Biochimica et Biophysica Acta 1429 (1998) 113–128

Statistical analysis of predicted transmembrane α-helices

Isaiah T. Arkin [1],*, Axel T. Brunger *

Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

Received 9 July 1998; received in revised form 23 September 1998; accepted 25 September 1998

Abstract

Statistical analyses were undertaken for putative transmembrane α-helices obtained from a database representing the subset of membrane proteins available in Swiss-Prot. The average length of a transmembrane α-helix was found to be 22–21 amino acids with a large variation around the mean. The transfer free energy from water to oil of a transmembrane α-helix in bitopic proteins, −48 kcal/mol, is higher than that in polytopic proteins, −39 kcal/mol, and is nearly identical to that obtained by assuming a random distribution of solely hydrophobic amino acids in the α-helix. The amino acid composition of hydrophobic residues is similar in bitopic and polytopic proteins. In contrast, the more polar the amino acids are, the less likely they are to be found in bitopic proteins compared to polytopic ones. This most likely reflects the ability of α-helical bundles to shield the polarity of residues from the hydrophobic bilayer. One half of all amino acids were distributed non-randomly in both bitopic and polytopic proteins. A preference was found for tyrosine and tryptophan residues to be at the ends of transmembrane α-helices. Correlated distribution analysis of amino acid pairs indicated that most amino acids are independently distributed in each helix. Exceptions are cysteine, tyrosine, and tryptophan which appear to cluster closely to one another and glycines which are preferentially found on the same side of α-helices.

Keywords: Membrane protein; α-Helix; Lipid bilayer; Protein database

1. Introduction

There is a large interest in obtaining structural information about membrane proteins. The importance of membrane proteins is illustrated by the fact that more than 75% of all pharmaceuticals are targeted to one family of membrane proteins: the G protein coupled receptors [2]. Open reading frames encoding proteins with predicted transmembrane α-helices are exceptionally abundant in sequence databases (20–50%) including the *Mycoplasma genitalium* [3], *Haemophilus influenza* [4], *Methanococcus jannaschii* [5] and *Saccharomyces cerevisiae* [6] genomes [7].

The basis of transmembrane α-helix prediction is that a stretch of 15–30 hydrophobic amino acids is likely to be an α-helix in a membrane bilayer. The underlying assumption is that formation of secondary structure hydrogen bonds is highly favorable in the lipid bilayer, as no hydrogen-bonding groups are present in the lipid environment [8]. Furthermore, the low dielectric environment increases the strengths of the hydrogen bonds considerably. A 20-residue transmembrane α-helix has a transfer

Abbreviations: TMbase25, transmembrane protein database representing the entire Swiss-Prot data base [1]

* Corresponding authors.

[1] Present address: Cambridge University, Department of Biochemistry, 80 Tennis Court Road, Cambridge CB2 1GA, UK.

0167-4838/98/$ – see front matter © 1998 Elsevier Science B.V. All rights reserved.
PII: S0167-4838(98)00225-8

free energy from an aqueous solution (irrespective of its secondary structure) to a lipid bilayer of approximately −30 kcal/mol [8]. The energetic cost of unfolding the α-helix in the lipid bilayer would be 120 kcal/mol and therefore highly improbable. The high free energy gain of inserting an α-helix into the lipid bilayer would therefore result in diminishing solubility of the protein in an aqueous environment.

A database of putative transmembrane α-helices can be generated in an automated way by hydropathy analysis from a protein sequence database. This database will contain sequences with variable degree of similarity to each other. In order to minimize statistical bias it is imperative to adjust the representation of frequently occurring sequences. This can be done by subjectively removing homologous sequences by inspection of the sequences [9]. In a database as large as the SWISS-PROT this approach is not practical. An automated procedure was developed by Hofmann and Stoffel [1] in order to compute a statistical weight for each sequence in the database.

Membrane proteins can be divided into two categories based on the number of transmembrane α-helices: bitopic and polytopic. Bitopic proteins contain a single transmembrane α-helix while polytopic proteins contain two or more α-helices. These two categories define the interactions that the transmembrane α-helices are involved in. Helices in bitopic membrane proteins interact with the lipid environment, whereas in polytopic proteins they may interact with other transmembrane α-helices as well. One caveat of this categorization is that no account is made of protein-protein contacts by oligomerization of bitopic transmembrane α-helices.

In comparing a large number of transmembrane α-helices, a fundamental problem arises: the length of the transmembrane α-helices shows significant variation. A lipid bilayer can be regarded simply as consisting of three domains: a hydrophobic core whose thickness varies as a function of the length of the lipid acyl chain and two polar head group regions on either side of the hydrophobic core. As a lipid bilayer thickness varies, so presumably do the hydrophobic α-helices embedded within them. Therefore, alignment of transmembrane α-helices of different lengths could map the lipid head group of one transmembrane α-helix to the hydrophobic core of another transmembrane α-helix. One way of reducing this problem is to generate sequences of a specified length [9]. However, truncation artifacts could occur resulting in misaligned segments.

Four different alignment schemes were averaged in an attempt to minimize truncation and misalignment artifacts. The four alignment schemes consist of aligning the amino termini, the carboxy termini, the middle of the sequences and the ends of the sequences. Test calculations showed that averaging the four alignment schemes yields a more accurate description of the actual distribution.

Here we present a statistical analysis of the amino acid composition and distribution of transmembrane α-helices in both bitopic and polytopic proteins. While the transmembrane α-helices are, similar in length, differences of amino acid composition exist between the two classes that may be important in terms of their structure and function. One half of all amino acids are distributed non-randomly in both polytopic and bitopic proteins. Several of these amino acids exhibit different non-random distributions for polytopic and bitopic membrane proteins. Correlations between residue positions along transmembrane α-helices are only found for a small subset of amino acids (tryptophan, tyrosine, cysteine, and glycine).

2. Methods

2.1. Database

The database used for statistical analysis was TMbase25 [1]. This database was generated from version 25 of the Swiss-Prot database and contained a total of 7490 putative transmembrane segments from a total of 3596 different putative membrane proteins. The size of the Swiss-Prot database is still significantly larger than any of the sequenced genomes.

2.2. Transmembrane α-helix length restriction

Before any statistical analysis can be made, the size of the database needs to be defined from which to extract information. As the TMbase25 contains hydrophobic stretches of varying lengths, we restricted the analysis to stretches with lengths between

Table 1
Percentage of proteins with specified number of transmembrane α-helices

α-Helices per protein	No. of proteins	No. of α-helices	Genomic databases
1	426.1, 55.7%	426.1, 16.9%	45.95%
2	50.1, 6.6%	100.2, 4.0%	15.28%
3	26.9, 3.5%	80.6, 3.2%	6.14%
4	54.6, 7.1%	218.5, 8.7%	4.32%
5	21.0, 2.7%	104.8, 4.2%	2.79%
6	25.8, 3.4%	155.1, 6.2%	2.25%
7	86.7, 11.3%	607.2, 24.1%	2.70%
8	14.4, 1.9%	115.4, 4.6%	1.83%
9	2.0, 0.3%	18.0, 0.7%	2.60%
10	10.0, 1.3%	99.8, 4.0%	1.88%
11	6.0, 0.8%	65.7, 2.6%	2.32%
12	35.1, 4.6%	421.6, 16.7%	1.58%
13	2.0, 0.3%	26.0, 1.0%	1.10%
14	1.0, 0.1%	14.0, 0.6%	1.10%
15	0.1, 0.0%	1.8, 0.1%	0.89%
17	0.2, 0.0%	2.6, 0.1%	1.21%
24	2.5, 0.3%	59.8, 2.4%	0.06%
Total	764.5, 100%	2517.2, 100%	94%

The middle two columns refer to the TMbase25. The last column is an average over the genomic databases of *Mycoplasma genitalium, Haemophilus influenza, Methanococcus jannaschii* and *Saccharomyces cerevisiae* [7].

15–30 residues [8]. This reduced the size of the database to 2517 transmembrane α-helices out of 765 statistically distinct protein families (PAM-80, see below).

The length restriction to 15–30 residues was justified by the following arguments. In order to traverse an 'average' lipid bilayer with a hydrophobic core of ca. 30 Å thickness [10] a 20-residue α-helix parallel to the bilayer normal is required (1.5 Å rise per residue). However, residue stretches shorter than 20 amino acids could also be membrane-associated. For example, the membrane retention signal domain of vesicular stomatitis virus G protein (VSVG) only becomes soluble after truncation of the transmembrane segment to a size of less than 14 residues [11]. Thus, although the length of the native transmembrane segment is 20 residues, a significantly shorter α-helix could maintain integral membrane retention. A similar result was obtained for the synaptic vesicle protein synaptobrevin: it remained anchored to the membrane with a 12-residue polyleucine stretch, although the native transmembrane domain is 22 residues long [12].

Reducing the lower length limit of putative transmembrane α-helices to less than 15 may lead to the identification of false positives, i.e. hydrophobic segments picked up the hydropathy analysis that are not membrane-associated [8]. The upper length limit was set to 30 which corresponds to an α-helix with a 48° tilt angle relative to the membrane normal in a 30 Å thick lipid bilayer. In addition to the length restriction, a high energy cut-off of −20 kcal/mol was used.

2.3. Statistical weights

In order to avoid bias resulting from sequence homology, putative transmembrane α-helical segments were statistically weighed [1]. Briefly, for each pair of proteins, a PAM score (number of assumed point mutations per 100 residues) was calculated. The proteins were then grouped into families where any member of the family has at least one additional member with a PAM score of less than 80 (i.e., at least 20% sequence identity). Consequently the size of the family can be regarded as inversely proportional to the 'uniqueness of a sequence'.

The size of a PAM80 family does not take into account similarity within the family and is therefore not suitable as a statistical weight. A more useful statistical weight can be obtained from the following expression:

$$\left(\sum_j \frac{1}{1+\mathcal{P}(i,j)}\right)^{-1}$$

where $\mathcal{P}$(i,j) is the PAM80 score between sequence *i* and *j*, where *j* is every sequence in the PAM80 family to which sequence *i* belongs to.

2.4. Statistical analysis of α-helices in soluble proteins

The composition of amino acids in α-helices in soluble proteins was determined by retrieval of a representative, non-homologous list of protein structures [13,14] from the Protein Data Bank [15]. Each of these proteins was subjected to secondary structure analysis using the program DSSP [16].

2.5. Statistical assignment of positions along the helix

Each putative transmembrane α-helical segment was divided in half. Residues after the mid-point (starting from the amino terminus) were defined as +1, +2, +3, ..., +*n* while preceding residues were numbered −1, −2, −3, ..., −*n*. Mid-point residues in odd number segments were defined as +1. Thus, a maximum of 30 different positions were defined: −15, −14, ..., −1, +1, +2, ..., +15.

2.6. Sequence alignment schemes

Four different schemes were used to align all of the sequences of varying lengths:

1. Sequences were aligned according to their amino termini. This alignment was therefore termed ‘amino alignment’.
2. Sequences were aligned according to their carboxy termini (termed ‘carboxy alignment’).
3. Sequences were aligned using their median (‘median alignment’).
4. Sequences were aligned according to their ends, by taking the first seven and last eight amino acids of every sequence and generating new sequences that can readily be aligned. Thus, sequences longer than 15 amino acids had their middle residues omitted in this alignment scheme. Accordingly this alignment scheme was termed ‘ends alignment’.

In schemes 1–3, the sequences are aligned according to the longest allowed sequence, that of 30 amino acids. In scheme 4, sequences are aligned according to the shortest allowed sequence, that of 15 residues. In order to compensate for the shorter resulting length from this alignment scheme, every residue was duplicated so that the final length would be 30 residues.

2.7. Distribution analysis

For each putative transmembrane α-helix, the number of times each amino acid appeared in a particular position was counted (multiplied by the statistical weights) using an automated Perl script [17]. Visual inspection of the resulting distributions was used to asses the character of the amino acid distributions. Six different kinds of distributions emerged: descending, ascending, concave, convex, spike, and random. The first two kinds of distributions define an amino acid that has a descending or ascending probability of occurrence from the amino terminus to the carboxy terminus. The third and forth kinds of distribution define an amino acid that is more probable at the end or middle of the sequence in a symmetrical fashion. The fifth kind of distribution is defined when a random distribution occurs in all but a very short segment of the sequence (length <4).

2.8. Correlated distributions

Correlations between hydrophobic amino acids at different positions were obtained by calculating the number of amino acid pairs at a particular distance (in sequence space) averaged over all instances occurring in the subset of transmembrane α-helices taken from TMbase25. The total number of amino acid pair instances corrected for sequence homology was 653 295. Thus, for each of the 190 possible amino acid pairs, an average of 3438 instances were observed. Normalized probability histograms were computed for each amino acid pair. A theoretical, non-interacting correlation plot was computed from a large number of transmembrane α-helices with lengths between 15 and 30 residues and random sequences. Correlations among polar amino acids were not analyzed due to poor statistical representation.

3. Results and discussion

3.1. Database

There are 765 statistically distinct proteins in the TMbase25 database with a total of 2517 transmembrane α-helices, or 3.3 transmembrane α-helices per protein on average. More than one half of the proteins in the database are bitopic membrane proteins, while the remaining 44.3% are divided into families of variable sizes (Table 1). One quarter of all polytopic membrane proteins contain seven transmembrane α-helices while the remainder contain 4, 2, 12, 3, 6, 5, 8 and 24 in decreasing frequency. Very few proteins contain 9, 11, 13, 14, 15 or 17 α-helices. No protein contains 16, 18–23 or >24 transmembrane α-helices. Although the total number of bitopic proteins was more than half of all of the proteins in the database, the total number of transmembrane α-helices from bitopic proteins was only 16.9%.

It is interesting to note that the relative abundance of the different families of membrane proteins (each containing a different number of transmembrane α-helices) is different from that found in the genomic databases [7]. The currently available genomic databases show a gradual, monotone decrease in the relative abundance of a particular family as a function of the number of transmembrane α-helices. In contrast, there is an abundance of particular families in the TMbase25 database (e.g. those containing 7 and 12 transmembrane α-helices). The genomic databases are from evolutionary primitive organisms while TMbase25 is representative of a broader range of organisms. Membrane proteins containing 7 or 12 transmembrane α-helices may be exceptionally abundant in higher organisms. It should be noted however, that the statistical weight employed in TMbase25 may not be sufficient to overcome the research community's bias toward studies of (and hence sequencing) members of the families containing 7 or 12 transmembrane α-helices.

3.2. Length distribution of transmembrane α-helices

Fig. 1 shows a histogram of the length of transmembrane α-helices found in the TMbase25 database for both bitopic and polytopic membrane proteins. As described in Section 2, transmembrane α-helices of lengths ranging between 15 and 30 amino acids were chosen for the analysis. The average length of a transmembrane α-helix in the TMbase25 is roughly 21 for polytopic proteins, and one to two residues longer in bitopic membrane proteins. This length coincides with the minimum length of an α-helix required to traverse a 30 Å thick lipid bilayer [8].

The variation in length of α-helices may result from the low energetic cost of deforming the lipid bilayer [18–20]. Furthermore, the thickness of a bilayer may vary as a function of the length of the acyl chains of the lipid. For example, erythrocytes may contain hundreds of different types of lipids with varying acyl length chains [21]. Subsequently, as the lengths of the acyl chains are correlated to the thickness of the bilayer, significant thickness differences can persist in a single bilayer [21]. The variation of the length of transmembrane α–helices may also play a role as a Golgi retention signal [22].

3.3. Hydrophobicity of transmembrane α-helices

Fig. 2 shows the averaged transfer free energy ($\Delta G_{\mathrm{Water}\Rightarrow\mathrm{Oil}}$) of all transmembrane α-helices in bitopic and polytopic proteins, as a function of its length. Transmembrane α-helices shorter than 20 amino acids in both bitopic and polytopic proteins have a nearly constant $\Delta G_{\mathrm{Water}\Rightarrow\mathrm{Oil}}$ of −26 kcal/mol. Therefore the shorter the α-helices are, the more hydrophobic on average the residues have to be. Helices longer than 20 residues become more hydrophobic with increasing length. This effect is more pronounced in bitopic compared to polytopic proteins. In bitopic proteins, each additional amino acid beyond 20 residues adds a $\Delta G_{\mathrm{Water}\Rightarrow\mathrm{Oil}}$ of 1.8 kcal/mol while in polytopic proteins, the value is roughly half that amount. The reason for the lower hydrophobicity of the longest α-helices in polytopic proteins may be functionally related. The average $\Delta G_{\mathrm{Water}\Rightarrow\mathrm{Oil}}$ of an α-helix is −48 kcal/mol and −37 kcal/mol, for bitopic and polytopic proteins, respec-

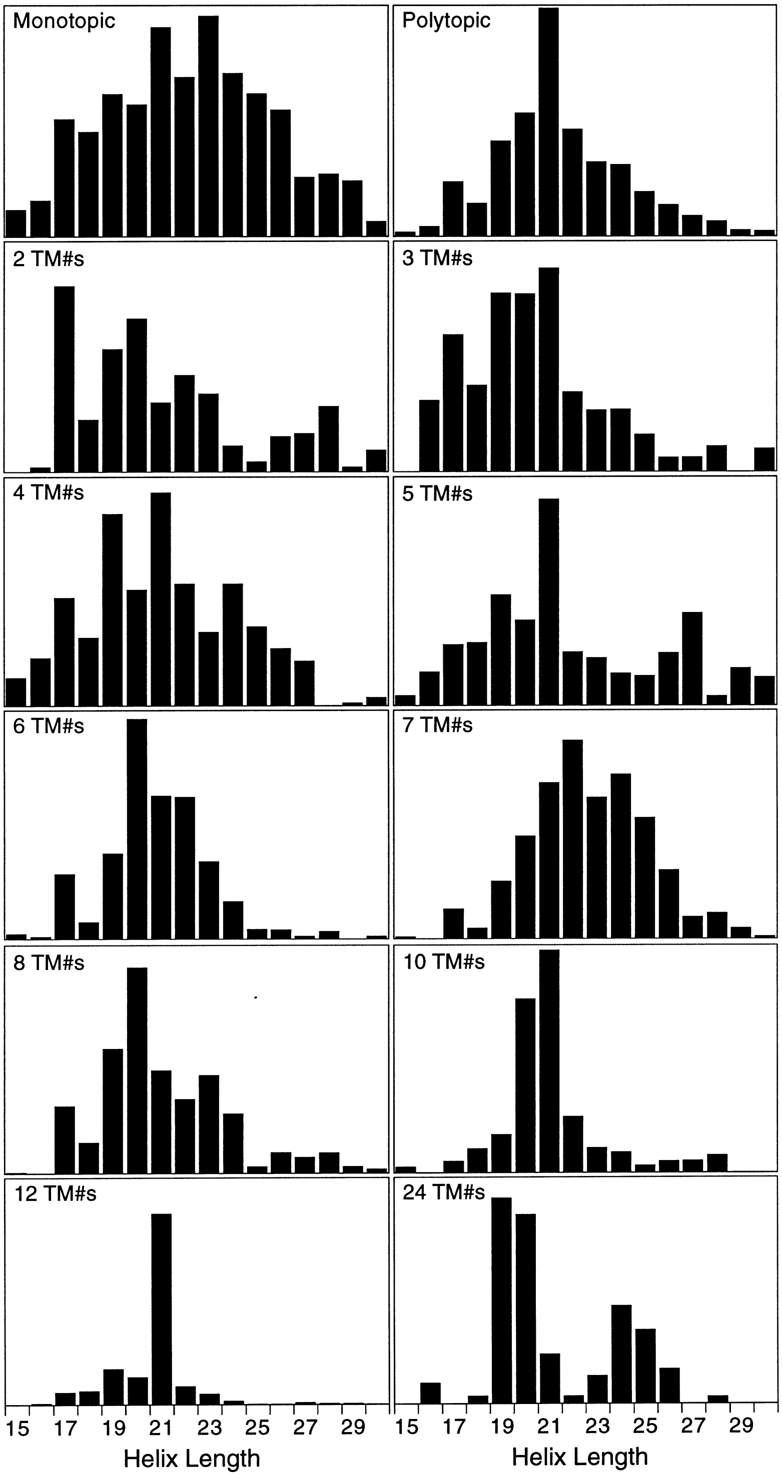


Fig. 1. Histogram of the length of transmembrane α-helices found in the weighted subset of TMbase25. The top two panels present the overall distribution for bitopic and polytopic proteins, while the other panels represent a breakdown by number of transmembrane segments.

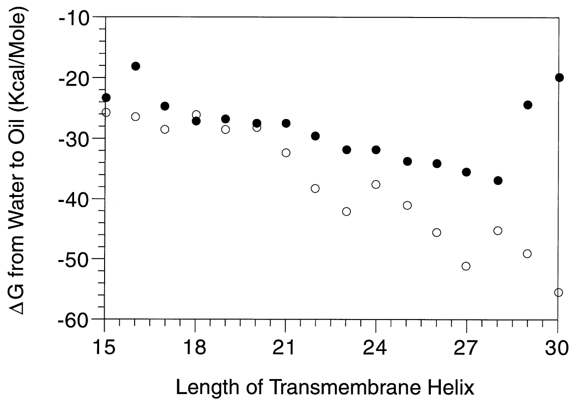


Fig. 2. Transfer free energy from water to oil ($\Delta G_{\text{Water}\Rightarrow\text{Oil}}$) as a function of α-helix length. Open circles represent transmembrane α-helices in bitopic membrane proteins while filled circles represent those occurring in polytopic membrane proteins.

tively. The increase in $\Delta G_{\text{Water}\Rightarrow\text{Oil}}$ in helices longer than 27 amino-acids may reflect the fact that such helices may represent two short helices grouped as one. Thus, the average hydrophobicity of a transmembrane α-helix is very high. A considerably smaller value is expected for membrane-associated soluble proteins that have been palmitilated, myristilated, farnesilated, geranylgeranylated or attached to phosphatidylinositol.

Why are these the transfer energies so high and are they required for biological function? One answer may be the need to accommodate a random mutation of a charged amino acid. For example, the most polar of all amino acids, arginine, has a $\Delta G_{\text{Water}\Rightarrow\text{Oil}}$ of 12.3 kcal/mol. Such a highly polar residue could be accommodated by increasing the average hydrophobicity of the other residues. This does not seem to be the case because there is no correlation between the sum of the contributions of all apolar amino acids and polar amino acids to the transfer free energy from water to oil (data not shown). Therefore, one has to conclude that the average residue of a transmembrane α-helix does not 'become' more hydrophobic when polar residues are present. Another consideration is that such high transfer free energies are used by the translocation machinery in the endoplasmic reticulum to ensure that only true transmembrane α-helices undergo translocation.

3.4. Amino acid composition of transmembrane α-helices

Among the hydrophobic amino acids ($\Delta G_{\text{Water}\Rightarrow\text{Oil}}$), no correlation is seen between the hydrophobicity and amino acid frequency in transmembrane helices. In other words, the hydrophobicity requirement of a transmembrane α-helix is readily served by a random distribution of hydrophobic amino acids. In such a random distribution the $\Delta G_{\text{Water}\Rightarrow\text{Oil}}$ of a 20-residue helix would be −47.8 kcal/mol, very close to the 'observed' average $\Delta G_{\text{Water}\Rightarrow\text{Oil}}$ of a transmembrane α-helix. The average $\Delta G_{\text{Water}\Rightarrow\text{Oil}}$ for an α-helix with a random distribution of both polar and apolar amino acids would be 27.4 kcal/mol. The maximum $\Delta G_{\text{Water}\Rightarrow\text{Oil}}$ (i.e. for a polyphenylalanine) would be −74.0 kcal/mol.

The relative frequencies of hydrophobic amino acids in soluble α-helices and transmembrane α-helices are very similar (see Fig. 3). This is a surprising result, since amino acids common in transmembrane α-helices such as isoleucine and valine have a high propensity for β-sheet formation [23,24].

Our study extends earlier work [9,25,26] by distinguishing between polytopic and bitopic membrane proteins. The more polar the amino acid is, the less likely it is found in bitopic proteins as opposed to polytopic proteins. This may be due to the ability of polytopic proteins to shield polar amino acids from the environment more efficiently than bitopic proteins. Note that the ionizable amino acids

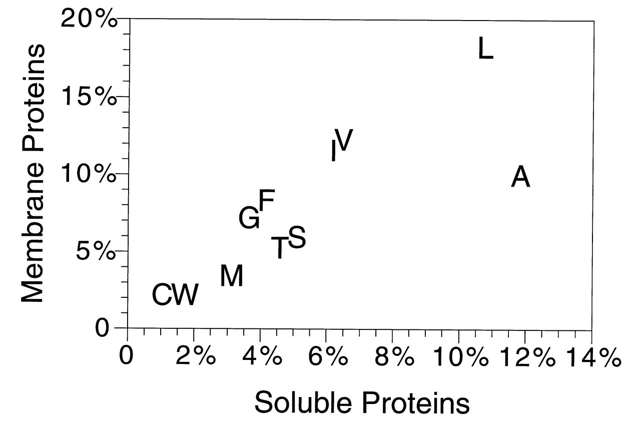


Fig. 3. Correlation plot of the amino acid frequency in transmembrane transmembrane α-helices and soluble proteins. Only hydrophobic amino acids ($\Delta G_{\text{Water}\Rightarrow\text{Oil}}$) were used in this correlation analysis.

in transmembrane domains are not expected to ionize [8].

3.5. Ionizable amino acid composition

Whenever a sequence of a transmembrane α-helix is inspected, the most conspicuous elements are polar residues. Even more outstanding are ionizable residues (having a very low frequency 2.5%) among all residues in transmembrane α-helices. This is not surprising as the energetic cost of inserting an ionizable amino acid in the hydrophobic environment of the membrane is very high. Nearly all membrane proteins with six or more transmembrane α-helices contain at least one ionizable residue.

Statistical analysis (data not shown) reveals that when taking into account all of the charges that would result if all ionizable residues ionize proteins would retain a positive charge. This result indicates that salt bridge formation is not a prevalent property of proteins containing ionizable residues. Note, that the ionizable amino acids in transmembrane domains are not expected to ionize [8].

Although the frequency of ionizable residues is low, they may play an important role in biological function and are most likely the functional groups responsible for protein activity. For example, in bacteriorhodopsin, protons are transferred from one carboxylate side chain to another [27]. One way that ionizable residues may be accommodated in membrane proteins would be to form salt bridges [8].

3.6. Distributions

In order to investigate the influence of the different alignment schemes, synthetic sequence data sets were created consisting of a normally distributed amino acid type within a large number of transmembrane α-helices of varying lengths L (15–30 residues),

$$F(x) = \frac{1}{(\sigma 30/L)\sqrt{2\pi}} e^{-(x-\mu)/2(\sigma 30/L)^2}$$

where $\sigma 30/L$ represents the width of the distribution (± one standard deviation). The relative abundance of a particular length L was set to that found in the TMbase25 database. In the first test case, the amino acid type is distributed symmetrically around the middle of the sequence ($\mu = L/2$, Fig. 4a). In the second test case, the amino acid type has a high probability at the left end ($\mu = 0$, Fig. 4e).

Fig. 4b,c,f,g demonstrates that depending on the alignment scheme, the distributions obtained from the synthetic data sets vary significantly. For the second case, when aligning sequences according to their carboxy termini, the result is a symmetrical distribution around the middle of the sequence, as opposed to the actual asymmetrical distribution (Fig. 4f).

Averaging of all four alignments yields a distribution which is significantly closer to the original distribution (Fig. 4d,h). However, small distortions produce localized artifacts (e.g. Fig. 4h). Thus, in categorizing distributions, we have differentiated between gradual distributions and small spikes. A small spike was ignored since there is no way to determine whether a spike is real effect or an artifact caused by the alignment scheme averaging.

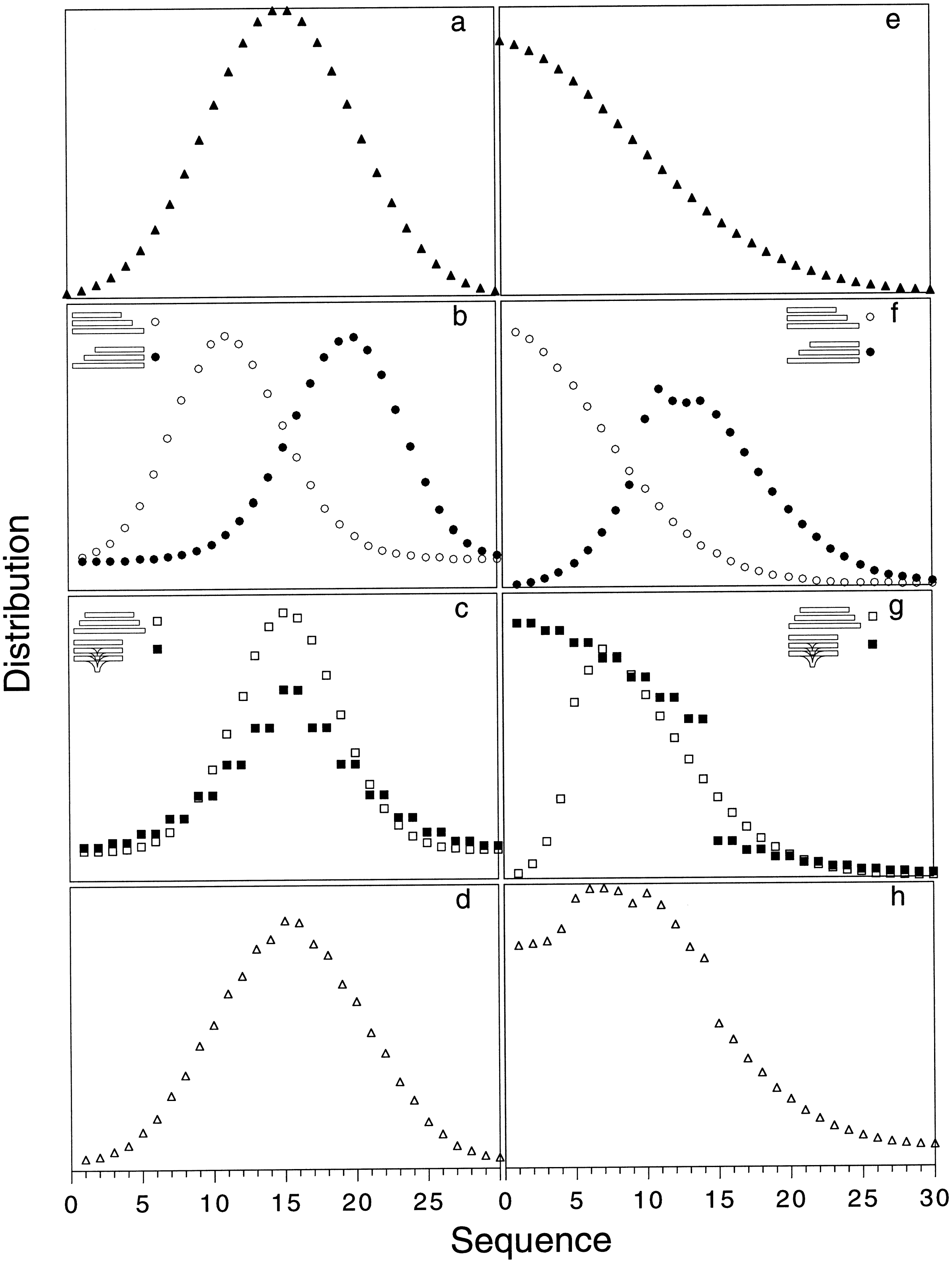
Fig. 5 shows the distributions for apolar and polar amino acids. There appear to be characteristic differences between bitopic and polytopic membrane proteins for some of the amino acids. These will be discussed in detail below.

3.7. Correlated distributions

Normalized probabilities of distances in sequence space were computed between every pair of amino

→

Fig. 4. Theoretical distributions for different alignment schemes. Normal distributions of an amino acid in transmembrane sequences of varying lengths (15–30 amino acids) were used to generate two synthetic datasets. Two different test cases are shown: that of an amino acid type distributed symmetrically (a), and that of an amino acid type distributed in a descending fashion from the beginning to the end of a sequence (e). The results of the four alignment schemes are shown in panels (b) and (c) for the first test case and in panels (f) and (g) for the second test case. Open circles represent amino alignment, closed circles carboxy alignment, open boxes median alignment and closed boxes ends alignment, respectively. The average of the four different alignments is depicted in panels (d) and (h).


a
e
b
f
c
g
d
h
Distribution
0
5
10
15
20
25
0
5
10
15
20
25
30
Sequence

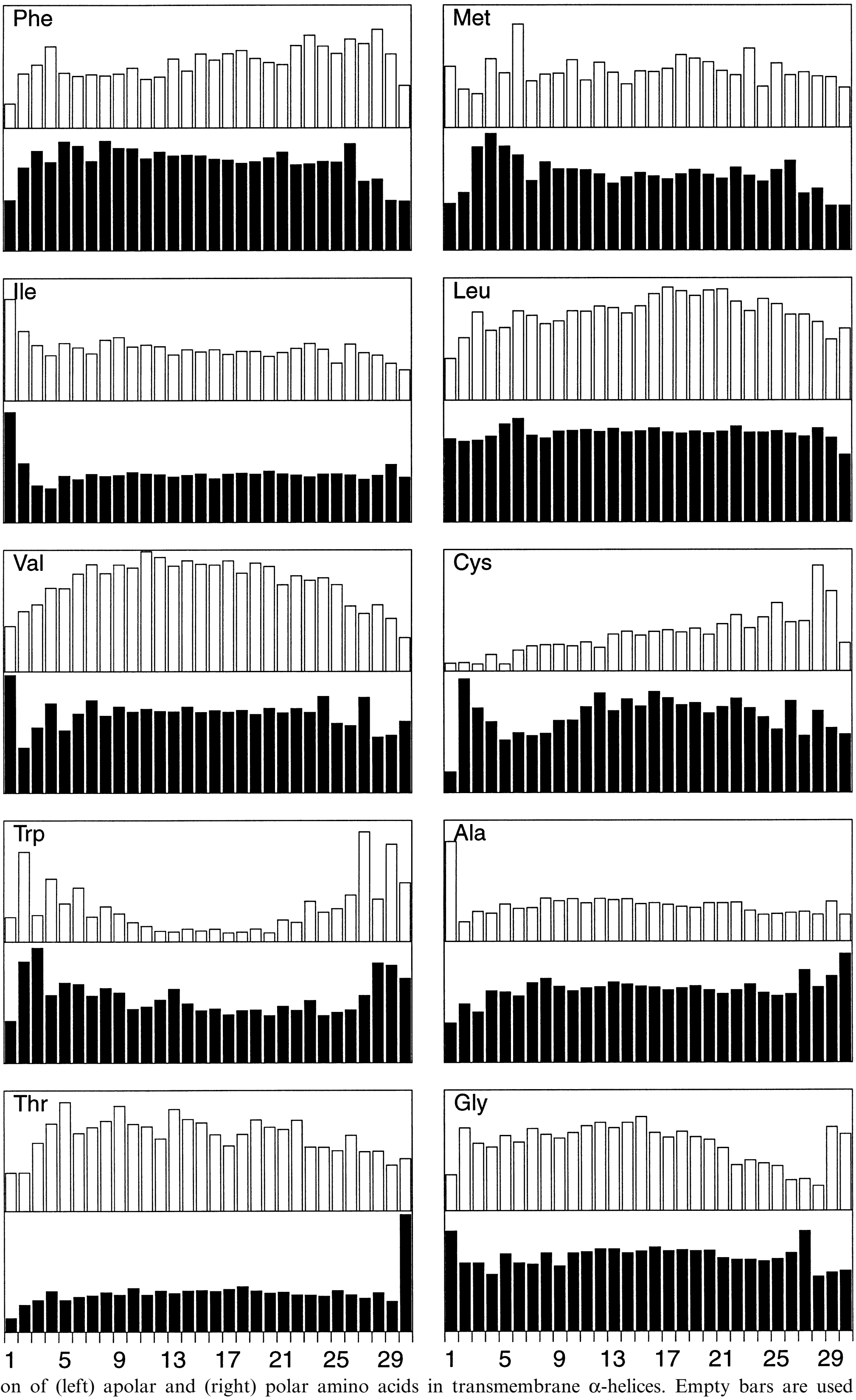


Fig. 5. Distribution of (left) apolar and (right) polar amino acids in transmembrane α-helices. Empty bars are used for bitopic proteins while filled bars are used for polytopic proteins.

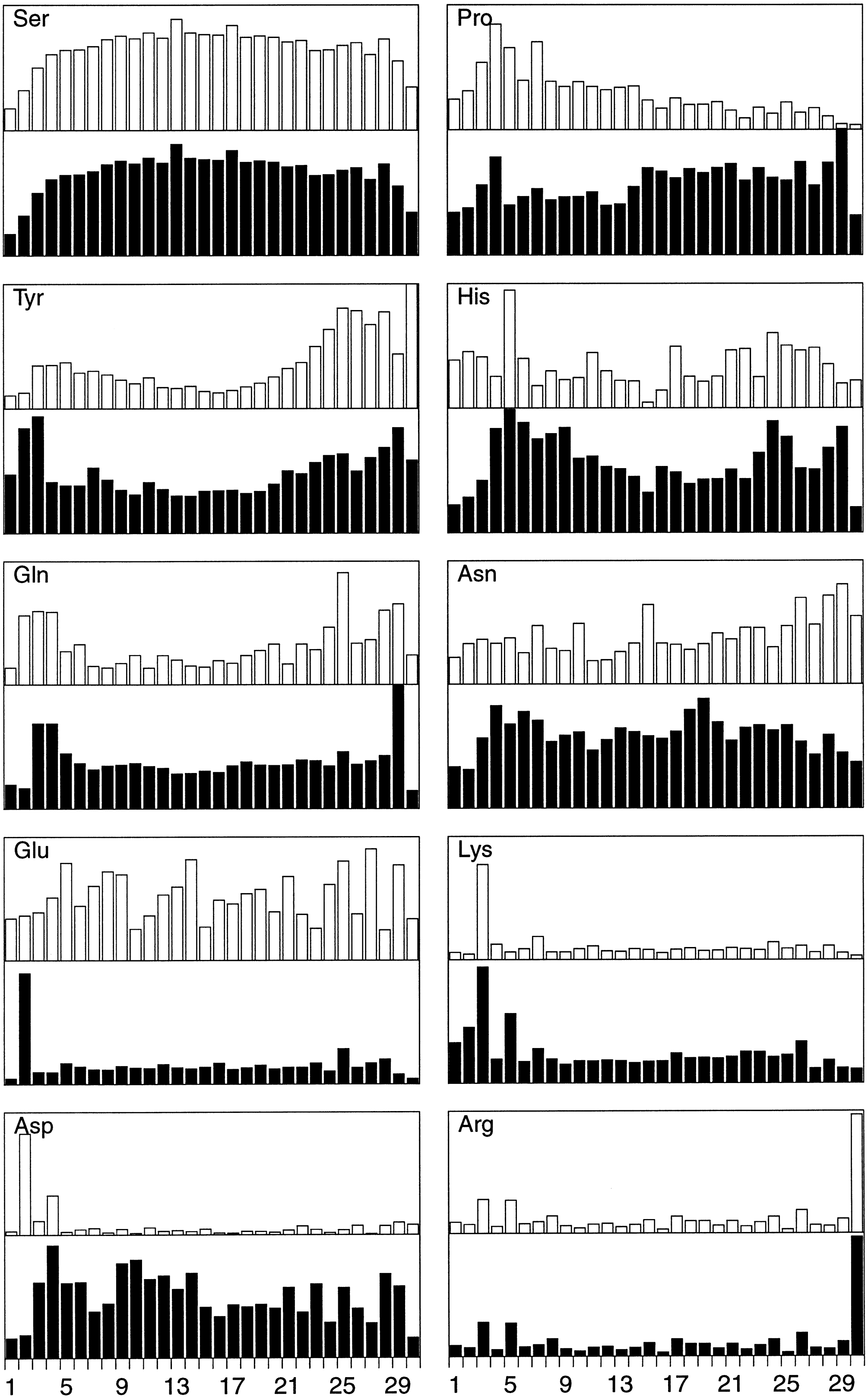


Fig. 5 (*Continued*).

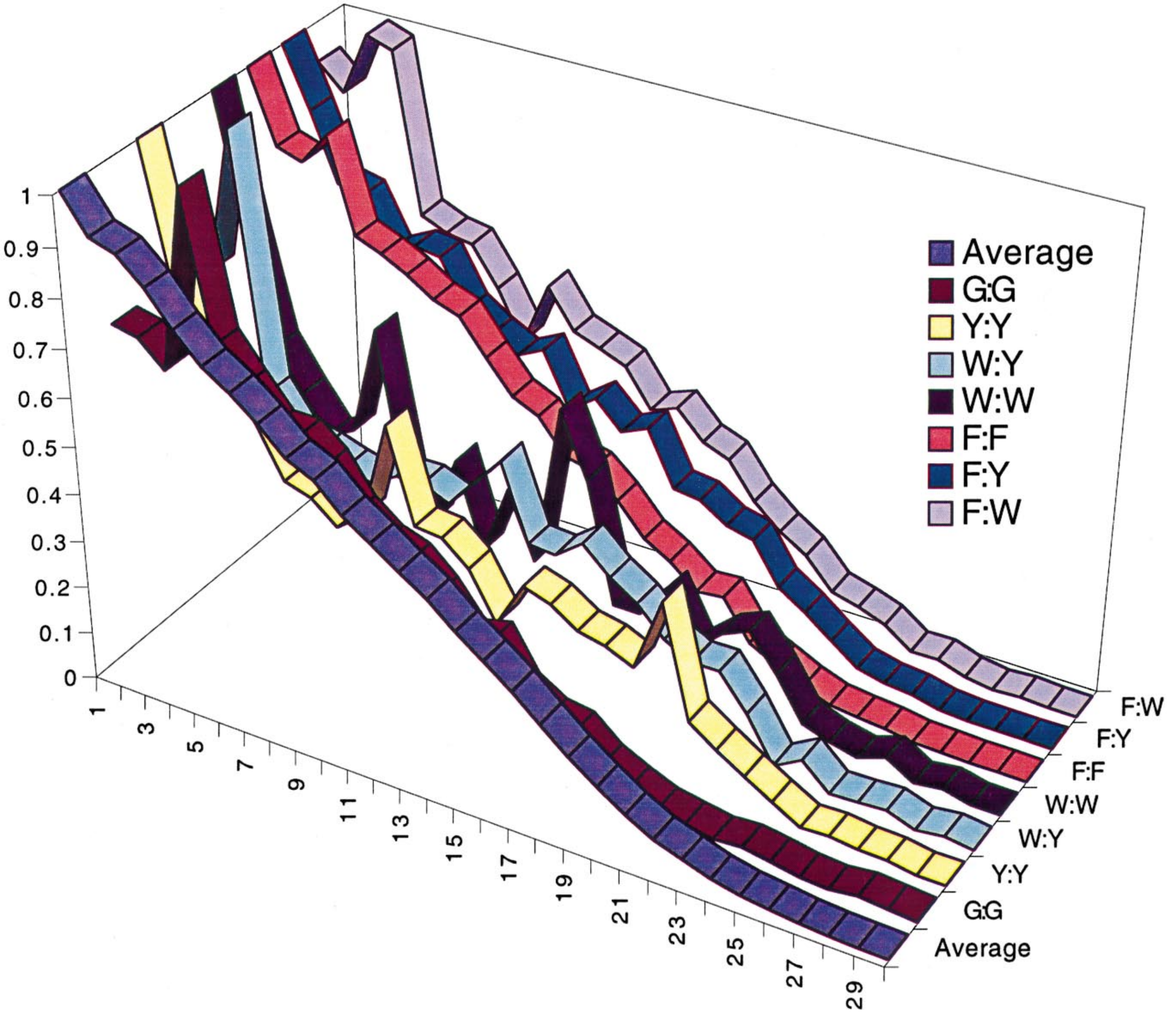


Fig. 6. Plots showing distance correlations in sequence space between particular pairs of amino acids that exhibited distributions differing significantly from a random distribution. The theoretical distributions expected for uncorrelated pairs of amino acids are indicated as an average distribution.

acids. Most amino acid pairs were distributed according to that expected for non-interacting residues except tryptophan, tyrosine, cysteine which showed a weak tendency to cluster with same type amino acids and glycine which shows a pronounced peak at a sequential distance of four residues apart (see Fig. 6).

3.8. Aliphatic amino acids

The aliphatic amino acids methionine, isoleucine, leucine, valine, and alanine represent 62% of all amino acids found in bitopic membrane proteins and 54% in polytopic membrane proteins (not shown). All aliphatic residues except isoleucine are randomly distributed in transmembrane α-helices (Fig. 5). Methionine is randomly distributed in bitopic membrane proteins as well. All other aliphatic amino acids are distributed in a non-random fashion in bitopic membrane proteins. The character of the non-random distribution of leucine and valine is the same, indicating that these amino acids are more prevalent in the middle of the bilayer than at the ends. The extent of the non-random distribution is most pronounced in the case of valine, where the probability of finding valine at the ends is 50% less than finding it in the middle. Isoleucine, on the other hand, exhibits a descending distribution. The correlated distributions of the aliphatic amino acids are the most uniform of all the amino acids, i.e. they are independently distributed (not shown).

3.9. *Aromatic amino acids*

While phenylalanine is distributed nearly randomly in both polytopic and bitopic membrane proteins, both tryptophan and tyrosine are distributed in a non-random fashion: they are found preferentially at the ends of the bilayer. Both of these distributions are more pronounced in bitopic membrane proteins compared to polytopic ones. The aromatic amino acids are the most favored of all amino acids to reside at membrane interfaces [28].

It is interesting to note that this non-random distribution of tryptophan and tyrosine occurs in membrane proteins with very different folds (Fig. 7). In OmpF [29], nearly every β-strand contains tryptophan and tyrosine only on one side (Fig. 7). In gramicidin [30] four of the last six residues are tryptophan. However, the active conformation of gramicidin is presumed to be a head-to-head dimer with tryptophan residues at both ends of the protein. Another example of clustering of aromatic residues is found in the photosynthetic reaction center [31].

The reason for the particular distribution of tryptophan and tyrosine residues could be explained by their ability to form hydrogen bonds as well as to exhibit hydrophobic character. These properties might target tryptophan and tyrosine to the interfacial region between water and the hydrophobic core of the bilayer. It should be noted, however, that aromatic residues have also been implicated in other roles in membrane proteins, such as providing a pathway for sugar [32] or drug [33] translocation.

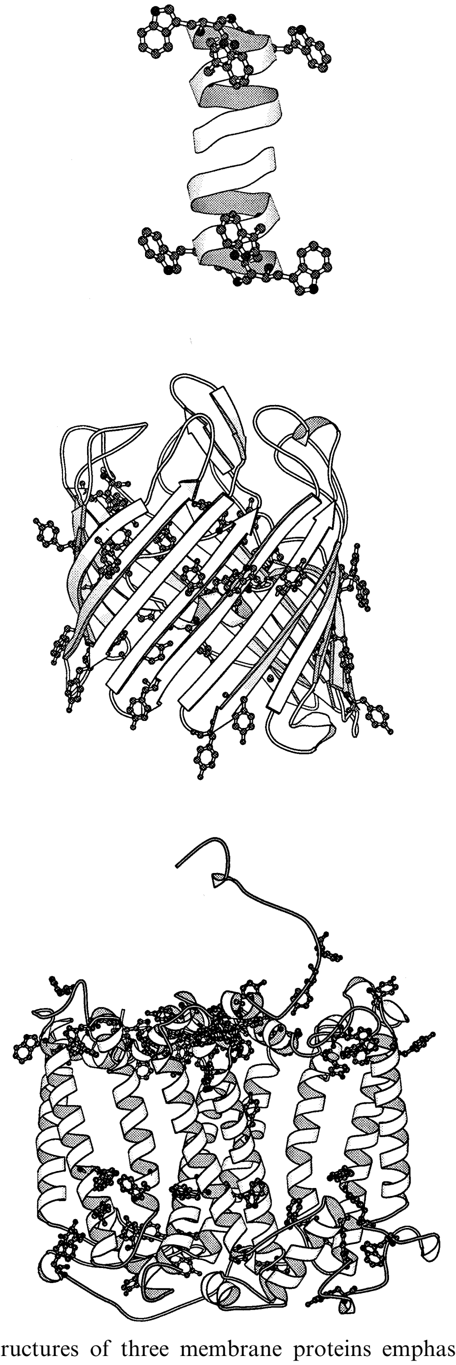


Fig. 7. Structures of three membrane proteins emphasizing the correlated distribution of tryptophan and tyrosine residues. Top panel: gramicidin, 1GRM [30], bottom panel: photosynthetic reaction center, 1PRC [31] and middle panel: OmpF, 2OMF [29]. Figure generated by MOLSCRIPT [42].

3.10. *Hydroxylic and sulfhydrylic amino acids*

The amino acids cysteine, threonine, and serine represent the most frequent polar residues in transmembrane α-helices (a combined total of 13.2%) (not shown). All of these amino acids are capable of partially satisfying their polar side chains by hydrogen-bonding to the carbonyl oxygen at position *i*-4 in the α-helix [34–36]. These amino acids exhibit convex distributions in both bitopic and polytopic membrane proteins (Fig. 5). The only exception is cysteine in bitopic proteins, exhibiting an ascending distribution.

The convex distribution of threonine and serine is surprising since one would have expected that these polar amino acids would prefer to reside in the polar head group region. It is difficult to assign a functional role for the particular distribution of these amino acids, since it may simply represent a compensation of the distribution of tryptophan and tyrosine residues. The ascending distribution of cysteines in bitopic proteins (Fig. 5) is unique among all amino acids.

3.11. Helix destabilizing amino acids

Both glycine and proline are known to be α-helix (and β-sheet) destabilizing amino acids in soluble proteins. Furthermore, both amino acids are frequently present in turns. When present in an α-helix, proline is known to induce kinks ranging from 20° to 30° [37], while glycines are prevalent in flexible linker regions. In membrane proteins, where the driving forces for α-helix formation are different than in soluble proteins, both proline and glycine are accommodated readily. Transmembrane α-helices are formed due to the necessity of satisfying backbone hydrogen bonds in the low dielectric environment of the lipid bilayer. Soluble α-helices on the other hand are thought to be formed in the process of hydrophobic collapse. Short polar stretches of amino acids rarely form stable α-helices in solution, due to the hydrogen-bonding partners present in the solvent. In transmembrane α-helices, kinks are found at several locations where prolines are found [27,31]. Both a dynamic and structural role has been suggested for proline residues in transmembrane α-helices based mostly on site-directed mutagenesis data [38].

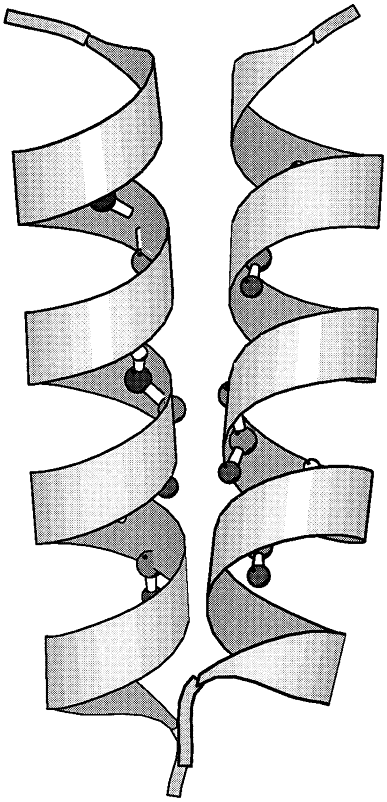


Fig. 8. Structure of the transmembrane domain of human glycophorin A determined by nuclear magnetic resonance spectroscopy in detergent micelles [39]. The glycine residues are depicted in ball-and-stick in order to emphasize their α-helical periodicity and the location in the protein-protein interface. Figure generated by MOLSCRIPT [42].

Proline and glycine are distributed randomly in polytopic membrane proteins (Fig. 5). On the other hand glycine exhibits a convex distribution and proline a descending distribution in bitopic membrane proteins. The glycine-glycine correlated distribution shows a pronounced peak at four residues apart (not shown). This could result from a preference of glycine to reside on the same face of an α-helix which has been observed in the structure of the dimerizing transmembrane α-helices of human glycophorin A [39] (Fig. 8). Glycine occurs at positions 84, 87 and 99, creating a groove in the α-helix that permits very close packing of the two helices. The statistical analysis may imply that such close positioning of α-helices at glycine interfaces may be a common feature of transmembrane α-helical packing.

4. Conclusions

The amino acid composition of transmembrane α-helices was analyzed in both bitopic and polytopic membrane proteins. In general, only small differences are present in both types of membrane proteins. However, these small differences may be attributed to the structural and functional differences between the two groups of proteins.

The observed small differences between bitopic and polytopic proteins are consistent with the two-stage model for membrane protein folding and oligomerization [40,41]. In this model, transmembrane α-helices in polytopic proteins are first inserted into the lipid bilayer as stable entities that later associate (in stage II of the model) to form the final folded protein. Therefore, even α-helices that are completely shielded from the lipid bilayer are predicted to be largely hydrophobic in nature.

The distribution analysis indicated that while the aromatic amino acids tyrosine and tryptophan are segregated to the ends of the helix, a large group of amino acids are distributed in a complementary fashion. It is difficult to infer from these results if one group compensates for the distribution of the other, or both require their own particular distribution for functionality. Several other amino acids exhibit other non-symmetrical distributions. This may imply an interaction with the α-helix dipole, or a possible role during folding, and/or translocation.

Analysis of correlated distributions between amino acid pairs indicated that most aliphatic amino acids are distributed independently from one another. Cysteine, tryptophan and tyrosine residues represent exceptions: if a pair of these amino acids is present in a transmembrane α-helix, then there is a high probability of finding them close together in sequence space. Glycine residues showed a preference for a distance of four residues from each other implying that residues prefer to reside on the same face of an α-helix.

Acknowledgements

The authors wish to thank Drs. D.M. Engelman, K.R. MacKenzie and P.D. Adams for helpful discussions and W.P. Russ for help in manuscript preparation. This work was funded in part by a grant from NIH (GM-54160-01) to A.T.B.

References

[1] K. Hofmann, W. Stoffel, A database of membrane spanning protein segments, Biol. Chem. Hoppe-Seyler 374 (1993) 166.

[2] J.M. Baldwin, The probable arrangement of the helices in G protein-coupled receptors, EMBO J. 12 (1993) 1693–1703.

[3] C.M. Fraser, J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley et al., The minimal gene complement of *Mycoplasma genitalium*, Science 270 (1995) 397–403.

[4] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick et al., Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, Science 269 (1995) 496–512.

[5] C.J. Bult, O. White, G.J. Olsen, L. Zhou et al., Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*, Science 273 (1995) 397–403.

[6] S.G. Oliver, From DNA sequence to biological function, Nature 379 (1996) 597–600.

[7] I.T. Arkin, A.T. Brunger, D.M. Engelman, Are there dominant membrane protein families with a given number of helices?, Proteins Struct. Func. Genet. 28 (1997) 465–466.

[8] D.M. Engelman, T.A. Steitz, A. Goldman, Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, Annu. Rev. Biophys. Biophys. Chem. 15 (1986) 321–353.

[9] C. Landolt-Marticorena, K.A. Williams, C.M. Deber, R.A. Reithmeier, Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins, J. Mol. Biol. 229 (1993) 602–608.

[10] H. Hauser, I. Pascher, R.H. Pearson, S. Sundell, Preferred conformation and molecular packing of phosphatidylethanolamine and phosphatidylcholine, Biochim. Biophys. Acta 650 (1981) 21–51.

[11] G.A. Adams, J.K. Rose, Structural requirements of a membrane-spanning domain for protein anchoring and cell surface transport, Cell 41 (1985) 1007–1015.

[12] P. Whitley, E. Grahn, U. Kutay, T.A. Rapoport, G. von Heijne, A 12-residue-long polyleucine tail is sufficient to anchor synaptobrevin to the endoplasmic reticulum membrane, J. Biol. Chem. 271 (1996) 7583–7586.

[13] U. Hobohm, M. Scharf, R. Schneider, C. Sander, Selection of representative protein data sets, Protein Sci. 1 (1992) 409–417.

[14] U. Hobohm, C. Sander, Enlarged representative set of protein structures, Protein Sci. 3 (1994) 522–524.

[15] F.C. Bernstein, T.F. Koetzle, G.J.B. Wiliams, E.F. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, The protein data bank: A computer-based archival file for macromolecular structures, J. Mol. Biol. 112 (1977) 535–542.

[16] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[17] L. Wall, R.L. Schwartz. Programming Perl, 1st edn., O'Reilly and Associates, Sebastopol, CA, 1991.

[18] L.J. Lis, M. McAlister, N. Fuller, R.P. Rand, V.A. Parsegian, Measurement of the lateral compressibility of several phospholipid bilayers, Biophys. J. 37 (1982) 667–672.

[19] V.A. Parsegian, R.P. Rand, Membrane interaction and deformation, Ann. NY Acad. Sci. 416 (1983) 1–12.

[20] R.P. Rand, Interacting phospholipid bilayers: measured forces and induced structural changes, Annu. Rev. Biophys. Bioeng. 10 (1981) 277–314.

[21] V.V. Kumar, Lipid molecular shapes and membrane architecture, Indian J. Biochem. Biophys. 30 (1993) 135–138.

[22] S. Munro, Localization of proteins to the Golgi apparatus, Trends Cell Biol. 8 (1998) 11–15.

[23] V. Munoz, L. Serrano, Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: comparison with experimental scales, Proteins 20 (1994) 301–311.

[24] M.B. Swindells, M.W. MacArthur, J.M. Thornton, Intrinsic ϕ, ψ propensities of amino acids, derived from the coil regions of known structures, Nature Struct. Biol. 2 (1995) 596–603.

[25] F.A. Samatey, C. Xu, J.L. Popot, On the distribution of amino acid residues in transmembrane alpha-helix bundles, Proc. Natl. Acad. Sci. USA 92 (1995) 4577–4581.

[26] G. von Heijne, Y. Gavel, Topogenic signals in integral membrane proteins, Eur. J. Biochem. 174 (1988) 671–678.

[27] R. Henderson, J.M. Baldwin, T.A. Ceska, F. Zemlin, E. Beckmann, K.H. Downing, Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy, J. Mol. Biol. 213 (1990) 899–929.

[28] W.C. Wimley, S.H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces, Nature Struct. Biol. 3 (1996) 842–848.

[29] S.W. Cowan, T. Schirmer, G. Rummel, M. Steiert, R. Ghosh, R.A. Pauptit, J.N. Jansonius, J.P. Rosenbusch, Crystal structures explain functional properties of two *E. coli* porins, Nature 358 (1992) 727–733.

[30] A.S. Arseniev, I.L. Barsukov, V.F. Bystrov, A.L. Lomize, Yu.A. Ovchinnikov, ^{1}H-NMR study of gramicidin A transmembrane ion channel. Head-to-head right-handed, single-stranded helices, FEBS Lett. 186 (1985) 168–174.

[31] J. Deisenhofer, H. Michel, Nobel lecture. The photosynthetic reaction centre from the purple bacterium *Rhodopseudomonas viridis*, EMBO J. 8 (1989) 2149–2170.

[32] T. Schirmer, T.A. Keller, Y.F. Wang, J.P. Rosenbusch, Structural basis for sugar translocation through maltoporin channels at 3.1 Å resolution, Science 267 (1995) 512–514.

[33] A.B. Pawagi, J. Wang, M. Silverman, R.A. Reithmeier, C.M. Deber, Transmembrane aromatic amino acid distribution in P-glycoprotein. A functional role in broad substrate specificity, J. Mol. Biol. 235 (1994) 554–564.

[34] T.M. Gray, B.W. Matthews, Intrahelical hydrogen bonding of serine, threonine and cysteine residues within alpha-helices and its relevance to membrane-bound proteins, J. Mol. Biol. 175 (1984) 75–81.

[35] L.M. Gregoret, S.D. Rader, R.J. Fletterick, F.E. Cohen, Hydrogen bonds involving sulfur atoms in proteins, Proteins 9 (1991) 99–107.

[36] I.T. Arkin, P.D. Adams, A.T. Brunger, S. Aimoto, D.M. Engelman, K.J. Rothschild, S.O. Schmit, Structure of the transmembrane cysteine residues in phospholamban, J. Membr. Biol. 155 (1997) 199–206.

[37] D.J. Barlow, J.M. Thornton, Helix geometry in proteins, J. Mol. Biol. 201 (1988) 601–619.

[38] K.A. Williams, C.M. Deber, Proline residues in transmembrane helices: structural or dynamic role?, Biochemistry 30 (1991) 8919–8923.

[39] K.R. MacKenzie, J.H. Prestegard, D.M. Engelman, A transmembrane helix dimer: structure and implications, Science 276 (1997) 131–133.

[40] J.L. Popot, S.E. Gerchman, D.M. Engelman, Refolding of bacteriorhodopsin in lipid bilayers. A thermodynamically controlled two-stages process, J. Mol. Biol. 198 (1987) 655–676.

[41] J.L. Popot, C. de Vitry, On the microassembly of integral membrane proteins, Annu. Rev. Biophys. Biophys. Chem. 19 (1990) 369–403.

[42] P.J. Kraulis, MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures, J. Appl. Crystallogr. 24 (1991) 946–950.