Membrane Protein Structure: Prediction versus Reality

Arne Elofsson and Gunnar von Heijne

Center for Biomembrane Research, Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden; email: arne@bioinfo.se, gunnar@dbb.su.se

Annu. Rev. Biochem. 2007. 76:125–40

First published online as a Review in Advance on January 11, 2007

The *Annual Review of Biochemistry* is online at biochem.annualreviews.org

This article's doi:
10.1146/annurev.biochem.76.052705.163539

0066-4154/07/0707-0125$20.00

Key Words

bioinformatics, membrane protein structure prediction, topology

Abstract

Since high-resolution structural data are still scarce, different kinds of theoretical structure prediction algorithms are of major importance in membrane protein biochemistry. But how well do the current prediction methods perform? Which structural features can be predicted and which cannot? And what can we expect in the next few years?

Contents

INTRODUCTION

Membrane proteins are crucial players in the cell and take center stage in processes ranging from basic small-molecule transport to sophisticated signaling pathways. Many are also prime contemporary or future drug targets, and it has been estimated that more than half of all drugs currently on the market are directed against membrane proteins (1). By contrast, it is still frustratingly hard to obtain high-resolution three-dimensional (3D) structures of membrane proteins, and they represent less than 1% of the structures in the Protein Data Bank (2). Even if the number of experimentally known membrane protein structures is on the rise (3, 4), methods to predict their topology (i.e., the transmembrane segments and their in-out orientation across the membrane) and fold type from the amino acid sequence will be needed for many years to come.

In this review, we discuss current topology and structure prediction methods against a background of knowledge that has been gleaned from membrane protein structures and from studies of protein insertion and folding in cellular membranes. We attempt to provide a realistic picture of what one may and may not expect from the various prediction schemes and to identify major issues yet to be resolved.

MEMBRANE PROTEIN STRUCTURES: THE BASIC FACTS

Integral membrane proteins come in two basic architectures: the α-helix bundle and the β-barrel. Helix-bundle proteins are found in all cellular membranes and represent an estimated 20% to 25% of all open reading frames (ORFs) in fully sequenced genomes (5). The number of β-barrel membrane proteins is more uncertain because they are more difficult to identify by sequence gazing; for bacteria, a rough estimate, based on the fact that all known β-barrel proteins are in the outer membrane and hence are made with an (easily predicted) N-terminal signal peptide, suggests that they account for no more than a few percent of all ORFs. The EcoCyc database (6) currently lists 58 outer membrane and 511 inner membrane proteins out of a total of 4332 proteins; considering that the number of inner membrane proteins in *Escherichia coli* has been estimated to be close to 1000 (5), one may guess at somewhere between 100 and 150 outer membrane proteins in total. This number of ~100 *E. coli* outer membrane proteins is consistent with the results from recent attempts to identify bacterial outer membrane proteins computationally (7–9).

Whether a helix bundle or β-barrel, all integral membrane proteins share common surface characteristics with a belt of hydrophobic (mainly aliphatic) amino acids flanked by two "aromatic girdles" composed of Trp and Tyr residues (10–12). This mirrors the structure of the surrounding lipid bilayer, with the lipid headgroup regions corresponding to the aromatic girdles and the hydrocarbon tail region to the hydrophobic belt, and ensures a seamless fit of the proteins to the membrane.

Even though their surface structures are similar, the two classes of proteins have completely different secondary structures and folds. As their names imply, helix-bundle proteins are built from long transmembrane α-helices that pack together into more or less complicated bundles, whereas β-barrel proteins are large antiparallel β-sheets rolled up into a barrel closed by the first and last strands in the sheet. In both cases, all backbone hydrogen bonds in the membrane-buried parts of the protein are internally satisfied within the helices or between the β-strands. Another fundamental difference between the helix-bundle and β-barrel proteins pertains to their biosynthesis and mechanism of membrane insertion; this is discussed in the next section.

Because all current membrane protein topology and structure prediction schemes first seek to identify the transmembrane segments, they are obviously quite different, and their variations depend on the class of protein for which they are designed. Generally speaking, long hydrophobic transmembrane helices are easier to recognize in an amino acid sequence than the much shorter and less hydrophobic transmembrane β-strands, and partly for this reason, much more bioinformatics work has been devoted to the helix-bundle proteins—another instance of the well-known dictum "always go for the easy problems."

MEMBRANE PROTEIN BIOSYNTHESIS, FOLDING, AND OLIGOMERIZATION

Membrane Targeting and Insertion

As do all other proteins, a membrane protein starts its life on the ribosome. But already at this early stage, helix-bundle and β-barrel proteins are handled differently (13, 14): ribosomes making helix-bundle proteins typically bind cotranslationally to translocons in a target membrane [the inner membrane in bacteria, the endoplasmic reticulum (ER) in eukaryotes], whereas bacterial β-barrel proteins are initially transferred from the ribosome to the soluble cytoplasmic SecB chaperone, **Figure 1**.

Translocon: a protein complex that assures the translocation of proteins across a cellular membrane

Endoplasmic reticulum (ER): organelle into which secretory and membrane proteins are delivered upon synthesis on the ribosome

The cotranslational membrane insertion of helix-bundle proteins has been studied intensely for many years, and it now appears that the transmembrane helices move laterally from the translocon channel into the surrounding lipid bilayer, either one at a time or in pairs, depending on their hydrophobicity and their ability to form stable helix-helix interactions. Furthermore, it appears that the molecular features that allow the translocon to recognize a stretch of a polypeptide in transit as a transmembrane helix are the same as those seen to mediate protein-lipid interactions in the known membrane protein structures (15), strongly suggesting that the translocon is designed such that it allows a translocating nascent chain to sample the surrounding bilayer. At its simplest, transmembrane helix insertion may thus be approximated as a thermodynamic partitioning between the aqueous milieu in the translocon channel and the lipid membrane.

The β-barrel proteins in the bacterial outer membrane are also translocated through the inner membrane translocon, but they do so posttranslationally with the aid of the SecA ATPase, and their short transmembrane β-strands are not sufficiently hydrophobic to get stuck across the inner membrane (16). Instead, they are chaperoned through the periplasmic space and finally insert into the outer

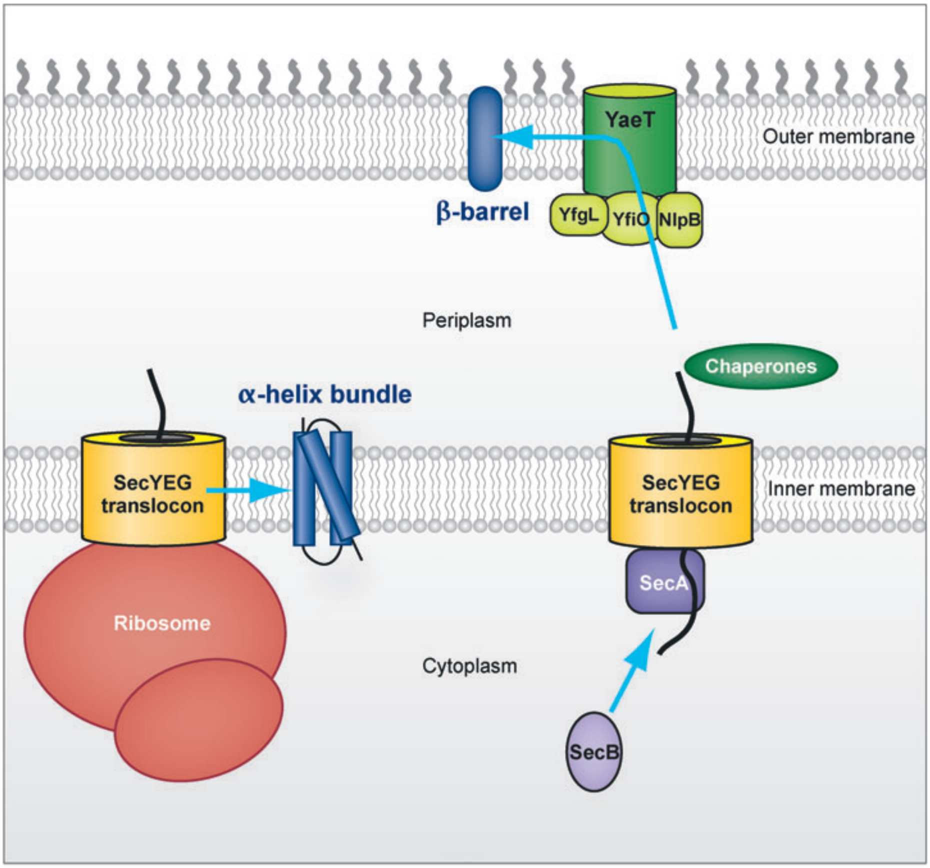

Figure 1

Biogenesis of α-helix bundle (*left*) and β-barrel (*right*) membrane proteins in *Escherichia coli*.

membrane with the aid of the resident YaeT hetero-oligomeric outer membrane integration complex (14).

Folding and Stability

Once inserted into the membrane, transmembrane helices pack into the typical helix-bundle folds, and many then go on to form homo- or hetero-oligomeric complexes. Membrane proteins form closely packed structures, and it is believed that an important driving force for folding is better shape complementarity between the transmembrane helices than between the helices and the lipid (17). Other factors that come into play are hydrogen bonding between polar side chains (18) and possibly even the formation of C_α–H–O hydrogen bonds (19, 20; but also see Reference 21).

Many membrane proteins, both helix bundle and β-barrel, form stable structures with little flexibility, whereas others undergo substantial rearrangements of their transmembrane domains as part of a reaction cycle. Proteins involved in proton and electron transfer typically coordinate a range of cofactors that need to be positioned relative to each other with Å-level precision and hence must be quite rigidly packed (22), whereas small-molecule transporters must flip between dramatically different conformations open either toward the external or the internal side of the cell (23, 24).

MEMBRANE PROTEIN BIOINFORMATICS: WHAT THE SEQUENCES TELL

For the helix-bundle membrane proteins, amino acid sequences told their story long before the first high-resolution structures were determined: the typical transmembrane segment is formed by a stretch of predominantly hydrophobic residues long enough to span the lipid bilayer as an α-helix (25–29). The early topology prediction methods were consequently little more than plots of the segmental hydrophobicity (averaged over 10–20 residues) along the sequence (30–32). With more sequences came the realizations that aromatic Trp and Tyr residues tend to cluster near the ends of the transmembrane segments (10, 33) and that the loops connecting the helices differ in amino acid composition, depending on whether they face the inside or outside of the cell (the "positive-inside" rule) (34–36). More recent analyses have focused on the higher-than-random appearance of sequence motifs, such as the GxxxG-motif in transmembrane segments (37, 38) as well as other periodic patterns within the membrane helices (39), with the aim of providing information that may help in predicting helix-helix packing and 3D structure.

MEMBRANE PROTEIN BIOINFORMATICS: WHAT THE STRUCTURES TELL

For a long time, the general view has been that membrane proteins form simple helix bundles, with their transmembrane helices crisscrossing the membrane in more or less perpendicular orientations. Indeed, many membrane proteins abide by this principle. However, some more recently solved membrane protein structures show that reality is not always this simple. This is illustrated by the structure of the glutamate transporter homolog from *Pyrococcus horikoshii* (40), shown in **Figure 2**. This protein has six typical transmembrane helices and two irregular helices with breaks inside the lipid bilayer. The structure also contains two reentrant loops that go only halfway through the membrane and then turn back to the side from which they

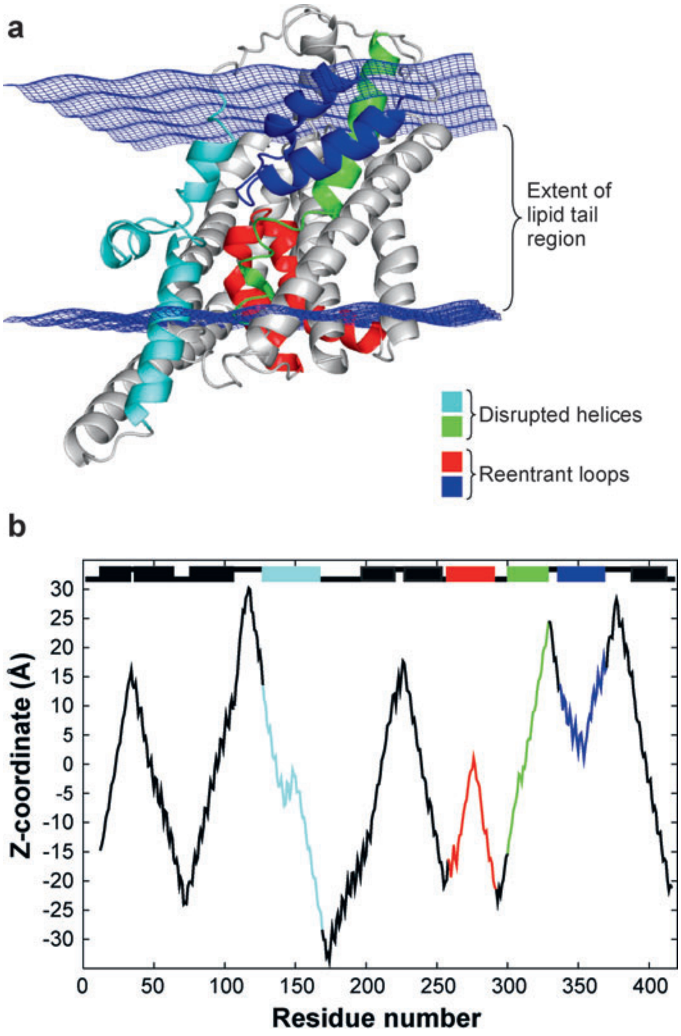

Figure 2

(*a*) The glutamate transporter homolog (1XFH) contains both disrupted transmembrane helices and reentrant loops. Disrupted helices are shown (*cyan* and *green*), and reentrant loops are also shown. The mesh indicates the approximate extent of the lipid tail region (± 15 Å). (*b*) Topology (*upper part*) and z-coordinate plot. The z-coordinate plot shows the distance from the center of the membrane for each residue. The coloring is the same as in panel *a*. Modified with permission of Oxford University Press (79).

Reentrant loop: a structural motif in which the polypeptide dips only partway across the membrane originate. The two reentrant loops meet in the middle of the membrane, a feature also seen in the aquaporin structures (41).

Other structural elements, largely ignored until recently in statistical studies of membrane protein structure, are found in those parts of the protein that are located in the membrane-water interface region. Here one finds irregular structure and interfacial helices running roughly parallel to the membrane surface, while β-strands are extremely rare (42–44). The average amino acid composition is different between the interfacial helices, the parts of the transmembrane helices located in the interface region, and the irregular structures. Hydrophobic and aromatic residues in this region tend to point toward the center of the membrane, whereas charged and polar residues tend to point away from the membrane. The interface region thus imposes different constraints on protein structure than do the central hydrocarbon core of the membrane and the surrounding aqueous phase.

For β-barrel membrane proteins, a number of structural rules have been deduced from the known structures (45): The number of β-strands is even; the N and C termini are at the periplasmic barrel end; the β-strand tilt is ~45°; all β-strands are antiparallel and connected locally to their next neighbors along the chain; and the β-barrel surface, contacting the nonpolar membrane interior, consists of a belt of aliphatic side chains lined by two girdles of aromatic side chains. These generalizations provide a framework for the development of topology prediction algorithms.

MEMBRANE PROTEIN STRUCTURE PREDICTION: FROM 2D TO 2.5D AND 3D

Over the years, many topology and structure prediction schemes have been developed for helix-bundle membrane proteins. In general, there has been a logical progression from simple to more complicated models. This has been enabled by the increase in available training data, a better understanding of membrane protein structure, and advances in machine-learning methods.

2D Predictions

The earliest topology prediction methods relied only on the fact that transmembrane helices are on average more hydrophobic than loop regions. Although these simple methods worked surprisingly well, they left a lot to be desired. Inclusion of the positive-inside rule led to a significant improvement in the predictions (46), and a further step was taken when hidden Markov models (HMMs), **Figure 3**, and other machine-learning techniques were employed to extract the relevant sequence features (5, 47–50). In addition, HMM-based methods that also include evolutionary and/or limited experimental information to improve topology predictions have been developed (51–53).

One particular problem faced by all topology predictors is to discriminate between signal peptides and transmembrane helices—two kinds of topogenic elements that look quite similar. This problem has recently been addressed with the development of Phobius (54), an HMM-based method that predicts both signal peptides and transmembrane segments simultaneously and thereby significantly decreases the confusion between them.

Much less work has been devoted to the development of topology prediction schemes for β-barrel membrane proteins, in part because the membrane-spanning β-strands are both considerably shorter and much less conspicuous in terms of amino acid sequence than the long, hydrophobic transmembrane helices in the helix-bundle proteins. The simplest methods attempt to identify bacterial outer membrane β-barrel proteins using only two criteria: the presence of an N-terminal signal peptide [predicted using a program such as SignalP (55, 56)] and the overall amino acid composition of the protein (9, 57). The more advanced methods also predict the individual β-strands and the topology of the protein, in most cases using HMMs (58–60).

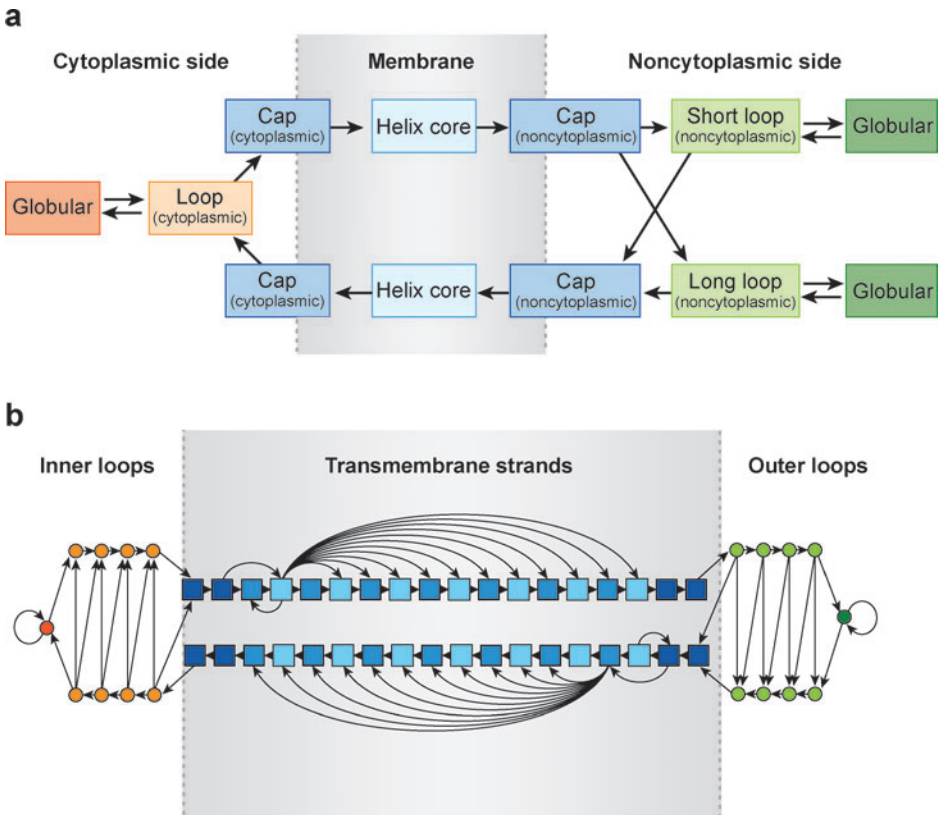

Figure 3

Typical hidden Markov models (HMMs) for topology predictions of (*a*) helix-bundle membrane proteins (5) and (*b*) β-barrel membrane proteins. Modified with permission of Oxford University Press (58).

Benchmarking

The topology of more than 400 helix-bundle membrane proteins has been determined experimentally by a variety of genetic, biochemical, and structural techniques, and such data has been used both to train and to benchmark the various topology prediction methods. However, much of the experimental data are of quite low resolution and not always correct. In addition, when benchmarking different methods, many diverse quality measures have been used. In principle, these can be divided into residue-based and topology-based instruments. In our experience, benchmarking with residue-based measures is not optimal as (*a*) all methods perform more or less equally well by such measures, and (*b*) most experimental topology information is not of high enough quality to exactly define the borders between membrane and nonmembrane parts of the proteins.

The different data sets used in benchmarking studies can also impact the results. An especially important point is if single-spanning membrane proteins are included or not. On average, transmembrane helices in single-spanning proteins are more hydrophobic than in multispanning (polytopic) membrane proteins (A. Bernsel & G. von Heijne, unpublished data). Including single-spanning proteins in training data may compromise the performance on multispanning proteins and vice versa. Ideally, single-spanning proteins should, therefore, be treated differently than multispanning proteins. However, to our knowledge no method has successfully included this distinction in a prediction scheme, and it is rarely taken into account in benchmarking studies.

Given these caveats, it is nevertheless clear from several recent benchmarking studies (61–63) that HMM-based methods that also include evolutionary information, e.g., polyPhobius (64), HMMTOP (50), and prodivTMHMM (52), perform best. These methods predict the correct topology (i.e., the correct number of transmembrane helices and the correct overall orientation of the protein in the membrane) for close to 70% of all membrane proteins, and this number can be improved even further when additional experimental information is available to constrain the predictions (51, 53, 65, 66).

A recent benchmarking of β-barrel membrane protein topology predictors concluded that HMM-based methods also perform better than other methods for these proteins (67). The best HMM methods predicted the correct topology for 14 out of 20 test proteins (70%), although the real accuracy is likely lower than this because many of the proteins in the test set were used also to develop the different methods.

It should be kept in mind that a correct topology prediction does not mean that the predicted starts and ends of the transmembrane α-helices or β-strands can be trusted; only the number of transmembrane helices and their approximate positions are correct. In fact, this part of the structure prediction problem has not yet been satisfactorily resolved. With the availability of more exact structural information, it should be possible to evaluate how well different methods can predict the exact helix locations. Our experience has shown that it is possible to predict the points of entrance and exit from the membrane environment with acceptable accuracy, whereas the prediction of helix start and termination points is very hard.

The rapid increase in high-resolution structural data for membrane proteins means that in the future both benchmarking and development of novel prediction methods should be based on structural data only. Luckily, the general conclusions from using structure-based benchmarks are similar to those of earlier studies (52, 68). One remaining problem with the structural data is that the precise location of the protein in the lipid membrane is often not immediately available because crystals are grown from detergent-solubilized proteins. Automatic methods that optimize the fit between a 3D protein structure and a model membrane are available (69, 70), although it is difficult to assess their accuracy.

Genome Annotation

An important application of topology prediction algorithms is to annotate genome sequencing data. It has been reported that algorithms such as TMHMM can discriminate helix-bundle proteins from other proteins with better than 95% sensitivity and specificity (5), meaning that the helix-bundle membrane proteome of an organism can be quite reliably predicted from its genome sequence.

It was initially observed that the distribution of helix-bundle membrane protein topologies in a genome seemed to follow a power law with respect to the number of transmembrane helices, i.e., that proteins with few transmembrane helices are more frequent than proteins with many transmembrane helices (71–73). As the topology predictors improved, several exceptions to this general trend were noted (5, 65, 66, 74). In particular, bacterial genomes encode large numbers of small-molecule transporters with 6 or (more often) 12 transmembrane helices, whereas mammalian genomes are strongly enriched for G protein–coupled receptors (GPCRs) with 7 transmembrane helices, as well as for small-molecule transporters. With the availability of more accurate predictors and genome-wide experimental topology data, it was also noted that there is a strong overrepresentation of proteins with an even number of transmembrane helices and with their N and C termini located on the cytoplasmic side of membrane, both in bacteria and eukaryotes (5, 65, 66).

2.5D Predictions

As noted above, the general view, until recently, has been that the basic structural feature of helix-bundle membrane proteins is the perpendicularly penetrating transmembrane helices. But as many 3D structures now show, membrane protein structures are often too complex to fit completely into such a simple topology model. To advance further, a more fine-grained definition of topology, "a two-and-a-half dimensional (2.5D) structure," is needed where structural elements, such as interfacial helices and reentrant loops, are taken into account. In addition, several other limitations of the current generation of predictors exist, e.g., it has been noted that such a trivial characteristic as the exact length of transmembrane helices is very difficult to predict using current methodologies.

Reentrant loops are a common feature in many membrane proteins. They were first seen in the aquaporin-1 water channel (75) and in the KcsA potassium channel (76). Detailed analysis suggests that reentrant loops can be divided into three distinct categories based on secondary structure content: long loops with a helix-coil-helix structure, loops of medium length with a helix-coil or coil-helix structure, and loops of short to medium length consisting entirely of an irregular secondary structure (77). Residues in reentrant loops are significantly smaller on average compared to other parts of the protein, and they can be detected in regions between the transmembrane helices with ~70% accuracy based on their amino acid composition. Reentrant loops often contain particular functional motifs that enable them to be detected (78). On the basis of a novel predictor for reentrant loops (TOP-MOD), it appears that more than 10% of all multispanning membrane proteins contain such loops (77). Reentrant loops seem to be most commonly found in ion and water channel proteins and least commonly in cell surface receptors.

Although the division of a membrane protein into different substructures is clearly useful, distinguishing different types of structural elements is not always straighforward. Reentrant loops can vary quite dramatically in their secondary structure and depth of penetration into the membrane, and the length of transmembrane helices varies significantly. An alternative approach to membrane protein 2.5D structure prediction is to directly predict the distance from the center of the membrane (i.e., the z-coordinate) for each residue in a protein, rather than the type of structural element of which it is a part. One recent algorithm of this kind correctly classified 88% of all residues in the test set proteins to be inside or outside the membrane, with an average error of 2.5 Å in the predicted residue distances from the center of the membrane (79). A similar z-coordinate predictor has also been developed for β-barrel membrane proteins (80).

An important characteristic of residues in transmembrane helices is their degree of lipid exposure in the folded structure. In contrast to globular proteins, membrane proteins do not show a large difference in hydrophobicity between the lipid-exposed and buried residues in the membrane-embedded region, and the prediction of surface accessibility becomes much harder. The major features distinguishing the lipid-exposed and buried residues are the polarity of the side chain (more hydrophobic residues tend to be more lipid exposed) and the degree of sequence conservation (less conserved residues tend to be more lipid exposed). Many attempts to predict lipid exposure have been published; one of the most recent studies reports the prediction of lipid-exposed surface patches in transmembrane helices that interface with lipid molecules with a per residue accuracy of 88% (81).

A final 2.5D characteristic that can be predicted with reasonable accuracy is the presence of proline-induced kinks in the transmembrane helices. Interestingly, such kinks can be preserved even when the Pro residue is mutated (82, 83), and a kink can confidently be predicted if proline is conserved in a particular position in a transmembrane helix in more than 10% of the sequences in a multiple alignment (83).

The prediction of 2.5D features of membrane proteins should not only be useful as a step toward 3D predictions. Such predictors will also help in the classification of membrane protein families because the different substructures provide unique sequence signatures separating different membrane protein families with the same topology. In addition, the identification of suitable peptide antibody epitopes may be facilitated if z-coordinates can be accurately predicted.

3D Predictions

Interestingly, 3D structure predictions of membrane proteins were attempted even before the first high-resolution structure of any membrane proteins was solved. Using information from low-resolution experiments, in particular electron microscopy, quite accurate models of bacteriorhodopsin (84) as well as GPCRs (85, 86) were made.

Despite these early successes, the field of nonhomology-based 3D structure prediction for membrane proteins has followed a similar trend as that seen in the globular protein structure prediction field, wherein the general experience is that most methods when tested in blind predictions show a much lower accuracy than first reported. However, through rounds of iterative refinement, the best methods can now predict the structure of small globular proteins quite accurately (87). As it turns out, one of the most important improvements to the methodology has been to base the 3D prediction on short sequence fragments extracted from known protein structures (88–90).

It is, however, not straightforward to apply similar schemes to membrane proteins because the different environment introduced by the membrane has to be modeled in some way, and because most membrane proteins are significantly larger than the globular proteins successfully predicted so far. Therefore the success to date has been quite limited even using the most advanced methods adapted from the globular protein field (91, 92). If experimentally derived distance constraints from techniques, such as Fourier transform infrared spectroscopy, electron paramagnetic resonance spectroscopy, and chemical cross-linking, or low-resolution models based on electron microscopy, are available, more reliable models can be built (93).

An interesting attempt to model all GPCRs of the human proteome was recently made by Skolnick and coworkers (94) using the TASSER algorithm. Although the accuracy of the predicted rhodopsin structure was quite good, the correctness of the GPCR structures can not be verified until more structures are available.

Many membrane proteins, in particular channels and transporters, undergo substantial structural changes during a reaction cycle. Often the structure of only one of the states is available, and methodology to reliably model structural changes will be needed for a long time to come. In a recent study, the ROSETTA membrane folding algorithm was used to model the closed and open states of a voltage-dependent potassium channel (95), generating a number of testable hypotheses that may guide further experimental work. For a more thorough review of ab initio structure prediction methods, see Reference 96.

While ab initio structure modeling can at best predict the overall fold of a protein, structure modeling based on a preexisting structure of a close homologue promises atomic-level structural detail. Homology modeling of membrane proteins is still in its infancy, however, because so few structures are known. A recent benchmarking study suggests that, when a template is available, homology models of membrane proteins are comparable in quality to those that can be made for globular proteins; i.e., when the sequence identity between the template and the target is >30%, one can expect the root mean-square deviation between the modeled and correct structure to be less than 2 Å in the transmembrane regions (97).

MEMBRANE PROTEIN CLASSIFICATION SCHEMES AND DATABASES

Hierarchically organized databases of protein structure have found many uses, both in studies of protein evolution, as a means to put together nonredundant sequence and structure collections for statistical studies, and as test beds for benchmarking fold recognition and structure prediction algorithms. For globular proteins, several well-established hierarchical, structure-based domain classification schemes, such as SCOP (98) and CATH (99), exist. However, the number of membrane protein structures is still too low for such classifications to cover an important part of the membrane protein universe (4). All currently known high-resolution membrane protein structures are listed at **http://blanco.biomol.uci.edu/MemPro_resources.html** and have also been organized into a database at **http://pdbtm.enzim.hu/** (100).

Sequence-based, nonhierarchical classifications of membrane protein domains are available in Pfam (101) and other similar databases, and specialized databases collecting families of membrane transporters (**http://www.tcdb.org/**), GPCRs (**http://www.gpcr.org/7tm/**), and potassium channel proteins (**http://www.receptors.org/KCN**) have been set up. Because these databases are not based on 3D structure information, they do not contain complete information about very distant evolutionary relationships (e.g., the SCOP class, fold, and superfamily levels).

The detection of distantly related globular proteins has seen great progress during the past decade through the use of fold-recognition methods, improved use of evolutionary relationships, and careful benchmarking. Because of the low incidence of polar and charged residues in transmembrane helices, the use of algorithms optimized for globular proteins for the detection of distantly related membrane proteins is problematic. To compound these difficulties, a major obstacle for the development of improved methods to identify distantly related membrane proteins is the lack of a structure-based "gold standard" such as SCOP.

Nevertheless, large, divergent membrane protein families, such as the GPCRs, can be used to benchmark fold-recognition algorithms for membrane proteins. As for globular proteins, it appears that HMM-based sequence family models, profile-profile similarity searches, and the inclusion of secondary structure information in the form of predicted topology models all help in the detection of distant homologues (102, 103). It has also been shown that the best alignment of related membrane proteins is obtained using profile-profile methods in combination with predicted secondary structures (97).

PROTEIN-PROTEIN INTERACTIONS

The final step on the structure prediction ladder is the prediction of quaternary structure, i.e., protein-protein interactions. This is especially pertinent for membrane proteins because membrane-integral protein domains in most cases seem to be encoded by separate polypeptides rather than as multidomain polypeptides as often found in globular proteins (104). Large-scale experimental protein-protein interaction studies tend to ignore membrane proteins, although some data are now starting to appear in the literature (105–107).

The current state of predicting interactions between membrane proteins may be summarized in a few words: much remains to be done. For example, a recent attempt to predict interacting proteins in the *Saccharomyces cerevisiae* membrane proteome by integrating data, such as amino acid sequence, annotated function, subcellular localization, mRNA and protein abundance, transcriptional coregulation, and gene knock-out phenotype, resulted in a predictor that could identify ~40% of 304 experimentally well-documented gold standard interactions while minimizing the number of false-positive predictions (108).

CONCLUSIONS AND OUTLOOK

If bioinformatics methods are evaluated by how they are received by the scientific community at large, it is clear that membrane protein structure prediction algorithms hold their ground; to give but one example, TMHMM (5, 47) has been cited well over 1200 times. But precisely what sort of information can one expect to get from the various prediction methods? And what sort of advances can we see on the horizon?

First and foremost, do not expect the computer to tell you the truth! Topology predictions are just predictions. True, high-scoring predictions are nearly always right (51, 63, 109), but this only means that the really clear-cut cases (i.e., those that can equally well be done by hand) are easy to predict. Still, if taken with a grain of salt, topology models, predicted lipid-exposed residues, potential reentrant loops, and lists of possibly interacting partner proteins can be invaluable guides for planning experiments and interpreting results. And large-scale computational studies of entire genomes can provide tantalizing clues to everything from the basic patterns of membrane protein evolution (110, 111) to differences in lifestyle between different organisms—show me your transporters, and I will tell you where you live.

Still, much remains to be done, both in perfecting the current arsenal of prediction methods and devising entirely new algorithms to do new things. Our current representation of membrane protein topology as a simple string of membrane-spanning α-helices or β-strands does not fully capture the structural diversity seen in membrane proteins; defining a fuzzy area between the 2D and 3D structure is in need of more exploration. The rapid growth in known membrane protein 3D structures improves the prospects for effective fold-recognition and homology modeling approaches, although the day when most of membrane protein fold space has been mapped experimentally seems desperately far off (4). Computational means to map out the membrane interactome will become an important complement to high-throughput (but error-prone) experimental studies, and here, as in so many other areas, tight integration between the "wet" and "dry" approaches is certainly the best way forward.

SUMMARY POINTS

1. Integral membrane proteins come in two basic architectures: α-helix bundles and β-barrels.
2. The lipid-facing surface of integral membrane proteins is composed of a central "hydrophobic belt" flanked by two "aromatic girdles."
3. In the helix-bundle proteins, nontranslocated loops are enriched in Lys and Arg compared to translocated loops (the positive-inside rule).
4. Helix-bundle membrane proteins are built from transmembrane α-helices, interfacial helices lying flat on the membrane, reentrant loops, and extramembraneous globular domains.
5. For the β-barrel protein, the number of β-strands is even, the N and C termini are at the periplasmic barrel end, the β-strand tilt is ~45°, and all β-strands are antiparallel and connected locally to their next neighbors along the chain.

6. The best topology prediction algorithms forecast the correct topology for ≤70% of all proteins but cannot accurately predict the start and end of a transmembrane segment.
7. Only a few recent prediction algorithms attempt to identify surface helices and reentrant loops.
8. Ab initio high-resolution 3D structure prediction is still not feasible for membrane proteins. Homology-based structure modeling of membrane proteins performs on a par with homology modeling of globular proteins.

ACKNOWLEDGMENTS

The authors' laboratories are supported by grants from the Swedish Foundation for Strategic Research, the Marianne and Marcus Wallenberg Foundation, the Swedish Cancer Foundation, the Swedish Research Council, and the European Commission (BioSapiens, Genefun, EMBRACE).

LITERATURE CITED

1. Klabunde T, Hessler G. 2002. *ChemBioChem* 3:928–44
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. 2000. *Nucleic Acids Res.* 28:235–42
3. White SH. 2004. *Protein Sci.* 13:1948–49
4. Oberai A, Ihm Y, Kim S, Bowie JU. 2006. *Protein Sci.* 15:1723–34
5. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. *J. Mol. Biol.* 305:567–80
6. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, et al. 2005. *Nucleic Acids Res.* 33:D334–37
7. Wimley WC. 2002. *Protein Sci.* 11:301–12
8. Casadio R, Fariselli P, Finocchiaro G, Martelli PL. 2003. *Protein Sci.* 12:1158–68
9. Garrow AG, Agnew A, Westhead DR. 2005. *BMC Bioinformatics* 6:56
10. Wallin E, Tsukihara T, Yoshikawa S, von Heijne G, Elofsson A. 1997. *Protein Sci.* 6:808–15
11. Seshadri K, Garemyr R, Wallin E, von Heijne G, Elofsson A. 1998. *Protein Sci.* 7:2026–32
12. Ulmschneider MB, Sansom MS, Di Nola A. 2005. *Proteins* 59:252–65
13. Luirink J, von Heijne G, Houben E, de Gier JW. 2005. *Annu. Rev. Microbiol.* 59:329–55
14. Ruiz N, Kahne D, Silhavy TJ. 2006. *Nat. Rev. Microbiol.* 4:57–66
15. White SH, von Heijne G. 2004. *Curr. Opin. Struct. Biol.* 14:397–404
16. MacIntyre S, Freudl R, Eschbach ML, Henning U. 1988. *J. Biol. Chem.* 263:19053–59
17. Popot J-L, Engelman DM. 2000. *Annu. Rev. Biochem.* 69:881–922
18. Curran AR, Engelman DM. 2003. *Curr. Opin. Struct. Biol.* 13:412–17
19. Senes A, Ubarretxena-Belandia I, Engelman DM. 2001. *Proc. Natl. Acad. Sci. USA* 98:9056–61
20. Arbely E, Arkin IT. 2004. *J. Am. Chem. Soc.* 126:5362–63
21. Yohannan S, Faham S, Yang D, Grosfeld D, Chamberlain AK, Bowie JU. 2004. *J. Am. Chem. Soc.* 126:2284–85
22. Brzezinski P, Adelroth P. 2006. *Curr. Opin. Struct. Biol.* 16:465–72
23. Guan L, Kaback HR. 2006. *Annu. Rev. Biophys. Biomol. Struct.* 35:67–91

24. Toyoshima C, Nomura H, Tsuda T. 2004. *Nature* 432:361–68
25. Henderson R, Unwin PNT. 1975. *Nature* 257:28–32
26. Tomita M, Marchesi VT. 1975. *Proc. Natl. Acad. Sci. USA* 72:2964–68
27. Ovchinnikov YA, Abdulaev NG, Feigina MY, Kiselev AV, Lobanov NA. 1977. *FEBS Lett.* 84:1–4
28. Khorana HG, Gerber GE, Herlihy WC, Gray CP, Anderegg RJ, et al. 1979. *Proc. Natl. Acad. Sci. USA* 76:5046–50
29. Ovchinnikov YA, Abdulaev NG, Feigina MY, Kiselev AV, Lobanov NA. 1979. *FEBS Lett.* 100:219–24
30. von Heijne G, Blomberg C. 1979. *Eur. J. Biochem.* 97:175–81
31. Engelman DM, Steitz TA. 1981. *Cell* 23:411–22
32. Kyte J, Doolittle RF. 1982. *J. Mol. Biol.* 157:105–32
33. Weiss MS, Kreusch A, Schiltz E, Nestel U, Welte W, et al. 1991. *FEBS Lett.* 280:379–82
34. von Heijne G. 1986. *J. Mol. Biol.* 189:239–42
35. von Heijne G. 1986. *EMBO J.* 5:3021–27
36. von Heijne G. 1989. *Nature* 341:456–58
37. Senes A, Gerstein M, Engelman DM. 2000. *J. Mol. Biol.* 296:921–36
38. Kim S, Jeon TJ, Oberai A, Yang D, Schmidt JJ, Bowie JU. 2005. *Proc. Natl. Acad. Sci. USA* 102:14278–83
39. Samatey FA, Xu C, Popot J-L. 1995. *Proc. Natl. Acad. Sci. USA* 92:4577–81
40. Yernool D, Boudker O, Jin Y, Gouaux E. 2004. *Nature* 431:811–18
41. Hedfalk K, Tornroth-Horsefield S, Nyblom M, Johanson U, Kjellbom P, Neutze R. 2006. *Curr. Opin. Struct. Biol.* 16:447–56
42. Orgel JPRO. 2004. *J. Struct. Biol.* 148:51–65
43. Granseth E, von Heijne G, Elofsson A. 2005. *J. Mol. Biol.* 346:377–85
44. Liang J, Adamian L, Jackups RJ. 2005. *Trends Biochem. Sci.* 30:355–57
45. Schulz GE. 2000. *Curr. Opin. Struct. Biol.* 10:443–47
46. von Heijne G. 1992. *J. Mol. Biol.* 225:487–94
47. Sonnhammer ELL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6:175–82
48. Jones DT, Taylor WR, Thornton JM. 1994. *Biochemistry* 33:3038–49
49. Tusnady GE, Simon I. 1998. *J. Mol. Biol.* 283:489–506
50. Tusnady GE, Simon I. 2001. *Bioinformatics* 17:849–50
51. Melén K, Krogh A, von Heijne G. 2003. *J. Mol. Biol.* 327:735–44
52. Viklund H, Elofsson A. 2004. *Protein Sci.* 13:1908–17
53. Bernsel A, von Heijne G. 2005. *Protein Sci.* 14:1723–28
54. Käll L, Krogh A, Sonnhammer ELL. 2004. *J. Mol. Biol.* 338:1027–36
55. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. 1997. *Protein Eng.* 10:1–6
56. Dyrløv-Bendtsen J, Nielsen H, von Heijne G, Brunak S. 2004. *J. Mol. Biol.* 340:783–95
57. Nakai K, Horton P. 1999. *Trends Biochem. Sci.* 24:34–35
58. Martelli PL, Fariselli P, Krogh A, Casadio R. 2002. *Bioinformatics* 18(Suppl. 1):S46–53
59. Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ. 2004. *Nucleic Acids Res.* 32:W400–4
60. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B. 2004. *Nucleic Acids Res.* 32:2566–77
61. Chen CP, Kernytsky A, Rost B. 2002. *Protein Sci.* 11:2774–91
62. Chen CP, Rost B. 2002. *Protein Sci.* 11:2766–73
63. Käll L, Sonnhammer ELL. 2002. *FEBS Lett.* 532:415–18
64. Käll L, Krogh A, Sonnhammer ELL. 2005. *Bioinformatics* 21(Suppl. 1):i251–57

65. Daley DO, Rapp M, Granseth E, Melén K, Drew D, von Heijne G. 2005. *Science* 308:1321–23
66. Kim H, Österberg M, Melén K, von Heijne G. 2006. *Proc. Natl. Acad. Sci. USA* 103:11142–47
67. Bagos PG, Liakopoulos TD, Hamodrakas SJ. 2005. *BMC Bioinformatics* 6:7
68. Cuthbertson JM, Doyle DA, Sansom MS. 2005. *Protein Eng. Des. Sel.* 18:295–308
69. Tusnady GE, Dosztanyi Z, Simon I. 2005. *Bioinformatics* 21:1276–77
70. Lomize AL, Pogozheva ID, Mosberg HI. 2004. *Protein Sci.* 13:2600–12
71. Wallin E, von Heijne G. 1998. *Protein Sci.* 7:1029–38
72. Liu J, Rost B. 2001. *Protein Sci.* 10:1970–79
73. Gerstein M. 1998. *Proteins: Struct. Funct. Genet.* 33:518–34
74. Lehnert U, Xia Y, Royce TE, Goh CS, Liu Y, et al. 2004. *Q. Rev. Biophys.* 37:121–46
75. Walz T, Hirai T, Murata K, Heymann J, Mitsuoka K, et al. 1997. *Nature* 387:624–27
76. Doyle D, Cabral J, Pfuetzner R, Kuo A, Gulbis J, et al. 1998. *Science* 280:69–77
77. Viklund H, Granseth E, Elofsson A. 2006. *J. Mol. Biol.* 361:591–603
78. Lasso G, Antoniw JF, Mullins JGL. 2006. *Bioinformatics* 22:e290–97
79. Granseth E, Viklund H, Elofsson A. 2006. *Bioinformatics* 22:e191–96
80. Diederichs K, Freigang J, Umhau S, Zeth K, Breed J. 1998. *Protein Sci.* 7:2413–20
81. Adamian L, Liang J. 2006. *BMC Struct. Biol.* 6:13
82. von Heijne G. 1991. *J. Mol. Biol.* 218:499–503
83. Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU. 2004. *Proc. Natl. Acad. Sci. USA* 101:959–63
84. Baldwin JM. 1993. *EMBO J.* 12:1693–703
85. Baldwin JM, Schertler GF, Unger VM. 1997. *J. Mol. Biol.* 272:144–64
86. Unger VM, Hargrave PA, Baldwin JM, Schertler GF. 1997. *Nature* 389:203–8
87. Bradley P, Misura KM, Baker D. 2005. *Science* 309:1868–71
88. Jones TA, Thirup S. 1986. *EMBO J.* 5:819–22
89. Simons KT, Bonneau R, Ruczinski I, Baker D. 1999. *Proteins* 3(Suppl.):171–76
90. Bowie JU, Eisenberg D. 1994. *Proc. Natl. Acad. Sci. USA* 91:4436–40
91. Pellegrini-Calace M, Carotti A, Jones DT. 2003. *Proteins* 50:537–45
92. Yarov-Yarovoy V, Schonbrun J, Baker D. 2006. *Proteins* 62:1010–25
93. Fleishman SJ, Unger VM, Ben-Tal N. 2006. *Trends Biochem. Sci.* 31:106–13
94. Zhang Y, Devries ME, Skolnick J. 2006. *PLoS Comput. Biol.* 2:e13
95. Yarov-Yarovoy V, Baker D, Catterall WA. 2006. *Proc. Natl. Acad. Sci. USA* 103:7292–97
96. Fleishman SJ, Ben-Tal N. 2006. *Curr. Opin. Struct. Biol.* 16:496–504
97. Forrest LR, Tang CL, Honig B. 2006. *Biophys. J.* 91:508–17
98. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. *J. Mol. Biol.* 247:536–40
99. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. *Structure* 5:1093–108
100. Tusnady GE, Dosztanyi Z, Simon I. 2005. *Nucleic Acids Res.* 33:D275–78
101. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. 2004. *Nucleic Acids Res.* 32:D138–41
102. Hedman M, DeLoof H, von Heijne G, Elofsson A. 2002. *Protein Sci.* 11:652–58
103. Wistrand M, Käll L, Sonnhammer EL. 2006. *Protein Sci.* 15:509–21
104. Liu Y, Gerstein M, Engelman DM. 2004. *Proc. Natl. Acad. Sci. USA* 101:3495–97
105. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. 2002. *Nature* 415:141–47
106. Stenberg F, Chovanec P, Maslen SL, Robinson CV, Ilag LL, et al. 2005. *J. Biol. Chem.* 280:34409–19

107. Lasserre JP, Beyne E, Pyndiah S, Lapaillerie D, Claverol S, Bonneu M. 2006. *Electrophoresis* 27:3306–21

108. Xia Y, Lu LJ, Gerstein M. 2006. *J. Mol. Biol.* 357:339–49

109. Nilsson J, Persson B, von Heijne G. 2000. *FEBS Lett.* 486:267–69

110. Shimizu T, Mitsuke H, Noto K, Arai M. 2004. *J. Mol. Biol.* 339:1–15

111. Rapp M, Seppälä S, Granseth E, von Heijne G. 2006. *Nat. Struct. Mol. Biol.* 13:112–16

Annual Review of Biochemistry

Volume 76, 2007

Contents

Mitochondrial Theme

Indexes

Errata

An online log of corrections to *Annual Review of Biochemistry* chapters (if any, 1997 to the present) may be found at http://biochem.annualreviews.org/errata.shtml