PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2006) **361**, 465–475
doi:10.1098/rstb.2005.1804
Published online 3 February 2006

Towards genome-scale structure prediction for transmembrane proteins

Naama Hurwitz[1], **Marialuisa Pellegrini-Calace**[2] **and David T. Jones**[1,*]

[1] *Bioinformatics Unit, Department of Computer Science & Department of Biochemistry and Molecular Biology, Darwin Building, University College London, Gower Street, London WC1E 6BT, UK*
[2] *Department of Biochemical Sciences, University of Rome 'La Sapienza', P.le Aldo Moro, 5, 00185 Rome, Italy*

In this paper we briefly review some of the recent progress made by ourselves and others in developing methods for predicting the structures of transmembrane proteins from amino acid sequence. Transmembrane proteins are an important class of proteins involved in many diverse biological functions, many of which have great impact in terms of disease mechanism and drug discovery. Despite their biological importance, it has proven very difficult to solve the structures of these proteins by experimental techniques, and so there is a great deal of pressure to develop effective methods for predicting their structure. The methods we discuss range from methods for transmembrane topology prediction to new methods for low resolution folding simulations in a knowledge-based force field. This potential is designed to reproduce the properties of the lipid bilayer. Our eventual aim is to apply these methods in tandem so that useful three-dimensional models can be built for a large fraction of the transmembrane protein domains in whole proteomes.

Keywords: bioinformatics; transmembrane proteins; protein structure prediction; protein sequence analysis; protein structure

1. INTRODUCTION

A wide range of fundamental biological processes such as cell signalling, transport of membrane-impermeable molecules, cell–cell communication, cell recognition and cell adhesion are mediated by membrane proteins. Not surprisingly, therefore, understanding the structure and function of membrane proteins is of great importance in biological and pharmacological research.

Analysis of the complete genomic sequences for several organisms indicates that 20–25% of all genes code for transmembrane proteins (Jones 1998; Wallin & von Heijne 1998). Despite their large number and their importance only less than 1% of all three-dimensional protein structures deposited in the Protein Data Bank (PDB) are of membrane proteins (Berman *et al.* 2000), probably because they are not easy to crystallize and are hardly tractable by nuclear magnetic resonance (NMR). It appears therefore of particular importance to develop efficient theoretical structure prediction methods for transmembrane proteins.

2. BIOLOGICAL MEMBRANES

To understand how transmembrane protein structures can be predicted, it is important to understand the properties of biological membranes, which are composed of a lipid bilayer. Membranes serve to separate different compartments of the cell or the cell from its environment, and to achieve this, the lipid bilayer is impermeable to polar (soluble in water) molecules and ions.

A space-filling model of a lipid bilayer is shown in figure 1 (Heller *et al.* 1993). Each phospholipid is composed of a negatively charged phosphate group and two tails, which are two highly hydrophobic hydrocarbon chains. The hydrophobic effect ensures that the tails of the phospholipids in each layer orient towards each other creating a highly hydrophobic environment within the membrane. The charged phosphate groups face out into the hydrophilic environment.

3. MEMBRANE PROTEIN CLASSES: STRUCTURES AND FUNCTIONS

All of the active processes involving membranes are carried out by proteins within the membrane environment. These proteins are usually classified as being either peripheral (membrane associated proteins) or integral on the basis of how readily they can dissociate from the membrane. Peripheral membrane proteins are loosely associated with the membrane and usually interact with the polar head groups of the membrane phospholipids. These proteins can therefore be solubilized under relatively mild conditions, such as an environment of high ionic strength. Integral membrane proteins, on the other hand, are found to interact extensively with the hydrocarbon chains of the membrane lipids (figure 2) and can therefore be solubilized only by using detergents or an organic solvent. They are embedded in the phospholipid

[*] Author for correspondence (dtj@cs.ucl.ac.uk).

One contribution of 15 to a Discussion Meeting Issue 'Bioinformatics: from molecules to systems'.

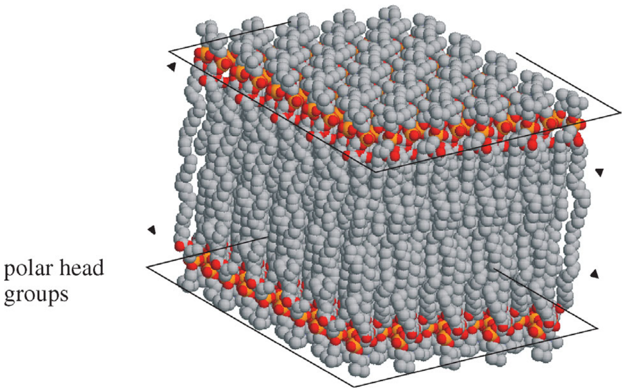

Figure 1. A space-filling model of a typical phospholipid bilayer.

bilayer, often membrane spanning, although some unilateral ones can be embedded in only one leaflet (see figure 2).

(a) *Integral membrane proteins*

The principle that underlies the structure and stability of membrane proteins is the high energetic cost of dehydrating the peptide bond during its transfer into a non-polar phase (White 2001). This has two consequences. First, and perhaps most obviously, most of the amino acid side chains found within transmembrane segments should be non-polar. Second, the polar groups of the polypeptide backbone of the transmembrane segments must participate in hydrogen bonds in order to lower the energetic cost. This second constraint is typically accomplished by exploiting two structural motifs: the membrane-spanning α-helix bundle and the β-barrel (White & Wimley 1999). In the α-helical structure, the peptide bonds are internally bonded with hydrocarbon bonds whereas β-strands form a closed structure termed the β-barrel.

(b) *β-Barrel integral membrane proteins*

The β-barrel proteins, also called porins, consist of β-strands spanning the membrane connected by short loops facing the periplasm and larger loops protruding outside the outer membrane (von Heijne 1996). The β-strands are amphiphilic, i.e. the side chains of the strand residues are alternately polar and hydrophobic with polar residues projecting into a central pore. Thus, the structure forms a pore with a polar environment. All β-barrel membrane proteins form oligomers (Seshadri *et al.* 1998).

The porins are found in the outer membrane of Gram-negative bacteria and in the outer membrane of chloroplasts and mitochondria. Their function is to facilitate diffusion of salts and polar compounds.

(c) *α-Helical integral membrane proteins*

In the α-helical proteins, the transmembrane segments are arranged in helices of 17–25 residues length and may cross the membrane once or several times (figure 3). Bi-topic proteins (or membrane-anchored proteins) are α-helical membrane proteins, which cross the membrane once (or sometimes twice), exposing globular domains on the extracellular and cytoplasmic surfaces. They typically act as cell surface markers, adhesion factors or receptors. The cytoplasmic domains often play a role in cell signalling (e.g. tyrosine kinases) or may connect to the cellular cytoskeleton. Polytopic (multi-spanning) α-helical membrane proteins have more than one α-helical transmembrane segment and the helices are arranged into a bundle (see figure 4). Possible driving forces for helix–helix association in the lipid bilayer are van der Waals interactions and inter-helical polar interactions, including hydrogen bonds and electrostatic interactions (Popot & Engelman 2000).

When polytopic α-helical membrane proteins are grouped according to their topology, differences between various species can be observed. In general, eubacteria, archaea, fungi and plants have large collections of membrane proteins built of 6 and 12 transmembrane segments, whereas in *Caenorhabditis elegans* and human proteins with seven transmembrane segments are preferred (Wallin & von Heijne 1998).

Perhaps the most biologically important example of polytopic proteins (at least with respect to pharmacology) is the superfamily of G-protein-coupled receptors (GPCRs), which includes receptors for hormones, neurotransmitters, growth factors, light and many other kinds of ligands (Dewji & Singer 1997). Other families of this superfamily function as channel and pore forming proteins involved in membrane transport (Singer 1990).

4. HELIX-BUNDLE INTEGRAL MEMBRANE PROTEIN FOLDING

The folding process of helix-bundle membrane proteins consists of two stages. The first stage involves formation of stable helices across the hydrophobic region of the membrane lipid bilayer. In the second stage, the helices interact to give a functional membrane protein (Popot & Engelman 1990). The assembly is carried out by a translocon apparatus involving the transient attachment of an active ribosome to a translocon embedded in the membrane. As soon as the protein is synthesized into the translocon and transferred into the membrane, the apparatus disassembles leaving the folded protein within the membrane (White & Wimley 1999).

5. EXPERIMENTAL STUDIES OF HELIX-BUNDLE INTEGRAL MEMBRANE; MEMBRANE HELIX LOCATIONS AND TOPOLOGY

As mentioned above, elucidation of the three-dimensional structure of membrane proteins is a difficult task. Therefore, other approaches for studying the structure of the proteins are being employed. One such approach is to determine the protein topology, i.e. the inside–outside location of the N and C termini relative to the membrane, and the number and positions of the membrane spanning regions. Knowing a protein's topology is a significant step towards determining both its structure and function. Topology assignments are also sometimes referred to as 'low resolution structures' (Kernytsky & Rost 2003), but there is clearly no actual three-dimensional information produced.

Several experimental approaches can be used to determine transmembrane topology:

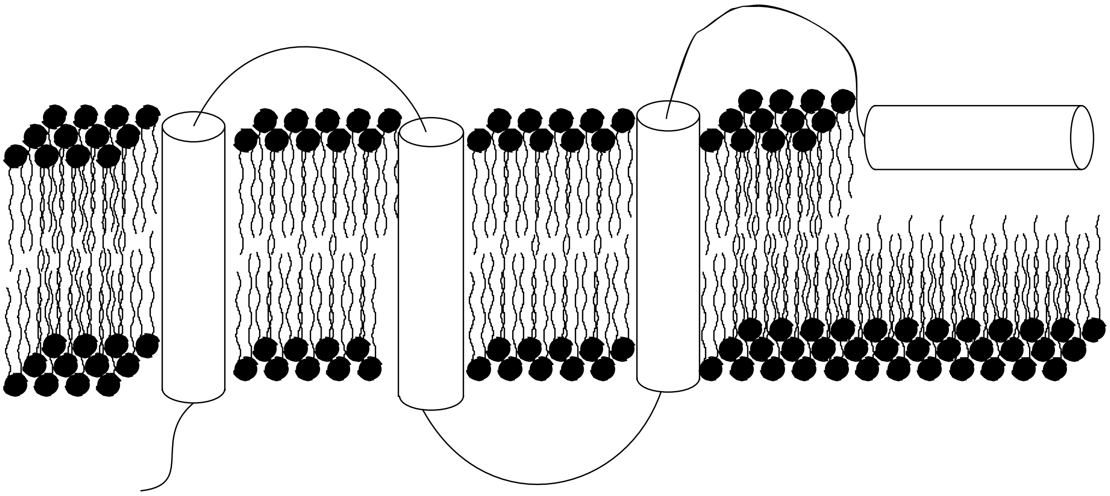

Figure 2. Diagrammatic representation of an integral membrane (transmembrane) protein. The first two helices fully span both leaves of the bilayer, but the third helix (typically an amphipathic helix) is shown not fully penetrating the bilayer.

(i) Fusion proteins: a protein segment that can be detected while it is translocated through the membrane is fused to the predicted loops of the tested protein (van Geest & Lolkema 2000).

(ii) Proteolytic digestion *in situ*: proteolytic enzymes can be used to cut the loops outside the membrane. It is then possible to run the segments remaining in the membrane by SDS-PAGE (Kuroiwa *et al.* 1996).

(iii) Antibody binding: antibodies specific to the loops are used to locate the loops outside the membrane (Amstutz *et al.* 2001).

6. PREDICTING TRANSMEMBRANE HELIX-BUNDLE PROTEIN TOPOLOGY

Given the amount of information that can be potentially obtained from topology assignment and the relative difficulty in obtaining this information experimentally, it is not surprising that a great deal of attention has been paid to prediction of transmembrane topology from sequence.

Transmembrane protein topology prediction methods rely on two major topological features. The first is that, as already discussed, transmembrane helices are ultimately formed by hydrophobic stretches, the second is primarily the bias towards positively charged residues in the regions flanking the hydrophobic stretches, especially on the intracellular side of the membrane. The feature is commonly known as 'the positive-inside rule' where short loops are found to be enriched with Lys and Arg residues on the intracellular side and depleted on the outside (Wallin & von Heijne 1998; von Heijne 1999).

More than 30 methods have been developed for predicting the topology of helix-bundle membrane proteins (Kernytsky & Rost 2003) and here there is only space to mention a few of them. Below is a brief summary of the main representative methods developed and the advances made over the last two decades.

In the earliest transmembrane prediction methods, simple hydrophobicity scales were used (e.g. Kyte & Doolittle 1982) to detect probable transmembrane segments. These scales classify amino acids according to their preference to be found in polar or non-polar environments. Thus, a high hydrophobicity value indicates a preference for a non-polar environment, i.e. the lipid bilayer. Kyte & Doolittle used a 'sliding window' approach to identify membrane segments where a fixed window of width 19 residues is moved along the protein sequence and the sum or average hydrophobicity is calculated for amino acids within the window. Using these mean hydrophobicity values, a threshold can be identified for deciding whether the centre of the window is within a membrane spanning membrane helix or not. The Kyte & Doolittle method, along with other similar approaches, predict the occurrence of transmembrane segments only and were not designed to predict the inside–outside location of the segments relative to the membrane.

The first major advance in transmembrane topology prediction was the TopPred method proposed by von Heijne (1992). TopPred still made use of hydrophobicity scales and a sliding window to predict transmembrane segments, but combined these predictions with a simple topological rule: the so-called 'positive-inside rule'. The observation that there was a strong bias for positively charged residues on the inside facing segments of a transmembrane protein provided a means for identifying which predicted topology is correct from a small number of alternatives. Even though the starting point to TopPred was a basic hydrophobicity plot, it was nonetheless the first transmembrane topology prediction method.

The MEMSAT method of Jones *et al.* (1994) was the first prediction method to fully integrate the prediction of transmembrane topology with the prediction of transmembrane segments. Rather than simply deciding between a few possible topological models, MEMSAT was able to calculate the most probable length, location and topological orientation for each transmembrane segment, guaranteeing a mathematically optimal solution. The method made use of statistical tables (log likelihood ratios) compiled from membrane protein data and a dynamic programming algorithm to search through

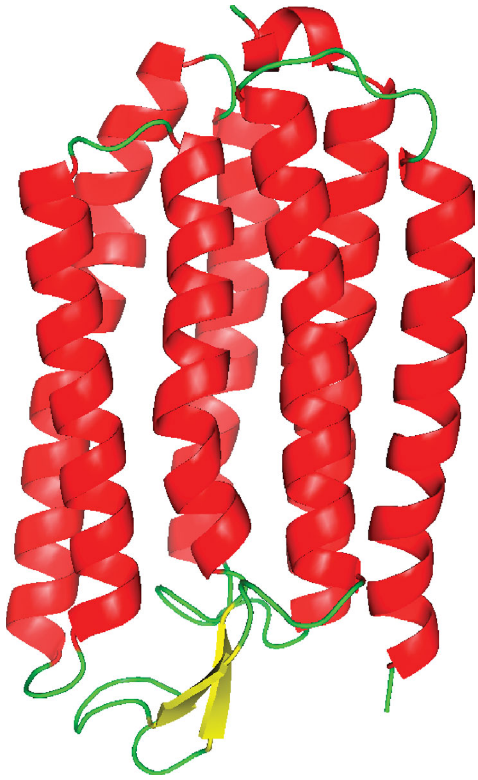

Figure 3. An example of an α-helical bundle integral membrane protein (halorhodopsin from *Halobacterium salinarum*).

all possible topological models by a process of expectation maximization. The propensity of each amino acid to be in one of five states (inside loop, outside loop, inside helix end, helix middle and outside helix end) was calculated from experimentally well-described membrane proteins and was represented as a log-likelihood ratio. This approach can clearly be seen as a forerunner of more recent approaches which are based on formal Hidden Markov Models.

PHDhtm (Rost *et al.* 1996) was the first method to use neural networks for prediction of transmembrane helix topology. It used multiple sequence alignments to do a consensus prediction of the target protein, and then predicted topology by using a neural network trained on proteins with experimentally characterized topologies.

TMHMM (Sonnhammer *et al.* 1998) and HMMTOP (Tusnady & Simon 1998) are methods which are both based on Hidden Markov models. TMHMM implements a cyclic model with seven states for transmembrane helix, whereas HMMTOP uses a Hidden Markov model which distinguishes between five structural states (helix core, inside loop, outside loop, helix caps (C and N) and globular domains). The states are connected by transition probabilities. As with the earlier MEMSAT approach, dynamic programming is used to match a sequence against the model in order to find the most probable match.

DAS-Tmfilter (Cserzo *et al.* 2004) is based on computing a dot plot between the query protein and a library of known transmembrane proteins. The result is a hydrophobicity profile.

Finally, in line with recent developments in methods for predicting globular protein structure, consensus methods have started to appear. The first such approach has been developed by Nilsson *et al.* (2002), and uses the consensus of five topology prediction methods (TMHMM, HMMTOP, MEMSAT, PHD, TopPred). They find that approximately 90% of partial consensus topologies are correctly predicted in membrane proteins from both prokaryotic and eukaryotic organisms, which is a higher accuracy than can be achieved by any single component method. They further go on to show that a consensus topology can be predicted for 70% of all membrane proteins in a bacterial genome and for *ca* 55% of all membrane proteins in eukaryotic genome. These accuracy estimates were surprisingly low compared to some of the values quoted in the original method papers, which was further confirmed by Melen *et al.* (2003). One possible reason for this is the low reliability of experimentally determined topology information available when the first prediction methods were developed. Another possibility is that earlier accuracy estimates were biased by the lack of proteins with unusual three-dimensional structures in the testing sets used. As more high resolution structures are determined for transmembrane proteins, more accurate benchmarking of methods should be possible and hopefully this will stimulate further developments in the field.

7. PREDICTING TERTIARY STRUCTURE OF HELIX-BUNDLE MEMBRANE PROTEINS

At present there is no general-purpose method for three-dimensional structure prediction for transmembrane proteins, though of course the same can be said for globular proteins. As with globular proteins, the most reliable method for deriving a three-dimensional model for a protein is that of comparative modelling. Unfortunately, the reputation of comparative modelling for membrane proteins suffered slightly from the efforts made in the early 1990s to model proteins in the seven-helix GPCR superfamily based on templates derived from bacteriorhodopsin. The subsequent crystal structure of rhodopsin showed that despite similarities in the overall topology and approximate positioning of the helices, the structure of bacteriorhodopsin was substantially different in terms of the helix packing arrangements. The very remote sequence similarity observed between the GPCR sequences and bacteriorhodopsin also contributed to the inaccuracy of the models which were built, and so in hindsight the attempts to build useful models of the GPCRs by comparative modelling must be seen as failures.

Even ignoring the difficulties in identifying remote evolutionary relationships between membrane proteins and the difficulties in arriving at accurate alignments, a much more fundamental problem is the lack of structural templates available. Given the difficulties in experimentally determining structures for integral membrane proteins, it is not surprising that so few structures are currently available in the structure databases. At the time of writing, there are 95 sequence-unique membrane proteins in the current

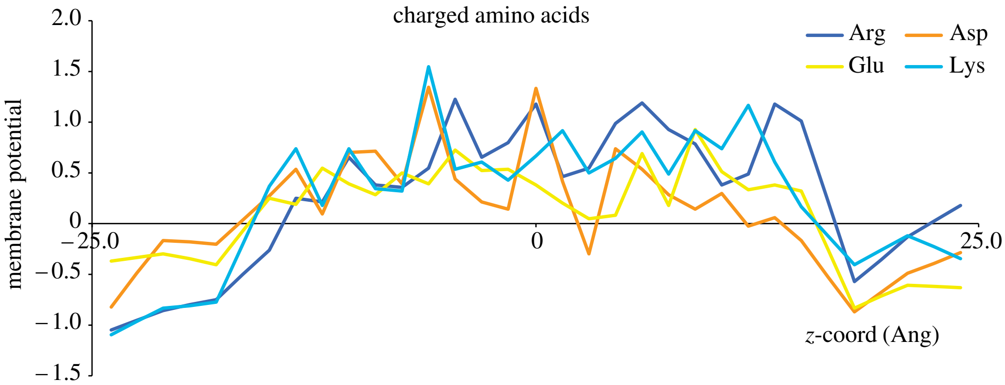

Figure 4. A plot of the potentials of the bilayer potentials (in units of kcal mol^{-1}) for the charged amino acids.

release of PDB (Berman *et al.* 2000), compared to 20 000 or so globular proteins. The chance of finding a suitable template structure for transmembrane proteins is consequently very small in comparison to that of globular proteins, though a little higher than might otherwise be expected due to the dominance of very large families such as the GPCRs.

In view of the difficulties in applying comparative modelling to transmembrane proteins, a number of groups have looked into the problem of structure prediction for these proteins without the requirement for a template structure, and several studies have been published on analysing the important structural features of helix-bundle membrane proteins that can be used for the prediction of their structure.

As discussed earlier, topology prediction methods give some spatial information, but as a first step towards full three-dimensional modelling of proteins with multiple transmembrane segments, information on the orientations of each transmembrane helix is required. After topology, the next most readily predicted feature of each transmembrane helix is the relative orientations of the helix around its own axis, i.e. the identification of which residues are exposed to the lipid phase and which are packed against the interior of the transmembrane bundle.

Early attempts were made to predict relative helix orientation by using the concept of the hydrophobic moment (Eisenberg *et al.* 1984; Rees *et al.* 1989). The hydrophobic moment is essentially a vector pointing from the helix axis to the most hydrophobic surface of the helix. In these methods, the angular orientations of transmembrane helices could be predicted by assuming that the helical hydrophobic moments should point out into the lipid phase. Later, however, it was found that hydrophobic moments are poor indicators of the angular orientation of the transmembrane helices due to the fact that hydrophobic residues often face both the core of the protein and the lipid (Stevens & Arkin 1999; Rees & Eisenberg 2000).

In later work, a statistical analysis was conducted on known high-resolution structures of integral membrane proteins in order to find the lipid exposure propensities of the different residues (Cronet *et al.* 1993; Donnelly *et al.* 1993). The work of Donnelly was particularly important in that it described very clearly the importance of sequence conservation in discriminating between lipid exposed and buried residues. Lipid exposed residues, while needing to be highly hydrophobic are also under no significant steric constraints and so are often seen to be evolutionarily quite variable. Buried residues, on the other hand, while also being typically hydrophobic are also subject to steric constraints and so are commonly seen to be highly conserved in sequence alignments. As an alternative to the evolutionary approach, Pilpel *et al.* (1999) proposed a knowledge-based scale for the propensity of residue orientation in the transmembrane segments. The authors made the assumption that residues which tend to be exposed to the membrane will be more frequent in the transmembrane segments of single spanning transmembrane proteins than in multi-spanning proteins, whereas residues that prefer to be buried in the transmembrane bundle interior would show the opposite trend.

Using this kind of knowledge, some attempts have been made to develop prediction methods for membrane protein three-dimensional structure. Taylor *et al.* (1994) adapted some programs originally developed for the prediction of globular protein structures to derive a method for predicting integral membrane protein structures. The method uses the ‘variphobicity’ (evolutionarily variable and hydrophobic) faces of transmembrane helices to predict the structure and was successfully applied to two protein family sequence alignments (bacteriorhodopsin and rhodopsin).

Nikiforovich *et al.* (2001) developed a modelling approach which was a combination of helical packing, based on the bacteriorhodopsin template, and selection of low-energy conformers for loops that are closest to the known X-ray structure of bacteriorhodopsin. Using this method, the authors were able to accurately reproduce the bacteriorhodopsin structure.

Fleishman & Ben-Tal (2002) used data on residue environment preferences to predict the likely arrangement of transmembrane helices, and this method was used to predict successfully the native structure of transmembrane protein glycophorin A. In the same year, Ledesma *et al.* (2002) produced a model for uncoupling protein 1 (UCP1), using a computational docking method, and in 2003, Chen & Chen used a Monte Carlo method for protein folding and successfully predicted the seven helix bundle structure of rhodopsin I.

Pellegrini-Calace *et al.* (2003) developed a method (FILM) for predicting small membrane protein structures based on a method previously developed for predicting tertiary structure of globular proteins. The method is based on the assembly of super-secondary structural fragments taken from a library of proteins with known structure, using a standard simulated annealing algorithm. The method was applied to small membrane proteins of known structure and was able to predict at reasonable accuracy level the helix topology and the conformation of a number of these proteins. We give a brief outline here of the FILM method.

8. FILM: MEMBRANE POTENTIAL DEFINITION

At the core of the FILM method was a simple approach to modelling the physicochemical constraints of the lipid bilayer. Globular protein structure prediction is simplified by the assumption that the water which surrounds the folding protein is essentially isotropic. Indeed, many successful globular protein prediction methods treat the effects of the solvent as an implicit term, e.g. by altering the dielectric constant in an application of Coulomb's Law.

The lipid bilayer (see figure 1) clearly cannot be treated as an isotropic environment, and so some means is needed to model its effects in a realistic way. Although in theory the lipid bilayer could be incorporated in atomic detail in a molecular dynamic simulation (e.g. Heller *et al.* 1993), this would require far too much computational power, and so FILM makes use of a very simplistic model of the bilayer properties in the form of membrane potentials.

Membrane potentials (MPs) were defined by a statistical analysis carried out on a set of 640 transmembrane helices, belonging to 133 membrane proteins extracted from the SWISSPROT database (Bairoch & Apweiler 1996) and having an experimentally defined topology at the very least. We would have preferred to limit our analysis to transmembrane proteins of known three-dimensional structure, but there are currently insufficient structures to allow this.

The membrane bilayer was modelled as an infinite slab 60 Å in thickness made of a 30 Å core and a 15 Å interface at both periplasmic and cytoplasmic sides. The z-axis was taken as the direction perpendicular to the Cartesian plane formed by the bilayer surface. A simplifying assumption here is that we assume that membrane physicochemical properties are variable only along the z-axis and are constant across each plane parallel to the membrane surface. The core region was then divided along the z-axis into 21 'layers' with a layer thickness of 1.5 Å (corresponding to the translation per residue of a right-handed α-helix), with $z=0$ being right at the centre of the membrane. Arbitrarily, we defined the cytoplasmic direction as being in the negative z-direction.

The relative frequencies of occurrence for each naturally occurring amino acid within each layer of the core region were calculated by analysing the sequences of the 640 transmembrane helices, giving a total dataset of 17 162 residues. The midpoint of each helix was considered to be positioned at the middle of the membrane ($z=0$), and sequence positions $(0, \dots, n)$ were transformed into z-coordinates (z) as follows:

$$z = \frac{c(i-\frac{1}{2}h)}{h},$$

where i is the position number, h is the helix length and c is the membrane core thickness.

Periplasmic and cytoplasmic interfaces were divided into three layers each, with a larger step size of 3 Å, as we assume that the backbone is more likely to be extended at that level (the observed span per α-carbon in an extended polypeptide is *ca* 3.6 Å). The three amino acids leading into both ends of the transmembrane helices were considered in calculating the frequencies of occurrence within the interface layers (sequence positions again have been transformed as above).

Finally, the inverse Boltzmann equation was applied so that, given an amino acid type a at a specified z-coordinate (z) inside of the slab, its membrane propensity (MP) was calculated according the equation:

$$\mathrm{MP}(z) = -RT \ln\frac{f_{\mathrm{a}}(z)}{f(z)},$$

where $f_{\mathrm{a}}(z)$ is the observed relative frequency of occurrence of amino acid type a at z, $f(z)$ is the observed relative frequency of occurrence of all amino acids found at z, and RT is taken to be 0.582 kcal mol^{-1}.

As an example, figure 4 shows the calculated bilayer potentials for the charged amino acids. The main features of the potentials concur with what is expected from our knowledge of amino acids and their occurrence in transmembrane segments. The well-known positive inside rule (von Heijne 1992) shows itself in the clear preference for the positively charged residues to be located at z-coordinates between -25 and -18, which correspond to the cytoplasmic facing helix cap regions. Other trends, such as the preference for aromatic residues to be located close to the polar head groups, also show up in their respective potentials.

9. FILM: PREDICTION METHOD

(a) *Energy function*

In addition to the membrane potentials, further terms are needed in the complete potential function. FILM was based on the FRAGFOLD method used for globular protein folding (Jones 1997) and so the core of the used objective function is a set of pairwise potentials of mean force representing both short range and long range interresidue interactions. These potentials were determined by a statistical analysis of highly resolved protein X-ray crystal structures and again the application of the inverse Boltzmann equation, along with a solvation potential (E_{solv}).

In FILM, the solvation energy is set to zero at the membrane core and smoothed at interfaces by a factor SF as follows:

$$\mathrm{SF} = \frac{|z|-L}{U-L},$$

where z is any interface z-coordinates (i.e. 15 Å $<|z|<$ 30 Å), L is the positive lowest interface z-coordinate (15 Å) and U is the positive higher one (30 Å). In this way, the traditional 'water' solvation potential has no effect within the bilayer, but does have an effect outside of the membrane environment.

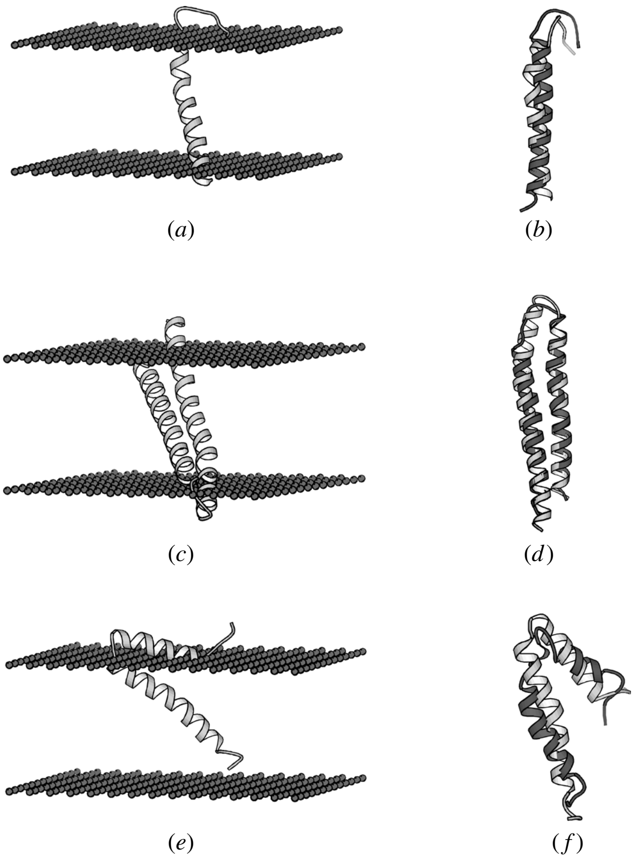

Figure 5. (*a*) FILM model for glycophorin A (predicted transmembrane helix from T93 to I118); (*b*) superposition of FILM model with NMR model (RMSD = 3.6 Å); (*c*) FILM model of subunit C of the F1Fo ATPase (predicted transmembrane helices from E2 to R41 and from L48 to A77); (*d*) superposition of FILM model with NMR model (RMSD = 4.2 Å); (*e*) FILM model of major fd coat protein (predicted transmembrane helix from W49 to T69 and predicted amphipathic helix from A32 to A41); and (*f*) superposition of FILM model with NMR model (RMSD = 4.8 Å).

All other terms in the FILM potential function (e.g. hydrogen bonding and steric terms) are the same as currently used by FRAGFOLD. The total potential value is obtained by summing the membrane potentials (E_{mem}) and then added to the other potential terms with an appropriate weight, e.g.

$$E_{\mathrm{tot}} = aE_{\text{short-range}} + bE_{\text{long-range}} + cE_{\mathrm{solv}} + dE_{\mathrm{mem}} + eE_{\mathrm{steric}} + fE_{\mathrm{hbond}},$$

a–f are the adjustable weights.

FILM was tested on a number of small transmembrane proteins. The requirement that targets be relatively small and have both known three-dimensional structure and known transmembrane topology limits the number of available targets to a small handful. Figure 5 shows the results of applying FILM to three such targets (glycophorin A, subunit C of the F1Fo ATPase and major fd coat protein). In all cases, FILM produced a final model with the correct topology and a reasonable approximation of the chain conformation (RMSDs between 3.6 and 4.8 Å). Unfortunately, there are almost no other targets available on which the method could be tested, and the lack of suitable small transmembrane targets is a critical restriction in testing and developing methods for folding transmembrane proteins. Although FILM appears to be a useful method, further benchmarking will be needed on newly solved targets before its performance can be accurately assessed.

10. FURTHER DEVELOPMENT OF THE FILM METHOD

Although the FILM method appears to be effective in predicting the fold and topological arrangement of small transmembrane proteins, the method was not able to predict the conformations of larger transmembrane proteins. The main limitation of FILM is that the potential function is not able to reproduce the compactness of transmembrane bundles. Transmembrane helix bundles are usually not optimally compact, although neighbouring helices are closely packed. Bacteriorhodopsin, for example, is an elongated two-layer bundle of seven helices and not a compact bundle with a circular cross-section. We have been trying to improve the prediction of these large bundles by incorporating the prediction of lipid exposure from variphobicity analysis into the FILM potential function.

In order to develop better potential functions for folding larger multihelical transmembrane proteins, we have recently started making use of a set of decoy structures. Decoy sets have been widely used to validate and optimize potential functions for folding globular proteins (e.g. Park & Levitt 1996), and a number of globular protein decoy sets can be downloaded from the Decoys ‘R’ US web site (http://dd.stanford.edu/). To date, however, no decoy sets have been developed for evaluating potential functions designed to fold transmembrane proteins.

We employed the polyhedral modelling approach of Taylor *et al.* (1994) to make a set of challenging decoy sets for transmembrane protein scoring functions. The approach is briefly summarized in figure 6. Firstly, for a given topology (number of transmembrane helices and direction with respect to inside/outside phasing) all possible windings are generated on a hexagonal close-packed lattice. Windings which violate topological rules (e.g. short crossing loops) are then eliminated and the remaining windings are transformed into regularized three-dimensional coordinates firstly by generating a ‘stick model’ of the structure where each helix is represented by a single vector. These vectors are rotated slightly to allow for helix packing and then an alpha-carbon trace is generated by alternating rounds of distance geometry and real space refinement.

Figure 7 shows the results of applying one of our simplified scoring functions to the decoys generated for bacteriorhodopsin and rhodopsin. The scoring function is the same as previously described for the FILM method, with the addition of a weighted term for variphobicity as defined by Taylor *et al.* (1994). As can be seen, this simple scoring function is reasonably effective in assessing the quality of the transmembrane models. In the case of bacteriorhodopsin, the best generated model (6 Å RMSD from native) has the lowest pseudo-energy sum out of all the decoys. One slight problem is that the model in fact scores slightly

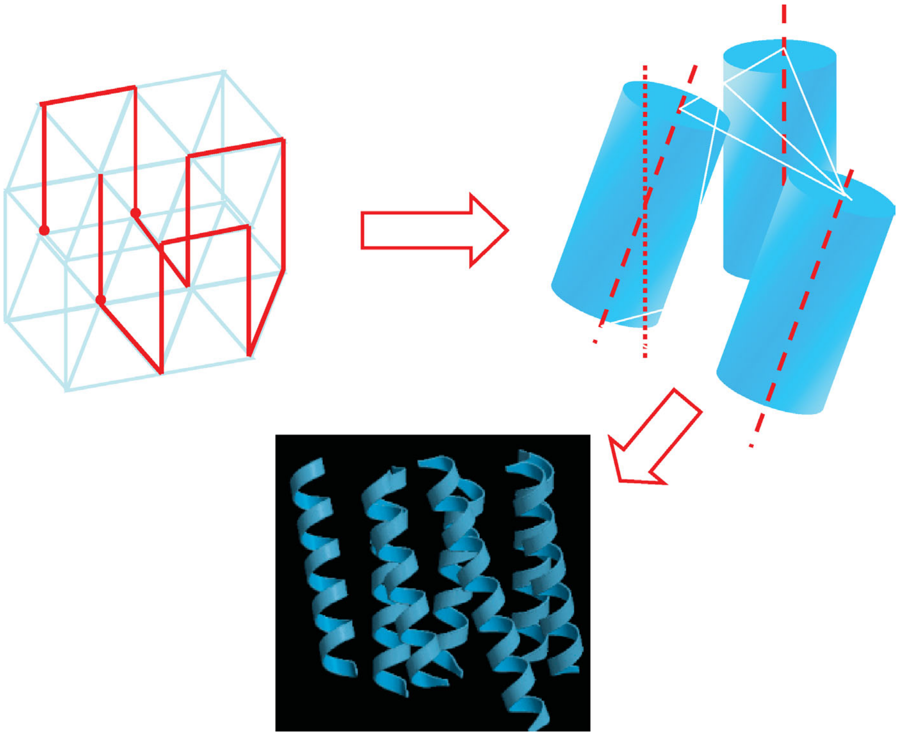

Figure 6. An outline of the decoy generation procedure using a polyhedral model of transmembrane helix bundles (Taylor *et al.* 1994). In the first step, different tracings are generated through a hexagonal close packed lattice (packed cylinder model). Tracings that violate obvious structural constraints (e.g. loops too short to make a particular connection) are eliminated. Given a tracing through the lattice, helix axes are generated taking into account the normal packing angles between helices in close packed helix bundles. Finally, using these axes, alpha-carbon coordinates are generated by means of distance geometry and real-space refinement.

better than the crystal structure in this case, which underlines the fact that the scoring function is still not fully accounting for all aspects of protein stability in a bilayer. In the case of rhodopsin, the crystal structure enjoys a much more pronounced gap between its own energy and that of the next closest decoy structure. However, in this case, the best model is not the model which produces the lowest pseudo-energy sum, though the lowest energy model is close to the native (7 Å RMSD) and the model with lowest RMSD is the next best.

11. CLASSIFYING TRANSMEMBRANE PROTEINS USING PREDICTED STRUCTURAL FEATURES

Over the past 15 years there has been a great deal of progress in the classification of proteins both by sequence (Sonnhammer *et al.* 1997; Mulder 2005) and by structure (Murzin *et al.* 1995; Orengo *et al.* 1997). In both cases, improvements have been driven by the rapid explosion in sizes of both sequence and structure data banks, and these improvements have led to better and more comprehensive annotations of genome sequences.

With the limited available structural data for transmembrane proteins and the large fraction of transmembrane families for which there is little or no functional information, it is not surprising that annotations for membrane-associated genome sequences are relatively sparse. Automatic sequence clustering techniques have a particular problem with transmembrane proteins because of the sequence constraints on the transmembrane helices. Put simply, one transmembrane protein sequence looks very similar to many others. Nevertheless, basic sequence analysis techniques are a good starting point. Liu & Engelman (2002) have classified polytopic membrane

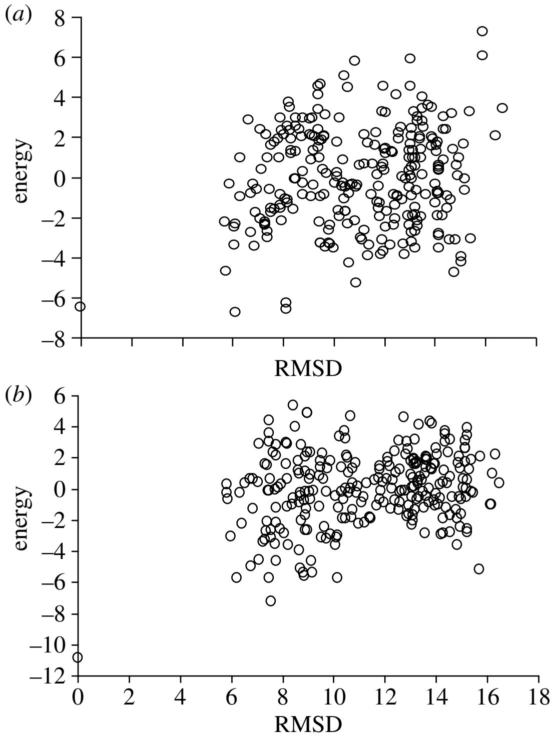

Figure 7. Validation results for FILM2 potential function applied to decoy sets based on (*a*) bacteriorhodopsin and (*b*) rhodopsin.

Table 1. Calculated transmembrane helix variphobicity scores for the five bacteriorhodopsin-like sequences, the seven opsin sequences and a sphingosine-1-phosphate receptor.

	TM1	TM2	TM3	TM4	TM5	TM6	TM7
BACA_HALS1	0.42	0.32	0.02	0.31	0.35	0.23	0.31
BACH_HALSP	0.20	0.24	−0.01	0.30	0.37	0.28	0.20
BACH_HALSS	0.28	0.24	−0.01	0.30	0.33	0.28	0.20
BACR_HALHA	0.37	0.33	0.04	0.30	0.40	0.25	0.29
BACS_HALHA	0.21	0.22	0.05	0.27	0.32	0.19	0.19
EDG1_HUMAN	0.29	0.01	0.30	0.64	0.67	0.43	0.22
OPS1_CALVI	0.72	0.05	0.18	0.46	0.06	0.50	−0.12
OPS2_DROME	0.70	0.09	0.22	0.40	0.14	0.53	−0.11
OPS3_DROME	0.68	0.12	0.19	0.17	0.49	0.50	0.08
OPS4_DROME	0.57	0.06	0.06	0.32	0.48	0.51	0.00
OPSB_HUMAN	0.61	0.06	0.19	0.31	0.27	0.43	0.03
OPSD_BOVIN	0.83	0.10	0.20	0.23	0.36	0.67	−0.07
OPSG_HUMAN	0.67	0.04	0.28	0.30	0.54	0.52	−0.12
OPSR_HUMAN	0.66	0.04	0.27	0.27	0.54	0.54	−0.05

proteins in 26 genomes, according to their number of transmembrane helices and sequence similarities, into 637 families. Classification based on similarity of amino acid sequences can be very informative when very significant sequence similarity exists; however, it is known that the impact of amino-acid sequence similarity on protein evolutionary or functional relationships is rather limited in the cases when the similarity between sequences is low. Moreover, we know from globular protein domains that sequence comparisons fail to identify many of the relationships that emerge from the comparison of protein structures.

With little prospect of vast numbers of transmembrane structures being solved experimentally over the next few years, we have to focus on *predicted* structural features to improve existing transmembrane classifications. The aim of some of our recent work on transmembrane protein sequence analysis has therefore been to classify proteins according to their predicted structural features, i.e. residue and helix orientation, as opposed to classifying them by sequence similarity alone. Examining predicted transmembrane topology is one very crude way in which uncharacterized transmembrane proteins can be grouped into meaningful families using structural clues (e.g. Jones 1997). However, we are already very aware that transmembrane proteins of identical topology can have dissimilar folds (e.g. the case of bacteriorhopsin and rhodopsin), and should therefore not be clustered into the same superfamily.

The aim of some of our recent work on transmembrane protein sequence analysis has been to classify proteins using a greater variety of predicted structural features. For example, by classifying membrane proteins according to their packing and orientation of their helices we hope to be able to improve on the classifications carried out to date, which have been based just on sequence and topology.

One powerful structural feature we have been examining is again the variphobicity signal which can be deduced from a multiple sequence alignment, and this has been found to provide a strong signal which can be used to correctly segregate transmembrane sequence clusters that would otherwise be grouped together.

As an example of using variphobicity analysis to distinguish topologically similar families of transmembrane proteins, we set up a simple experiment to see whether we can correctly cluster a mixture of bacteriorhodopsin and rhodopsin-like protein sequences using variphobicity alone. Of course, a great deal of work remains to be done in constructing an optimal clustering method based around variphobicity scoring, but we show here an example which is intended to serve as a minimal proof of principle.

As a first step, the MEMSAT method (Jones *et al.* 1994) was used to predict the topology of each protein along with the transmembrane segment locations and lengths. As expected, all of the proteins were predicted to have the same overall topology (seven transmembrane segments with the N-terminus on the outside). For each of the transmembrane segments, we calculated a measure of variphobicity (Taylor *et al.* 1994). As discussed by Taylor *et al.*, transmembrane segments having a high variphobicity score suggests that the helix has a high lipid exposure, whereas a low score suggests that the helix is not likely to be exposed to lipid (and is consequently like to be packed into the core of the protein).

Table 1 shows the resulting variphobicity scores for the seven transmembrane segments in each of the target proteins. Five of the proteins (BAC1_HALS1, BACH_HALSP, BACH_HALSS, BACR_HALHA, BACS_HALHA) come from the bacteriorhodopsin family, eight from the family of opsins and one from a distant member of the GPCR superfamily (sphingosine 1-phosphate receptor; EDG1_HUMAN). For each pair of proteins we derive a metric of similarity by calculating the (Pearson) correlation coefficient for the variphobicity scores. Our expectation is that proteins with the same approximate ranking of variphobicity across their transmembrane helices are more likely to share a common folding pattern. A distance matrix was formed by subtracting each pairwise correlation coefficient from 2.0, and then weighted pair group clustering was used to cluster the proteins. Figure 8 shows the results of this clustering in the form of a dendrogram.

The results obtained are very close to the results that would be obtained from clustering these proteins by

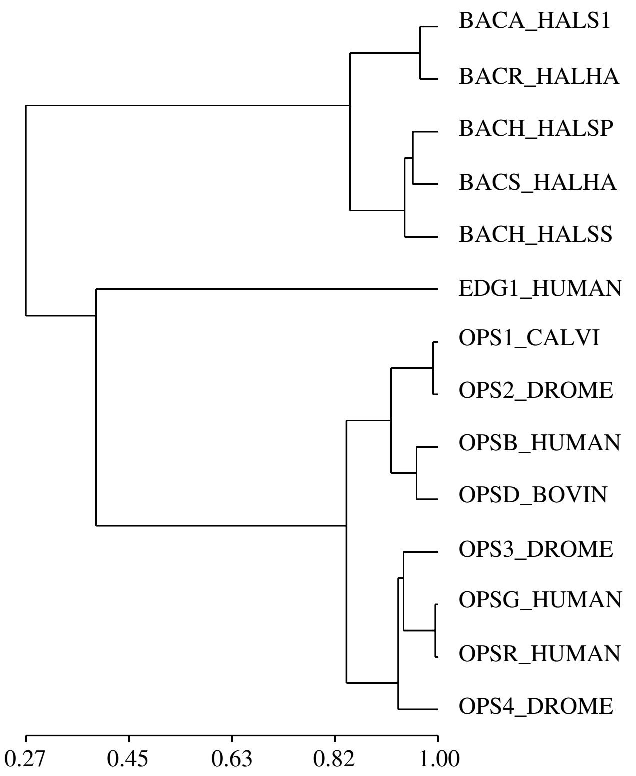

Figure 8. Dendrogram of the sequences shown in table 1 based solely on the pairwise correlation coefficients of the variphobicity scores.

sequence similarity. However, in this case, the proteins are represented merely by a vector of seven variphobicity values and the similarity by a crude calculation of correlation. Based just on these variphobicity patterns, the proteins form three distinct clusters (with EDG1_HUMAN forming a singleton group) suggesting that the relative variphobicity of the transmembrane segments is indeed a useful 'fingerprint' for a structurally similar family of transmembrane proteins.

We expect that a robust method for clustering transmembrane proteins of unknown structure will need to take into account a number of different features. For example, the lengths and physicochemical characteristics of connecting loops might be very informative. With the right combination of features it should be possible to get a clearer picture of the as yet uncharacterized transmembrane proteins found in completed genomes than can be obtained by using sequence similarity alone.

12. CONCLUSIONS

Despite a significant amount of progress in recent years in the prediction of globular protein structure from amino acid sequence, particularly in the areas of fold recognition and distant homology modelling, technology for the prediction of transmembrane protein structure is clearly lagging some way behind. In this paper we have described a number of recently developed methods for transmembrane protein structure prediction which we hope will eventually form parts of a 'pipeline' for automatically building structural models for all of the transmembrane proteins in a genome.

The main bottleneck to progress in improving transmembrane modelling methods is clearly the lack of experimentally determined structures for integral membrane proteins. This not only limits our ability to calculate reliable statistics for our knowledge-based approaches, but also limits our ability to test them. A common feature of all the methods discussed in this paper is that they have been designed to extract maximum value from what little experimental data is available. Were there to be a lot more available structural data then we would have made different decisions in the design of our algorithms.

Although there has clearly been some progress in increasing the efficiency of structure determination for membrane proteins, it is apparent that the lack of data will be the limiting factor for some considerable time to come. However, we prefer to look on this as a challenge for bioinformatics rather than an explanation for failure. The true value of bioinformatics has always been in maximizing the potential for exploitation of limited amounts of data, and the structural characterization of membrane proteins has to be seen as a perfect example of doing just this. We hope, therefore, that this paper will at the very least serve as a rallying call to both the experimental and bioinformatics community to develop imaginative new approaches to predicting the structure of transmembrane proteins which effectively and efficiently combine theory with experimental data.

REFERENCES

Amstutz, P. *et al.* 2001 *In vitro* display technologies: novel developments and applications. *Curr. Opin. Biotechnol.* **12**, 400–405. (doi:10.1016/S0958-1669(00)00234-2)

Bairoch, A. & Apweiler, R. 1996 The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Res.* **24**, 21–25. (doi:10.1093/nar/24.1.21)

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. 2000 The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242. (doi:10.1093/nar/28.1.235)

Chen, C. M. & Chen, C. C. 2003 Computer simulations of membrane protein folding: structure and dynamics. *Biophys. J.* **84**, 1902–1908.

Cronet, P., Sander, C. & Vriend, G. 1993 Modeling the transmembrane seven helix bundles. *Protein Eng.* **6**, 59–64.

Cserzo, M., Eisenhaber, F., Eisenhaber, B. & Simon, I. 2004 TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics* **20**, 136–137. (doi:10.1093/bioinformatics/btg394)

Dewji, N. & Singer, S. J. 1997 The seven-transmembrane spanning topography of the Alzheimer disease-related presenilin proteins in the plasma membranes of cultured cells. *Proc. Natl Acad. Sci. USA* **94**, 14 025–14 030. (doi:10.1073/pnas.94.25.14025)

Donnelly, D., Overington, J. P., Ruffle, S. V., Nugent, J. & Blundel, T. L. 1993 Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid facing residues. *Protein Sci.* **2**, 55–70.

Eisenberg, D., Weiss, R. M. & Terwillger, T. 1984 The hydrophobic moment detects periodicity in the protein hydrophobicity. *Proc. Natl Acad. Sci. USA* **81**, 140–144.

Fleishman, S. J. & Ben-Tal, N. 2002 A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J. Mol. Biol.* **321**, 363–378. (doi:10.1016/S0022-2836(02)00590-9)

Heller, H., Schaefer, M. & Schulten, K. 1993 Molecular dynamics simulation of a bilayer of 200 lipids in the gel and in the liquid-crystal phases. *J. Phys. Chem.* **97**, 8343–8360. (doi:10.1021/j100133a034)

Jones, D. T. 1997 Successful *ab initio* prediction of the tertiary structure of NK-Lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins*-Suppl. 1, 185–191. (doi:10.1002/(SICI)1097-0134(1997)1+<185::AID-PROT24>3.0.CO;2-J)

Jones, D. 1998 Do transmembrane protein superfolds exist? *FEBS Lett.* **423**, 281–285. (doi:10.1016/S0014-5793(98)00095-7)

Jones, D. T., Taylor, W. R. & Thornton, J. M. 1994 A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**, 3038–3049. (doi:10.1021/bi00176a037)

Kernytsky, A. & Rost, B. 2003 Static benchmarking of membrane helix predictions. *Nucleic Acids Res.* **31**, 3642–3644. (doi:10.1093/nar/gkg532)

Kuroiwa, T., Sakaguchi, M., Omura, T. & Mihara, K. 1996 Reinitiation of protein translocation across the endoplasmic reticulum membrane for the topogenesis of multispanning membrane proteins. *J. Biol. Chem.* **271**, 6423–6428. (doi:10.1074/jbc.271.11.6423)

Kyte, J. & Doolittle, R. F. 1982 A simple method for displaying the hydrophathic character of proteins. *J. Mol. Biol.* **157**, 105–132. (doi:10.1016/0022-2836(82)90515-0)

Ledesma, A., de Lacoba, M. G., Arechaga, I. & Rial, E. 2002 Modeling the transmembrane arrangement of the uncoupling protein UCP1 and topological considerations of the nucleotide-binding site. *J Bioenerg. Biomembr.* **34**, 473–486. (doi:10.1023/A:1022522310279)

Liu, Y., Engelman, D. & Gerstein, M. 2002 Analysis of membrane protein families: genomic prevalence and conserved motifs. *Genome Biol.* **3**, 00 054.1–00 054.12.

Melen, K., Krogh, A. & von Heijne, G. 2003 Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.* **327**, 735–744. (doi:10.1016/S0022-2836(03)00182-7)

Mulder, N. J. 2005 InterPro, progress and status in 2005. *Nucleic Acids Res.* **33**, D201–D205. (doi:10.1093/nar/gki106)

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. 1995 SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540. (doi:10.1006/jmbi.1995.0159)

Nikiforovich, G. V., Galaktionov, S. & Marshal, G. R. 2001 Novel approach to computer modeling of seven-helical case of bactriorhodopsin. *Acta Biochim.* **48**, 53–64.

Nilsson, J., Person, B. & von Heijne, G. 2002 Prediction of partial membrane protein topologies using a consensus approach. *Protein Sci.* **11**, 2974–2980. (doi:10.1110/ps.0226702)

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. 1997 CATH-A hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108. (doi:10.1016/S0969-2126(97)00260-8)

Park, B. & Levitt, M. 1996 Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **258**, 367–392. (doi:10.1006/jmbi.1996.0256)

Pellegrini-Calace, M., Carotti, A. & Jones, D. T. 2003 Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3-D structures. *Proteins: Struct. Funct. Genet.* **50**, 537–545. (doi:10.1002/prot.10304)

Pilpel, Y., Ben-Tal, N. & Lancet, D. 1999 kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J. Mol. Biol.* **294**, 921–935. (doi:10.1006/jmbi.1999.3257)

Popot, J.-L. & Engleman, D. M. 1990 Membrane protein folding and oligomerization: the two-stage model. *Biochemistry* **29**, 4031–4037. (doi:10.1021/bi00469a001)

Popot, J.-L. & Engleman, D. M. 2000 Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.* **69**, 881–922. (doi:10.1146/annurev.biochem.69.1.881)

Rees, D. C. & Eisenberg, D. 2000 Turning a reference Inside-out: commentary on an article by Stevens and Arkin Entitled: are membrane proteins 'inside-out' proteins? *Proteins* **38**, 121–122. (doi:10.1002/(SICI)1097-0134(20000201)38:2<121::AID-PROT1>3.0.CO;2-M)

Rees, D. C., DeAntonio, L. & Eisenberg, D. 1989 Hyrophobic organization of membrane proteins. *Science* **245**, 510–513.

Rost, B., Casadio, R. & Fariselli, P. 1996 Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **4**, 521–533.

Seshadri, K., Garemyr, R., Walling, E., von Heijne, G. & Elofsson, A. 1998 Architecture of {beta}-barrel membrane proteins: analysis of trimeric porin. *Protein Sci.* **7**, 2026–2032.

Singer, S. J. 1990 The structure and insertion of integral proteins in membranes. *Ann. Rev. Cell Biol.* **6**, 247–296.

Sonnhammer, E. L., Eddy, S. R. & Durbin, R. 1997 Pfam: a comprehensive database of protein families based on seed alignments. *Proteins* **28**, 405–420. (doi:10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L)

Sonnhammer, E. L. *et al.* 1998 A hidden Markov model for predicting transmembrane helices in protein sequences. *Intell. Syst. Mol. Biol.* **6**, 175–182.

Stevens, T. & Arkin, I. T. 1999 Are membrane proteins 'inside–out' proteins? *Proteins Struct. Funct. Genet.* **36**, 135–143. (doi:10.1002/(SICI)1097-0134(19990701)36:1<135::AID-PROT11>3.0.CO;2-I)

Taylor, W. R., Jones, D. T. & Green, N. 1994 A method for alpha-helical integral membrane protein fold prediction. *Proteins* **18**, 281–294. (doi:10.1002/prot.340180309)

Tusnady, G. E. & Simon, I. 1998 Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **283**, 489–506. (doi:10.1006/jmbi.1998.2107)

van Geest, M. & Lolkema, J. 2000 Membrane topology and insertion of membrane proteins: search for topogenic signals. *Microbiol. Mol. Biol. Rev.* **64**, 13–33. (doi:10.1128/MMBR.64.1.13-33.2000)

von Heijne, G. 1992 Membrane protein structure prediction. *J. Mol. Biol.* **255**, 487–494. (doi:10.1016/0022-2836(92)90934-C)

von Heijne, G. 1996 Principles of membrane protein assembly and structure. *Progr. Biophys. Mol. Biol.* **66**, 113–139. (doi:10.1016/S0079-6107(97)85627-1)

von Heijne, G. 1999 Recent advances in the understanding of membrane protein assembly and structure. *Quart. Rev. Biophys.* **32**, 285–307. (doi:10.1017/S0033583500003541)

Wallin, E. & von Heijne, G. 1998 Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7**, 1029–1038.

White, S. H. 2001 How membranes shape protein structure. *J. Biol. Chem.* **31**, 32 395–32 398. (doi:10.1074/jbc.R100008200)

White, S. H. & Wimley, W. C. 1999 Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Struct.* **28**, 319–365. (doi:10.1146/annurev.biophys.28.1.319)