Dissecting membrane protein architecture: an annotation of structural complexity

Jaime Arce, James N. Sturgis and Jean-Pierre Duneau

LISM/CNRS Aix-Marseille Université
Laboratoire d'Ingénierie de Systèmes Macromoléculaires (UPR9027), Institut de Biologie Structurale et Microbiologie, 31 Chemin Joseph Aiguier, 13402 Marseille cedex 20, France
Phone : 33 4 91 16 44 95; Fax 33 4 91 71 21 24; Email : duneau@ibsm.cnrs-mrs.fr

Additional title page Footnotes : Dissecting membrane protein architecture

© 2009 Wiley Periodicals, Inc.
Abstract

α-helical membrane proteins exist in an anisotropic environment which strongly influences their folding, stability and architecture, which is far more complex than a simple bundle of transmembrane helices, notably due to helix deformations, prosthetic groups and extramembrane structures. However, the role and the distribution of such heterogeneity in the supra molecular organization of membrane proteins remains poorly investigated. Using a non redundant subset of α-helical membrane proteins, we have annotated and analyse the statistics of several types of new elements such as incomplete helices, intramembrane loops, helical extensions of helical transmembrane domains, extracellular loops and helices lying parallel to the membrane surface.

The relevance of the annotation scheme was studied using residue composition, statistics, physical chemistry and symmetry of their distribution in relation to the immediate membrane environment.

Calculation of hydrophobicity using different scales show that different structural elements appear to have affinities coherent with their position in the membrane. Examination of the annotation scheme suggests that there is considerable information content in the amino-acid compositions of the different elements suggesting that it might be useful for structural prediction.

More importantly, the proposed annotation will help to decipher the complex hierarchy of interactions involved in membrane protein architecture.
Introduction

Biological membranes are crucial in cell biology where they function as a physical barrier both protecting the cell from its environment and confining the cellular machinery. However, cells need to transfer both material and information across this barrier, material in the form of substrates and waste products and information on the environment to ensure adaptation. These transfers are performed by membrane proteins, which regulate the passage of the membrane barrier ensuring that the transfers do not compromise the functions of protection and confinement. In addition to these roles the ubiquity of chemiosmotic coupling ensures that biological membranes play an important role in cellular energetic metabolism\textsuperscript{1-4}. The diversity of functions have evolved in a diversity of architectures, ranging from single pass transmembrane domain (including toxin like mellitin or surfactant proteins, the structures of which are not necessarily permanently associated to membranes) to proteins with a complex assembly of multiple subunits including extrinsic polypeptide chains.

Membrane protein complexity is also related to their heterogeneous environment that contains three more or less well defined zones: a hydrophobic core composed mainly by the apolar lipid chains; an interface that contains the polar lipid head-groups and is partially hydrated; and beyond the membrane an aqueous medium\textsuperscript{5}. The transfer of charged or polar compounds into the hydrophobic membrane core is energetically unfavourable. This has two consequences for the structure of membrane proteins: the amino-acid composition of the membrane embedded portion of membrane proteins strongly favours apolar residues; and the hydrogen bonding potential of the polar peptide bond oxygen and hydrogen atoms must be satisfied. To satisfy the hydrogen bonding potential of polar atoms, they are involved in a hydrogen bonding network that gives rise to extended secondary structures\textsuperscript{6}. It is for this reason that membrane proteins cross the membrane using one of two different architectures, $\alpha$ helical bundles or $\beta$ barrels. The $\beta$ class of membrane proteins appears to be restricted to rather special membranes such as bacterial or mitochondrial
outer membranes. This type of protein appears to be much less common than \( \alpha \) helical membrane proteins.

A conceptual approach to the folding of this second type of membrane proteins is the two stage model by Popot & Engelman. In this model there are two thermodynamically distinguishable steps considered, the insertion of \( \alpha \) helices into the membrane and the association of membrane embedded \( \alpha \) helices (TM). In a further extension, the three stage model includes as a final step the folding of loops, incorporation of prosthetic groups, insertion of peripheral domains and the formation of quaternary structure.

Today the main challenges to understand folding of \( \alpha \)-helical membrane protein bundles lie in the second or third stage of the three stage model. In particular understanding the rules that govern the second, association, step seems particularly important, all the more so as the third step that includes the folding of loops and the incorporation of prosthetic groups is probably dependent on the association of helices.

One approach to understanding the association of transmembrane helices is to analyse the organisation and associations that are present in proteins of known structure. For this it is necessary to locate and annotate transmembrane domains from known structures. This requires definition of the membrane edge, from which the protein should be considered, including the distinct orientation of single helices and the nature of the host membrane. Solutions for this problem have been proposed on the basis of physico-chemical analysis and the results are now included in databases like OPM and PDBTM.

A consequence of these updated annotations is to emphasise that transmembrane domains do not recapitulate the whole architecture of such proteins. Indeed, a quick survey of PDBTM annotation show that only 27\% of residues of membrane proteins of known structures are within the transmembrane domains. As the number of subunits rises, the chances of including extrinsic subunits or large extracellular domains becomes greater. How do these non transmembrane
structures interfere with the two or three stage model of folding? To what extent, in such conditions, do transmembrane domains still play their predominant role?

Non-transmembrane elements like these large extramembrane domains may resemble soluble proteins. Also, many other structures like short loops, helical extensions of TM, or helices lying in the membrane surface are frequently encountered. Furthermore, within the slice that corresponds to the most hydrophobic core of membrane proteins non transmembrane elements must also be considered.

To address the role of such elements in the assembly of membrane proteins it is desirable to define, classify and locate them on available membrane protein structures. Then, their pertinence for the description of membrane protein architecture can be considered relatively to residue composition and physical chemistry. Finally, their individual role relatively to that of TM can be investigated.

In this article we carefully and manually annotated such elements from a selection of non-redundant membrane protein structures and have made a detailed statistical and physico-chemical analysis. The results provide an approach for further analysing the complexity of the folding of these proteins. It provide new characters whom role in establishing and maintaining membrane protein architecture could be now considered. The statistics about the analysed structural elements give rise to the possibility of developing tools for the prediction of these elements in membrane protein structures.

Results

Preliminary analysis

As the result of a selection procedure (Figure 1; see Methods), a set of 54 non redundant membrane proteins where retained (Table 1) for structural annotation. The different annotations are depicted on diagrams of parts of the spinach aquaporin SoPIP2 (Figure 2; panel A) and the spinach major
light-harvesting complex LHCII (panel B). The polypeptide chains are shown as ribbons and the cofactors as space-filling models. The different annotations have been color-coded to aid the eye. The section of the aquaporin shows the red TM with in certain cases orange helix extensions (EH), the membrane penetrating loops (LM) are shown in blue and the incomplete membrane embedded helices (IH) are depicted in purple. The yellow extramembrane loops (LO) connect the different elements. In addition, the helices that lie in the vicinity of the predicted membrane plane (HH) are green. The light-harvesting complex shown in panel B illustrates numerous dark grey prosthetic groups (PG) and the terminal regions (NC) in pink.

In Table 2 we distinguish the content of the protein selection in terms of the assembly units which are distinct from the crystallographic asymmetric units. First the biological units, that correspond to the functional entities in the native membrane provide the appropriate ensemble to study molecular interactions and second the collection of unique polypeptides which is the proper set for the majority of statistical analyses. This distinction is important because many of the proteins included in the database are multimeric. From Table 2, it is clear that the abundance of the structural elements varies considerably with numerous helix extension (EH) and relatively few intramembrane helices (IH) and intramembrane loops (LM).

For the various structural elements, Figure 3 shows the cumulative length distributions of the different components in unique polypeptides. This representation highlights that most structural components are considerably smaller than transmembrane domains. At least 65% of each of the newly annotated elements are less than 9 residues long. Moreover, even 80% of loops and extremities are smaller than 20 residues.

The intramembrane helices (IH) distribution (violet) corresponds to half of the length of the TM which suggests that this structure crosses on average a single leaflet of the membrane. This structure is often associated to membrane loops (LM) for the return to the membrane surface, as fewer residues are necessary the corresponding trace (blue) is shifted toward smaller element size. Membrane loops (LM) show a small plateau in the distribution which indicates that the data set has
no examples containing 8 or 9 residues, but smaller and larger loops. This may be due to the existence of two different types of LM, short and long however we can not exclude that it simply derives from poor sampling.

**Representativeness of the protein collection**

Our selection of proteins contains only 54 proteins and 122 unique polypeptide chains and will not adequately sample a genome wide distribution of membrane proteins. However at the level of the 387 unique transmembrane helices there is a chance that the it becomes more representative.

This is illustrated in Figure 4 where we compare the distribution of the number of transmembrane domains per polypeptide in the unique polypeptides from the membrane protein set, and the distributions observed in the *E.coli* and human genomes\(^{15}\). Overall the observed distribution of TM number per polypeptide is not far from what might be expected from a random mixture of eukaryote and prokaryote α-helical integral membrane proteins. Thus, at least on this criterion it would seem relatively representative of membrane proteins in general. This may also be the case for other well populated structural elements like helix extension (EH), extramembrane loops (LO), termini (NC) and horizontal helices (HH) that are represented by more than 100 distinct structures in unique polypeptides. However, with only 21 intra-membrane helices (IH) and 19 intra-membrane loops (LM) conclusions concerning these annotations must be made cautiously.

**Composition**

The graphic in Figure 5 shows the composition of the different components and their enrichment relative to Swiss-Prot (y axis) and transmembrane domains (x axis). The graph is divided in 2 panels for clarity. The proportion of the residue within each structural element is encoded in the size of the residue one letter code. In these scales expressed in ln of an odds ratio,
the residues in transmembrane segments (TM; red symbols) are by construction aligned at x=0. It is obvious that all extramembrane elements (bottom; LO, EH, HH and NC) tends to spread along the y=0 axis (black, orange, green and pink symbols) showing that their general feature resemble that of Swiss-Prot. In contrast the membrane embedded element (top; IH and LM (violet and blue symbols) appear very dispersed. This is largely due to the poor statistics on most of the corresponding residues (see supporting information, figure S1 and S2 for error bars).

However, the location of the big blue G on the upper right corner of the graphic is highly significant and represents the 21% of glycine in the composition of intra-membrane loops (LM). The position indicates that glycine is over-represented when compared to the composition of Swiss-Prot (the upper section of the graphic) and it is also over-represented when compared to the composition of the TMs (the right section of the graphic). Other features for LM are also meaningful at the 95% confidence level: the over representation of Asn and the under representation of Trp relatively to TM. On the contrary the statistics are not sufficient for all the other residues including the apparent under-representation of Cys relatively to both Swiss-Prot or TM. The enormous over abundance of Gly in loops is possibly linked to the need for a particular flexibility or conformation in this type of loop. It is conceivable that the high affinity of Trp to the hydrophilic-hydrophobic interface opposes the membrane embedding of such loops. As shown in a later section, the presence of Asn at the LM in the vicinity of IH may be the signature of capping motifs that could stabilize these particular structures.

The intra-membrane helices (IH) are also characterized by poor statistics for most residues. However the enrichment in Ala, Gly, Leu, Phe and probably Trp is meaningfully very close to that of TM. In this kind of structure Pro and Asp appear over represented relatively to TM. So in contrast to membrane loops the abundance of Trp apply for IH. The abundance of Pro is probably linked with the helix breaking function of this residue that can stabilise a short helix. A remarkable characteristic of IH is its deficiency in His when compared to transmembrane helices. The same
deficiency relatively to both Swiss-Prot and TM is observed for Cys. We can conclude from the composition of these different elements that they are characteristic to the element in question.

The extramembrane elements (EH, LO, NC and EH) have a composition similar to that of Swiss-Prot but several variations discriminate among them. For example loops (LO-black) and terminal domains (NC-pink) have very well defined statistics showing a globally similar behaviour. However Gly are twice as abundant in loops and Cys half as abundant. Other smaller but meaningful differences concern charged and hydrophilic residues including Gln, Arg and Glu. Horizontal helices (HH in green) have a composition similar to that of EH, as perhaps is to be expected for non-membrane embedded \( \alpha \)-helical segments. However their differences are also very meaningful. For example the Gly content of horizontal helices (HH) mimics that of Swiss-Prot, but it appears very under-represented in helix extensions (EH): flexibility is probably an unwanted feature in such structures. Trp have the same marked tendency to be be over-represented in both structures, however the green W in the upper right corner is the most biased residue of the whole analysis. Trp is known to serve as a membrane anchor\(^{17} \), so this compositional anomaly reinforces the belief that these helices play a role in the anchorage and organisation of membrane proteins\(^{18} \).

Another aspect that is particularly visible in Figure 5 is the extent that certain residues are biased, this is visible as the length of the diagonal line containing particular letters. Thus we see enormous variations in the content of Trp and His, with ln(odds ratios) varying by about 2.5 between different components and very minor variations in for example Ser or Leu content between components.

The above description shows how each annotated structure may have a distinct composition, however it is limited to major features. From a more general point of view we now address the physical chemistry and the information content of the corresponding sequences.
Hydrophobicity

The differences in the overall composition of the structural elements will also change such physico-chemical properties as their hydrophobicity. Three different experimental hydrophobicity scales were used to measure the average hydrophobicity in each component, a side-chain scale and the two whole residue scales using POPC and octanol as the hydrophobic medium. Since the different components have different environments we might expect differences in membrane affinity and thus hydrophobicity. The results of these calculations are collected in Table 3. It is important to remember that the first scale is a side chain scale thus the differences rather than absolute values are important, while in the whole residue scales the zero has a significance, in addition these latter scales can take into account other effects beyond those of the side chain, for instance the hydrogen bonding effect of the helix. For the whole residue scales, the POPC and octanol scales are believed to reflect slightly different tendencies, with the POPC better representing the membrane interface and octanol the hydrophobic membrane core. In general the three scales give very consistent results.

The differences between the POPC and octanol scales are thus particularly pleasing with transmembrane elements partitioning into the octanol, the membrane core, while horizontal helices prefer POPC, the membrane interface, in complete agreement with their structural location. Intra-membrane loops which penetrate more deeply than horizontal helices, but do not cross the membrane, are more or less equivalent in the two whole residue scales. Equally reassuring is that the loops (LO) and terminal domains (NC) are not hydrophobic and have no tendency to partition into the interface. The helical extensions (EH) again have a very reasonable profile and appear to have an affinity for the interface region as indeed one would expect from their location. Thus all these elements seem to have a hydrophobicity entirely in keeping with their native environment. The unique exception is the intra-membrane loops, the composition of which results in a
hydrophilic profile using octanol and a weak affinity for POPC, and yet these elements penetrate the membrane as deeply as the intra-membrane helices (IH), it would thus seem that the formation and positioning of these loops is perhaps dependant on more subtle physico-chemical attributes than hydrophobicity and might be driven by constraints coming from other structures.

Information

An alternative method to examine the variations of composition between different elements is to use the Kullback-Leibler divergence (D_{KL}) also termed relative entropy of Shannon\textsuperscript{21-23}. This is a measure of the difference between two probability distributions (see methods). This method has an advantage over the more usual approach of using the ln(odds ratio) in that abundance effects are taken into account. In Table 4, we show the D_{KL} of the different elements using the composition of the 54 membrane proteins as the reference. Though it is not straightforward to evaluate the absolute numbers in Table 4 it is pertinent to realize that the prediction of transmembrane helices is relatively efficient even using a first principle approach without sequence alignment's (up to 94\% in the SCAMPI methods\textsuperscript{24}). Then, a divergence of 0.109 over about 15 residues, (Figure 3), could be sufficient for predictive purposes.

Surprisingly the highest divergence is found in the composition of intra-membrane loops (LM), these as we have seen above are particularly rich in glycine and have other anomalies in their composition. This compositional bias suggests that despite their small average size of 7 residues efficient automatic prediction might be possible. Furthermore such high divergence strongly suggests that composition bias has functional implications for the folding of these proteins. However we must keep in mind that with only 19 representatives in the database the results for LM substructures need to be confirmed. This will be achieved by incorporating results from new (non redundant) membrane protein structures using an automatic procedure that we are working on.

In table 4, the IH substructures have a divergence of 0.076 similar to that of transmembrane helices.
(0.109). It would seem given this divergence and their size that the composition might be sufficient for automatic prediction. However, as with intra-membrane loops the small number of these elements in our records means that this result needs confirmation. The remaining elements have lower divergence, thus the composition of these different elements are probably too similar to be easily distinguished.

In Table 5 we have calculated the Kullback-Leibler divergence for each element using the other elements as references. It is clear that most of the divergences are considerably larger than those calculated using the total composition as the reference. This comparison is pertinent to the use of Markov models for prediction rather than simpler predictive methods. Here, divergences fall in the range 0.41 to 0.0, as opposed to 0.183 to 0.0 in Table 4. Indeed 20 of the 42 comparisons have a relative entropy greater than 0.2, two thirds are greater than 0.15 and 86% greater than 0.1.

The most informative changes in composition were those associate with the transitions between LM and HH, with values near 0.4. This high value contains a large contribution from the glycine contents of these elements. In the context of a protein structure these two elements could follow each other, however there are no examples of it in our data set. Particularly high information content is also associated with the transitions for N-terminal domains (NC) to transmembrane helices (TM) and loops (LO) to transmembrane domains (TM). These elements are often sequential, and are also clearly separated by the $D_{KL}$ in their sequence. It is important to note that there is more information in one direction of the evaluation of the relative entropy than the other, the values in the top-right and lower-left of Table 5 are different, this follows from the formula because of the weighting of the ln(odd ratios). This means that it should be easier distinguishing NCs and LOs among TMs than finding TMs among LOs and NCs.

The relatively high $D_{KL}$ of the different comparisons indicates that all the different elements that we describe have distinguishable compositions and that it is possible that the annotation scheme that we propose could usefully be used to develop a predictive method. Indeed the only frequently observed transitions that does not involve a marked change in sequence composition,
D_{KL} near 0.08, are those between extensions of membrane embedded helices (EH) and N or C terminal regions (NC). The transition NC-EH is particularly common in the annotation, not surprisingly, about 60% of the N terminal regions are followed by a helical extension, while slightly less than 10% of C terminal regions are preceded by a helical extension.

The transitions with very low D_{KL} (~ 0.02) concern the transitions between loops and terminal domains and the transitions between helical extension and horizontal helices. Transitions between loops and terminal domains are impossible in our annotation scheme thus this low information content is unimportant. Direct transitions between helical extensions and horizontal helices are represented once in each direction, however it is not surprising that they are hard to distinguish simply on composition since the comparison concerns elements with the same secondary structure (α-helix) in the same type of environment (membrane surface). However application of additional filtering like hydrophobic moment analysis may help to discriminate among them.

**Homogeneity**

If the annotation scheme that we have developed is to be useful we need to critically assess its quality. We would like the annotation scheme to divide the structures into meaningful and distinguishable segments and we would like each annotation to give a homogeneous population of components and not to include for example two different types of object with the same label. This assessment requires two different measurements: first are all the differently annotated components really different, and second are the components homogeneous?

This assessment requires a statistical evaluation of the homogeneity of the populations, we have done this at the amino-acid level to ensure reasonable significance. Our measurement based on the Kulkarni-Shah test\textsuperscript{25} as described by Krishnamoorthy et al.\textsuperscript{26} (see Methods). Briefly we attempt to evaluate if each structural element (for example each individual transmembrane helix) could
reasonably be expected to derive from a population with the average composition. For example if
the probability of a Trp in a particular structural element is 0.02 then in such elements with a length
of 20 amino-acids we expect to frequently find zero, one or two Trp (67%, 27% and 11%
respectively) and very rarely 5 Trp (0.4%), if the distribution is homogeneous. However, if the
distribution is inhomogeneous the same average proportion could arise from a mixture of elements
with zero (82%) or five (8%) Trp. This test was performed for the distribution of each amino-acid in
all of the different elements, using the annotations proposed above, sub-divisions and merged sets
of these annotations. Through these different tests we have tried to examine to what extent similarly
annotated segments resemble each other and if we can improve the annotation scheme.

In Table 4 are also shown the inhomogeneities observed in the 7 structural annotations that
we have used, only those inhomogeneities significant at the 1% level are shown. Since the tests
have been made 20 times, one for each amino acid, for each of seven elements, a total of 140 tests,
by chance one or two inhomogeneities at the 1% level can be expected. For the different
inhomogeneities detected a number of possible origins were examined in an attempt to assess if a
sub-division of the annotation would be useful or informative. Among all tests performed only 9
inhomogeneities where found at the 1% level. For most of these, structural or functional reasons can
be proposed. For example, over representation of His in TM is associated with coordination of
cofactors in the protein and functionally related to oxydo-reductase activity. Also, Lys
inhomogeneities detected within some NC may be a logical consequence of the known asymmetry
of the basic residue composition across the membrane (positive inside rule). Gly inhomogeneities in
loops are associated with requirement for flexibility in smaller loops, whereas the His excess
involve cofactor coordination as for TM. The presence of excess of N in some loops could be
associated with helix capping propensities of such residues in this kind of structure. Particularly,
this residue appears overrepresented in smaller loops. The M inhomogeneity in NC corresponds to
the translation initiation codon. Other inhomogeneities incompatible with a simple distribution
involve Ala in some HH or Phe in few TM. Explanations for these are less obvious: it is known that
A has a weak but distinct preference for the interface region\textsuperscript{27}. It is possible that this role as a typically interfacial component is at the origin of this over-abundance in certain helices. An unusual enrichment of Phe in TM essentially concerns proteins containing many cofactors but also the notable exception in the first transmembrane helix from the lactose permease which contains 10 Phe! No clear explication was associated to these observations.

Thus a possible way to achieve better composition homogeneity would consist of separating short and large loops. Unfortunately the variations in composition with size were slow and continuous and it proved difficult to find a reasonable method to annotate "small" and "large" loops. Thus, though in a predictive scheme it would seem useful to distinguish large loops often containing soluble domains and short loops that serve to send the polypeptide chain back across the membrane, we have not yet found a reliable method for applying this distinction.

To consider if all the different annotations that we used are justified we also examined how the heterogeneity of amino acid distribution in the components was affected by merging different blocks. The majority of the attempted fusions resulted in more inhomogeneous groups. Only the fusion of NC and LO maintained reasonably low heterogeneity, except for the G and N distribution previously associated with small loops that in the merged context appear much less homogeneous. As expected from the composition analysis larger loops and NC are very similar, but they are trivial to discriminate so that we decided to maintain them separated especially in view of our concern about loop heterogeneity.

**Relation to the membrane**

Crucial to our annotation is the position of the protein with respect to the membrane. Unfortunately in most structures the membrane is not actually visible, and only occasionally are several lipid molecules visible on the surface of the protein. Although the structure of AQP0 was determined with part of its membrane by cryo-electron microscopy\textsuperscript{28}; the set of non redundant membrane
protein sequences presented here the membrane is approximated as a flat hydrophobic region as determined by TMDET. The organisation of the different annotations of the proteins was examined in relation to the membrane normal (Figure 6).

Although the TMDET algorithm uses only physico-chemical constraints to delineate the membrane core slice, the residue population is equally and symmetrically distributed on each side of the membrane. However, this global symmetry is not maintained for 4 of our annotations. Helical extension (EH) are found at 61% in the inside whereas loops (LO and LM) are more common (65%) in the outside leaflet. Importantly in our sample of proteins 85% of the extrinsic units (UN) are on the inside surface of the membrane. Overall this finding is probably important for the architecture of membrane proteins showing that the inside is more structured with EH that may support anchor of extrinsic subunits.

Examination of asymmetry for various residues (Figure 7) shows that the well known asymmetry of basic residues\textsuperscript{15, 29, 30} which probably plays a role in stop-transfer sequences during insertion\textsuperscript{31}. Our annotation shows that this asymmetry is not only associated with the extensions to transmembrane helices (EH) but also with HH and LO. However, the positive inside rule seems not to apply to NC extremities.

Secondly (Figure 7 middle), in apparent contradiction with previous findings of a “negative outside” trend\textsuperscript{32-34}, we find that acidic residues also show a weak “negative inside” tendency. However, this feature appears very contrasted among the substructures and the layer considered (the z axis). Loops do show a “negative outside” tendency and are enriched in D+E in the interfacial layers (-/+ 11-30 Å) and in the outer layers (-/+ 30-45 Å). In these regions, the enrichment reach 1,7 and 1,30 respectively. An even greater enrichment, 2,4 fold, is observed for the outer layer NC regions. Intriguingly, this contrasts to what happens with the interfacial NC regions which are strongly enriched in acidic residues on the inside by a factor of 1,5. The “negative inside” asymmetry is also marked for horizontal helices (HH) and for EH which are inside enriched by a factor of 1,9 and 1,4 respectively. Thus strong opposing trends compensate to result in the small
overall tendency. Therefore, it is not very surprising that depending on the protocol for membrane proteins selection and the definition of their topology from the sequences, some contradictory results have emerged.

As a final example, Asn (N) (Figure 7 bottom) shows a strong tendency to be associated with loops close to the outside surface of the membrane. Three previously suggested hypotheses seem plausible: a role in helix capping; a special role in the extra cellular loops of eukaryotic proteins or the presence of glycosylation sites. More interesting is the over representation of Asn in membrane loops near the centre of the membrane slab. This also positions the residue at the vicinity of the adjacent IH thus re-enforcing a role for the capping of those particular helices. This feature may be of particular interest for prediction of this kind of structure.

**Discussion**

We have annotated a non-redundant set of $\alpha$-helical membrane protein structures to describe the relationship between local structure elements and the membrane environment. The composition of these elements reflects their surroundings during folding as described by the three-stage model, and the structures they adopt. This description opens perspectives for the study of relationships between elements in the complex membrane protein architectures and will help to understand their role in the assembly and or in the function.

We have used TMDET to determine the portion of the protein surface most likely to be in the hydrophobic membrane core. This automatic method, which is the one used by PDBTM to extract membrane protein structures from the PDB, is based on the burial of non-polar residue surface area by insertion into the membrane. It should be noted that several other methods exist for this purpose. Including Garlik, IMPALA or PPM.

In a comparison of the different methods Lomize et al. showed that different estimates of bilayer
thickness may be obtained. In a set of TM α helical proteins the differences correspond to a standard error of 5 Å which corresponds to less than one helix turn on each side. This is roughly the same error as manually attributing the membrane boundaries. This uncertainty raises a concern with our findings: how do our statistics on residue composition depend on the choice of TMDET. To answer this question we have specifically analysed the residue compositions of the 4 residues of EH and TM that straddle the TMDET boundaries. Reassuringly, these sub-populations strongly resemble the population from which they were taken (EH or TM) and differ from the neighbouring population.

Previous works have mainly focused on the transmembrane domains of α-helical membrane protein. The important contribution of proper packing and assembly of these structures in the folding of membrane protein has been conceptualizes in the “two-stage” model\textsuperscript{11} and illustrated in many experimental studies concerning both bitopic and polytopic protein as well as model peptides (reviewed in\textsuperscript{41}).

However, a first level of complexity has emerged from considering that transmembrane domains are tilted and not idealized α-helical structures\textsuperscript{42}. This severely limits our ability to predict the real packing\textsuperscript{43}. Moreover more than a third of the unique polypeptides present in our protein subset also contain re-entrant loops and/or incomplete helices in the membrane core. These structures should also be considered for the proper description of the folding of membrane proteins.

Last, such membrane embed structures only represents 1/4 of our dataset as estimated by the number of residues. There is no reason why they should be the sole contributor to the membrane protein assembly. Indeed, the important role of transmembrane helices has been investigated mostly for small integral membrane proteins like bacteriorhodopsin or for intra-subunit folding but much less work has been done to study their role in the organization of complexes made of multiple subunit assemblies that often include extrinsic polypeptides.

This analysis of the different sources of membrane protein complexity demonstrates the need for
more details in the description of their structure by considering more substructure including extramembrane elements. However, environment or sequence constraints should support the pertinence of such substructures. Statistical differences linked to physical chemistry rather than empirical clustering should render them more practicable for predictive purposes.

Indeed, most of the structural elements proposed in our annotation meet these requirements: they have compositions in agreement with the physico-chemical properties of the region where they find themselves in during folding and the analysis of the compositional specificity of the different elements that we identify suggests that they convey sufficient information to be distinguishable, and could thus form the basis for future progress in structure prediction scheme specific for membrane proteins. For example, the distinction of helix extensions as a separate substructure may alleviate the problem of finding such a trivial characteristic as the exact length of transmembrane helices which remain very difficult to predict using current methodologies.

The paucity of information concerning incomplete helices and membrane loops is a consequence of the low number of non redundant high quality membrane protein structures. However, in spite of poor overall statistics some very significant features may be also pertinent for distinguishing them from transmembrane sequences. For example the enrichment of glycine and the role of asparagine, appear to be very significant. Other studies have particularly drawn attention to such “reentrant loops” and predict that based on residue composition they may be detected at an accuracy of approximately 70% and genome wide analyses suggests that 10% of membrane proteins contain such alternate structures which is consistent with the enrichment found in our subset of analysed structures.

Ideally the annotation should be available for the whole database of membrane protein structures. Our initial and manually annotated subset of structures give the rules, and the preliminary statistics a control, for an automated annotation procedure which is currently in progress in our group. This will open the possibility to improve the statistics by re-incorporating some of the information removed from the non redundant subset. Such better statistics are needed to improve algorithms for
topology and structure prediction that should include new structural element pertinent for membrane protein folding and the architecture of complexes.

The finding of new inside-outside enrichment patterns for acidic residues, in respect to the substructure and the local physical chemistry, illustrates how the rise of complexity might help in better understanding the constraints on membrane proteins and ultimately the prediction of more detailed and accurate topologies for membrane proteins. Beyond this, the rise in complexity we propose will be also of interest for describing the interactions that stabilize the fold of membrane proteins. It opens the way for the evaluation of the respective role of substructures in context of various membrane protein architecture. The analyses of network of interactions would help to answer fundamental questions like the relative contribution of transmembrane domains for the monomolecular intrachain folding and for the promotion of the aggregation in multiple subunits complexes.

Materials and Methods

Generation of an annotated non-redundant membrane protein structure set

The selection of membrane protein structures, adequate for exploring the different interactions that lead to membrane protein folding, should consider quality of the experimental structures, a maximum of diversity and should allow their assessment with respect to natural environment in the membrane. Therefore, the redundancy and the quality of the available structures was used to strike a balance in the selection with the aim of maintaining a reasonable number of molecules but representative of the full database. The generation of the data set involved three different phases: first selection of a non-redundant set of membrane protein structures; second improvement and completion of these structures; and finally annotation of these structures.
Selection of proteins.

The selection procedure of the structures is outlined in Figure 1. We chose PDBTM as the starting point of our study and as a main source of data, including list of proteins and a first level of structural annotations. The choice of PDBTM was driven by its great completeness and its high level of maintenance which assure an automatic and weekly based update since its creation on September 2003. Also, PDBTM provides structural annotations that make use of previous works of TMDET to discriminate between transmembrane and extra membrane domains.

Using the list of protein found in PDBTM (pdb identifier), we wanted to select our own set of non-redundant protein that meet our criteria for quality.

This choice was carried out in several steps, as shown in Figure 1. First, poor quality structures are eliminated, then in 2 steps a non-redundant subset is selected based on the clustering of primary sequences.

In order to analyse the organization within proteins it is important to use reliable, high-quality, structures. Therefore in a first step PDBTM entries are filtered to remove low resolution and structures containing only backbone atoms. Structures determined by crystallography were considered to be of poor quality if their resolution was greater than 3.5 Å, this cut-off was selected as it retains 99% of the PDB as a whole, and is thus not too stringent, and corresponds approximately to a resolution adequate for correctly positioning side-chains. All-atom NMR structures were retained as it is more difficult to accurately assess their quality. However structures containing only backbone atoms were removed. After this initial filter 391 structures were retained.

The selection of a non-redundant subset was performed in two sequential steps. First the protein-sequences extracted from the structure files were clustered using cdhit with a 90% identity threshold. This step serves to cluster very similar structures together and facilitates the next step, in all 112 clusters were formed from the 391 structures. In the next step blastclust was used with a single representative of each cluster to create larger groups with greater than 35% identity.
Blastclust uses a score-based single-linkage clustering algorithm. The identity threshold was set at 35% in this step after examining the merger and break-up of clusters in the 20-50% region. This threshold kept cytochromes bc₆ and bc₆ in separate groups, but put all water channels in the same group. This procedure gave 67 different groups of structures among which 13 were dismissed, including monotopic proteins, β-barrel proteins or proteins such as mistic. Each group was then examined and for each a single representative structure was chosen. The selection was based on the completeness, nativeness and resolution of the different structures in the group. Specifically, structures with missing polypeptide chains were eliminated from the group, then those with many mutations were eliminated, then any structures of excited or meta-stable states were eliminated, finally, the structure with the best resolution among those remaining in the group was selected.

The final set contains 54 α-helical membrane proteins, including NMR structures and structures with only one TM for the sake of completeness. These are listed in Table 1.

**Annotation.**

We wished to annotate our selected set of membrane proteins both with respect to the internal structural elements, their location within the membrane, and their possible role in assembly of the protein. The usual simple classification into: prosthetic groups, transmembrane helices, and inter-helical loops was extended and elaborated.

The structural annotation depends on knowledge of the position of the membrane plane. We have used TMDET to determine the portion of the protein surface most likely to be in the hydrophobic membrane core. This automatic method, which is the one used by PDBTM to extract membrane protein structures from the PDB, is based on the burial of non-polar residue surface area by insertion into the membrane.

In Figure 2 the position of the hydrophobic core of the membrane plane is shown by the light grey
lines. We have also annotated the orientation of the different proteins in the membrane. This was determined using available literature for each protein, with the cytoplasmic, mitochondrial matrix or chloroplast stroma considered to be “inside” while the extracellular, periplasmic, granal lumen or intramembrane space were considered to be “outside”. For each protein the structure of the biological unit was taken and oriented so that the z-axis was normal to the presumed membrane plane and negative z-coordinate corresponds to the inside (cytoplasmic, mitoplasmic or stromal) compartment, i.e. the surface of the membrane from which the protein is inserted.

On the basis of TMDET results, the transmembrane helices are divided, if necessary, into the transmembrane part (TM) and the hydrophilic extensions of these helices (EH). The idea of inter-helical loops was refined to separate horizontal helices (HH) often found at the surface of the membrane; incomplete helices (IH) that penetrate the membrane plane but do not cross it; membrane loops (LM) that penetrate the bilayer plane and are frequently associated with incomplete helices; the polypeptide extremities (NC); and the associated but not membrane crossing polypeptides (UN). Most of these different structural elements are illustrated in the structures shown in Figure 2. For this annotation α-helices play a particularly important role, STRIDE49 was therefore used to detect secondary structure elements.

Indeed, when annotating the proteins manually, we tried to integrate the element secondary structure, the position relative to the membrane and the preceding and following blocks to come up with a coherent annotation scheme. Again, at this stage our aim was not to apply strong constraints for annotation but to establish a manually checked gold standard that will serve in the future to optimize strategy for automatic annotation that could be applied to the entire database of membrane protein structures.

Statistics.

The Kullback–Leibler divergence21-23, also termed relative entropy or information gain, is used as an indicator of how much information has the composition of a component j with respect to a
reference k.

\[ D_{KL_i} = \sum_{i=0}^{i=20} p_{ij} \ln \left( \frac{p_{ij}}{q_{jk}} \right) \]

Homogeneity

We would like the annotation scheme to divide the structures into meaningful and distinguishable segments and we would like each annotation to give a homogeneous population of components and not to include for example two different types of object with the same label. For this purpose, we have tracked inhomogeneities for all the residues within each individual segment. We call inhomogeneity, an observed population which can not be explain by the proportion of the corresponding component at a given confident interval.

We used a Monte Carlo method based in the Kulkarni-Shah test \(^{25}\) for estimating the equality of proportions to a specified standard as described in \(^{26}\).

Here the hypothesis of interest are:

\[ H_0 : p_1 = \ldots = p_m = p_0 \]

\[ H_1 : p_i \neq p_0 \]

where \( p_1, \ldots, p_m \) are given proportions in m samples (distinct elements) and \( p_0 \) is the reference proportion (global composition).

Inhomogeneities were evaluated on the basis of the kulkarni-shah test:

\[ T_k = \sum_{i=1}^{m} \frac{X_i - n_i p_0 q_0}{n_i p_0 q_0} \]

with \( q_0 = (1-p_0) \) whose null distribution is asymptotically chi square with m degrees of freedom, therefore the null hypothesis \( H_0 \) will be rejected with risk \( \alpha \) whenever the p-value:

\[ P \left( \chi^2_m > T_k \right) \leq \alpha \]

In the Monte Carlo procedure adapted from \(^{26}\) we used the pooled sample proportions to randomly
create samples for every amino acid. These samples had the same length as observed experimentally, and for this collection of samples we calculated Tk. The proportion of Tk's that are greater than or equal to the Tk calculated for the original data are then a Monte Carlo estimate of the p-value.

This procedure was iterated 100000 times thus the maximum simulation error would be

\[ 2 \times \frac{0.5 \times 1 - 0.5}{100000} = 0.0032 \]

which seems good enough.

This p-value can be interpreted as an indication of how likely the observed distribution for the given amino acid is to stem by random choice from a universe with the proportion of the pooled sample.

**Acknowledgements**

This work was supported by grants from the CNRS especially on the programmes Nanosciences-2004 (NR206) and Protéomique et Génie des Protéines (PGP). JAL was supported by a studentship from the French Ministry of Education Research and Technology (MENRT).
References

Figure Legends

**Figure 1.** Schema used to select proteins for inclusion in the data set. Starting from PDBTM the different steps in the selection procedure and the number of different structures, or groups of structures, at each stage are shown. On the right are shown the various cut-offs and parameters used for the selections. See Methods for details.

**Figure 2.** Illustration of the different structures annotated. Two different structures are shown, in panel A a section through the spinach aquaporin (SoPIP2) and in panel B the spinach major light harvesting protein LHCII, the polypeptide chains are shown as ribbons and prosthetic groups as space-filling models. The different annotated structural components are shown using a colour code. The position of the hydrophobic core of the membrane, determined by TMDET, is delimited by two black lines. Within the membrane are the transmembrane helices are in red (TM), the incomplete helices that do not cross the membrane in purple (IH) and the intra-membrane loops (LM) are in blue. Beyond this hydrophobic core we can see the extension of the transmembrane helices in orange (EH) and other element such as the N or C termini in pink (NC), the loops between the transmembrane helices in yellow (LO) or the horizontal helices that run approximately parallel to the membrane surface(HH) in green. There are also numerous the prosthetic groups (PG) in grey.

**Figure 3.** Cumulative length distributions of the different components. For each structural annotation, the number of such components of different size was calculated. Here we show the cumulative distributions, which rise from 0 to 1 and illustrate the differences in length. For clarity the graphic shows only the distribution up to 30 residues. The lines are colour coded, using the same code as in Figure 2 namely: transmembrane helices (TM) in red; intra-membrane helices (IH) purple; intra-membrane loops (LM) in blue; transmembrane helix extensions (EH) in orange; N and C termini in (NC) pink; loops (LO) in yellow; and horizontal helices (HH) in green.

**Figure 4.** Number of transmembrane helices per polypeptide. Comparison of the distribution of the number of transmembrane helices per polypeptide in polypeptides extracted from our non-
redundant data set (open bars), and two sets of predicted α-helical transmembrane proteins, those from the *E. coli* genome (light gray) and the *H. sapiens* genome (dark gray).

**Figure 5,** Amino acid composition of different structural components. The amino acid composition of each component was calculated, and from this the log odds ratio of finding each amino acid in a component, compared to that of finding the same amino acid in either the non-redundant Swiss-Prot database – ln odds ratio vs Swiss-Prot, or compared to transmembrane segments – ln odds ratio vs TM. Each residue in each structural component is then plotted using the single letter residue codes with a size encoding its composition within the component, the position indicating its proportion compared both to all proteins and transmembrane segments, and a color encoding the component. The color codings are: transmembrane helices (TM) in red; intra-membrane helices (IH) purple; intra-membrane loops (LM) in blue (top panel); transmembrane helix extensions (EH) in orange; N and C termini in (NC) pink; loops (LO) in black; and horizontal helices (HH) in green (bottom panel). Note that in this representation each amino acid is found on a diagonal line. The upper panel shows the amino acid compositions of membrane embedded substructures (TM, IH and LM) while the lower panel illustrate the composition of the “soluble substructures (EH, HH, LO and NC). The inset in the upper panel illustrates the KL divergence of the different substructures from transmembrane domains (horizontal axis) and the Swiss-Prot database (vertical axis).

**Figure 6,** Depth distribution of different structural components in the membrane. The distribution of amino-acids along the membrane normal axis was calculated. Bins 2 Å wide centred on the middle of the predicted membrane were used, negative positions correspond to the inside (intracellular, cytoplasm, stroma) and positive to the outside (extracellular, periplasm, lumen). The fraction of all residues in each bin was calculated and color coded to indicate the distribution between different structural components. The color codings are as in Figure 2, namely: transmembrane helices (TM) in red; intra-membrane helices (IH) purple; intra-membrane loops (LM) in blue; transmembrane helix extensions (EH) in orange; N and C termini in (NC) pink; loops (LO) in yellow; and horizontal helices (HH) in green, and non-membrane spanning polypeptides
(UN) in grey.

**Figure 7**, Distribution of several amino acids in the membrane and between structural elements. The distribution of residues was calculated and displayed as for Figure 6. We show the distributions of: top, basic amino acids; middle, acidic residues; bottom, asparagine.
### Table 1, Proteins included in the data set.

<table>
<thead>
<tr>
<th>PDB code</th>
<th>Resolution</th>
<th>Protein name</th>
<th>Protein Origin</th>
<th># TMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2CPS</td>
<td>NMR</td>
<td>major coat protein</td>
<td><em>Coliphage m13</em></td>
<td>1</td>
</tr>
<tr>
<td>1ZZA</td>
<td>NMR</td>
<td>Stannin</td>
<td><em>Homo sapiens</em></td>
<td>1</td>
</tr>
<tr>
<td>1SPF</td>
<td>NMR</td>
<td>pulmonary surfactant-associated polypeptide SP-C</td>
<td><em>Sus scrofa</em></td>
<td>1</td>
</tr>
<tr>
<td>1SKH</td>
<td>NMR</td>
<td>N-terminal of bovine prion protein</td>
<td><em>Bos taurus</em></td>
<td>1</td>
</tr>
<tr>
<td>1RKL</td>
<td>NMR</td>
<td>oligosaccharyltransferase subunit Ost4p</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>1</td>
</tr>
<tr>
<td>1JI</td>
<td>NMR</td>
<td>neu/erbB-2 transmembrane segment</td>
<td><em>Rattus norvegicus</em></td>
<td>1</td>
</tr>
<tr>
<td>1JDM</td>
<td>NMR</td>
<td>sarcophilin</td>
<td><em>Homo sapiens</em></td>
<td>1</td>
</tr>
<tr>
<td>1AFO</td>
<td>NMR</td>
<td>glycofohin a</td>
<td><em>Homo sapiens</em></td>
<td>2</td>
</tr>
<tr>
<td>1A9I</td>
<td>NMR</td>
<td>subunit c of the F1F0 ATP synthase</td>
<td><em>Escherichia coli</em></td>
<td>2</td>
</tr>
<tr>
<td>1WU0</td>
<td>NMR</td>
<td>subunit c of the F1F0 ATP synthase</td>
<td><em>Bacillus sp. ps3</em></td>
<td>2</td>
</tr>
<tr>
<td>1P49</td>
<td>2.6</td>
<td>human placental estrone/DHEA sulfatase</td>
<td><em>Homo sapiens</em></td>
<td>2</td>
</tr>
<tr>
<td>1WAZ</td>
<td>NMR</td>
<td>bacterial mercury transporter, MerF</td>
<td><em>Morganella morganii</em></td>
<td>2</td>
</tr>
<tr>
<td>1ZLL</td>
<td>NMR</td>
<td>phospholamban</td>
<td><em>Oryctolagus cuniculus</em></td>
<td>5</td>
</tr>
<tr>
<td>1Q16</td>
<td>1.9</td>
<td>nitrate reductase a, narGHI</td>
<td><em>Escherichia coli</em></td>
<td>5</td>
</tr>
<tr>
<td>1OKC</td>
<td>2.2</td>
<td>mitochondrial ADP/ATP carrier</td>
<td><em>Bos taurus</em></td>
<td>6</td>
</tr>
<tr>
<td>1XIO</td>
<td>2</td>
<td>sensory rhodopsin</td>
<td><em>Anabaena sp.</em></td>
<td>7</td>
</tr>
<tr>
<td>1U19</td>
<td>2.2</td>
<td>rhodopsin</td>
<td><em>Bos taurus</em></td>
<td>7</td>
</tr>
<tr>
<td>1R3J</td>
<td>1.9</td>
<td>potassium channel KcsA</td>
<td><em>Mus musculus</em></td>
<td>8</td>
</tr>
<tr>
<td>2AHY</td>
<td>2.4</td>
<td>Na+ complex of the NaK Channel</td>
<td><em>Bacillus cereus</em></td>
<td>8</td>
</tr>
<tr>
<td>1XL4</td>
<td>2.6</td>
<td>K+ channel kirbac3.1</td>
<td><em>Magnetospirillum magnetotacticum</em></td>
<td>8</td>
</tr>
<tr>
<td>1H2S</td>
<td>1.93</td>
<td>sensory rhodopsin II-transducer complex</td>
<td><em>Natronomonas pharaonis</em></td>
<td>9</td>
</tr>
<tr>
<td>1RWT</td>
<td>2.72</td>
<td>major light-harvesting complex</td>
<td><em>Spinacia oleracea</em></td>
<td>9</td>
</tr>
<tr>
<td>1MSL</td>
<td>3.5</td>
<td>MscL homologue</td>
<td><em>Mycobacterium tuberculosis</em></td>
<td>10</td>
</tr>
<tr>
<td>1QLA</td>
<td>2.2</td>
<td>fumarate reductase</td>
<td><em>Wolinella succinogenes</em></td>
<td>10</td>
</tr>
<tr>
<td>PDB code</td>
<td>Resolution</td>
<td>Protein name</td>
<td>Protein Origin</td>
<td># TMS</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>-------------------------------------</td>
<td>---------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>1SU4</td>
<td>2.4</td>
<td>calcium ATPase</td>
<td>Oryctolagus cuniculus</td>
<td>10</td>
</tr>
<tr>
<td>1AIJ</td>
<td>2.2</td>
<td>photosynthetic reaction center</td>
<td>Rhodobacter sphaeroides</td>
<td>11</td>
</tr>
<tr>
<td>1RH5</td>
<td>3.2</td>
<td>protein conducting channel</td>
<td>Methanococcus jannaschii</td>
<td>12</td>
</tr>
<tr>
<td>1ZCD</td>
<td>3.45</td>
<td>Na⁺/H⁺ antiporter NhaA</td>
<td>Escherichia coli</td>
<td>12</td>
</tr>
<tr>
<td>1PV6</td>
<td>3.5</td>
<td>lactose permease</td>
<td>Escherichia coli</td>
<td>12</td>
</tr>
<tr>
<td>1PW4</td>
<td>3.3</td>
<td>glycerol-3-phosphate transporter</td>
<td>Escherichia coli</td>
<td>12</td>
</tr>
<tr>
<td>1ORQ</td>
<td>3.2</td>
<td>voltage-dependent K⁺ channel</td>
<td>Mus musculus</td>
<td>12</td>
</tr>
<tr>
<td>2A65</td>
<td>1.65</td>
<td>LEUTAA</td>
<td>Aquifex aeolicus</td>
<td>12</td>
</tr>
<tr>
<td>1YEW</td>
<td>2.8</td>
<td>particulate methane monoxygenase</td>
<td>Methylococcus capsulatus</td>
<td>13</td>
</tr>
<tr>
<td>1XME</td>
<td>2.3</td>
<td>cytochrome ba₃ oxidase</td>
<td>Thermus thermophilus</td>
<td>15</td>
</tr>
<tr>
<td>1KQF</td>
<td>1.6</td>
<td>formate dehydrogenase</td>
<td>Escherichia coli</td>
<td>15</td>
</tr>
<tr>
<td>2A79</td>
<td>2.9</td>
<td>Shaker Kᵥ1.2 potassium channel</td>
<td>Rattus norvegicus</td>
<td>16</td>
</tr>
<tr>
<td>1LGH</td>
<td>2.4</td>
<td>light-harvesting complex II</td>
<td>Rhodospirillum molischianum</td>
<td>16</td>
</tr>
<tr>
<td>1NKZ</td>
<td>2</td>
<td>light-harvesting complex II</td>
<td>Rhodopseudomonas acidophila</td>
<td>18</td>
</tr>
<tr>
<td>1L7V</td>
<td>3.2</td>
<td>ABC transporter</td>
<td>Escherichia coli</td>
<td>20</td>
</tr>
<tr>
<td>1OTS</td>
<td>2.51</td>
<td>CIC chloride channel</td>
<td>Escherichia coli</td>
<td>20</td>
</tr>
<tr>
<td>1EI2</td>
<td>1.8</td>
<td>halorhodopsin</td>
<td>Halobacterium salinarum</td>
<td>21</td>
</tr>
<tr>
<td>1M0L</td>
<td>1.47</td>
<td>bacteriorhodopsin</td>
<td>Halobacterium salinarum</td>
<td>21</td>
</tr>
<tr>
<td>1YCE</td>
<td>2.4</td>
<td>F-type Na⁺-ATPase</td>
<td>Ilyobacter tartaricus</td>
<td>22</td>
</tr>
<tr>
<td>1Z98</td>
<td>2.1</td>
<td>spinach aquaporin SoPIP2</td>
<td>Spinacia oleracea</td>
<td>24</td>
</tr>
<tr>
<td>1PP9</td>
<td>2.1</td>
<td>cytochrome bc₁</td>
<td>Bos taurus</td>
<td>24</td>
</tr>
<tr>
<td>2F2B</td>
<td>1.68</td>
<td>aquaporin AqpM</td>
<td>Methanobacterium thermoautotrophicum</td>
<td>24</td>
</tr>
<tr>
<td>1FX8</td>
<td>2.2</td>
<td>glycerol facilitator (GlpF)</td>
<td>Escherichia coli</td>
<td>24</td>
</tr>
<tr>
<td>1XTH</td>
<td>3.5</td>
<td>glutamate transporter</td>
<td>Pyrococcus horikoshii</td>
<td>24</td>
</tr>
<tr>
<td>1JB0</td>
<td>2.5</td>
<td>photosystem I</td>
<td>Cyanobacteria</td>
<td>32</td>
</tr>
<tr>
<td>1U7G</td>
<td>1.4</td>
<td>ammonia channel AmtB</td>
<td>Escherichia coli</td>
<td>33</td>
</tr>
<tr>
<td>1SSL</td>
<td>3.5</td>
<td>photosynthetic oxygen evolving center</td>
<td>Thermosynechococcus elongatus</td>
<td>34</td>
</tr>
<tr>
<td>1OYE</td>
<td>3.48</td>
<td>AcrB multidrug</td>
<td>Escherichia coli</td>
<td>36</td>
</tr>
<tr>
<td>2BL2</td>
<td>2.1</td>
<td>V-type ATPase</td>
<td>Enterococcus hirae</td>
<td>40</td>
</tr>
</tbody>
</table>
For each protein are indicated the code of the PDB file used, the resolution of the structure, if it is a crystallographic structure, the name of the protein, its origin. The last column shows the number of transmembrane segments in the protein.

<table>
<thead>
<tr>
<th>PDB code</th>
<th>Resolution</th>
<th>Protein name</th>
<th>Protein Origin</th>
<th># TMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1V54</td>
<td>1.8</td>
<td>cytochrome c oxidase</td>
<td>Bos taurus</td>
<td>56</td>
</tr>
</tbody>
</table>
Table 2: Content of the selected membrane proteins

<table>
<thead>
<tr>
<th>Element</th>
<th>Biological Units</th>
<th>Asymmetric units</th>
<th>Unique polypeptides</th>
<th>Residue count in biological units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins</td>
<td>54</td>
<td>54</td>
<td>-</td>
<td>52305 (100%)</td>
</tr>
<tr>
<td>Monomers</td>
<td>140</td>
<td>150</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Polypeptides</td>
<td>254</td>
<td>272</td>
<td>122</td>
<td>-</td>
</tr>
<tr>
<td>Amino acids</td>
<td>52305</td>
<td>48018</td>
<td>26910</td>
<td>-</td>
</tr>
<tr>
<td>Transmembrane helices (TM)</td>
<td>750</td>
<td>768</td>
<td>387</td>
<td>14589 (27.9%)</td>
</tr>
<tr>
<td>Helix extensions (EH)</td>
<td>1041</td>
<td>1107</td>
<td>549</td>
<td>6255 (12.0%)</td>
</tr>
<tr>
<td>Horizontal helices (HH)</td>
<td>237</td>
<td>231</td>
<td>132</td>
<td>2335 (4.5%)</td>
</tr>
<tr>
<td>Intramembrane helices (IH)</td>
<td>68</td>
<td>54</td>
<td>21</td>
<td>729 (1.4%)</td>
</tr>
<tr>
<td>Intramembrane loops (LM)</td>
<td>63</td>
<td>43</td>
<td>19</td>
<td>473 (0.9%)</td>
</tr>
<tr>
<td>Extramembrane loops (LO)</td>
<td>779</td>
<td>797</td>
<td>422</td>
<td>10144 (19.4%)</td>
</tr>
<tr>
<td>Termini (NC)</td>
<td>337</td>
<td>360</td>
<td>158</td>
<td>6001 (11.5%)</td>
</tr>
<tr>
<td>Extramembrane chain (UN)</td>
<td>39</td>
<td>29</td>
<td>18</td>
<td>11779 (22.5%)</td>
</tr>
<tr>
<td>Prosthetic group (PG)</td>
<td>525</td>
<td>652</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Each structure was examined and the different structural elements that it contained were counted. As many biological units of membrane proteins are oligomers we counted the number of components for biological units, crystallographic asymmetric units and the different unique polypeptides. We also added the residue count of the components in biological units to give an idea of the overall size of these elements.
Table 3: Hydrophobicity of different components.

<table>
<thead>
<tr>
<th>Element</th>
<th>Pliska</th>
<th>Octanol</th>
<th>POPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM</td>
<td>-5.64</td>
<td>-1.54</td>
<td>-1.00</td>
</tr>
<tr>
<td>IH</td>
<td>-4.67</td>
<td>-0.70</td>
<td>-0.61</td>
</tr>
<tr>
<td>HH</td>
<td>-3.64</td>
<td>-0.34</td>
<td>-0.57</td>
</tr>
<tr>
<td>EH</td>
<td>-3.44</td>
<td>-0.04</td>
<td>-0.39</td>
</tr>
<tr>
<td>LM</td>
<td>-3.20</td>
<td>0.58</td>
<td>-0.18</td>
</tr>
<tr>
<td>LO</td>
<td>-2.52</td>
<td>0.86</td>
<td>0.04</td>
</tr>
<tr>
<td>NC</td>
<td>-2.51</td>
<td>0.81</td>
<td>0.08</td>
</tr>
<tr>
<td>Total</td>
<td>-3.82</td>
<td>-0.19</td>
<td>-0.42</td>
</tr>
<tr>
<td>Swiss-Prot</td>
<td>-2.41</td>
<td>0.80</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Three different hydrophobicity scales (Pliska\textsuperscript{19}, Octanol and POPC\textsuperscript{20}) were used to measure the average hydrophobicity of the different components. The given values are the per residue free energy change for transfer from a hydrophilic to a hydrophobic environment, expressed in kJ mol\textsuperscript{-1}. 


Table 4: Kullback-Leibler divergence and inhomogeneities of individual element.

<table>
<thead>
<tr>
<th>Element</th>
<th>$D_{KL}$</th>
<th>Inhomogeneities found</th>
<th>Possible structural or functional role</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>0.167</td>
<td>none</td>
<td></td>
</tr>
<tr>
<td>TM</td>
<td>0.109</td>
<td>F (0.646%), H (0.008%)</td>
<td>Helices interacting with cofactor</td>
</tr>
<tr>
<td>IH</td>
<td>0.076</td>
<td>none</td>
<td>Cofactor coordination</td>
</tr>
<tr>
<td>NC</td>
<td>0.064</td>
<td>A (0.430%), K (0.002%), M (0.998%)</td>
<td>Positive inside rule, N-terminal residue</td>
</tr>
<tr>
<td>HH</td>
<td>0.055</td>
<td>A (0.119%)</td>
<td>Specific role in interfacial regions</td>
</tr>
<tr>
<td>LO</td>
<td>0.055</td>
<td>G (0.078%), H (0.371%), N (0.038%)</td>
<td>Flexibility in smaller loops, Cofactor coordination, Helix capping properties</td>
</tr>
<tr>
<td>EH</td>
<td>0.030</td>
<td>none</td>
<td></td>
</tr>
</tbody>
</table>

The Kullback-Leibler divergence ($D_{KL}$) was calculated for each component using the overall composition as a reference. Inhomogeneities found in the distribution of individual amino acids within components were detected using a Monte Carlo procedure based on the Kulkarni-Shah test\textsuperscript{25} as described by Krishnamoorty \textit{et al.}\textsuperscript{26}. Shown for each structural component are those amino acids that appear to be non-randomly distributed amongst the different segments and the risk associated with this inhomogeneity.
The Kullback-Leibler divergence of the composition of each component using the other components as reference was calculated. This gives a measure of how easily one element can be distinguished from a subsequent element based on the changing amino acid composition. Since the formula is asymmetric there are small differences between the values in the upper right and lower left parts of the table.
Figure 1

PDPTM
627 structures

Elimination of poor structures

Resolution > 3.5 Å, Cα only

391

Clustering CDHIT

> 90% identity

112

Clustering Blastclust

> 35% identity

67

Elimination of unwanted structures

Peripheral proteins, β barrel structures

54 structures
Figure 2
Figure 6
Figure 7