Biochimica et Biophysica Acta 1512 (2001) 1–14

BIOCHIMICA ET BIOPHYSICA ACTA

www.bba-direct.com

Rapid report

Amino acid distributions in integral membrane protein structures

Martin B. Ulmschneider, Mark S.P. Sansom *

Laboratory of Molecular Biophysics, Rex Richards Building, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

Received 5 December 2000; received in revised form 8 February 2001; accepted 13 February 2001

Abstract

Advances in structure determination of membrane proteins enable analysis of the propensities of amino acids in extramembrane versus transmembrane locations to be performed on the basis of structure rather than of sequence and predicted topology. Using 29 available structures of integral membrane proteins with resolutions better than 4 Å the distributions of amino acids in the transmembrane domains were calculated. The results were compared to analysis based on just the sequences of the same transmembrane α-helices and significant differences were found. The distribution of residues between transmembrane α-helices and β-strands was also compared. Large hydrophobic (Phe, Leu, Ile, Val) residues showed a clear preference for the protein surfaces facing the lipids for β-barrels, but in α-helical proteins no such preference was seen, with these residues equally distributed between the interior and the surface of the protein. A notable exception to this was alanine, which showed a slight preference for the interior of α-helical membrane proteins. Aromatic residues were found to follow saddle-like distributions preferring to be located in the lipid/water interfaces. The resultant 'aromatic belts' were spaced more closely for β-barrel than for α-helical membrane proteins. Charged residues could be shown to generally avoid surfaces facing the bilayer although they were found to occur frequently in the transmembrane region of β-barrels. Indeed detailed comparison between α-helical and β-barrel proteins showed many qualitative differences in residue distributions. This suggests that there may be subtle differences in the factors stabilising β-barrels in bacterial outer membranes and α-helix bundles in all other membranes. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Amino acid distribution; Membrane protein; α-Helix; β-Barrel

1. Introduction

Integral membrane proteins play a central role in many of the biological activities of cells. It is estimated that approx. 30% of genes may encode membrane proteins [1,2]. However, difficulties in overexpression and crystallisation of membrane proteins [3] may hinder structural studies. This presents both a challenge and a problem to attempts to predict membrane protein structure. The difficulties of experimental structure determination make successful structure prediction an important and pressing need. The relative paucity of structural data has hindered the development of knowledge-based potentials that have been successfully applied in prediction of globular protein structures [4]. Instead, a range of methods

* Corresponding author. Fax: +44-1865-275-182; E-mail: mark@biop.ox.ac.uk

PII: S0005-2736(01)00299-1

of increasing levels of sophistication have employed either physicochemical considerations or analysis of databases of experimentally confirmed membrane protein transbilayer topologies to devise methods for predicting the location of transmembrane (TM) α-helices within the sequences of integral membrane proteins [5–15]. There have been similar attempts to devise methods for the somewhat rarer β-barrel class [16,17] of membrane proteins [18–21].

Using such prediction methods to identify TM α-helices, there have been a number of studies on the residue distributions between extramembrane and transmembrane regions within TM helices. For example Landolt-Marticorena et al. [22] restricted their studies to the members of the single helix human type I single span membrane protein family, which have a clearly defined TM region, with the amino terminus located on the extracellular side. More recently Arkin and Brunger [23] screened a large database for TM α-helices and aligned their sequences to calculate the residue distributions in the TM domain. Other studies have focused on the role of individual residues such as glycine [24]. Analyses of residue distributions have also been used in attempts to aid prediction of the packing of TM helices within membrane proteins, e.g. [25]. TM helix sequences motifs associated with certain modes of helix packing have been investigated [26,27]. Few comprehensive studies of residue distributions for β-barrel proteins have been undertaken, mainly because of problems in prediction of TM β-strands.

A number of studies have focused on available structures of membrane proteins, both α-helical and β-barrel. For example, Stevens and Arkin [28] analysed the known structures of TM α-helices while Bowie [29] investigated the helix packing of 45 TM α-helices. Seshadri et al. [30] investigated the structures of trimeric porins and derived the TM distribution functions for different types of residues. The study confirmed the existence of aromatics rings or belts at the interfacial regions that was previously found in the structure of *Rhodobacter capsulatus* [31].

Recent years have seen a dramatic increase in the number of membrane protein structures available at atomic resolution [32,33] (a useful summary of current structures is provided by White: http://blanco.biomol.uci.edu/). Consequently a new survey is timely. Two types of integral membrane proteins were considered, membrane proteins containing α-helices and β-barrels. The present study involved 15 α-helical membrane proteins containing a total of 129 TM α-helices and 14 β-barrel membrane proteins with a total of 220 TM β-strands. The individual proteins included in our survey are listed in Appendix A and Appendix B.

2. Methods

Two different methods were employed to derive the amino acid distribution in the TM domain. The first was based on the sequences of all structurally known α-helices and β-strands (see Appendix A and Appendix B), the second was based on the C_α coordinates of the amino acids with respect to the centre of the presumed lipid bilayer.

2.1. Sequence-based method

The sequence method calculates the distribution of amino acids in a TM α-helix or β-strand, i.e. the frequency of each amino acid occurring at a certain position in a TM α-helix/β-strand. First the sequences of all non-redundant TM α-helices and β-strands were extracted from the Protein Data Bank (PDB) [34] for the proteins listed in Appendix A and Appendix B. DSSP [35] (as implemented in Rasmol) was used to identify the secondary structure elements. Identical helices and strands were discarded to avoid biasing the distributions. Since the sequence data for this study were based on structurally known helices and strands it proved best to align their sequences from the termini inwards in the following fashion.

1. Identify the termini of the TM α-helices and β-strands (using DSSP).
2. Determine which helix/strand terminus faces the 'outside' (extracellular or periplasmic space) and which terminus faces the 'inside' (intracellular space or the matrix).
3. Divide the sequences of each helix or strand into an outside and an inside part by cutting it in half in the middle.
4. Align the two parts separately at their respective termini.

The number of residues at each position in the sequence was summed for all helices/strands. For α-helical proteins the total number of sequence positions was therefore limited to 22, while for β-barrel proteins this number was only 12, due to the much shorter average length of a β-strand. Helices and strands with sequences longer than 22 and 12 residues are thus truncated in the centre after positions ±11 and ±6 respectively. In order to obtain comparable results the frequency of each amino acid was normalised such that

$$f_{\mathrm{ij}} = \frac{n_{\mathrm{ij}}}{N_{\mathrm{i}}}$$

where N_{i} is the total number of residues of type i (i.e. i = Ala, Arg, etc.) given by

$$N_{\mathrm{i}} = \sum_{\mathrm{all}\ j}^{j} n_{\mathrm{ij}}$$

and n_{ij} is the number of residues of type i at sequence position j (i.e. for α-helices $j = -1, -2, ..., -11, +11, +10, ..., +2, +1$). The sum is over all histogram positions j.

Note that there are two assumptions implicit in this analysis. The first is that all TM segments in a given class (i.e. α or β) are of the same length; the second is that all TM segments are centred on the centre of the bilayer. The first is approximately true; the second assumption cannot at present be tested in the absence of structural data for the lipid environment of each membrane protein.

2.2. Structure-based method

The structure-based method evaluates the distribution function of each residue along the bilayer normal z. Taking the centre of the bilayer as the origin $z = 0$, the normalised distribution function of an amino acid is given by

$$f_{\mathrm{i}}(z)\Delta z = \frac{n_{\mathrm{i}}(z)}{N}\Delta z$$

where i is the amino acid type, $n_{\mathrm{i}}(z)\Delta z$ is the number of amino acids of type i in the interval z to $z+\Delta z$ (with the location of an amino acid being defined by that of its C_α atom) and N is the normalisation constant defined by

$$N = \int_{\substack{\text{over}\\ \text{bilayer}\\ \text{width}}} n_{\mathrm{i}}(z)\mathrm{d}z$$

2.3. Aligning the protein

Before calculating the frequency of each residue at some distance from the centre of the bilayer $z = h$, the protein needs to be properly placed in the membrane. Unfortunately most crystal structures of membrane proteins do not indicate the location of the lipids. TM helices, however, are good indicators of the hydrophobic membrane region. Each protein was aligned by rotating its Cartesian co-ordinates until the sum of the tilt angles of all TM helices, with respect to an arbitrary z-axis, was at a minimum. The z-axis now coincides approximately with the bilayer normal. Comparison of the thus aligned X-ray crystallographic structure [36] with the structure of bacteriorhodopsin derived from electron microscopy analysis of 2D crystals containing lipid bilayers [37] showed a deviation of the calculated z-axis from the real bilayer normal of less than 1°.

Subsequently the helix termini were used to determine the centre of the bilayer, defined at $z = 0$. This was done by calculating the centre of each helix with respect to the bilayer normal and taking the mean over all helices in a protein as the centre of the bilayer. It is important to keep track which side of the protein faces the extracellular region and which side faces the cytoplasm. The proteins were rotated by 180° about the x-axis, if necessary, so that the bilayer region facing the inside (i.e. cytoplasm or matrix) has negative and the outside face positive z-co-ordinates.

This method was repeated for all proteins. Once all proteins are aligned and correctly positioned in the bilayer the distribution of each residue with respect to the bilayer centre can be calculated. β-Barrel proteins were aligned in a similar fashion. However, the alignment along the z-axis caused problems for some β-barrels which had to be aligned by eye. Unlike α-helical proteins the β-barrels were not aligned in the centre of the TM region but on the periplasmic side. The latter was chosen because the small loops on the periplasmic side of β-barrels present a well defined alignment point which was set to have the *z*-co-ordinate −15 Å.

2.4. Residue distributions

The height *h* of the backbone carbon α-atom was calculated for each residue. It is defined as the *z*-co-ordinate of the carbon α-atom. Residues in the TM region facing the extracellular side have a positive height and residues facing the cytoplasm have a negative height, while the modulus of the height represents the normal distance of the residue from the centre of the bilayer.

2.5. Solvent and lipid accessible residues

Not all residue side chains in the TM region face the lipid bilayer, some are buried within the TM domain of the membrane protein itself while others line the pores of ion channels or β-barrels. A method was devised to investigate how the residue distribution in the TM domain varies between buried and exposed residues.

First the accessible surface area of each residue was calculated, using Quanta (Biosym/MSI) and a probe radius of 1.4 Å. The accessibility fraction was calculated dividing the accessible surface area reached by the sphere by the total surface area of the respective residue. If this fraction is greater than 10% the residue is located at the surface of the protein.

The surface ratio f_{TM} is defined as the fraction of residues in the TM region which are located at the surface

$$f_{\mathrm{TM,i}} = \frac{s_{\mathrm{TM,i}}}{n_{\mathrm{TM,i}}}$$

where $n_{\mathrm{TM,i}}$ is the total number of residues of type *i* (e.g. *i* = Ala) in the TM region and $s_{\mathrm{TM,i}}$ is the number of residues of type *i* that are located on the surface of the TM domain.

2.6. Statistics

As part of the structural analysis statistical information about the lengths and tilt angles of α-helices and β-strands was calculated. A helix or strand vector was defined by the co-ordinates of the two terminal residue's C_{α} atoms. The tilt angle is then defined as the angle of this vector with respect to the 'bilayer normal'.

3. Results and discussion

3.1. Overall statistics

It is useful to examine the overall lengths (in residues) and transmembrane heights (in Å) of the secondary structure elements (Table 1). For α-helices, these results are very close to those of Bowie [29]. Using 45 TM α-helices, Bowie gave a mean number of 26.4 residues per TM helix compared to our figure of 27.1 (±5.4). For an ideal α-helix exactly parallel to the bilayer normal this would correspond to a TM helix height (i.e. helix length projected onto the bilayer normal; see Section 2 for details) of approx. 40 Å. However, the mean height of the TM helices was found to be 35.0 (±7.4) Å. The difference lies in the mean tilt angle of 22° (cf. 21° in Bowie's analysis). Thus, even if one takes into account the tilting of TM α-helices, they appear to be somewhat longer than the 'canonical' length of 20 residues spanning a 30 Å bilayer [5]. It is noteworthy that the mean height of the bilayer spanning β-strands (approx. 28 Å) is a little shorter than that of TM α-helices. This may reflect the special lipid composition of bacterial outer membranes from that of membranes in general.

Table 1
Statistical data from the analysis of both α-helical and β-barrel proteins

	Proteins	Helices/strands	Length of helices/strands (residues)	Height of helices/strands (Å)	Tilt angle (°)
α	15	129	27.1 ± 5.4	35.0 ± 7.4	22.0 ± 11.6
β	14	220	11.9 ± 3.1	27.5 ± 8.6	36.9 ± 7.7

The height represents the projection of the helix or strand vector onto the bilayer normal (see Section 2 for details).

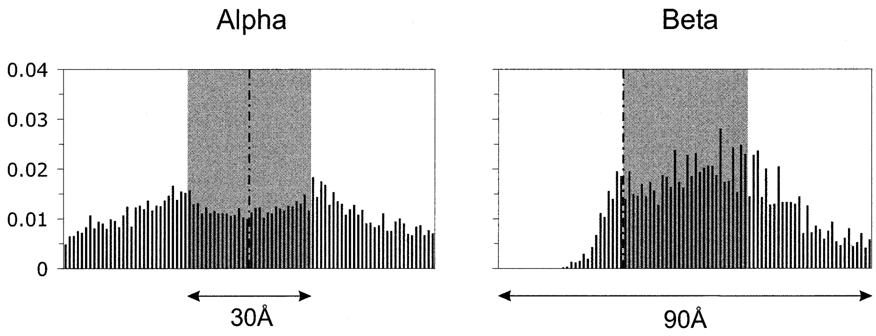

Fig. 1. Total number of residues (vertical axis) for each height (horizontal axis) with respect to the centre of the bilayer. α-Helical proteins are shown in the left panel and β-barrels on the right. The bilayer region (shaded) was assumed to have a width of 30 Å. The inside (cytoplasm or matrix for α-helical proteins and periplasmic side for β-barrels) is always on the left of the distribution graphs and the *z*-positions at which the proteins were aligned are indicated by a dash-dotted line. All graphs are normalised.

It is also informative to look at the overall distributions of residue frequencies (for all amino acid types) along the (presumed – see Section 2) bilayer normal (Fig. 1). The distributions for α-helical and β-barrel proteins exhibit distinct differences. While α-helical proteins show a saddle-like distribution with two peaks at the interfacial regions, caused by loops connecting helices, β-barrels show only a single broad peak. The periplasmic side of the peak is very well defined due to the short loops at this side. For β-barrel proteins the 30 Å wide TM region contains 58% of all the residues in these proteins, compared to 38% of residues for α-helical proteins that fall within the TM region. These overall distributions were used to normalise the distributions of the individual amino acids.

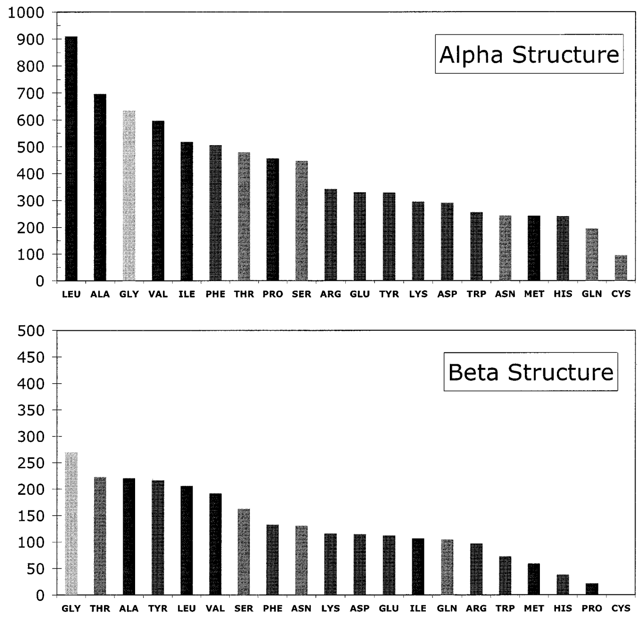

Fig. 2. Amino acid composition in the TM domain for α-helical (upper panel) and β-barrel (lower panel) proteins. The total number of residues of each type (vertical axis) is plotted against residue type (horizontal axis). For β-barrels plug domains were excluded.

3.2. TM residue composition

The amino acid compositions in the TM domain for both membrane protein types have been examined (Fig. 2). Note that plug domains from β-barrels were excluded (since they contain many water accessible charged and polar residues). As previously reported [38] the hydrophobic residues Ala, Ile, Leu

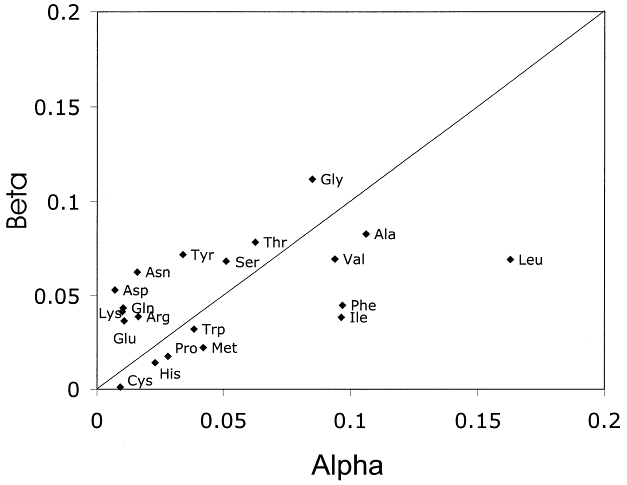

Fig. 3. Normalised frequency of amino acids in the TM domains for α-helical versus β-barrel membrane proteins. Residues below the diagonal line occur more frequently in α-helical TM proteins and residues above this diagonal prefer the TM domains of β-barrels.

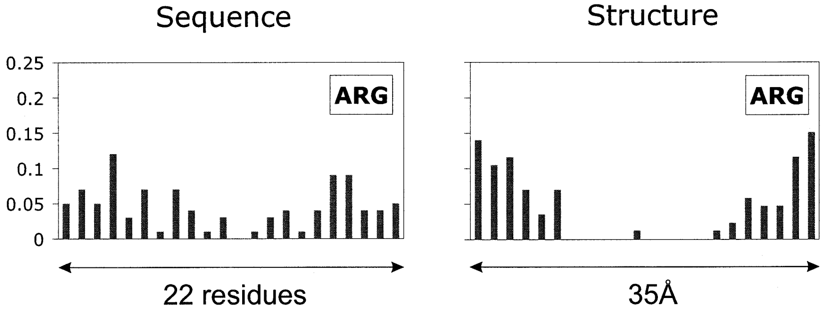

Fig. 4. Comparison of sequence-based and structure-based methods for evaluating the amino acid distribution in the TM domain of α-helical proteins, illustrated for arginine in α-helical membrane proteins. The cytoplasmic side is on the left. By dividing the TM region into 22 sections of 1.5 Å each it is possible to calculate the distribution function by counting the frequency of all residue types for each section. 1.5 Å was chosen as the width of each section as this coincides with the α-helical rise of one residue allowing direct comparison of the sequence-based and structure-based methods. The vertical axes show normalised frequencies; the horizontal axes the relative position of a residue in the TM helix sequence or its location projected onto the bilayer normal axis.

and Val make up the bulk of the amino acids in the TM domain accounting for one third (34%) of all residues in α-helical proteins and 28% in β-barrels. There are evidently differences in composition of the two classes of membrane protein, despite their similar environments. These differences are highlighted in Fig. 3. It can be seen that while charged residues occur much more frequently in the TM domain of β-barrels hydrophobic residues have higher propensities in the TM domain of α-helical proteins. For example Leu appears more than twice as often in α-helical proteins than in β-barrels. This preference can be explained by taking into account that Leu also acts as a strong helix former [39]. It is also notable that the hydrophobic β-branched residues (Val, Ile) occur at a higher frequency in the TM regions of α-helical proteins, in agreement with earlier sequence-based studies [40] and with experimental studies on peptides in a membrane-mimetic environment [41,42]. Also significant is the high frequency of glycine in TM segments. Although sometimes considered as a 'helix breaker' it has been reported that glycine residues occur frequently at helix-helix interfaces and crossing points [24] and it has been suggested that this may facilitate closer packing of TM helices [27], especially in motifs combining Gly and β-branched side chains [26].

3.3. Charged residues

Energetic considerations [5] and the positive-inside rule [6] suggest that charged amino acids should generally be excluded from TM segments. Thus, charged residues provide a convenient way of assessing three different approaches to analysis of residue distributions across a membrane, namely approaches based on analysis: (i) of TM segment sequence; (ii) of structure taking into account all residues in TM segments; and (iii) of structure taking into account only those residues which are on the lipid-exposed surface

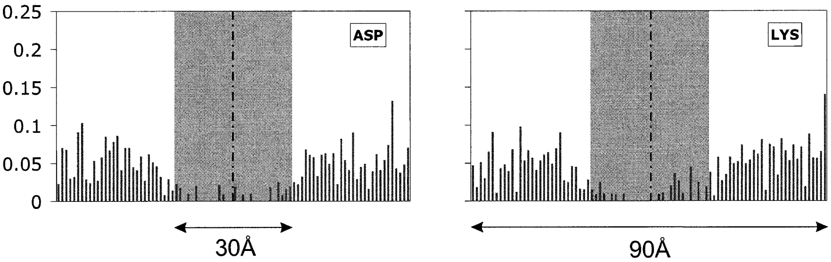

Fig. 5. Distribution of the charged residues Asp and Lys for α-helical proteins, using the structure-based method. The vertical axes show normalised frequencies; the horizontal axes the location of a residue projected onto the bilayer normal axis. The grey band represents the (presumed) location of the 30 Å hydrophobic core of the lipid bilayer.

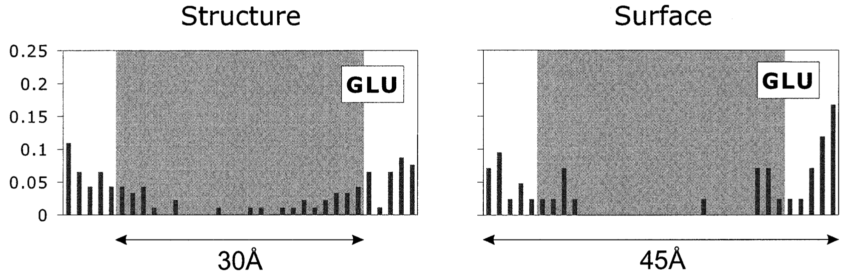

Fig. 6. Comparison of structure-based and lipid surface-exposed methods for glutamate in α-helical membrane proteins. The total range of the histogram along the bilayer normal is 45 Å.

(see Section 2 for details). The first two approaches are compared in Fig. 4 for arginine residues in α-helical membrane proteins. It is clear that taking into account the structure (i.e. the actual location of the residues along the bilayer normal relative to the bilayer centre) provides a much more distinct distribution than that obtained by simply measuring the position of residues in the TM helix sequences. Indeed, in our analysis (using (ii) above) there is only a single arginine in the centre of the TM region (as opposed to the two ends where such a residue might be able to e.g. interact with lipid head groups), this being from the purple bacteria light harvesting protein B-800/850 [43] where it forms part of the retinal binding site. The sequence method (i.e. using (i) above) on the other hand has a much lower resolution. Arg shows two pronounced peaks near the termini of the helices but occurs even near the centre of TM helices. This apparent discrepancy is due to the alignment of the helices, which cannot account for the different lengths and tilt angles. Lysine, aspartate and glutamate are distributed in a similar fashion in the structure-based analysis (Figs. 5 and 6), although the exclusion from the TM region is not as extreme as for arginine. The structure-based analysis can be extended by comparing all transmembrane residues with those only present on the lipid-exposed surface. Such a comparison for glutamate in α-helical membrane proteins (Fig. 6) reveals just one Glu on a TM lipid-exposed surface. This residue (Glu 32, chain J) is from the chicken ubiquinol cytochrome *c* oxidoreductase [44]. It is part of a slightly tilted TM helix and has an accessible surface fraction of 25%. It is conceivable that it may interact with a lipid head group (see discussion in [45]).

For β-barrels the sequence- and structure-based analyses do not give such markedly different results. Note that for β-barrels the sequence analysis spans a smaller residue range since the average length of a TM β-strand is only 12 residues (see Table 1). The two distributions for lysine (Fig. 7) are rather similar, both indicating that Lys prefers to be located on the extracellular side of the membrane. Extending the structure-based analysis to all four charged residues (Fig. 8) reveals that their distribution functions do not drop to zero towards the middle of the membrane region. In general, charged residues are much more frequent in β-barrels, either forming part of a pore lining or fixing the plug domain within a barrel.

From this analysis of charged residues it is evident that the structure-based analysis provides improved distributions over the sequence-based analysis. We

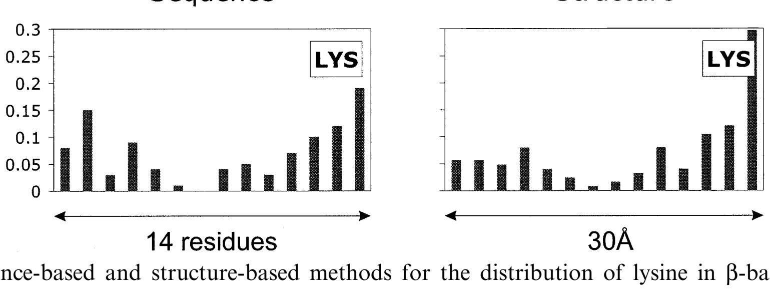

Fig. 7. Comparison of sequence-based and structure-based methods for the distribution of lysine in β-barrel membrane proteins. The sequence-based analysis (left panel) extends over 14 residues. The right panel shows the structure analysis histogram over 30 Å.

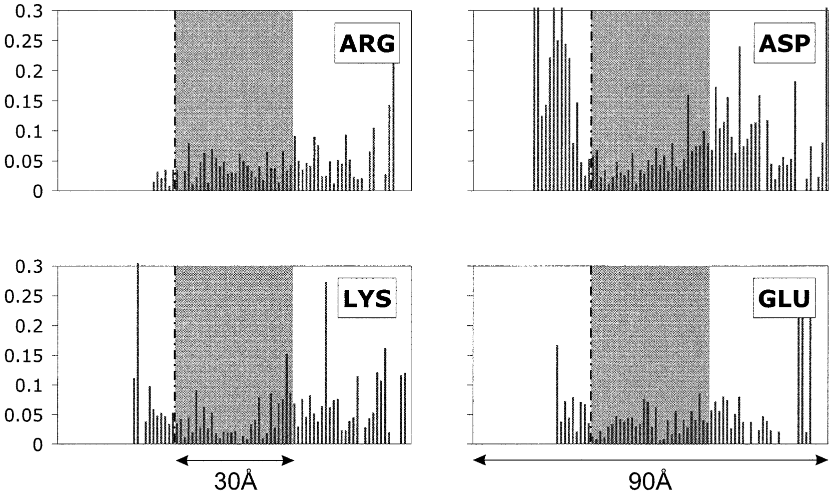

Fig. 8. Structure-based distributions of charged residues for β-barrel proteins.

therefore will restrict our attention to structure-based distributions for the remainder of the paper.

3.4. Hydrophobic residues

Turning to the hydrophobic aliphatic residues, as expected all four show a clear preference for the transbilayer region in both α-helical proteins and β-barrels (Figs. 9 and 10). This is equally true of the β-branched residues. For both classes of membrane protein, leucine shows the highest propensity to be in the TM region. Leucine is very hydrophobic and hence expected to prefer the lipid-exposed surface of a TM region. However, it should be noted that not all proteins in the study have large extramembranous domains so that the distributions are slightly biased. The bias can, however, be eliminated by calculating the fraction of Leu residues located at the surface for each domain (see Table 2). In the TM domains this fraction is 54% for α-helical proteins and 69% for β-barrels while for the domains exposed to the water on either side of the bilayer the fractions are only 38% and 37% respectively. For β-barrels this means that a Leu residue is almost twice as likely to be located on the surface in the TM domain compared to the rest of the protein while for α-helical proteins the preference for the TM surface is not so obvious.

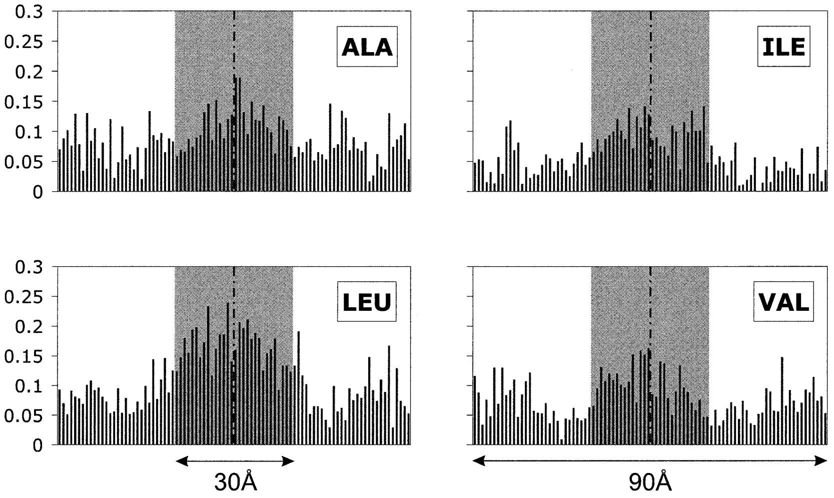

Fig. 9. Structure-based distribution functions of aliphatic hydrophobic residues for α-helical proteins.

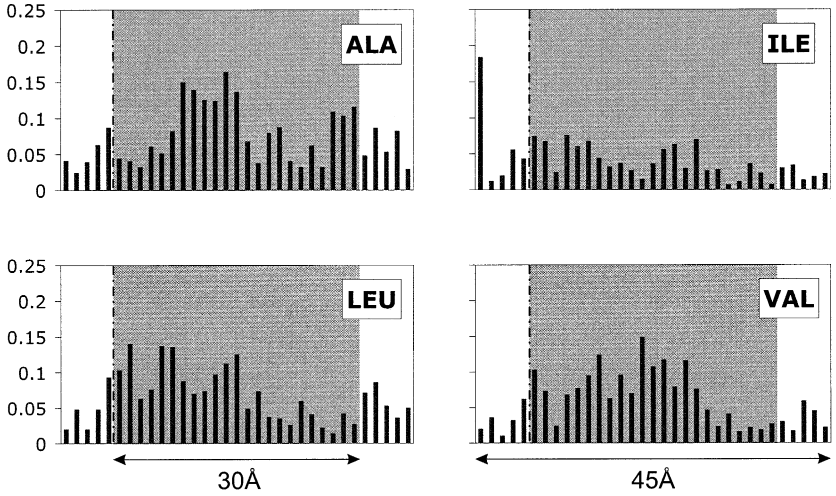

Fig. 10. Structure-based distribution functions of aliphatic hydrophobic residues for β-barrel proteins.

Table 2 shows the fraction of residues located at the surface of the TM domain for Phe, Leu, Ile, Val, Ala and Gly. For α-helical proteins the values for the large hydrophobic residues (i.e. Phe, Leu, Ile and Val) are around 50% indicating that there is no preference for these residues to be either buried or located on the surface of the TM domain. This result is in agreement with a previous study of seven α-helical membrane proteins [28] that found no correlation between hydrophobicity and accessibility in the TM domain. It also suggests that the notion sometimes expressed that the TM surface of an α-helical membrane protein is more hydrophobic than its core cannot be justified on the basis of the current data set of structures (see [28,46,47] for an extended discussion). In contrast, the small hydrophobic residues (alanine and glycine) show a preference (more marked for glycine) to be buried when in an α-helical TM domain. It has been suggested that these residues play an important role in helix-helix packing due to their short side chains [26,27] thus explaining their preference for the core of the TM region. This suggestion is in agreement with our analysis.

Table 2
Surface fraction f_{TM} (see Section 2 for definition) of hydrophobic residues in TM domains of α-helical and β-barrel membrane proteins

Hydrophobic residue	f_{TM} (%)	
	α	β
Phe	47	68
Leu	54	69
Ile	55	61
Val	47	63
Ala	41	47
Gly	31	46

β-Barrels show an altogether different picture. Here the hydrophobic residues Phe, Leu, Ile and Val show a clear preference for the TM surface with surface ratios around 65%. This suggests a much more hydrophobic exterior compared to the core of the protein. Indeed even non-porin β-barrel proteins like OmpA [48], FhuA [49,50] and OMPLA [51] contain water molecules flanked by charged and polar residues inside the barrel. Ala and Gly again show a different behaviour from the other hydrophobic residues, exhibiting no clear preference to be either buried or on the surface.

3.5. Aromatic residues

Aromatic residues have been suggested to play a special role in membrane proteins (see e.g. [45,52]). They are believed to anchor the proteins into the membrane through an interaction of their aromatic rings with the lipid head groups. A preferred localisation of aromatic residues in the interfacial regions has previously been noted for both the photosynthetic reaction centre [53] and bacterial porins [54]. Such anchoring has been explored via molecular dynamics simulations [55] and by experimental studies of model transmembrane peptides [56]. In addition it

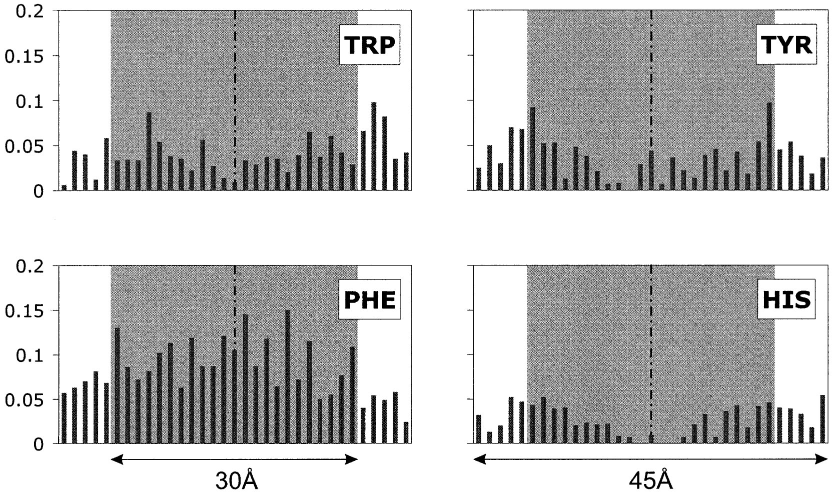

Fig. 11. Structure-based distribution functions of aromatic residues for α-helical proteins.

has been observed that while all aromatic residues in the TM domain clearly prefer being buried they show a higher propensity to face the lipid head groups at either or both TM termini [25].

In the structure-based distributions for α-helical membrane proteins (Fig. 11), tyrosine and histidine were found to have two pronounced peaks at both interfacial regions. For tryptophan the peak is much stronger at the non-cytoplasmic interface. In contrast, phenylalanine is distributed throughout the transbilayer region. These results are in general agreement with the kPROT analysis of all predicted α-helical membrane proteins in the SWISS-PROT database [25] and with the earlier analysis of [22]. Thus on the basis of an enlarged database of structures it is clear that an interfacial location requires *both* aromaticity *and* the ability to form a H-bond, i.e. an amphipathic aromatic side chain.

The distributions of aromatics in the bilayer region for β-barrels are rather different from those for α-helical membrane proteins (Fig. 12). In general it seems that for β-barrels the 'aromatic belts' are closer together, with a spacing of about 20 Å for β-barrels compared with 30 Å for α-helical proteins. This feature was seen in the earlier analysis of trimeric porins [30]. Thus, if aromatic residues are located at the bilayer/water interface, this implies a thinner bilayer in Gram-negative outer membrane than in

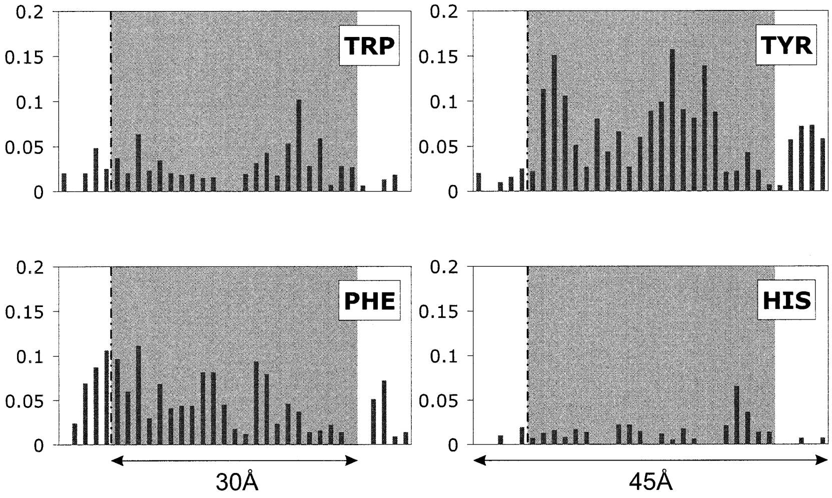

Fig. 12. Structure-based distribution functions of aromatic residues for β-barrel proteins.

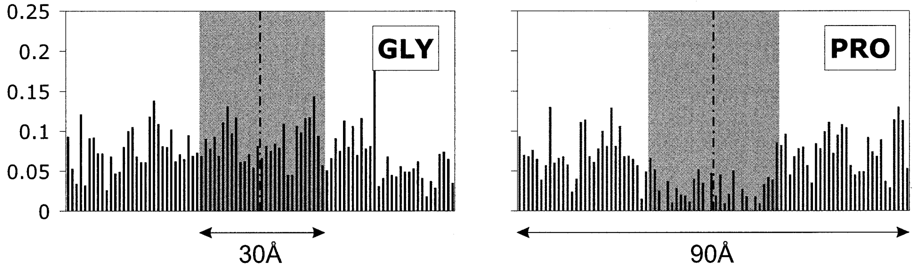

Fig. 13. Structure-based distribution functions of Gly and Pro for α-helical proteins.

other membranes. Alternatively, the unusual lipid composition of the outer leaflet of the Gram-negative outer membrane may lead to a subtly different mode of lipid head group/protein interaction. This difference merits further investigation.

3.6. Proline and glycine residues

Proline and glycine are often considered to be 'helix breaking' residues. Proline has often been suggested to play a special role in TM α-helices [57–60] associated with its ability to generate a helix kink. Glycine may mediate helix flexibility and also provides an opportunity for close packing of TM helices (as noted above). Structure-based distributions for α-helical membrane proteins (Fig. 13) show that both residues can occur in the TM region. Although proline is, as expected, predominant in the loop regions outside the TM domain, a peak is detectable towards the centre of the bilayer which is associated with kinked helices in the TM domain. A nice example of this is the strongly kinked G chain helix in the chicken ubiquinol cytochrome *c* oxidoreductase [44], where two neighbouring Pro residues (Pro 50 and Pro 51 chain G) cause a large bend in the middle of the helix. This feature was not found in an analysis of single spanning TM helices [22], suggesting that prolines may play a structural role in membrane proteins with multiple TM helices, perhaps increasing the stability of the TM domain by 'interlocking' the helices, or by providing molecular hinges that enable conformational transitions in more complex membrane proteins [60,61]. Indeed, it has been shown that in several TM helices bends caused by proline are observed when more than one proline or a combination of proline and glycine residues are found spaced four residues apart [24]. Glycine residues have the highest packing values of any amino acid and are believed to play an important role in TM helix packing [27]. Although it has been pointed out that Gly residues occur more than twice as often in membrane proteins than in soluble proteins [27] no clear preference for the TM region can be seen for either α-helical (Fig. 13) or β-barrel proteins (not shown).

3.7. Polar residues

The uncharged polar residues of α-helical membrane proteins display two different types of behaviour (data not shown) with asparagine and glutamine following the distribution pattern of charged residues avoiding the TM region. This presumably reflects their need to form multiple H-bonds. In contrast serine and threonine show no preference for either the transmembrane or extramembrane region. It has been noted [62] that serine and threonine side chains in a helix can form H-bonds to the carbonyl oxygen of the preceding turn of the helix, thus enabling such side chains to occur in a TM region. Furthermore, as noted by Eilers et al. [27] serine and threonine may be associated with tight packing of TM α-helices.

4. Conclusion

Previous analyses of amino acid distributions in the TM domain of α-helical membrane proteins were chiefly based on sequence analysis of putative TM helices. This was due to the lack of a sufficient number of solved structures for a statistically significant analysis. The number of membrane proteins solved at atomic resolution is now sufficient for a statistical derivation of the amino acid distribution functions as well as properties such as average helix/strand lengths and tilt angles [29]. A comparison of sequence- and structure-based methods showed the latter to be better resolved, aiding insights into the specific roles of amino acids in TM domains. Distribution functions were found to vary significantly not only between different types of residues but also between residues located at the protein surface as compared to those buried inside the TM domain. Hydrophobic residues with the notable exception of alanine showed a clear preference for the surfaces of the TM domain of β-barrel proteins while amphipathic aromatic and charged residues appear to prefer the protein surfaces near the lipid/water interfaces. Aromatic residues proved to have a saddle type distribution with increased occurrence at the extra- and intracellular interfacial regions. For α-helical proteins this seems to follow the overall amino acid distribution. However, since all graphs were normalised according to the overall distribution the aromatic belts represent a genuine effect rather than just following the overall distribution. Charged residues were found to likewise have a strong preference for the interfacial as well as the water accessible regions and plug domains. Distributions for α-helical and β-barrel proteins were seen to differ significantly in some respects, but many key results are comparable. The differences may reflect differences between the environment (e.g. thickness of hydrophobic core) provided by bacterial outer membranes and by other membranes.

In summary, these studies support the re-analysis of residue propensities for membrane proteins based on three-dimensional structures rather than predicted topologies. As more structures emerge, statistics will improve and more detailed analyses will be possible. In particular, it may be possible to derive empirical potentials to aid prediction of membrane protein structures in a similar manner to the derivation of such potentials for water soluble proteins [63]. Given the increasing number of medium resolution electron microscopy structures for membrane proteins (e.g. [64]), empirical potentials based on residue propensities (as in the paper) combined with analysis of TM helix packing [29] should provide tools for a hybrid approach to modelling membrane protein structures.

Acknowledgements

MBU is an MRC student. Research in MSPS's group is supported by the Wellcome Trust. We thank the Oxford Supercomputing Centre for use of their facilities. Our thanks to Phil Biggin for his comments on this manuscript, and to Manuel Than for an early copy of the coordinates of *1ehk*.

Appendix A. α-Helical membrane

PDB	Name
1a91	F_0F_1 ATPase subunit c (*Escherichia coli*)
1afo	Glycophorin A (human)
1ap9	Bacteriorhodopsin (*Halobacterium salinarium*)
1ar1	Cytochrome *c* oxidase (*Paracoccus denitrificans*)
1bcc	Ubiquinol cytochrome *c* oxidoreductase (chicken)
1bgy	Cytochrome bc_1 complex (bovine)
1bl8	Potassium channel (KcsA) (*Streptomyces lividans*)
1ehk	ba(3)-Cytochrome *c* oxidase (*Thermus thermophilus*)
1fum	Fumarate reductase respiratory complex (*E. coli*)
1kzu	Light harvesting protein B-800/850 (*Rhodopseudomonas acidophila*)
1lgh	Light harvesting complex II (*Rhodospirillum molischianum*)
1msl	Mechanosensitive ion channel (*Mycobacterium tuberculosis*)
1occ	Cytochrome *c* oxidase (bovine)
1prc	Photosynthetic reaction centre (*Rhodopseudomonas viridis*)
4rcr	Photosynthetic reaction centre (*Rhodobacter sphaeroides*)

Appendix B. β-Barrel membrane proteins

PDB	Name
1a0s	Sucrose-specific porin (ScrY) (*Salmonella typhimurium*)
1af6	Maltoporin (LamB) (*E. coli*)
1bxw	Outer membrane protein A (OmpA) (*E. coli*)
1by3	Ferric hydroxamate uptake receptor (FhuA) (*E. coli*)
1fep	Ferric enterobactin receptor (FepA) (*E. coli*)
1mpr	Maltoporin (LamB) (*S. typhimurium*)
1opf	Matrix porin (OmpF) (*E. coli*)
1osm	Osmoporin (OmpK36) (*Klebsiella pneumoniae*)
1pho	Phosphoporin (PhoE) (*E. coli*)
1prn	Porin (*Rhodopseudomonas blastica*)
1qd5	Outer membrane phospholipase A (OMPLA) (*E. coli*)
1qj8	Outer membrane protein X (OmpX) (*E. coli*)
2por	Porin (*R. capsulatus*)
7ahl	α-Hemolysin (α-toxin) (*Staphylococcus aureus*)

References

[1] E. Wallin, G. von Heijne, Protein Sci. 7 (1998) 1029–1038.

[2] D.T. Jones, FEBS Lett. 423 (1998) 281–285.

[3] R. Grisshammer, C.G. Tate, Q. Rev. Biophys. 28 (1995) 315–422.

[4] T. Lazaridis, M. Karplus, Curr. Opin. Struct. Biol. 10 (2000) 139–145.

[5] D.M. Engelman, T.A. Steitz, A. Goldman, Annu. Rev. Biophys. Biophys. Chem. 15 (1986) 321–353.

[6] G. von Heijne, J. Mol. Biol. 225 (1992) 487–494.

[7] K. Hofmann, W. Stoffel, Biol. Chem. Hoppe-Seyler 374 (1993) 166.

[8] D.T. Jones, W.R. Taylor, J.M. Thornton, Biochemistry 33 (1994) 3038–3049.

[9] B. Persson, P. Argos, J. Mol. Biol. 237 (1994) 182–192.

[10] B. Persson, P. Argos, J. Protein Chem. 16 (1997) 453–457.

[11] B. Rost, R. Casadio, P. Fariselli, C. Sander, Protein Sci. 4 (1995) 521–533.

[12] B. Rost, P. Fariselli, R. Casadio, Protein Sci. 5 (1996) 1704–1718.

[13] R. Casadio, P. Fariselli, C. Taroni, M. Compiani, Eur. Biophys. J. 24 (1996) 165–178.

[14] M. Cserzo, E. Wallin, I. Simon, G. von Heijne, A. Elofsson, Protein Eng. 10 (1997) 673–676.

[15] E.L.L. Sonnhammer, G. von Heijne, A. Krogh, in: J. Glasgow, T.L., F. Major, R. Lathrop, D. Sankoff, C. Sensen (Ed.), Proc. of Sixth Int. Conf. on Intelligent Systems for Molecular Biology, AAAI Press, 1998, pp. 175–182.

[16] S.K. Buchanan, Curr. Opin. Struct. Biol. 9 (1999) 455–461.

[17] G.E. Schulz, Curr. Opin. Struct. Biol. 10 (2000) 443–447.

[18] D. Jeanteur, J.H. Lakey, F. Pattus, Mol. Microbiol. 5 (1991) 2153–2164.

[19] T. Schirmer, S.W. Cowan, Protein Sci. 2 (1993) 1361–1363.

[20] M.M. Gromiha, R. Majumdar, P.K. Ponnuswamy, Protein Eng. 10 (1997) 497–500.

[21] K. Diederichs, J. Freigang, S. Umhau, Z.K., J. Breed, Protein Sci. 7 (1998) 2413–2420.

[22] C. Landolt-Marticorena, K.A. Williams, C.M. Deber, R.A.F. Reithmeier, J. Mol. Biol. 229 (1993) 602–608.

[23] I. Arkin, A. Brunger, Biochim. Biophys. Acta 1429 (1998) 113–128.

[24] M.M. Javadpour, M. Eilers, M. Groesbeek, S.O. Smith, Biophys. J. 77 (1999) 1609–1618.

[25] Y. Pilpel, N. Ben-Tal, D. Lancet, J. Mol. Biol. 294 (1999) 921–935.

[26] A. Senes, M. Gerstein, D.M. Engelman, J. Mol. Biol. 296 (2000) 921–936.

[27] M. Eilers, S.C. Shekar, T. Shieh, S.O. Smith, P.J. Fleming, Proc. Natl. Acad. Sci. USA 97 (2000) 5796–5801.

[28] T.J. Stevens, I.T. Arkin, Protein Struct. Funct. Genet. 36 (1999) 135–143.

[29] J.U. Bowie, J. Mol. Biol. 272 (1997) 780–789.

[30] K. Seshadri, R. Garemyr, E. Wallin, G. von Heijne, A. Elofsson, Protein Sci. 7 (1998) 2026–2032.

[31] M.S. Weiss, A. Kreusch, E. Schiltz, U. Nestel, W. Welte, J. Weckesser, G.E. Schulz, FEBS Lett. 280 (1991) 379–382.

[32] V. Kuhlbrandt, E. Gouaux, Curr. Opin. Struct. Biol. 9 (1999) 445–447.

[33] J.U. Bowie, Curr. Opin. Struct. Biol. 10 (2000) 435–437.

[34] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Res. 28 (2000) 235–242.

[35] W. Kabsch, C. Sander, Biopolymers 22 (1983) 2577–2637.

[36] E. PebayPeyroula, G. Rummel, J.P. Rosenbusch, E.M. Landau, Science 277 (1997) 1676–1681.

[37] N. Grigorieff, T.A. Ceska, K.H. Downing, J.M. Baldwin, R. Henderson, J. Mol. Biol. 259 (1996) 393–421.

[38] G. von Heijne, Y. Gavel, Eur. J. Biochem. 174 (1988) 671–678.

[39] P.Y. Chou, G.D. Fasman, Biochemistry 13 (1974) 211–222.

[40] S.C. Li, C.M. Deber, FEBS Lett. 311 (1992) 217–220.

[41] C.M. Deber, S.-C. Li, Biopolymers 37 (1995) 295–318.

[42] C.M. Deber, N.K. Goto, Nat. Struct. Biol. 3 (1996) 815–818.

[43] S.M. Prince, M.Z. Papiz, A.A. Freer, G. McDermott, A.M. HawthornthwaiteLawless, R.J. Cogdell, N.W. Isaacs, J. Mol. Biol. 268 (1997) 412–423.

[44] Z.L. Zhang, L.S. Huang, V.M. Shulmeister, Y.I. Chi, K.K. Kim, L.W. Hung, A.R. Crofts, E.A. Berry, S.H. Kim, Nature 392 (1998) 677–684.

[45] J.A. Killian, G. von Heijne, Trends Biochem. Sci. 25 (2000).

[46] D.C. Rees, L. DeAntonio, D. Eisenberg, Science 245 (1989) 510–513.

[47] D.C. Rees, D. Eisenberg, Proteins Struct. Funct. Genet. 38 (2000) 121–122.

[48] A. Pautsch, G.E. Schulz, Nat. Struct. Biol. 5 (1998) 1013–1017.

[49] A.D. Ferguson, E. Hofmann, J.W. Coulton, K. Diederichs, W. Welte, Science 282 (1998) 2215–2220.

[50] K.P. Lochner, B. Rees, R. Koebnik, A. Mitschler, L. Moulinier, J. Rosenbusch, D. Moras, Cell 95 (1998) 771–778.

[51] H.J. Snijder, I. Ubarretxena-Belandia, M. Blaauw, K.H. Kalk, H.M. Verheij, M.R. Egmond, N. Dekker, B.W. Dijkstra, Nature 401 (1999) 717–721.

[52] W.M. Yau, W.C. Wimley, K. Gawrisch, S.H. White, Biochemistry 37 (1998) 14713–14718.

[53] J. Deisenhofer, H. Michel, Science 245 (1989) 1463–1473.

[54] S.W. Cowan, T. Schirmer, G. Rummel, M. Steiert, R. Ghosh, R.A. Pauptit, J.N. Jansonius, J.P. Rosenbusch, Nature 358 (1992) 727–733.

[55] L.R. Forrest, M.S.P. Sansom, Curr. Opin. Struct. Biol. 10 (2000) 174–181.

[56] M.R.R. de Planque, J.A.W. Kruijtzer, R.M.J. Liskamp, D. Marsh, D.V. Greathouse, R.E. Koeppe, B. de Kruijff, J.A. Killian, J. Biol. Chem. 274 (1999) 20839–20846.

[57] C.J. Brandl, C.M. Deber, Proc. Natl. Acad. Sci. USA 83 (1986) 917–921.

[58] G. von Heijne, J. Mol. Biol. 218 (1991) 499–503.

[59] D.N. Woolfson, R.J. Mortishire-Smith, D.H. Williams, Biochem. Biophys. Res. Commun. 175 (1991) 733–737.

[60] M.S.P. Sansom, H. Weinstein, Trends Pharmacol. Sci. 21 (2000) 445–451.

[61] S. Subramaniam, R. Henderson, Nature 406 (2000) 653–657.

[62] T.M. Gray, B.M. Matthews, J. Mol. Biol. 175 (1984) 75–81.

[63] M.J. Sippl, J. Mol. Biol. 213 (1990) 859–883.

[64] B.L. de Groot, J.B. Heymann, A. Engel, K. Mitsuoka, Y. Fujiyoshi, H. Grubmüller, J. Mol. Biol. 300 (2000) 987–994.