Journal of Computer-Aided Molecular Design, **15:** 533–552, 2001.
KLUWER/ESCOM

A sequence and structural study of transmembrane helices

Robert P. Bywater[a,*], David Thomas[b] & Gerrit Vriend[b,c]
[a]*Biostructure Group, Novo Nordisk A/S, Novo Nordisk Park, DK-2760 MÅLØV, Denmark;* [b]*Biocomputing Group, EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany;* [c]*Center for Molecular and Biomolecular Informatics, Katholieke Universiteit Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands*

Received 20 September 2000; accepted 26 March 2001

Key words: helix-helix crossing angle, membrane proteins, peptide bond angles, residue preferences, rotamers, supercoiling

Summary

A comparison is made between the distribution of residue preferences, three dimensional nearest neighbour contacts, preferred rotamers, helix-helix crossover angles and peptide bond angles in three sets of proteins: a non-redundant set of accurately determined globular protein structures, a set of four-helix bundle structures and a set of membrane protein structures. Residue preferences for the latter two sets may reflect overall helix stabilising propensities but may also highlight differences arising out of the contrasting nature of the solvent environments in these two cases. The results bear out the expectation that there may be differences between residue type preferences in membrane proteins and in water soluble globular proteins. For example, the β-branched residue types valine and isoleucine are considerably more frequently encountered in membrane helices. Likewise, glycine and proline, residue types normally associated with 'helix-breaking' propensity are found to be relatively more common in membrane helices. Three dimensional nearest neighbour contacts along the helix, preferred rotamers, and peptide bond angles are very similar in the three sets of proteins as far as can be ascertained within the limits of the relatively low resolution of the membrane proteins dataset. Crossing angles for helices in the membrane protein set resemble the four helix bundle set more than the general non-redundant set, but in contrast to both sets they have smaller crossing angles consistent with the dual requirements for the helices to form a compact structure while having to span the membrane. In addition to the pairwise packing of helices we investigate their global packing and consider the question of helix supercoiling in helix bundle proteins.

Abbreviations: Standard codes are used for amino acid residue types throughout: three-letter codes in text and single-letter codes in tables and figures and in sequences. TM – transmembrane.

Introduction

The database of membrane protein structures is growing steadily but it still only small – all-helix proteins include 14 bacteriorhodopsins, 1 halorhodopsin, 21 photosynthetic reaction centres, 7 cytochrome c oxidases, 5 ATP synthase subunits, 1 potassium channel protein, 1 cytochrome bc1 transmembrane subunit, 1 fumarate reductase, 1 gated mechanosensitive channel, 1 calcium ATPase plus 3 all-beta porins and two other all-beta proteins, according to the SCOP database [1]. In contrast, there is only one member of the very important superfamily of G-protein coupled receptors (GPCRs) whose structure has been determined to atomic resolution [2].

Furthermore there are many other families of proteins predicted to have transmembrane regions. When the database of known genomes is searched for transmembrane regions, many families of proteins are found [3–5]. Membrane proteins most often have only one transmembrane helix but families having up to about 15 (for prokaryotes) and 20 or more (for the

*To whom correspondence should be addressed.
E-mail: byw@bysse.novo.dk

eukaryote *C. elegans*) putative membrane-spanning helical segments have been found. For the prokaryotes (and even for *S. cerevisiae* which does possess a small set of GPCR genes) there was no noticeable peak at the value 7 for the number of transmembrane helices but the results [5] for *C. elegans* and *H. sapiens* were dramatically different; here there were prominent peaks at the value 7 (all predicted to be oriented with the N-terminus facing 'outwards' i.e. towards the exterior of the cell).

The critical importance of GPCRs in receptor research and ligand design within the pharmaceutical industry has led to intensive activity in modelling of these structures [6–19] (see also models deposited in http://www.gpcr.org/7tm/models/). The constructors of most of these models endeavour to take into account published experimental data such as mutation data [20, 21], ligand binding data [22] and analyses of mutation data, so-called correlated mutations [23, 24] (see also: http://www.gpcr.org/7tm/ligands/), site-specific labelling information [25–28] and low-resolution electron crystallography data [29]. Recently, these sources of structural information have been dramatically enhanced, if not entirely supplanted, by the publication of a 2.8 Å resolution crystal structure of bovine rhodopsin [1]. This structure is, however, not necessarily representative of all GPCRs and is anyway, as in the case of the electron crystallography structures, a dark (i.e., **in**active) structure, and we need to be able to predict, or determine the active structures of GPCRs. Hence there is still a requirement for modelling of GPCRs, and we wish to establish a set of guidelines for building such models. When it comes to the finer details of protein architecture such as side chain conformations and helix-helix crossover angles, it seems to have been taken for granted either that the other constraints applied in the model building will automatically adjust these correctly or, in the event that attention is paid to these details, it is assumed that the situation that prevails in globular proteins can be transferred to the case of the membrane proteins without any modification.

The case for the need for further modelling is not confined to the GPCR superfamily. There are families of proteins possessing many more helices than the seven that are traditionally associated with GPCRs. Ion channels in eucaryotes can typically have a dozen or more TM helices, although these are often in multiples of four or five. Drug transporters, a family of membrane proteins of enormous importance to the field of drug metabolism, consist of twelve or more TM helices in the same polypeptide chain. Nothing at all is known about the structures of the latter superfamily.

Many speculations have been made about the physicochemical characteristics of TM helices. It seems likely that the occurrence of residue types Gly and Pro in TM helices will influence helix geometry and the way in which a helix can interact with its environment, although there may be other explanations for the presence of Pro residues in some cases (see Results and discussion). In the case of membrane bundles the presence of multiple charged and polar residues in the core of the bundle will undoubtedly have an effect. However, no systematic studies have been made so far that relate characteristics of helices to these differences in amino acid residue contents and distribution. We present here a systematic study of a series of characteristics of helices. We compare the small set of helix-containing transmembrane proteins available in the PDB [30] with a set of 478 representative proteins solved at high resolution by crystallography; this set does not contain transmembrane proteins. To see if any of the observed differences resulted from the fact that transmembrane helices only occur in bundles in our dataset whereas the helices in the representative subset of the PDB occur in all possible environments we also analysed a set of water soluble, four-helix bundle containing proteins. We call these three datasets MEMBRANE, TOTAL, and BUNDLE respectively. The contents of these sets are described below.

1. *Membrane*: This dataset consists of all known (meaning deposited in the PDB and not 'on hold') transmembrane proteins. This dataset contains a few proteins, such as the porins, that do not contain helices, and those proteins were not included in this study. The dataset is otherwise not 'carefully selected' but contains some redundancy. We decided that accepting redundancy was in this case a lesser evil than having too small a dataset. It should be kept in mind that these membrane proteins are solved from much lower resolution data than the proteins in the other two datasets. Also, although some of the proteins in this set are very similar their structures were solved by different groups using slightly different methods, so that the redundancy is somewhat alleviated by the fact that the freedom within the boundaries of the data may have been sampled differently. We are aware that, for example, the structures for the photosynthetic reaction center proteins, PDB i.d.'s 1pps, 1pcr, 1prc, 4rcr and 1yst, are very similar. Where the

advantage accrues in considering all of them is that a slightly greater expanse of protein sequence variability space is thereby explored. For any fold type it is of interest to know which residue types can map to that fold, which positions tolerate variability and which require conservation of residue type. In that sense the redundancy that we have included enriches the data set.

2. *Total*: This dataset consist of 478 proteins that are solved by crystallography, at $\leq$2.1 Å and with an R factor $\leq$20.0, do not have more than two chain breaks, and are not obviously wrong or poorly solved according to the WHAT_CHECK software[31]. No two proteins in this dataset have a pairwise sequence identity $\geq$30%. Although no special action was taken to filter out membrane proteins from this database, there are no membrane protein structures of sufficiently high quality to meet the above stringent selection criteria.
3. *Bundle*: This dataset contains all PDB entries found in the FSSP file [32] for the PDB entry 1ROP [33]. The ROP 4 helix bundle was selected because its four helices superimpose on the helices 1, 2, 3 and 7 of the bacterial heptahelical membrane protein bacteriorhodopsin, although the connectivity is different. Note that the '4 helix bundle' protein families contain some proteins, which contain more helices than those that make up the bundle itself.

Residue preferences

The issue of residue preferences in membrane spanning helices has been addressed for the case of synthetic peptides [34, 35] and certain differences were observed, in particular, there was the surprising finding that the β-branched residues valine and isoleucine were amongst the most predominant residues in these peptides although they are not normally regarded as having a high helix-forming propensity [36]. It would seem likely, based on the data of Deber *et al.* [34–35], that residue preferences are different in water-soluble globular proteins and in membrane proteins. We wished to see whether this pattern of residue preferences was typical for membrane proteins.

Three dimensional nearest neighbour contacts along the helix

An α-helix is characterised by a rise of 1.5 Å per residue along the helix axis representing 3.6 residues per complete turn. This means that the side chain of the $(i + 4)$th and to some extent the $(i + 3)$th residue will in 3D be located in the vicinity of the side chain of the ith residue. It would therefore be anticipated that these residues would be physically compatible with one another and are likely to be in physical contact. The clustering of residue types in this way can play many roles such as stabilising the helix, establishing contacts with other helices, embedding and orienting the helix correctly in the membrane, and there may well also be important functional requirements for mutual arrangements of side chains that are dictated by the helical geometry of the backbone.

Table 1a. Favoured interhelical crossing angles and their class assignments according to Chothia et al. [41]. The 1, 3 and 4 refer to the displacement in units of residues which defines the groove into which one helix docks onto the other

Class	(Corresponding secondary structure type)	Ideal packing angle (°)	Range of angles observed (°)
4–4		−52	−80 to −20
3–4		+23	0 to +40
1–4		−105	−110 to −100
3–3	(3_{10} helices)	−109	∼110

Rotamer analysis

Experience gained with the use of rotamer databases in globular protein modelling would lead us to suggest that assumption, mentioned above, that the best rotamers will be found automatically represents an inadequate strategy if the aim is to construct the best quality models. But the second assumption, may be dangerous too because there is no *a priori* way of knowing whether the immersion of polypeptide chains in a lipid environment will preserve the same pattern of rotamer distributions as is observed in water soluble globular proteins. In this paper we use the currently available set of published structures of membrane proteins in order to ascertain whether or not this pattern is preserved. Because of the low resolution at which membrane protein structures normally are solved, rotamers of residues with small side chains are considerably less precise than those with larger side chains that are more clearly defined in electron density maps.

Table 1b. Favoured interhelical crossing angles and their class assignments derived from the data in Figure 7 of Walther et al. [42]

Frequency	Packing angle (°) range observed
Frequent	+110 to +140
Frequent	−70 to −20
Intermediate	−170 to −150
Intermediate	−110 to −90
Rare	+10 to +60
Rare	+160 to +180

Peptide bond and helix distortion

It has been shown [37] that solvent accessible peptide bonds in water-soluble globular proteins undergo a 1°–3° compression of the OCN bond angle and a corresponding expansion of the NCC_α bond angle relative to buried peptide bonds. In amphiphilic helices this will lead to a curvature of the helix, with the hydrophobic face on the inside of the toroidal section that results. The implications, *inter alia* for model building, of these departures of bond angles from the supposed ideal geometry have also been discussed by Karplus [38]. While membrane spanning peptides are not amphiphilic as such, they may not always have a uniform environment around their entire surface; many of the helices that pack together, e.g., in GPCRs, have an uneven distribution of hydropathic character on their surfaces and can bury hydrophilic side chains in a folded membrane protein [13, 14]. These helices may therefore need to be curved as a result of this.

Helix-helix crossover analysis

Helix-helix interactions play a large role in stabilising membrane proteins and many membrane-spanning peptides also show a tendency to aggregate [39, 40]. It is therefore important to understand how helices associate with one another. One of the parameters of interest is the helix-helix crossing angle. This topic has received wide coverage in the literature [41, 42]. Chothia et al. [41] developed a 'ridges-into-grooves' model while Walther et al. [42] reverted to the more classic 'knobs-into-holes' model [43]. The results of these methods are somewhat different and are summarised in Table 1a and Table 1b respectively.

Bowie [43] examined a small database of helical membrane proteins and showed that a crossing angle of $+20°$ was most common, in contrast to the globular proteins where packing angles of around $-35°$ are most prevalent. The transmembrane helix crossing angle falls into the '3–4' class of Chothia et al. [41] (Table 1a) and corresponds to the sparsely populated '+10 to +60' class of Walther et al. [42] (Table 1b). We wish to see how well this pattern is preserved in a larger dataset of membrane proteins.

Bowie [43] also considered the helix tilt angle and the lengths of helices. These are almost the same thing since the angle of tilt will influence the length of helix required to completely span the membrane, unless the helix is interrupted in some way or highly over- or underwound or 'kinked'. In the membrane protein database there is no case of an α-helix converting into an extended, turn or coiled structure inside the membrane, although there are examples of an α-helix switching briefly to a π-helix in TM6 and stretches with $\mathbf{3}_{10}$ geometry in TM5 and TM7 respectively of bovine rhodospin (PDB i.d. 1f88). The number of residues required to span a membrane of 35 Å thickness (bacterial membranes are typically thinner than that, and eucaryotic membranes can range from 35 to 50 Å) perpendicular to the bilayer surfaces would be 23 if the rise per residue is the standard value of 1.5 Å. Helices shorter than that would not span the membrane, yet there are a few, according to the data of Bowie [44]. In the histogram of helix length (see http://www.cmbi.kun.nl/gv/service/helices) the 20–25 and 25–30 bins are most populated. In the crystal structure of bovine rhodopsin [1] the helices are 29, 29, 33, 32, 27, 32 and 20 residues long, and the most tilted ones are expected also to be the longer ones by virtue of the need to span the membrane. A *caveat* is needed here, since helices may extend beyond the membrane, e.g., TM3 in the 1f88 structure protrudes about 2 residues on the extracellular side and 5 residues on the cytosolic side, depending on exactly where the limits of the lipid bilayer are located.

We examine this set of membrane protein helical structures to extract the helix-helix crossover angles, and contrast these with the crossover angles in water-soluble 4-helix bundle containing proteins.

Overall helix packing

As stated earlier, there have been several excellent studies [41, 42] of helix-helix crossing angles. However in all of these studies, helix-helix crossing was

considered as an event occurring at a single site on each helix, as if the helices were straight and rigid. In reality it is very unlikely that helices pack in this way. Studies on fibrous proteins with a predominantly helical content by Astbury [45] showed the appearance of strong X-ray reflections at 5.1 Å instead of 5.4 Å. This was the first experimental demonstration that right-handed α-helices in bundles form a left-handed supercoil and since that time supercoiling has been demonstrated in numerous helical bundle proteins. The rules for α-helix packing have been established by Crick [43]. With supercoiled helix packing there is an extended contact area rather than a contact point. The most familiar everyday supercoiled object is rope. Each strand that is wound together to make the rope is in contact with another strand over their entire length. It is easy to understand this intuitively and the concept is easy to visualise graphically by inspecting the shape of helices in membrane proteins, especially when they have been surface-mapped or presented as splined, curved cylinders. A bundle of helices touching at only one point near their centres would splay out at the ends and generate a hyperboloid-like object, whereas what one observes in most helix bundles is something more akin to a prismatic ellipse with a twist as one proceeds along the major axis of the molecule. It is this twist that we wish to investigate.

We would expect supercoiling to allow not only better helix packing and core formation but also minimise the surface area of the protein. Globular proteins are well-packed and present minimal surfaces[1] to their surroundings. There is no *a priori* reason to expect membrane proteins to behave differently. Helix bundle proteins may be regarded as truncated fibrous proteins in which the ends of the helices happen to be connected by loops, in which case the supercoiling principle can be directly 'borrowed' from the fibrous protein field. In fact, supercoiling is already well known in globular proteins such as leucine zippers, tRNA synthases and hemaglutinin. The observed pitch angle values[2] have been shown [46] to be dependent on the number of helices in the bundle, varying from values near 150° for two-stranded coiled coils to values closer to 200° for three- and four-stranded bundles. The consequences of these observations should not be lost on model builders who construct models of helical membrane proteins.

It has been shown [47] that the contacts between adjacent helices in membrane proteins conforms to a 'knobs-into-holes' packing arrangement [48] with extended contacts along considerable stretches of the surfaces of helix pairs in a manner characteristic of coiled coils. Sansom et al. [49] carried out multiple *ab initio* calculations of a 7TM membrane protein structure using simulated annealing via restrained molecular dynamics and found recurrent examples of stable structures characterised by a '4–1' core, in which four helices form a distorted left-handed supercoil around a central, buried helix while the remaining two helices pack onto the outside of the core in a manner very reminiscent of GPCR structures as we understand them to be constructed.

Methods

Database searches

The WHAT IF program [50] provides two means to scan sets of PDB files. The simplest method is to use the WHAT IF relational database generation module, and make a database that contains just the files of interest. This allows the user to perform all queries[3] provided by the WHAT IF software.

The queries not foreseen in the design of the WHAT IF relational database query module can be performed by automatically running WHAT IF scripts over a set of files, or by hard-coding the query in the HTML page generator. Generation of a complete relational database with all query possibilities enabled takes between 1 and 36 hours CPU for the databases discussed in this article.

Residue preferences

Two kinds of residue preferences were analysed using the three relational databases that were generated using the WHAT IF database generation module.

[1] 'Everything tries to become round' according to Black Elk, medicine man of the Oglala Sioux tribe [45].

[2] Pitch angle is calculated [46] as the average of $P_{AB} = 2\pi$ [**R**/tan (Θ_{AB})] where Θ_{AB} denotes the dihedral angle between the A and B chains, **A** and **B** are parallel vectors along A and B respectively spanning over seven residues and **R** is the orthogonal vector joining their mid-points. $P_{12} = 2\pi$ [**D**/Θ_{12}] where Θ_{12} denotes the dihedral angle between the normals to A and B chains at their crossover point and **1** and **2** are parallel vectors across A and B respectively at a spacing of seven residues and **D** is the orthogonal vector joining their mid-points. Ref. 46 should be consulted for further details.

[3] Additional code required specially for this paper to deal with helix crossover and nearest neighbour queries is available on request.

First we simply looked at preference parameters for residues being in the helical state. The output modules of the WHAT IF relational database automatically provide preference parameters relative to a model in which the normalised frequency of each residue type is used as the random model. The frequencies of the four secondary structure states in the entire PDB are used as the random models for the preference parameter calculations. Preference parameters are calculated using

$$\text{Pref} = \ln (F_{observed}/F_{predicted})$$

in which Pref is the preference parameter, $F_{observed}$ is the number of hits found by the database query and $F_{predicted}$ is the number of hits that was predicted based on the random model assumptions.

Three dimensional nearest neighbour contacts along the helix

The search for $i-i+3$ and $i-i+4$ contact neighbours was performed using a small module written in the WHAT IF HTML page generation menu. We simply scan for helices and look at the frequency of occurrence of all 400 possible $i-i+3$ and $i-i+4$ pairs in helices. The results are converted into preference parameters using the occurrence frequency of each residue type in the helical state as the random model.

Rotamer analysis

We used position specific rotamer distributions as these have proven to be the best way of analysing rotamer preferences. This method was originally proposed by Jones and Thirup [51], and has been used extensively for modelling, structure validation and in proposing mutations in protein engineering [52, 53]. This methodology has been described in detail elsewhere [51, 53], and will only be described briefly here.

A rotamer distribution for a certain residue type at a certain position – called a position specific rotamer distribution – is determined by extracting from a protein structure database all suitable fragments of 5 residues. Suitable fragments are those that have a local backbone conformation similar to the one around the evaluated position, and have the same residue type at the central position. In the present study, the RMS deviation of the backbone α carbons is maximally allowed to be 0.5 Å.

Peptide bond angles

WHAT IF contains a module that analyses three atom angles in the backbone of peptides. This module lists the four backbone angles N-Cα-C, Cα-C-N, Cα-C-O and O-C-N. This module was used to analyse the backbone angles in the helices in the three datasets as function of the solvent accessibility of the residue. Three accessibility ranges were used: $0.0 \rightarrow 1.0$ Å^2 to represent buried residues, $10.0 \rightarrow 999.9$ Å^2 to represent accessible residues and $1.0 \rightarrow 10.0$ Å^2 as the intermediate class for which no clear-cut decision can be made as to whether to classify them as buried or accessible.

Helix-helix crossover analysis

The program WHAT IF contains a module that determines the cross-over angles between helix axes. Helix axes are defined as the straight lines that connect the centers of the helical backbone cylinders at the height of the N-cap and the C-cap position. The cross-over angle is the torsion angle between those straight lines.

Not all alpha helices are straight however. If there is a curvature spread along the entire helix (as is the case typically in supercoiled helices as discussed elsewhere) this still allows an average axis to be computed which is too all intents and purposes identical to that in the correspondingly oriented 'ideal' alpha helix.[4]

However, some helices are kinked, or bent. By this is meant that the helix is interrupted at one or more residues by stretches of polypeptide not having typical torsional angles for an α-helix. In such a case the helix axis of the two stretches of helix flanking the kinked or distorted region can be determined (an example of this is shown in Table 2c). The issue of kinks is helices is complex and is discussed fully in.

Results and discussion

Overall helix packing

We examine four-helix bundles which we define as a group of four helices of which each of the six pairwise combinations show at least six inter-residue contacts whereby a contact is defined by a distance of $\leq$2.5 Å. However, one of the six possible helix-helix pairs in the bundle is allowed to have only

[4]In the case of rope, already mentioned, the average helical axis of the individual super-wound strands is co-linear with the axis of the rope itself.

three residue-residue contacts. Helices can be part of multiple four-helix bundles, the only condition for accepting a bundle was that at least one (in practice this turns out to be at least two, of course) helix is different from any subset of four helices that was analysed previously.

A superhelical model is fitted to each four-helix bundle using the following algorithm. First, each helix is modelled individually using the established method of expanding as a series of Chebyshev polynomials orthogonal on the C_α positions [54]. This gives a continuously differentiable mathematical model of the helices, including a smooth line, which fits down the centre of each helix. This enables the points of closest approach, the so-called contact points, to be determined by straightforward geometry. A check is made to ensure that the contacts do not occur at the ends of one or more helices. This is normally not the case, and the angle with which the two smoothly curved fitted lines cross is calculated and called the crossing angle. Then the first helix in the bundle is chosen as a reference. Helices which subtend less than a right-angle to it are all classified as 'up', and all of the others are classified as 'down'. Making use of this classification, an estimate of the mean line of the putative superhelical bundle is calculated as a signed average of the lines of the helices. The crossing angles of the helices to this superhelical axis are worked out and averaged. This gives a mean twist angle of the superhelix. This and the mean radius of the superhelix together define the repeat length of the superhelix, which are plotted as a scatterplot. Because of the relationship between the quantities plotted on the two axes, the scatterplot is expected to have a hyperbolic shape, with scatter away from the hyperbola representing differences in radii of different bundles.

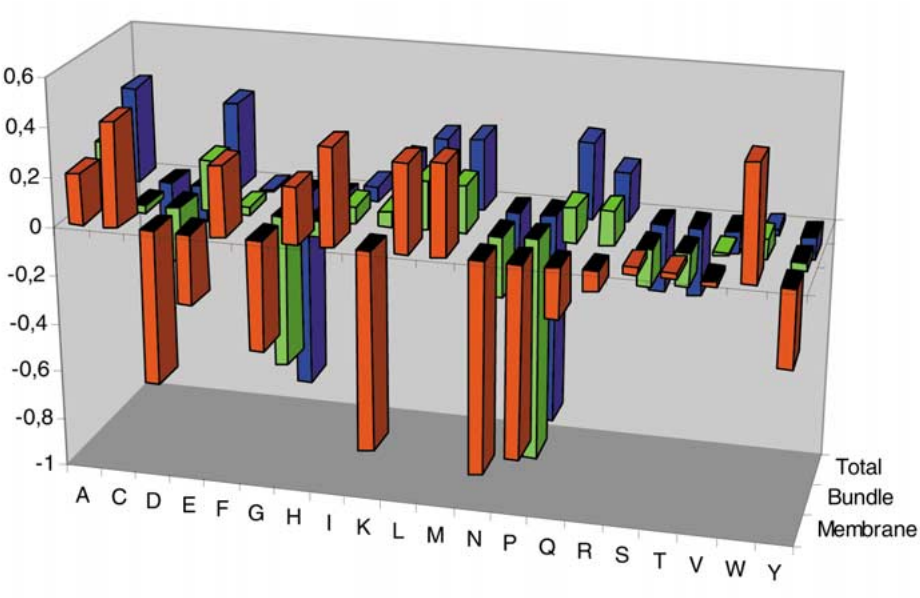
The full data can not be presented here for space reasons. The reader is referred to the http://www.cmbi.kun.nl/gv /service/helices/ for the original data, and only a summary is provided here of those data which warrant a fuller discussion. One general *caveat* that must be borne in mind throughout this and any other discussion on membrane protein structure is the fact that the resolution of the crystallographic structures is rarely better than 2.5 Å (e.g., that of 1f88 is 2.8 Å). Ideally, this should of course be better. An excellent structure with resolution 1.55 Å is however available for bacteriorhodopsin (1c3w) [55].

Residue preferences

Figure 1 shows the results of the determination of helix preferences of the 20 amino acids in the three datasets. Reduced preferences are observed in the MEMBRANE dataset for the charged residues and Asn and Gln as can be expected from the fact that every helix bundle has more residues pointing outwards than residues pointing into the internal cavities formed between the helices. Similarly, it is not surprising that the hydrophobic residues Cys, Phe, Ile and Trp occur more often in the MEMBRANE helices than expected from experience in water-soluble globular proteins. The fact that the weak hydrophilic residues Ser, Thr and His have positive preference parameters in the MEMBRANE set can be explained by the fact that transmembrane bundles often contain cofactors or active sites which utilize the functionality (ligand recognition/binding for example) of these residue types, or simply, that these residue types can participate in hydrogen bond networks that serve to stabilize the protein.

The preference parameters for Gly and Pro are somewhat less negative in the MEMBRANE set than in the two other sets. The differences between the TOTAL dataset and the BUNDLE dataset are much smaller than the differences between the MEMBRANE dataset and the other two sets. Therefore it can be concluded that these differences reflect the location of the MEMBRANE proteins inside the membrane more than that they reflect the differences between helices in helix bundle proteins and helices in other proteins. Some of the differences can also be attributed to the frequently observed cofactors and ligand binding sites located between the transmembrane helices.

It is difficult to conjure up any good measure to describe any preference parameter for residue types in a helix, since there is no obvious *a priori* way to agree about the proper normalisations. We just list the numbers in the database and their interpretation is left largely to the reader. Other studies, described as 'systematic' are no more realistic in that these have not been normalised either, but they can be referred to for comparison with our data. For example, Deber et al. [34, 35] found that for synthetic membrane-spanning peptides the helix propensity correlates with hydropathy, and that this was the dominant factor in determining helicity, overcoming the tendency found in globular proteins for the β-branched residues Val and Ile to eschew the helical option for secondary



Original data in tabular form.

Residue type	Bundle	Membrane	Total
A	0.25	0.21	0.39
C	−0.03	0.43	−0.29
D	−0.22	−0.63	−0.15
E	0.21	−0.29	0.36
F	0.03	0.29	0.01
G	−0.62	−0.45	−0.81
H	−0.07	0.23	−0.14
I	0.07	0.4	0.06
K	0.06	−0.82	0.18
L	0.2	0.36	0.28
M	0.19	0.37	0.29
N	−0.25	−0.87	−0.29
P	−0.91	−0.79	−0.87
Q	0.14	−0.2	0.31
R	0.14	−0.08	0.2
S	−0.15	0.03	−0.27
T	−0.14	0.02	−0.28
V	0.01	−0.02	−0.09
W	0.08	0.47	0.02
Y	−0.03	−0.31	−0.09

Figure 1. Relative frequency of residue types in the three datasets. Red: membrane set; green: bundle set; blue: total set.

structure. A similar conclusion was arrived at by Thorgeirsson et al. [56], who measured the membrane-insertion propensities of 14 amino acid residue types inserted individually at the same position in a 25 residue membrane-spanning peptide and found that they correlated with peptide hydrophobicity. This is perhaps not surprising, since what these authors are measuring is the relative affinities for different amino acid residue types to exist in a membrane environment, rather than the frequency of residue types that actually turn up in membrane-spanning peptides. Indeed, hydropathy is evidently not the only stabilising factor since other residue types such as Gly, Pro, Ser, Thr and Asn are encountered in membrane helices. That they are tolerated in this environment can in most instances be ascribed to the fact that there is little water present to initiate unfolding, e.g., at the C = O group of the $(i-4)$th residue, rendered naked by the blocked N atom of the Pro residue at the ith position. Their conservation in these helices may well imply functional roles for these residue types, as discussed below.

Considering now preferences for all residue types, these are, for the MEMBRANE set, in order:

$$\mathrm{W} > \mathrm{C} > \mathrm{I} > \mathrm{M} > \mathrm{L} > \mathrm{F} > \mathrm{H} > \mathrm{A} >$$

$$S > T > V > R > Q > E > Y > G > D >$$

$$P > K > N$$

whereas Deber et al. [34, 35] found (only 13 residue types studied):

$$\mathrm{I} > \mathrm{L} > \mathrm{V} > \mathrm{M} > \mathrm{F} > \mathrm{A} > \mathrm{Q} > \mathrm{Y} > \mathrm{T}$$

$$> S > N > G > P$$

The ranking based on affinity for the membrane environment (14 residue types) [56] is:

$$\mathrm{W} > \mathrm{L} > \mathrm{F} > \mathrm{I} > \mathrm{V} > \mathrm{M} > \mathrm{Y} > \mathrm{A} > \mathrm{T} >$$

$$G > S > Q > N > P$$

These rankings can be contrasted with e.g. the Chou-Fasman [36] data, which measures α-helix propensities in water-soluble globular protein structures:

$$\mathrm{E} > \mathrm{M} > \mathrm{A} > \mathrm{L} > \mathrm{K} > \mathrm{F} > \mathrm{Q} > \mathrm{W} \sim \mathrm{I} >$$

$$V > D > H > R > T > S > C > Y >$$

$$N > P \sim G$$

Of course these rankings are not based upon the same measure, but they indirectly highlight the differences in behaviour between amino acid residues in the membrane and the aqueous environments. In an early study of the prediction of secondary structure in membrane proteins [57] it was shown that the methods that work well for globular proteins did not give accurate predictions for the small database of membrane proteins that was studied.

Our data for helix preferences in the TOTAL set is:

$$\mathrm{A} > \mathrm{E} > \mathrm{Q} > \mathrm{M} > \mathrm{L} > \mathrm{R} > \mathrm{K} > \mathrm{I} > \mathrm{W} >$$

$$F > V \sim Y > H > D > S > T > C \sim N >$$

$$G > P$$

Although our database is not large, it consists of 49 helices in membrane proteins, while the Deber et al. [34, 35] work is based on a database of only 13 synthetic peptides. It may be that synthetic peptides, even if they span the membrane, are not good models for membrane proteins, where in addition to membrane spanning there are other constraints on the latter such as the need to mutually pack and form a protein core. Rather, they are designed with membrane-spanning in mind, and not for any specific biological function. Furthermore, membrane proteins may have functional requirements for certain residue types that may not otherwise be especially compatible with the membrane environment. As an example of this we conducted a search for the frequency of occurrence of all residue types in a set of 262 GPCR sequences from the biogenic amine subset of the rhodopsin family, selecting only the putative[5] TM helices [58] from these sequences. The following frequency of occurrence of residue type was obtained (raw data):

$$\mathrm{V} > \mathrm{L} > \mathrm{I} > \mathrm{A} > \mathrm{T} > \mathrm{S} > \mathrm{F} > \mathrm{M} > \mathrm{G} >$$

$$C > Y > N > R > P > K > Q > W >$$

$$E > H > D$$

Of course these data can be normalised to take account of residue-type frequency, but this normalisation can be carried out in several different ways in order to illustrate, for example, frequency in membrane helices as against helices in general, or compared with overall frequency in all secondary structures in the entire database.

We now consider some of the residue types that might be considered to be unexpected in transmembrane helices.

[5]The assignment of these sequences to transmembrane helices is supported by experiment through sequence alignment to the transmembrane helical regions in the lf88 crystal structure [1].

- *Glycine* is a residue type that is normally regarded as being conducive to turn formation, yet it is not uncommon in membrane-spanning helices. An analysis of the torsion angles for the entire MEMBRANE set shows that by far the majority of glycine residues sit comfortably in the α-helix with no abrupt departure from the standard helical Ramachandran (ϕ,φ) angles. There are certainly some exceptions such as the bend in TM2 of bovine rhodopsin (PDB i.d. 1f88) at a site where there are two adjacent glycines.

  Even where there is no apparent sign of any serious 'kinking' of the helix where glycines are located, this does not rule out the possibility that the peptide chain conformation may change at these positions, e.g., during activation [59] or as a stabilizing influence [60] in many enzymes. We can only report that we do not see any evidence for this at this stage. The data we have analysed is from crystal structures, which are necessarily static in nature (and in the case of the MEMBRANE set of restricted resolution).

  Glycines are relatively more frequent in GPCRs than in the membrane set as a whole and this could be of importance in connection with the activation of these receptor proteins, but this is only surmise. Several of the TM helices in GPCRs are flanked by conserved Gly residues, e.g., the 1f88 crystal structure of bovine rhodopsin has glycine residues at the ends of helices TM2 & 4 (N & C termini), TM3 (N terminus only) and TM5 & 7 (C terminus only). This could be part of the activation mechanism, allowing the TMs to move about these pivoting points.

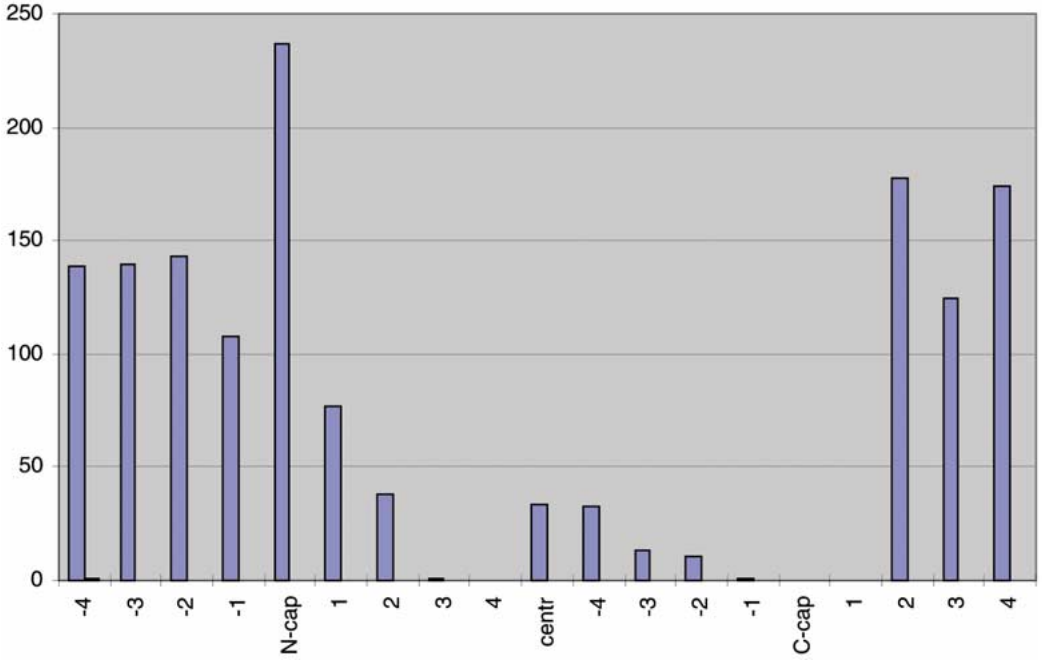
  There could be other important functional reasons for incorporating these residues into α-helices in GPCRs, which may not apply to other classes of membrane protein. Having no side chain, Gly allows for bulkier groups, whether other side chains or ligands, to be accommodated in the interior of the molecule and to dock into the cavity generated by the absent side-chain of the glycine in a 'knobs-into-holes' [42, 43, 48] fashion. This might be important e.g., for helix-helix packing, for dimerization or for recognition of other membrane proteins.
- *Proline* residues, while not abundant, do turn up in membrane helices (see Figure 1 and in the preference data for the MEMBRANE set above). The frequency is at the present time too low to be able to discuss preferred locations within the helix, but this is shown for the TOTAL set in Figure 2.

  In GPCRs, where we have a lot of sequences, prolines seem to be overrepresented in the transmembrane regions, coming now in the middle of the range of observed frequencies for the different residue types (this is discussed further below).

  Pro is regarded by many authors [34, 35, 61–66] as being *necessarily* 'helix breaking'. But Pro can in fact be accommodated in a helix. In a database of accurate water-soluble globular protein structures, 5% of Pro residues are found within α-helices (i.e., not counting the ones that quite frequently turn up as the first residue in a helix). They are not noticeably 'kinked"'at the site where the Pro itself is, as judged e.g., by the secondary structure determining program DSSP [67]. There is at least one report [68]in which an 'i $-$ 4' kink in a Pro-containing α-helix is *not* necessarily due to the presence of the Pro, since when this residue is mutated to Ala, the kink remains. Further, the bulge in helix G in the high-resolution structure (1.55 Å) of bacteriorhodopsin [55] is not due to proline but due to bound waters inside the helix bundle.

  There is no standard definition of what constitutes a kink. Some authors [34, 35, 61, 63, 65, 66] focus on the dihedral angles of the backbone residues at and prior to the Pro residue while others think in terms of a change in the direction of the helix [62, 64] without clearly defining what is meant. Presumably it is the angle between axes of the disjoint helices, although to be correct one must contend with the fact that it will usually require two angles to characterize a 'kink' defined in this way. It is important to have a proper definition of what constitutes a kink, and also to study the angle data very carefully; secondary structure algorithms such as DSSP [67] can appear to be rather too 'forgiving' in their assignment of secondary structures because of the smoothing algorithm that is employed. We use as a definition for a kink: a site at which (a) hydrogen bonding along the helix backbone is disrupted and (b) departure from linearity of the helix where the crossing angle of the mutual projection of the axes of the disjoint helices is greater than 15°.

  In many cases there is a discernible distortion in the helix at the $(i-3)$th and $(i-4)$th residues preceding a Pro. This is in accord with observations in synthetic peptides [61] and in the fungal membrane-spanning peptide alamethicin [69, 70].



Original data for Figure 2 in tabular form.

−4	−3	−2	−1	N-cap	1	2	3	4	centr	-4	−3	−2	−1	C-cap	1	2	3	4
139	140	143	108	237	77	38	1	0	34	33	13	11	1	0	0	178	125	174

Figure 2. Distribution of prolines over helices in the total database.

Pro is of course severely constrained in the ϕ angle but its φ angle is essentially as free as that of any other residue, maybe more so than most in view of the inability for proline to hydrogen bond back onto the carbonyl oxygen of the $(i-4)$th residue. The set of Ramachandran (ϕ, φ) angles for all Pro residues for water soluble globular proteins shows strong clustering around the α-helix region, another strong clustering in the region typical for collagen and widely spread between these extremes (all regions with approximately the same narrow range of ϕ angles). For every member of the MEMBRANE set, all torsion angles were calculated and those corresponding to the 4 residues preceding each Pro (again, not including the first Pro in a helix where this occurs) examined in detail. Of 36 cases examined, 23 could be passed as being α-helical and 13 were somewhat kinked. Typical for the former category is that the sequence prior to the Pro was largely hydrophobic (residue types Val, Ile, Leu, Phe, Met in particular). For the kinked category there is always some hydrophilic residue among the preceding four residues, with a preponderance of cases where the residue immediately before the Pro is an Arg (but it must be recalled that we are dealing with a database with a fair degree of redundancy). Typically in these cases the ϕ angle of the $(i-3)$th residue was in the range from $-100°$ to $-135°$ while that of the $(i-4)$th residue was usually normal but could be deformed up to $-100°$. The ω angle of the $(i-1)$th residue is mostly normal but for those sequences with Arg as the residue prior to the Pro there was a departure from planarity of up to $-10°$. Similar observations have been made in MD studies [69, 70] of the membrane spanning peptide alamethicin in which a Pro residue can be observed to cause a similar distortion of the preceding peptide group and of the ϕ or ι angles of the $(i-3)$th and $(i-4)$th residues.

In the MEMBRANE set, there are a very few cases where prolines seriously disrupt the helix but these all have either an asparagine or glycine residue among the preceding 4 residues. In bovine

rhodopsin (PDB i.d. 1f88) the bend in TM5 is a 'classical' Pro-induced kink situated 3 residues downstream of a Pro (see further discussion below). In TM7 the 'bend' is further downstream from the conserved NP site at about 5 residues, and in fact takes more the character of a $\mathbf{3}_{10}$ helix. TMs 1 and 6 both contain Pro, but they are gently curved along their entire length, not abruptly kinked. According to DSSP [67] these two helices are both classified as typical α-helices, except the TM6 has a slight distortion at the very bottom (I255). There may be some mechanistic significance to this slight distortion. A mutation in the closely neighbouring residue M257X (X = Y or N) causes constitutive activation of bovine rhodopsin [71]. The Gly induced kink in TM2 has been referred to above.

It is also a common occurrence that a helix starts with Pro, thus it could be claimed that Pro is a 'loop-breaking' residue. On the basis of these rather conflicting findings we affirm that the issue of proline kinking or flexibility has not been unequivocally settled but it is obviously important to know more, especially in the context of membrane proteins.

What has been said up to now refers only to 'static' structures, but Pro has otherwise been attributed a special role as a center of flexibility important for GPCR activation by many authors [72, 73], and there is no reason to eliminate this possibility, given that its φ angle, and the ϕ (and possibly ω) angle of the $(i-4)$th reside, are less constrained in a helix than for all other residue types. But we stress that there is no experimental evidence for this at present. Rather, there seems to be a lack of involvement of Pro residues in the light-induced activation of the GPCR rhodopsin [74] and in bacteriorhodopsin [75, 76]. In the case of the latter, activation appears to be more connected with rearrangements of the hydrogen-bonding network involving in particular a number of Thr residues (see discussion of the presence of Ser and Thr below). This latter observation bears out the theoretical studies in which changes in the hydrogen-bonding network were considered to be critical for GPCR activation [73, 77]. Experimental support for the notion that hydrogen-bonding networks are critical for stabilising proteins generally [78] and membrane proteins [79] has recently been published.

It may not be necessary to invoke such complicated explanations to account for the relatively higher prevalence of Pro in transmembrane helices in GPCRs. An alternative explanation might be that the free carbonyl oxygen on residue $i-4$ is required for ligand binding through hydrogen bonding. In a homology model of β-2-adrenergic receptor based on the 1f88 template, there is a Pro residue, 100% conserved in the family, 98% conserved in the entire biogenic amine family, in the middle of TM5, and the naked carbonyl at the fourth residue position prior to that points directly into the ligand binding pocket (further discussion of this will be deferred to a separate paper on GPCRs).

- The β-branched residue types *valine* and *isoleucine* feature prominently in the MEMBRANE set, despite the fact that they are normally not so frequently encountered in α-helices in water soluble globular proteins. They clearly fit into the hydrophobic ambience of the membrane. It could be surmised that any strain caused by the existence of the extra methyl group on the Cβ atom will be more than compensated for by the enhanced stability of the hydrogen bond in the nonaqueous environment. It has been suggested [80] that the constrained rotameric freedom of the side chains of these residue types might reduce the entropic cost of folding in transmembrane proteins.
- *Serine* and *threonine* would not be expected to be common in membrane helices, yet they turn up in the middle of the range. There may be functional reasons for the presence of some of these residues, but very possibly they have a role in stabilizing the protein through hydrogen-bonding networks [78, 79], or switching between different arrangements of hydrogen-bonding, as has been suggested [73, 77].

  Ser or Thr residues within the TM region should participate in an internal hydrogen-bonding network, otherwise they can sometimes 'back-bond' onto the backbone *via* a hydrogen bond linking the side-chain oxygen atom to an atom in the preceding or following peptide bond [81–83]. In order to predict the hydrogen-bonding status of a given Ser or Thr residue, whether this 'back-bonding' is in operation, or whether it is necessary to fit the side-chain hydroxyl group into a hydrogen-bonding network, it is recommended to study the sequence alignment at that residue position. If Ser (or Thr) is 100% conserved, or close to fully conserved, it

is likely that an involvement in a hydrogen-bond network is mandatory. If an alignment reads something like AASASSAA then the likely conclusion is that the requirement at that site is for a small residue type.

In GPCRs, there are indications that these residue types might play an important role in activation. This is GPCR-specific, so further discussion of this will be deferred to another work in preparation which deals with GPCR activation.

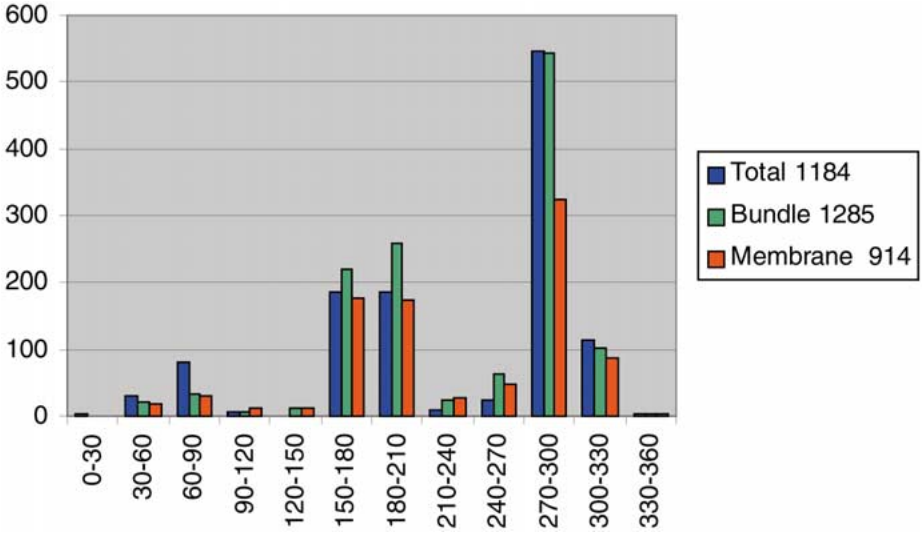
- *Charged residues* deserve special mention. They would not normally be expected to be frequent in membrane helices. This is indeed the case, both for the MEBRANE set and for the GPCR superfamily. The conditions under which the presence of charged residues are allowed, or alternatively destabilise, a transmembrane helix has been studied in detail [84]. There is almost certainly an explanation in terms of function for the presence of these residue types, for example Glu residues are critical for rhodopsin action. Lys or Arg residue types cluster at the cytosolic ends of transmembrane helices in accordance with 'positive inside rule' [85].
- *Asparagine* is the least abundant residue type in the MEMBRANE set, yet it is encountered quite frequently in the GPCRs. As for Ser and Thr, this is known to be due to functional requirements for this residue type and further discussion of this will be deferred to work in preparation. Nevertheless, the 'membrane protein engineer' and especially the GPCR researcher must be prepared when necessary to admit this residue type albeit with its acknowledged 'helix-breaking' character into a membrane helix.
- *Tryptophan* occupies a special position in our consideration of residue-type preferences. It is the most abundant residue type in the MEMBRANE set, yet it comes at the end of the list in the GPCR subset. Despite this relative infrequency this residue type has been assigned a pivotal role [86] in ensuring compatibility between membrane helices and the lipid environment. Further, Trp residues are typically found in the 'core' of GPCRs and in the 'lining' on the ligand-binding pocket. A fully conserved Trp is found on transmembrane helix IV and a very highly conserved CWXP motif on helix VI of GPCRs of the rhodopsin family, for example (http://www.gpcr.org/7tm/seq/001/001.ALI.html).

One general feature which may explain, or at least allow, the relative prevalence of Gly, Pro, Ile and Val (and even Asn, when it occurs) in membrane helices almost certainly has to do with the extent to which the two different solvent environments favour or disrupt the hydrogen bonding required to stabilise the helices. In the case of Gly, steric freedom in both the ϕ and φ torsion angles gives this residue-type access to a much wider region of torsion space than for residues with side chains. The stabilising effect of hydrogen bonding is not sufficient for always maintaining a helix structure at sites occupied by Gly in water soluble globular proteins. However, in membrane proteins, the relative effect of hydrogen bonding is much greater since there is no competition from water for the hydrogen bonding which stabilizes α-helices. Similar remarks may be made about the other residue types, especially Pro, although they are confined to more restricted regions of the ϕ**,** φ torsion space.

Three dimensional nearest neighbour contacts along the helix

In a helix there are often contacts between the residue at position i and the residues at positions $i+3$ or $i+4$. These contacts put restraints on the possible combinations of residue types and rotamers. We made tables for the preference parameters for the pairs $i-i+3$ and $i-i+4$. In the TOTAL dataset we observe several pairs that are strongly preferred. For example: the asymmetry seen for pairs involving a Pro is explained by the fact that this residue type is distributed very unevenly over helices.

Salt bridge-forming pairs and hydrophobic pairs are favourable, and several pairs that potentially form hydrogen bonds are observed, in agreement with the possibilities offered by the geometry of helices. However, we observe most outliers for residue pairs that contain at least one residue that has a low frequency among all residues used. It should be kept in mind when considering the preference parameters that we are looking at frequencies which are too low to allow proper determination of the counting statistics. The distributions observed in the MEMBRANE dataset of course suffer most acutely from low frequencies. If we disregard for the moment the significance of the numbers and the inherent redundancy in the MEMBRANE set, this set looks more similar to the TOTAL set than to the BUNDLE set. We do not offer any interpretation of this observation, preferring to defer this discussion



Original data for Figure 3 in tabular form.

	0–30	30–60	60–90	90–120	120–150	150–180	180–210	210–240	240–270	270–300	300–330	330–360
Total 1184	2	29	80	6	1	185	187	8	23	546	115	2
Bundl 1285	0	20	34	6	12	220	257	23	63	544	102	4
Membr 914	1	17	29	13	11	178	174	27	49	325	87	3

Figure 3. Rotamer angle distributions in the three datasets.

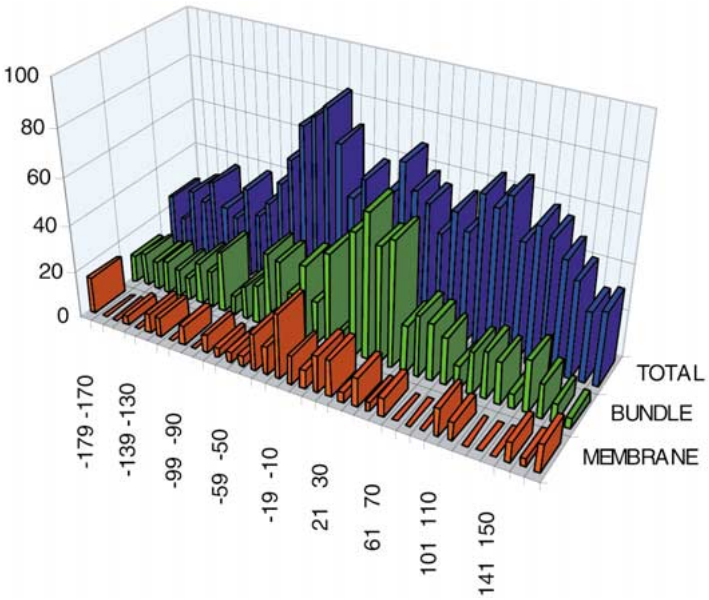
until such time as more abundant data is available for the MEMBRANE set.

Rotamer analysis

Inspection of the results (http://www.cmbi.kun.nl/gv/-service/helices/) show that on the whole, there are no major differences between the rotamer distributions for the set of membrane proteins and for the other two sets. It seems that, for helices at least, the g+ rotamer (60°) is not favored except weakly for some polar and aromatic residue types and this pattern is retained in the membrane peptides.

Examples of rotamer distributions are shown for residue types Cys, Phe, Lys, Pro and Thr in Figure 3 for TOTAL, BUNDLE and MEMBRANE. We compare with the rotamer distributions stored in the three relational databases to look for rotamer preferences given a 'perfect' helical backbone. Even though we only look at the rotamer distribution on a perfect helix, and thus cannot draw conclusions about distributions on strongly bent helices or at positions where the helix is locally deformed, the numbers in Figure 3 indicate that rotamer distributions show only very little dependence on the molecular class. And that is good news for homology modelling membrane proteins, or for completing membrane protein structures in cases where only the backbone atoms are seen clearly in the electron density.

It should be kept in mind, however, that the membrane protein structures are solved at a lower resolution than those of the proteins in the other two datasets. Therefore the fraction of the side chains for which the rotamer type is determined more by the refinement force field and less by the experimental data is much larger in the membrane dataset than in the other two datasets, and it cannot be excluded that the small differences that are observable will get larger when more high resolution membrane structures become available. It has earlier been shown [86] that the average χ_1 torsion angle for individual residue types can differ



Original data for Figure 4 in tabular form.

Angles		Total	Membrane	Bundle
−179	−170	26	4	6
−169	−160	17	0	7
−159	−150	31	0	6
−149	−140	28	1	7
−139	−130	38	0	6
−129	−120	28	2	5
−119	−110	25	2	9
−109	−100	39	0	8
−99	−90	29	2	13
−89	−80	36	0	4
−79	−70	46	2	7
−69	−60	58	1	8
−59	−50	73	1	16
−49	−40	78	1	15
−39	−30	83	4	9
−29	−20	70	3	16
−19	−10	49	7	9
−9	−0	58	3	20
1	10	34	2	4
11	20	55	4	26
21	30	69	4	31
31	40	59	1	25
41	50	55	3	27
51	60	44	1	10
61	70	55	2	13
71	80	48	0	12
81	90	65	0	10
91	100	61	0	5
101	110	67	3	9
111	120	50	2	10
121	130	58	0	9
131	140	55	0	3
141	150	48	0	11
151	160	41	2	7
161	170	30	1	4
171	180	32	3	2

Figure 4. Distribution of helix-helix crossing angles. Red: membrane set (multiplied 4x for clarity). Green: bundle set (multiplied 2x). Blue: total set.

by up to 10° if datasets solved around 1 Å and around 3 Å respectively are compared. Given this *caveat*, it is apparent that there are no major differences between the sets, and certainly nothing that amounts to any clear trend. There is at least no support at this time for the idea that a rotamer library, specific for membrane proteins need be used in preference to the existing rotamer library derived from accurate structures of water-soluble globular proteins.

Peptide bond and helix distortion

We could not discern any noticeable pattern in the distribution of bond angles and conclude that observed angle deviations published by others [41] may have arisen as a result of the procedure for selecting the datasets. Equally, our dataset has been selected with other criteria in mind, and the MEMBRANE set is not as accurate as the other two.

Helix-helix crossover analysis

The same three databases were analysed for helix crossover angles, using a procedure, which in addition to that of Chothia et al. [41], takes account of helix direction. This allows considerations of handedness to be taken into account. The results are shown in Figure 4. In TOTAL the helix crossing angles have a wide distribution over the entire torsion angle space but there is a maximum in the region of $-40°$ with smaller maximum at $+100°$. For BUNDLE the maximum appears at +30°, close to the ideal packing angle $+23°$ ('3–4' class [41]) with secondary peaks in the range $-20°$ to $-50°$. The MEMBRANE set is thinly populated but shows a minor peak close to $+23°$ and values that are clustered on either side of the angles 0° and 180°, compatible with the need for helices to pack alongside each other and yet span the membrane. Neither of these values correspond quite to the '3–4' class (ideal packing angle 23°).

Helix crossing angles show some differences between the three datasets although MEMBRANE resembles BUNDLE more than TOTAL. We would expect this latter result since BUNDLE was selected on the basis of having a topology resembling that of helical membrane proteins. The BUNDLE set does, however, have a number of helix contacts in the -50 and $+50°$ region. This former is very close to the ideal crossing angle for helices in the '4–4' class which was calculated [41] to be the most favoured helix crossing arrangement. These authors then analysed experimental structures and found crossing angles for the '4–4' class to fall in the range -80 to $-20°$. Although they used a database of only 50 helix-to-helix packings in 10 globular protein structures, we can conclude that the angles found for the MEMBRANE set fall within an 'allowed' region, albeit at the extremity that is most compatible with the membrane-spanning constraint.[6]

We do not expect to see any '4–4', '1–4' (ca. $-105°$) or '3–3' crossings in MEMBRANE, nor indeed do we, although this region of the crossing angle space is quite well populated in both BUNDLE and TOTAL. Lemmon et al. [40] described an exceptional case of synthetic peptides where crossover angles in the '4–4' class were observed, and ascribed this to the existence of supercoiling (see below for further discussion of supercoiling).

Since GPCRs are a very major focus of interest in biochemical and biomedical research it is of interest to check how the 'best' GPCR model compares with the overall data for membrane proteins. The values of the helix crossing angles for the crystal structure of bovine rhodopsin [1] are given in Table 2a. For comparison, we show in Table 2b the corresponding angles for bacteriorhodopsin [55]. Note that while the crossing angles mostly fall into the 'allowed' range,

Table 2a. Unsigned helix crossing angles for bovine rhodopsin (Baldwin, Schertler and Unger structure ([27]))

	TM1	TM2	TM3	TM4	TM5	TM6	TM7
TM1		27	40				23
TM2			17	21			26
TM3				23	24	31	29
TM4					23	24	
TM5						24	
TM6							5

Table 2b. Unsigned helix crossing angles for bacteriorhodopsin (crystel structure PDB i.d. 1c3w ([55])

	TM1	TM2	TM3	TM4	TM5	TM6	TM7
TM1		22	48				30
TM2			14	4			12
TM3				12	21	26	61
TM4					22	33	
TM5						17	
TM6							5

[6]Note that these authors ignored the direction of the individual helices while we take handedness into account.

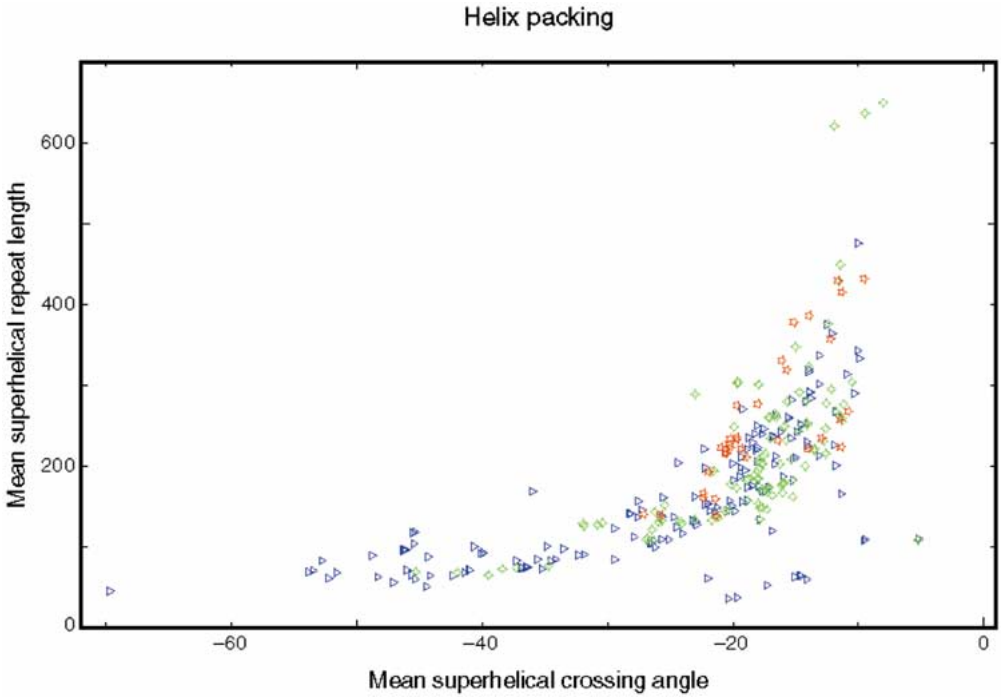


Figure 5. The mean superhelical repeat length as function of the mean superhelical crossing angle. Red: membrane set; green: bundle set; blue: total set.

the pattern is different to that of GPCRs, if indeed bovine rhodopsin is typical of the latter. Also, the angles are, if anything, somewhat smaller than the typical values for bovine rhodopsin suggesting that the 'membrane spanning' constraint is exerting a stronger effect on bacteriorhodopsin than on bovine rhodopsin. The interhelical loops are generally shorter in bacteriorhodopsin than in GPCRs but this is unlikely to be a major constraint. In four-helix bundle containing proteins, the loops have been shown to have no significant effect on helix packing [87] and there is no reason why seven-helix bundle proteins should not behave in an analogous fashion.

In modelling membrane helical proteins we would advocate the use of Figure 4 as a 'database' for selecting helix crossover angles. Angles in the region of $\pm 50°$ or $100°$, while favoured in the wider universe of helical protein structure, are not compatible with membrane spanning helical structure.

Overall helix packing

We searched for all possible four-helix bundles in the three datasets and found that all four-helix bundles show a left-handed superhelical twist. We analysed the superhelical crossing angle and the superhelical repeat length in these four-helix bundles. Of course, one expects the superhelical repeat length to be larger when the crossing angle gets closer to zero degrees. One also expects that transmembrane bundles will have relatively small superhelical crossing angles, simply because if the angles would get significantly less than $-25°$, very long helices would be needed to cross the membrane, and large holes would exist between the ends of the helices. Figure 5 shows a plot of the mean superhelical repeat length versus the superhelical crossing angle. Although there is a significant amount of variation and there are a few outliers in the TOTAL dataset (mainly caused by flaws in the automatic four-helix bundle detection as was determined by visual inspection), it can be seen that there are no principal differences between the helix bundle parameters for the three datasets beyond slight tendency towards a greater superhelical crossing angle in the MEMBRANE set compared with the BUNDLE set. As expected, the superhelical crossing angle is always in the 10–30° range.

Table 2c. Unsigned helix crossing angles for bovine rhodopsin (X-ray crystal structure PDB i.d. 1f88 ([1]). Where there is pronounced beinding in the helix axis, the C-terminal portion of the helix is indicated by a prime (TMs 2′, 4′ and 5′)

	TM1	TM2	TM2′	TM3	TM4	TM4′	TM5	TM5′	TM6	TM7	TM7′
TM1		37	16	45						19	43
TM2				10	32	33				45	30
TM2′				41	34	61				34	50
TM3					36	25	25	22	38	51	28
TM4							24	34	10		
TM4′									15	6	34
TM5										19	
TM5′										21	15
TM6										16	24

Conclusions

We have examined α-helices in three different classes of proteins, a general non-redundant set of water soluble globular proteins containing representatives of all well-known protein folds, a set of four-helix bundle containing proteins and a set of membrane-spanning helices from membrane proteins. Here we summarise the main differences between the membrane helix set and the other two sets.

Residue preferences for helices show considerable differences between the databases, to the extent that standard secondary prediction tools would probably not give a reliable prediction of membrane helix propensity. For example when the PHD program [88] is run on a typical transmembrane helical peptide from a GPCR in the PHDsec mode (secondary structure prediction without allowance for propensities for membrane location) they are not predicted as helical throughout their whole length (i.e., not at the ends, and they can even be interrupted) whereas in the PHDhtm mode, where careful attention has been paid to these propensities and prediction is directed towards finding transmembrane helices, accurate predictions are made. Residue types not typically associated with helix stabilization, such as Gly, Pro, Ser, Thr, Asn, Ile, Val, are relatively common in transmembrane helices.

Helix-helix crossover angles for the MEMBRANE set fall within the allowed ranges for both of the BUNDLE and TOTAL set, with a distinct preference for the narrower range on angles consistent with the need to cross the lipid bilayer at an angle that does not require a very long stretch of helix, nor cause problems with peptide-lipid packing.

We have further drawn attention to the fact that transmembrane helices are subject to deformations which are sometimes local and sometimes extend over almost the full extent of the helix. The issue of supercoiling, encountered in fibrous and some globular proteins but not considered before in the context of membrane proteins, and its importance in helix bundle packing was proposed as a way to account for these observations.

In contrast to these findings, there is no appreciable difference between membrane proteins studied so far and water-soluble globular proteins in respect of three dimensional nearest neighbour contacts along the helix, side chain rotamer angles and peptide bond angles.

These conclusions might be regarded a set of thumb-rules for use by the protein-modelling community when constructing models of membrane proteins and in chain-tracing from low-resolution electron density data.

References

1. Murzin, A.G., Brenner, S. E., Hubbard, T. and Chothia, C. J. Mol. Biol., 247 (1995) 536.
2. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Tromg, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M. and Miyano, M., Science, 289 (2000) 739–745.
3. Arkin, I.T., Brünger, A.T. and Engelman, D.M., Proteins, 28 (1997) 465–466.
4. Frishman, D. and Mewes, H.W., Nat. Struct. Biol., 4 (1997) 626–628.
5. Jones, D.T., FEBS Lett., 423 (1998) 281–285.
6. Findlay, J.B.C. and Eliopoulos, E., TIPS, 11 (1990) 492–499.
7. Dahl, S.G., Edvardsen, Ø. and Sylte, I., Proc. Natl. Acad. Sci. USA, 88 (1991) 8111–8115.

8. Hibert, M.F., Trumpp-Kallmeyer, S., Bruinvels, A. and Hoflack, J., Mol. Pharmacol., 40 (1991) 8–15.
9. Lewell, X.G., Drug Design Discovery, 9 (1992) 29–48.
10. MaloneyHuss, K. and Lybrand, T.P., J. Mol. Biol., 225 (1992) 859–871.
11. Kontoyanni, M. and Lybrand, T.P., Med. Chem. Res., 3 (1993) 407–418.
12. Cronet, P., Sander, C. and Vriend, G., Prot. Eng., 6 (1993) 59–64.
13. Taylor, W.R., Jones, D.T. and Green, N.M., Proteins, 18 (1994) 281–294.
14. Donnelly, D., Findlay, J.B.C. and Blundell, L.T., Receptors and Channels, 2 (1994) 61–78.
15. Lin, S.W., Biochemistry, 33 (1994) 2151–2160.
16. Herzyk, P. and Hubbard, R.E., Biophys. J., 69 (1995) 2419–2442.
17. Shieh, T., Han, M., Sakmar, T.P. and Smith, S.O., J. Mol. Biol., 269 (1997) 373–384.
18. Perez, J.J., Filizola, M. and Cariteni-Farina, M., J. Math. Chem., 23 (1998) 229–238.
19. Frimurer, T.M. and Bywater, R.P., Proteins, 35 (1999) 375–386.
20. Kristiansen, K., Dahl, S. G. and Edvardsen, Ø., Proteins, 26 (1996) 81–94.
21. Edvardsen, Ø. and Kristiansen, K., 7TM J., 6 (1997) 1.
22. Seeman, P., Receptor Tables, Vol. 2, SZ Research, Toronto,1993.
23. Kuipers, W., Oliveira, L., Vriend, G. and Ijzerman, A.P., Receptors Channels, 3 (1997) 159.
24. Horn, F., Bywater, R., Krause, G., Kuipers, W., Oliveira, L., Paiva, A.C.M., Sander, C. and Vriend, G., Receptors Channels, 5 (1998) 305.
25. Farahbakhsh, Z.T., Ridge, K.D., Khorana, H.G. and Hubbell, W.L., Biochemistry, 34 (1995) 8812.
26. Yang, K., Farrens, D.L., Altenbach, C., Farahbahksh, Z.T., Hubbell, W.L. and Khorana, H.G., Biochemistry, 35 (1996) 14040.
27. Yang, K., Farrens, D.L., Hubbell, W.L. and Khorana, H.G., Biochemistry, 35 (1996) 12464.
28. Farrens, D.L., Altenbach, C., Yang, K., Hubbell, W.L. and Khorana, H.G., Science, 274 (1996) 768.
29. Baldwin, J.M., Schertler, G.F.X. and Unger, V.M., J. Mol. Biol., 272 (1997) 144.
30. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., Arch. Biochem. Biophys., 185 (1978) 584.
31. Hooft, R.W., Sander, C. and Vriend, G. J., Appl. Cryst., 29 (1996) 714.
32. Holm, L. and Sander, C., Nucl. Acids Res., 26 (1998) 316.
33. Banner, B.W., Kokkinidis, M. and Tsernoglou, D. J., Mol. Biol., 196 (1987) 657.
34. Li, S.C. and Deber, C.M., Nat. Struct. Biol., 1 (1994) 368.
35. Deber, C.M. and Li, S.C., Bioploymers, 37 (1995) 295.
36. Chou, P.Y. and Fasman, G.D., Biochemistry, 13 (1974) 211
37. Chakrabarti, P., Bernard, M. and Rees, D.C., Biopolymers, 25 (1986) 1087.
38. Karplus, P.A., Prot. Sci., 5 (1996) 1406.
39. Mingarro, I., Elofsson, A. and Von Heijne, G.J., Mol. Biol., 272 (1997) 633.
40. Lemmon, M.A., MacKenzie K.R., Arkin I.T. and Engelman, D., in von Heijne, G. (Ed.), 'Membrane Protein Assembly'. Springer-Verlag, New York/Landes Austin TX, 1997, pp. 3–23.
41. Chothia, C., Levitt, M. and Richardson, D. J., Mol. Biol., 145 (1981) 215.
42. Walther, D., Eisenhaber, F. and Argos, P. J., Mol. Biol., 255 (1996) 536.
43. Crick, F.H.C., Acta Crystallogr., 6 (1953) 689.
44. Bowie, J.U., J. Mol. Biol., 272 (1997) 780.
45. Bragg, L., in Phillips, D.C. and Lipson, H. (eds), 'The Development of X-ray Analysis', Bell, London, 1975.
46. Zerger, M.J., Mathematical Intelligencer, 20 (1998) 5.
47. Seo J. and Cohen, C., Proteins, 15 (1993) 223.
48. Langosch, D. and Heringa, J., Proteins, 31 (1998) 150.
49. Sansom, M.S., Son, H.S., Sankararamakrishnan, R., Kerr, I.D. and Breed, J., Biophys. J., 68 (1995) 1295.
50. Vriend, G., J. Mol. Graph., 8 (1990) 52.
51. Jones, T.A. and Thirup, S., EMBO J., 5 (1986) 819.
52. De Filippis, V., Sander, C. and Vriend, G., Prot. Eng., 7 (1994) 1203.
53. Chinea, G., Padron, G., Hooft, R.W.W., Sander, C. and Vriend, G., Proteins, 23 (1995) 415.
54. Thomas, D.J., J. Mol. Biol., 222 (1991) 805.
55. Luecke, H., Schobert, B., Richter, H.T., Cartailler, J.P. and Lanyi, J.K. Structure of bacteriorhodopsin at 1.55 A resolution. J. Mol. Biol., 291 (1999) 899.
56. Thorgeirsson, T.E., Russell, C.R., King, D.S. and Shin, Y.K.,
57. Wallace, B.A., Cascio, M. and Miele, D.L., Proc. Natl. Acad. Sci. USA, 83 (1986) 9423.
58. Oliveira, L., Paiva, A.C.M., Sander, C. and Vriend, G., TIPS, 15 (1994) 170.
59. Peters, G.H. and Bywater, R.P., Prot. Eng., 12 (1999) 747.
60. Serrano, L., Neira, J.L., Sancho, J. and Fersht, A.R., Nature, 356 (1992) 453.
61. Piela, L., Némethy, G. and Scheraga H.A., Bioploymers, 26 (1987) 1587.
62. Barlow, D.J. and Thornton, J.M., J. Mol. Biol., 201 (1988) 601.
63. Deber, C.M., Glibowicka, M. and Woolley G.A., Bioploymers, 29 (1990) 149.
64. MacArthur, M. W. and Thornton, J. M., J. Mol. Biol., 218 (1991) 397.
65. Von Heijne, G., J. Mol. Biol.,218 (1991) 499.
66. Ballesteros, J.A. and Weinstein H., Biophys. J., 2 (1992) 07.
67. Kabsch,W. and Sander, C., Biopolymers, 22 (1983) 577.
68. Yuan, H.S., Wang, S.S., Yang, W.Z., Finkel, S.E., and Johnson, R.C., J. Biol. Chem., 269 (1994) 28947.
69. Jacob, J., Duclohier, H. and Cafiso, D.S., Biophys. J., 76 (1999) 1367.
70. Bak, M., Bywater, R.P., Hohwy, M., Thomsen, J.K., Adelhorst, K., Jakobsen, H.J., Sørensen, O.W. and Nielsen, N.C., Biophys. J. (2000) (submitted).
71. Han M., Smith S.O. and Shakmar T.P. Biochemistry, 37 (1998) 8253.
72. William, K.A. and Deber, C.M., Biochemistry, 30 (1991) 8919.
73. Lomize, A.L., Pogozheva, I.D. and Mosberg, H.I., J. Comput. Aid. Mol. Des., 13 (1999) 325.
74. Borhan, B., Souto, M.L., Imai, H., Schichida, Y. and Nakanishi, K., Science, 288 (2000) 2209.
75. Israelewitz, B., Izrailev, S. and Schulten, K., Biophys. J., 73 (1997) 2972.
76. Kandori, H., Kinoshita, N., Yamazaki, Y., Maeda, A., Shichida, Y., Needleman, R., Lanyi, J.K., Bizounok, M., Herzfeld, J., Raap, J. and Lugtenburg, J., Proc. Natl. Acad. Sci. USA., 97 (2000) 4643.

77. Pogozheva, I.D., Lomize, A.L. and Mosberg, H.I., Biophys. J., 72 (1997) 1963.
78. Cooper, A., Biophys. Chem (2000) 25.
79. Zhou, F.X., Cocco, M.J., Russ, W.P., Brunger, A.T. and Engelman, D.M., Nat. Struct. Biol., 7 (2000) 154.
80. Senes, A., Gerstein, M. and Engelman, D.M., J. Mol. Biol., 296 (2000) 921.
81. Aubry, A., Ghermani, N. and Marraud, M., Int. J. Peptide Protein Res. 23, (1984) 113.
82. Dey, S., Kaur, P. and Singh, T.P., Int. J. Peptide Protein Res., 48 (1996) 299.
83. Vijayakumar, M., Qian, H. and Zhou, H.X., Proteins, 34 (1999) 497.
84. Lew, S., Ren, J. and London, E., Biochemistry, 39 (2000) 9632.
85. Von Heijne, G., J. Mol. Biol., 225 (1992) 487.
86. Rippmann, F., 7TM J., 4 (1994) 1.
87. Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton J.M., Proteins, 12 (1992) 345.
88. Brunet, A.P., Huang, E.S., Huffine, M.E., Loeb, J.E., Weltman, R.J. and Hecht, M.H., Nature, 364 (1993) 355.
89. Rost, B. and Sander, C., Proc. Natl. Acad. Sci. USA, 90 (1993) 7558.