at first to be superior to the simple hydrophobicity-based methods, with success rates of 90% and above [1]. However, a fundamental difficulty in the validation of statistical methods is to obtain sufficiently disparate datasets for training and validation. Indeed, when Rost and co-workers recently revisited the problem of TM sequence prediction [11] using datasets that were carefully constructed with the aim of decreasing redundancy, they found that the success of the statistical approaches was overrated, and they in fact achieved results that were not much better than those that were obtained by some of the hydrophobicity-based methods. In this respect it is important to emphasize that an overlap of only three amino acids between the predicted and observed helices is considered sufficient for being an accurate prediction [11]. Thus, in a recent survey it was demonstrated that, on average, the best-performing prediction methods were in error by a little more than two turns at the helix termini [12]. Because most structural modeling approaches rely on the correct identification of the helical segments in the sequence (see below), these large errors are likely to propagate in subsequent modeling stages, requiring manual intervention. A more alarming conclusion made in this survey concerned the inability of current prediction methods to identify 'irregular' structures, such as half helices and re-entrant loops, as those seen in the structure of the potassium channel (Figure 2) [13] and the aquaporin family [14]. Hopefully, with the likely increase in the number of proteins exhibiting such irregularities over the next few years, some unifying principles will emerge from their sequences, enabling prediction of these features.

Recently, the hydrophobicity-based approach to detecting membrane-embedded segments was given another boost from the experimental studies by von Heijne and co-workers [15••]. The authors reported a series of experiments that attempted to obtain a hydrophobicity scale using an experimental setup that is far closer to the physiological system than previous experimental reports, including the translocon protein-conducting channel and membranes from the endoplasmic reticulum (ER). Concerns were raised regarding the possibility that some of the measured partitioning energies encompass contributions from interactions between the probe sequence segments and other protein components in the system, thus limiting the generality of the scale produced by these measurements [16•]. Nevertheless, this experimental

Figure 2



The potassium channel [13] is one of the several structures of membrane proteins that show structural 'irregularities', such as half helices (blue) and re-entrant loops. These irregularities cannot be identified from the sequence by current methods [12]. For clarity, only three out of four of the subunits comprising the potassium channel are shown. Figure generated with MolScript [70] and rendered with Raster3d [71]. Figure reproduced with permission from [37].

approach is promising, raising hope that the prediction of the location of TM helices in the sequence of membrane proteins will eventually be based on algorithms that account for the various factors that affect protein translocation in biological systems.

Topology

Determining the topology of a membrane protein is a crucial preliminary step to modeling its structure as it constrains the way individual TM segments could associate within the membrane, as well as subunits within complexes. The positive-inside rule (i.e. the observation that the segments in the cytoplasmic loops and the TM segments that are adjacent to the cytoplasm are often enriched in the positively charged lysine (K) and arginine (R) residues when compared with the extracellular loops (Figure 1) [17]) has remained the most powerful tool for predicting the topology of a protein from its sequence for almost two decades. The factors contributing to the (K + R) bias are under intense study, and it is still unclear whether the bias originates from properties of the translocon [18] or the cytoplasmic membrane [19], but a recent statistical survey of 107 genomes reconfirmed the validity of this empirical rule [20]. The (K + R) bias can serve as a rule for predicting topology, by requiring that more positively charged residues face the cytoplasm [1].

Recently, von Heijne and co-workers have conducted a whole-proteome experimental analysis of the topology of TM proteins in the *Escherichia coli* inner membrane [21••]. They used two reporter proteins that were linked to the C-terminus of each putative membrane-integral protein in *E. coli*. One of these reporters is only active in the cytoplasm, whereas the other is exclusively activated in the periplasm. By measuring the activities of the reporters, the authors assigned the topology of 601 out of 700 predicted TM proteins in the *E. coli* genome. Comparing these data to the predictions of a widely used algorithm that is based on a hidden Markov model called TMHMM [2], the authors found that roughly 80% of the predictions were in accord with the experimentally determined topologies. This correlation shows that the major aspects affecting protein topology are captured by contemporary computational methods, but that these still have significant room for improvement. These experimental results can serve as a much-needed large-scale benchmark for validation and comparison of future topology prediction algorithms.

The vast majority of proteins in von Heijne and co-workers' analysis exhibited unique topology [21••], whereby their C-terminus was found to be either cytoplasmic or periplasmic. However, for five out of 601 proteins both reporters were activated, implying that for each of these five proteins, some of the protein copies inserted with one topology, and the others with the reverse topology [21••,22]. The five proteins with dual topology are relatively small in size, comprising ~100 amino acid residues and are predicted to contain four TM domains. Furthermore, as expected, all five exhibit very small (K + R) biases. For at least one of these proteins, the prototypical small multidrug resistance antiporter EmrE, the suggestion of dual topology was already made in the past on the basis of structural data and the lack of clear (K + R) bias [23]. Nevertheless, it is important to note that a previous study based on a different biochemical assay reported a unique topology for this protein [24]. This conflict between two lines of experimental evidence still needs to be resolved, but the suggestion that some TM proteins insert with opposite topology has significant implications for understanding structures and functions of these proteins.

Threading and *ab initio* structure prediction

On the one hand, integral membrane proteins exhibit much higher uniformity of secondary structure (mostly α-helical bundles) than soluble proteins, and are highly constrained in their conformations because of the presence of the membrane [25]. It could therefore be expected that *ab initio* structure prediction, whereby the protein structure is predicted without resorting to homology with other proteins or to experimental data, should be a more feasible goal for TM than for soluble proteins. On the other hand, as sampling significant portions of conformation space remains a very challenging aspect of *ab initio* structure prediction [26], success in soluble protein structure prediction has been restricted to small proteins, consisting of approximately 80 amino acid residues [27]. Membrane proteins are usually much larger; for instance, visual rhodopsin, which serves as a prototype for the large family of 7-TM GPCRs, consists of more than 300 amino acid residues.

Two similar methods, MembStruk [28–31] and PREDICT [32,33], were specifically tailored to predict the structures of GPCRs on the basis of physicochemical principles. For both methods, a full-atom model of the GPCR is automatically obtained, based on the amino acid sequence of the protein alone. In the first step, the boundaries of the seven TM helices are predicted by means of hydrophobicity scales. A preliminary (tentative) coarse-grained model of the packing of these helices into a compact and closed structure is constructed, and various conformations in the vicinity of this state are sampled at random, favoring conformations in which hydrophobic residues face the lipid. Full-atom models of the TM domains of these structures are built and subjected to several cycles of optimization using molecular dynamics (MD) simulations. The outcome is a full-atom model of the entire protein, including the extra-membrane loops. The methods produced 3D models of bovine rhodopsin, the only GPCR structure available in the PDB, with ~3 Å root-mean-square deviation (RMSD) from the native structure in the TM region. Further validation of this approach includes *in silico* docking of known drug-like compounds to the receptors. Model structures of several GPCRs, including the β2 adrenergic [30] and D2 dopamine [28] receptors, were built this way and used successfully for drug design [32]. This suggests that important structural aspects of the ligand-binding site were accurately captured by these methods. However, it was not shown unambiguously that the remainder of the structure is correct too.

Another potentially promising approach utilizes the two-step TASSER method that threads the sequence on parts of solved protein structures, and then refines the resulting template [34•]. Validation on a set of 38 nonhomologous TM protein structures yielded 17 structures for which the RMSD to native was less than 6.5 Å, but many others with RMSD to native greater than 10 Å. When applied to predicting the structure of bovine rhodopsin, TASSER produced a model with a low 2.1 Å RMSD from native on the C^{α} coordinates of the TM domain. Subsequently, the method was applied to model the structures of most of the ~900 human GPCRs, and a few of these models were examined and appeared to be consistent with the available experimental data. It is important to note that although the method's success in modeling rhodopsin is promising, only a few other GPCRs showed substantial similarity (>30% sequence identity) to bovine rhodopsin [7,34•], and it is therefore uncertain that the other models are as faithful to the native state as the model of rhodopsin. Also, it is not known yet whether TASSER's GPCR models are likely to be closer to the receptors' inactive or active form, the latter of which is pharmaceutically more interesting [7]. Nevertheless, the models generated by TASSER might provide an important resource for probing structure–function relationships in this important class of receptors, as many of the current approaches to modeling GPCR structures rely on homology to bovine rhodopsin [6], despite the low sequence identity.

Recently, the Rosetta algorithm for structure prediction, which has been successful in the free-modeling category of the community-wide experiment on critical assessment of structure prediction (CASP) [35], was adopted and implemented for TM protein structures [36•]. Inter-residue contact potentials were derived from a set of solved protein structures, and enriched with their sequence homologues. Validation on a set of solved TM protein structures showed that the performance of this implementation of Rosetta (below 4 Å for 51–145 of the superimposed residues) is comparable to that of Rosetta for soluble proteins in the same size range. Although full-atom prediction was shown to produce significant improvements in prediction accuracy of soluble proteins [27], it was not tested in this implementation of Rosetta, partly because of the prohibitive computational load associated with full-atom prediction for large proteins.

Structure prediction based on experimental constraints

One potential venue for obtaining novel structures, which has been explored by several groups in recent years, is the exploitation of functional and low-resolution structural data on TM proteins to constrain models [37•]. Such data could involve site-specific mutagenesis, chemical cross-linking, intermediate-resolution structures and biophysical data, such as NMR, EPR and FTIR. These heterogeneous data are interpreted as constraints on the positions of individual amino acid residues or on the structural relationships among them. For instance, positions that are intolerant to substitution are likely to be packed inside the protein core, and positions that cross-link are likely to be vicinal. In addition to these experimental data, the modeling methods assume that the hydrophobic sequence segments form α-helices that traverse the membrane.

The pioneering work of Herzyk and Hubbard [38] employing such disparate data sources produced very promising results, with a model of bacteriorhodopsin matching the native-state structure by a low 1.87 Å RMSD. However, further modeling attempts that relied primarily on mutation and crosslinking data demonstrated that it is difficult to interpret many of these data in a structurally unequivocal way [37•]. Recent implementations of this approach have therefore relied on more limited data sources. For instance, a method was suggested recently that employs data that can be interpreted as distance constraints between amino acid residues from EPR, FTIR and chemical crosslinking [39]. Models consisting of α-helices were sampled using a Monte Carlo strategy. The conformations were scored according to the extent to which they satisfied the experimental distance constraints and structural parameters derived from a set of solved TM proteins, including preferred helix-packing angles and distances, pairwise amino acid contact preferences and overall structural compactness. Encouragingly, this method was shown to produce a model of rhodopsin, which was 3.2 Å RMSD from the native-state structure, based on only 27 experimentally derived distance constraints (taken from published studies), demonstrating that it might be possible to obtain close-to-native models of large membrane proteins on the basis of a limited set of experimental constraints.

Several groups have recently suggested methods that employ data from cryo-electron microscopy (cryo-EM) intermediate-resolution structures, together with data on hydrophobicity, evolutionary patterns and the lengths of the loops that connect neighboring TM segments [37•]. For several proteins, cryo-EM structures are available at in-plane resolutions of 5–10 Å (e.g. the gap junction [40] and EmrE [23]). At this resolution, it is impossible to either position individual amino acid residues, or even unambiguously identify the assignment of TM segments to the helices observed in the cryo-EM structure. Hence, structure prediction based on cryo-EM is typically comprised of helix assignment, followed by orientation of the helices around their principal axes.

To solve the helix assignment problem, various studies used biochemical data on the functional roles of individual TM segments [41,42]. A complementary approach relies on the fact that some of the loops that connect TM helices are quite short (less than eight amino acid residues). Such short loops constrain the distance between the helix termini that they connect. Based on this constraint, an algorithm was recently suggested, which, for a given cryo-EM structure and the lengths of each of the interconnecting loops, scans all possible assignments (potentially $n!$ permutations, where n is the number of helices in the map), and ranks them by their compatibility with the cryo-EM structure [43]. The performance of the algorithm was found to be sensitive to the exact delineation of the helix start and end points, which are difficult to predict with accuracy. Another proposed method that suffers less from such sensitivity ranks each TM sequence segment according to its overall hydrophobicity and evolutionary conservation [44]. Highly conserved and hydrophilic segments were ranked as helices that are likely to be buried within the protein core, and more variable and hydrophilic segments were assigned to lipid-exposed positions.

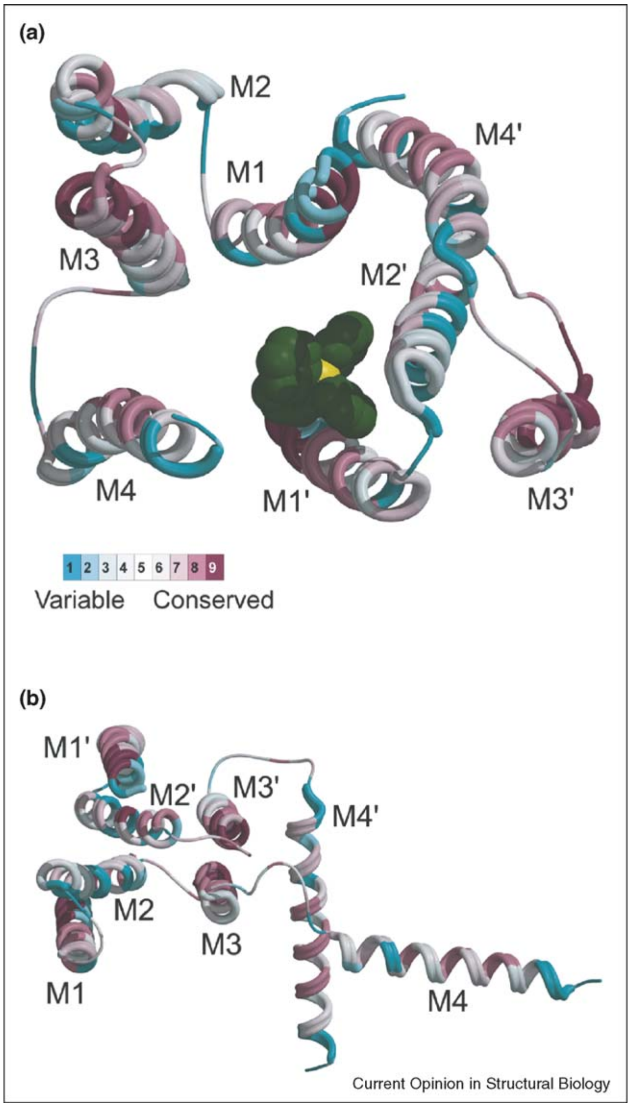
Once the helix assignment problem is solved for a given protein, canonical α-helices are constructed to fit the data in the cryo-EM map, and are rotated around their principal axes to identify the native state conformation. Following the work of Baldwin *et al.* [45] on the prediction of the structure of the TM domain of rhodopsin based on its cryo-EM structure and sequence analysis, recently two similar methods [46,47] were independently suggested. It was shown that the cores of many TM protein structures are much more evolutionarily conserved than their peripheries, and tend to pack the most polar residues [48]. These observations can be framed as predictive rules, according to which orientations that pack conserved and hydrophilic positions in the helix bundle are more favored than others. One of the methods generates only C^{α} models [47], whereas the other adds sidechains and uses manual refinements and minimization to generate full-atom models [46]. It should be noted, however, that often the energy landscape for full-atom models is extremely rugged and even 1 Å differences in the atom positions from the native-state structure can result in large energy penalties [26]; thus, it still remains to be seen whether the addition of sidechains improves the resulting models. The two methods were applied to intermediate-resolution structures of TM proteins, for which atomic-resolution data were not available: the oxalate transporter OxlT [46] and the gap junction [49••]. Because the evolutionary-conservation pattern on two of the helices of the gap-junction forming protein, connexin, was not informative enough to constrain their orientations, another sequence analysis method [50] was employed that identified correlated amino acid positions, thus predicting which pairs of amino acid residues could interact. Part of the attractiveness of an approach to structure prediction, which uses information from sequences and cryo-EM structures, lies in the fact that it does not necessarily rely on large amounts of previously published functional data. Hence, it is possible to subsequently use these data for validation. In the modeling of the gap junction TM domain, for instance, it was shown that, although the model was not constrained by clinical data, it placed almost 30 disease-causing but physicochemically mild mutations in the core of the helix bundle, where they would disrupt folding, whereas two physicochemically radical polymorphisms were placed in more spacious regions of the protein structure [49••]. Similarly, the model structure of OxlT placed residues that were found to crosslink in experimental assays in proximal positions [46].

Kinks in TM proteins are known to have important functional roles [51,52] but, until recently, could not be predicted from sequence information. Recently, it was shown that, in many cases where a kink is present in a TM protein structure, prolines are observed in the multiple-sequence alignment, even if the solved protein structure does not contain a proline at that position [53•]. The direction and magnitude of the kink might also be predicted from local sequence features [54]. Accordingly, it might be possible to model kinks where these have been observed in low-resolution structures, as in EmrE [23], or to bias the *ab initio* predictions to produce kinks and, thus, generate more native-like models.

Computational validation of structures

Recently, a small number of atomic resolution structures of membrane-integral proteins were suggested to represent conformations that are distorted with respect to the native-state structure [55,56]. Atomic resolution structures inspire a large amount of (usually very productive) work aimed at understanding structure–function relationships. Conversely, physiologically irrelevant structures might cause much work to be done in vain, on top of supplying a wrong view of the protein. Usually, the ultimate test for the physiological relevance of a structure is its compatibility with carefully crafted biochemical and biophysical analysis. However, such analyses are often difficult to conduct. Because some of the computational analyses described above can be used to predict the structures of membrane-integral proteins, it is reasonable to expect that they might provide grounds for doubting structures that have not been sufficiently supported by biochemical data. As an example of this approach, Figure 3 shows two structures of the bacterial multidrug resistance protein EmrE obtained by X-ray crystallography at 3.8 Å and 3.7 Å resolution [57,58]. Both structures

Figure 3



Two recently solved structures of homodimers of the multidrug resistance protein EmrE from *E. coli* are shown, which are incompatible with the observation that amino acid residues at the core of many membrane-integral proteins tend to be evolutionarily conserved, whereas those on the periphery are variable. **(a)** The structure of substrate-bound EmrE [58] exhibits highly variable residues on helix M2 forming tight contacts with M3, whereas highly conserved positions on M1, M2, M3, M3' and M4' are exposed to lipid. The substrate tetraphenylphosphonium molecule is shown in space-fill mode, with the phosphate colored in yellow, and carbon atoms in green. The structure is viewed perpendicular to the proposed membrane plane. **(b)** Similarly, the structure of EmrE without bound substrate [57] locates highly variable residues in the tight interface formed between M2 and M2', and highly conserved residues on M1, M4, M1', M3', and M4' in lipid exposed positions. The incompatibility between the conservation pattern and the burial of amino acid residues parallels the observation that both structures have many features that are in contradiction with biochemical data on EmrE [61]. Evolutionary conservation was computed using a multiple-sequence alignment of 99 small multidrug resistance proteins with the ConSurf webserver [72]. Figure generated with MolScript [70] and rendered with Raster3d [71].

are clearly at odds with the observation made on many TM protein structures that evolutionarily conserved positions tend to be packed in the core of the α-helix bundle, whereas the variable residues face the lipid environment [46,47,59,60]. The discrepancy between the conservation pattern and the packing of residues parallels an analysis, reported in this issue of *Current Opinion in Structural Biology* [61], that compares these structures with the known biochemical and biophysical data on EmrE, concluding that they most likely do not represent the physiological native state of the protein.

Future directions

In recent years, computational methods have been implemented for the prediction of TM protein structures. However, the roles of different energetic factors in contributing to TM protein folding are still poorly understood [25,62] and therefore difficult to predict. For instance, it was proposed that in low-dielectric environments polar bonds would make a large contribution to protein stability [10]. Indeed, in engineered systems, hydrogen bonds were shown to drive the interaction between TM helices [63,64], but recent measurements of the strengths of polar interactions in membrane proteins have yielded smaller magnitudes [65,66] than anticipated by computations on ideal hydrogen bonds [67,68]. Based on these and other measurements of the energetics of helix association in the membrane, it has been suggested that the primary contribution to helix interactions in the membrane comes from van der Waals packing and originates from buried surface area as in soluble proteins [69••]. This suggestion, which requires additional experimental support, is crucial because it implies that the major factors that are currently embodied in *ab initio* methods for structure prediction in soluble proteins, such as steric packing [27], might be equally useful in membrane-integral proteins. It is likely that the relative contributions of polar and van der Waals interactions to membrane protein stability will continue to be a matter of intense experimental investigation over the next few years, and that the lessons learned from these studies will be incorporated into the force fields of *ab initio* and threading algorithms for membrane proteins [34•,36•]. The use of these lessons could reduce, in part, the need for deriving pairwise contact potentials from the small number of solved TM protein structures.

One impediment on the way to the application of *ab initio* techniques to membrane proteins is the fact that these proteins are very large in comparison with soluble proteins, to which these methods were successfully applied, thus making full-atom prediction impractical [36•]. However, as modeling approaches that make use of experimental information, such as cryo-EM low-resolution structures and distance constraints, have been clearly successful in identifying near-native although coarse-grained conformations of TM proteins [38,39,45–47], a synergy might be attainable from combining these methods with full-atom predictions. This would result in reliable atomic models at a computationally feasible cost.

With the advent of new structures and the application of novel biochemical assays to membrane-integral proteins, the last few years have seen a large increase in the qualitative understanding of TM protein folds. This improved understanding has gone hand-in-hand with more sophisticated prediction and modeling attempts. Undoubtedly, the new structures and structure–function analyses that will be conducted over the next few years will teach us many lessons on the possible architectures of TM proteins and their governing thermodynamic principles, further increasing our predictive capabilities.

Update

Recently, the Rosetta membrane methodology [36•] was adapted and applied to study the voltage-induced conformational changes in the voltage-dependent potassium (Kv) channels [73]. Open and closed conformations were computed for the eukaryotic Kv1.2 channel and for the bacterial KvAP on the basis of the published methodology, the homology to X-ray structures of these channels and several experimental constraints. The computed open conformation of Kv1.2 was close to its crystal structure, thus serving as partial validation for the approach. Interestingly, the results suggest that the conformational changes in the voltage-sensor domain of the bacterial protein are larger than the changes in Kv1.2, which could explain the large inconsistencies between functional studies of the bacterial and eukaryotic channels.

Acknowledgements

The authors thank SE Harrington, JU Bowie, J Skolnick, O Kalid and CG Tate for critical reading, and B Honig, LR Forrest, L Adamian, J Liang, CG Tate and DT Jones for providing manuscripts before publication. This study was supported by a grant 222/04 from the Israel Science Foundation to N B-T. SJF was supported by a doctoral fellowship from the Clore Israel Foundation.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Rost B, Fariselli P, Casadio R: **Topology prediction for helical transmembrane proteins at 86% accuracy**. *Protein Sci* 1996, **5**:1704-1718.

2. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *J Mol Biol* 2001, **305**:567-580.

3. White SH: **The progress of membrane protein structure determination**. *Protein Sci* 2004, **13**:1948-1949.

4. Petrey D, Honig B: **Protein structure prediction: inroads to biology**. *Mol Cell* 2005, **20**:811-819.

5. • Forrest LR, Tang CL, Honig B: **On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins**. *Biophys J* 2006, in press.

An evaluation of homology modeling applied to TM proteins that segregated the available TM protein structures into families according to sequence homology, and attempted to predict the structures of proteins using their homologues as templates. It was concluded that those methods that were shown to work well for soluble proteins work equally well for TM proteins.

6. Fanelli F, De Benedetti PG: **Computational modeling approaches to structure–function analysis of G protein-coupled receptors**. *Chem Rev* 2005, **105**:3297-3351.

7. • Oliveira L, Hulsen T, Lutje Hulsik D, Paiva AC, Vriend G: **Heavier-than-air flying machines are impossible**. *FEBS Lett* 2004, **564**:269-273.

An extensive evaluation of modeling approaches applied to GPCRs, particularly to the use of rhodopsin's structure as a template.

8. Popot JL, Engelman DM: **Membrane protein folding and oligomerization: the two-stage model**. *Biochemistry* 1990, **29**:4031-4037.

9. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein**. *J Mol Biol* 1982, **157**:105-132.

10. White SH, Wimley WC: **Membrane protein folding and stability: physical principles**. *Annu Rev Biophys Biomol Struct* 1999, **28**:319-365.

11. Chen CP, Kernytsky A, Rost B: **Transmembrane helix predictions revisited**. *Protein Sci* 2002, **11**:2774-2791.

12. Cuthbertson JM, Doyle DA, Sansom MS: **Transmembrane helix prediction: a comparative evaluation and analysis**. *Protein Eng Des Sel* 2005, **18**:295-308.

13. Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R: **The structure of the potassium channel: molecular basis of K^+ conduction and selectivity**. *Science* 1998, **280**:69-77.

14. Fu D, Libson A, Miercke LJ, Weitzman C, Nollert P, Krucinski J, Stroud RM: **Structure of a glycerol-conducting channel and the basis for its selectivity**. *Science* 2000, **290**:481-486.

15. •• Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G: **Recognition of transmembrane helices by the endoplasmic reticulum translocon**. *Nature* 2005, **433**:377-381.

This article reports the use of an experimental system to probe the energetics of the transfer of peptides between translocated and membrane-inserted forms, using an experimental setup very close to physiological conditions. Thus, the authors derive a hydrophobicity scale.

16. • Shental-Bechor D, Fleishman SJ, Ben-Tal N: **Has the code of protein translocation been broken?** *Trends Biochem Sci* 2006, **31**:192-196.

A critique of the thermodynamic quantities obtained by Hessa *et al.* [15••] in their analysis of peptide insertion into the membrane. It is argued that the more polar peptides might be stabilized by other protein components in the experiment, causing the energetic penalty on the transfer for polar amino acid residues to appear lower than it actually is.

17. von Heijne G, Gavel Y: **Topogenic signals in integral membrane proteins**. *Eur J Biochem* 1988, **174**:671-678.

18. Goder V, Junne T, Spiess M: **Sec61p contributes to signal sequence orientation according to the positive-inside rule**. *Mol Biol Cell* 2004, **15**:1470-1478.

19. van Klompenburg W, Nilsson I, von Heijne G, de Kruijff B: **Anionic phospholipids are determinants of membrane protein topology**. *EMBO J* 1997, **16**:4261-4266.

20. Nilsson J, Persson B, von Heijne G: **Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes**. *Proteins* 2005, **60**:606-616.

21. •• Daley DO, Rapp M, Granseth E, Melen K, Drew D, von Heijne G: **Global topology analysis of the *Escherichia coli* inner membrane proteome**. *Science* 2005, **308**:1321-1323.

A whole-proteome analysis of the topology of proteins in *E. coli* that are predicted to be transmembrane. The data could serve as a benchmark for future studies and evaluations of topology prediction algorithms. Five out of 601 proteins were identified as having putative dual topology, with some of the protein copies inserting into the membrane with one topology and others with the reverse topology.

22. Rapp M, Granseth E, Seppala S, von Heijne G, Daley DO, Melen K, Drew D: **Identification and evolution of dual-topology membrane proteins**. *Nat Struct Mol Biol* 2006, **13**:112-116.

23. Ubarretxena-Belandia I, Baldwin JM, Schuldiner S, Tate CG: **Three-dimensional structure of the bacterial multidrug transporter EmrE shows it is an asymmetric homodimer**. *EMBO J* 2003, **22**:6175-6181.

24. Ninio S, Elbaz Y, Schuldiner S: **The membrane topology of EmrE — a small multidrug transporter from *Escherichia coli***. *FEBS Lett* 2004, **562**:193-196.

25. Bowie JU: **Solving the membrane protein folding problem**. *Nature* 2005, **438**:581-589.

26. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D: **Progress in modeling of protein structures and interactions**. *Science* 2005, **310**:638-642.

27. Bradley P, Misura KM, Baker D: **Toward high-resolution de novo structure prediction for small proteins**. *Science* 2005, **309**:1868-1871.

28. Kalani MY, Vaidehi N, Hall SE, Trabanino RJ, Freddolino PL, Kalani MA, Floriano WB, Kam VW, Goddard WA III: **The predicted 3D structure of the human D2 dopamine receptor and the binding site and binding affinities for agonists and antagonists**. *Proc Natl Acad Sci USA* 2004, **101**:3815-3820.

29. Trabanino RJ, Hall SE, Vaidehi N, Floriano WB, Kam VW, Goddard WA III: **First principles predictions of the structure and function of G-protein-coupled receptors: validation for bovine rhodopsin**. *Biophys J* 2004, **86**:1904-1921.

30. Freddolino PL, Kalani MY, Vaidehi N, Floriano WB, Hall SE, Trabanino RJ, Kam VW, Goddard WA III: **Predicted 3D structure for the human β2 adrenergic receptor and its binding site for agonists and antagonists**. *Proc Natl Acad Sci USA* 2004, **101**:2736-2741.

31. Vaidehi N, Floriano WB, Trabanino R, Hall SE, Freddolino P, Choi EJ, Zamanakos G, Goddard WA III: **Prediction of structure and function of G protein-coupled receptors**. *Proc Natl Acad Sci USA* 2002, **99**:12622-12627.

32. Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A, Kalid O, Bar-Haim S, Warshaviak D, Fichman M, Noiman S: **G protein-coupled receptors: *in silico* drug discovery in 3D**. *Proc Natl Acad Sci USA* 2004, **101**:11304-11309.

33. Shacham S, Marantz Y, Bar-Haim S, Kalid O, Warshaviak D, Avisar N, Inbal B, Heifetz A, Fichman M, Topf M *et al.*: **Predict modeling and *in-silico* screening for G-protein coupled receptors**. *Proteins* 2004, **57**:51-86.

34. • Zhang Y, Devries ME, Skolnick J: **Structure modeling of all identified G protein-coupled receptors in the human genome**. *PLoS Comput Biol* 2006, **2**:e13.

An adaptation of the TASSER algorithm for threading and refinement of protein structures to membrane proteins. The algorithm was validated on several proteins of solved structure, and then applied to predicting the structure of most human GPCRs. The resource of predicted structures is available at http://cssb.biology.gatech.edu/skolnick/files/gpcr/gpcr.html.

35. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D: **Free modeling with Rosetta in CASP6**. *Proteins* 2005, **61**:128-134.

36. • Yarov-Yarovoy V, Schonbrun J, Baker D: **Multipass membrane protein structure prediction using Rosetta**. *Proteins* 2006, **62**:1010-1025.

An adaptation of the Rosetta algorithm for *ab initio* protein structure prediction to membrane proteins. The quality of the predicted models was similar to that obtained for soluble proteins. Full-atom prediction was not attempted because of the computational cost of such implementations in large proteins.

37. • Fleishman SJ, Unger VM, Ben-Tal N: **Transmembrane protein structures without X-rays**. *Trends Biochem Sci* 2006, **31**:106-113.

A review of approaches for modeling TM protein structures based on intermediate resolution data. Some experimental data, particularly from crosslinking, are sometimes found to bias models away from the native state structures.

38. Herzyk P, Hubbard RE: **Automated method for modeling seven-helix transmembrane receptors from experimental data**. *Biophys J* 1995, **69**:2419-2442.

39. Sale K, Faulon JL, Gray GA, Schoeniger JS, Young MM: **Optimal bundling of transmembrane helices using sparse distance constraints**. *Protein Sci* 2004, **13**:2613-2627.

40. Unger VM, Kumar NM, Gilula NB, Yeager M: **Three-dimensional structure of a recombinant gap junction membrane channel**. *Science* 1999, **283**:1176-1180.

41. Hirai T, Heymann JA, Maloney PC, Subramaniam S: **Structural model for 12-helix transporters belonging to the major facilitator superfamily**. *J Bacteriol* 2003, **185**:1712-1718.

42. Baldwin JM: **The probable arrangement of the helices in G protein-coupled receptors**. *EMBO J* 1993, **12**:1693-1703.

43. Enosh A, Fleishman SJ, Ben-Tal N, Halperin D: **Assigning transmembrane segments to helices in intermediate-resolution structures**. *Bioinformatics* 2004, **20**:I122-I129.

44. Adamian L, Liang J: **Prediction of buried helices in multispan α helical membrane proteins**. *Proteins* 2006, **63**:1-5.

45. Baldwin JM, Schertler GF, Unger VM: **An α-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors**. *J Mol Biol* 1997, **272**:144-164.

46. Beuming T, Weinstein H: **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**:119-125.

47. Fleishman SJ, Harrington S, Friesner RA, Honig B, Ben-Tal N: **An automatic method for predicting the structures of transmembrane proteins using cryo-EM and evolutionary data**. *Biophys J* 2004, **87**:3448-3459.

48. Hurwitz N, Pellegrini-Calace M, Jones DT: **Towards genome-scale structure prediction for transmembrane proteins**. *Philos Trans R Soc Lond B Biol Sci* 2006, **361**:465-475.

49. •• Fleishman SJ, Unger VM, Yeager M, Ben-Tal N: **A C-α model for the transmembrane α-helices of gap-junction intercellular channels**. *Mol Cell* 2004, **15**:879-888.

A cryo-EM map of the gap junction was used together with evolutionary-conservation and correlated-mutations analyses to predict a model structure of the TM domain. The model puts disease-causing point mutations in structurally packed regions of the model.

50. Fleishman SJ, Yifrach O, Ben-Tal N: **An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels**. *J Mol Biol* 2004, **340**:307-318.

51. Ubarretxena-Belandia I, Engelman DM: **Helical membrane proteins: diversity of functions in the context of simple architecture**. *Curr Opin Struct Biol* 2001, **11**:370-376.

52. Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S: **Structure and mechanism of the lactose permease of *Escherichia coli***. *Science* 2003, **301**:610-615.

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

This analysis finds that in most cases where a proline is not observed in a kinked region of a TM protein structure, the multiple-sequence alignment exhibits a proline in several sequence homologues. This observation provides an approach for predicting the locations of kinks in protein structures.

54. Deupi X, Olivella M, Govaerts C, Ballesteros JA, Campillo M, Pardo L: **Ser and Thr residues modulate the conformation of pro-kinked transmembrane α-helices**. *Biophys J* 2004, **86**:105-115.

55. Lee SY, Lee A, Chen J, MacKinnon R: **Structure of the KvAP voltage-dependent K^+ channel and its dependence on the lipid membrane**. *Proc Natl Acad Sci USA* 2005, **102**:15441-15446.

56. Davidson AL, Chen J: **Structural biology. Flipping lipids: is the third time the charm?** *Science* 2005, **308**:963-965.

57. Ma C, Chang G: **Structure of the multidrug resistance efflux transporter EmrE from *Escherichia coli***. *Proc Natl Acad Sci USA* 2004, **101**:2852-2857.

58. Pornillos O, Chen YJ, Chen AP, Chang G: **X-ray structure of the EmrE multidrug transporter in complex with a substrate**. *Science* 2005, **310**:1950-1953.

59. Donnelly D, Overington JP, Ruffle SV, Nugent JH, Blundell TL: **Modeling α-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues**. *Protein Sci* 1993, **2**:55-70.

60. Briggs JA, Torres J, Arkin IT: **A new method to model membrane protein structure based on silent amino acid substitutions**. *Proteins* 2001, **44**:370-375.

61. Tate CG: **Comparison of three structures of the multidrug transporter EmrE**. *Curr Opin Struct Biol* 2006, **16**: this issue.

62. Mottamal M, Zhang J, Lazaridis T: **Energetics of the native and non-native states of the glycophorin transmembrane helix dimer**. *Proteins* 2006, **62**:996-1009.

63. Zhou FX, Cocco MJ, Russ WP, Brunger AT, Engelman DM: **Interhelical hydrogen bonding drives strong interactions in membrane proteins**. *Nat Struct Biol* 2000, **7**:154-160.

64. Choma C, Gratkowski H, Lear JD, DeGrado WF: **Asparagine-mediated self-association of a model transmembrane helix**. *Nat Struct Biol* 2000, **7**:161-166.

65. Arbely E, Arkin IT: **Experimental measurement of the strength of a C^{α}–H. . .O bond in a lipid bilayer**. *J Am Chem Soc* 2004, **126**:5362-5363.

66. Yohannan S, Faham S, Yang D, Grosfeld D, Chamberlain AK, Bowie JU: **A C^{α}–H. . .O hydrogen bond in a membrane protein is not stabilizing**. *J Am Chem Soc* 2004, **126**:2284-2285.

67. Vargas R, Garza J, Dixon D, Hay B: **How strong is the C^{α}–H. . .O=C hydrogen bond?** *J Am Chem Soc* 2000, **122**:4750-4755.

68. Scheiner S, Kar T, Gu Y: **Strength of the C^{α}–H. . .O hydrogen bond of amino acid residues**. *J Biol Chem* 2001, **276**:9832-9837.

69. Faham S, Yang D, Bare E, Yohannan S, Whitelegge JP, Bowie JU:
•• **Side-chain contributions to membrane protein structure and stability**. *J Mol Biol* 2004, **335**:297-305.

An analysis of the contributions to stability of individual amino acid residues on helix B from bacteriorhodopsin. It is found that the contribution correlates with the amount of buried surface area rather than the ability to provide hydrogen-bonding interactions, roughly as seen for soluble proteins. Surprisingly, a mutation of a kink-inducing proline to alanine did not decrease stability significantly, and only elicited minor changes in secondary structure.

70. Kraulis PJ: **MolScript: a program to produce both detailed and schematic plots of protein structures**. *J Appl Cryst* 1991, **24**:946-950.

71. Merritt EA, Bacon DJ: **Raster3D: photorealistic molecular graphics**. *Methods Enzymol* 1997, **277**:505-524.

72. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N: **ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information**. *Bioinformatics* 2003, **19**:163-164.

73. Yarov-Yarovoy V, Baker D, Caterall WA: **Voltage sensor conformations in the open and closed states in structural models of K^+ channels**. *Proc Natl Acad Sci USA* 2006, **103**:7292-7297.