doi:10.1006/jmbi.2001.4908 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **311**, 891–907

JMB

Helix-Helix Packing and Interfacial Pairwise Interactions of Residues in Membrane Proteins

Larisa Adamian and Jie Liang*

Department of Bioengineering University of Illinois at Chicago, 851 S. Morgan Street RM 218, MC-063 Chicago IL 60607, USA

Helix-helix packing plays a critical role in maintaining the tertiary structures of helical membrane proteins. By examining the overall distribution of voids and pockets in the transmembrane (TM) regions of helical membrane proteins, we found that bacteriorhodopsin and halorhodopsin are the most tightly packed, whereas mechanosensitive channel is the least tightly packed. Large residues F, W, and H have the highest propensity to be in a TM void or a pocket, whereas small residues such as S, G, A, and T are least likely to be found in a void or a pocket. The coordination number for non-bonded interactions for each of the residue types is found to correlate with the size of the residue. To assess specific interhelical interactions between residues, we have developed a new computational method to characterize nearest neighboring atoms that are in physical contact. Using an atom-based probabilistic model, we estimate the membrane helical interfacial pairwise (MHIP) propensity. We found that there are many residue pairs that have high propensity for interhelical interactions, but disulfide bonds are rarely found in the TM regions. The high propensity pairs include residue pairs between an aromatic residue and a basic residue (W-R, W-H, and Y-K). In addition, many residue pairs have high propensity to form interhelical polar-polar atomic contacts, for example, residue pairs between two ionizable residues, between one ionizable residue and one N or Q. Soluble proteins do not share this pattern of diverse polar-polar interhelical interaction. Exploratory analysis by clustering of the MHIP values suggests that residues similar in side-chain branchness, cyclic structures, and size tend to have correlated behavior in participating interhelical interactions. A chi-square test rejects the null hypothesis that membrane protein and soluble protein have the same distribution of interhelical pairwise propensity. This observation may help us to understand the folding mechanism of membrane proteins.

Keywords: membrane protein; helix-helix packing; pairwise propensity; contact potential; alpha shape

**Corresponding author*

Introduction

Integral membrane proteins play essential cellular roles, including signal transduction, proton pumping, ion transport, and light harvesting. They are abundantly found and account for about 20-30 % of the open reading frames of a typical genome.[1,2] Although the transmembrane (TM) region of a helical-bundle membrane protein can be predicted reliably using bioinformatics tools,[1,3–7] knowledge of the three-dimensional structures of membrane proteins is still limited. It has long been recognized that helical-helical interactions play a key role in stabilizing membrane proteins.[8,9] Dimerization experiments combined with extensive replacement mutagenesis[10–13] and insertion mutagenesis[14,15] are powerful tools that can define the critical interfacial packing regions on TM helices. Several recent studies further pointed out the importance of tight packing and specific residues in facilitating helical association of membrane proteins.[16,17] For example, residue G is found to be important for helix association in model peptides,[18] in single-pass membrane proteins,[19] and in poly-

Abbreviations used: TM, transmembrane; MHIP, membrane helical interfacial pairwise; PDB, Protein Data Bank; API, application program interface.

E-mail address of the corresponding author: jliang@uic.edu

0022-2836/01/040891–17 $35.00/0

topic membrane proteins, where it serves as molecular notches for orienting multiple helices and for mediating helix-helix interactions.[16] In addition, side-chains in the TM helices are found to be shorter at helix-helix interfaces.[20] Recent experimental data suggest that specific interhelical interactions between basic residues and W residues through cation-pi electron interactions may play an important role in the folding of proteins into membranes.[21] The propensity values of the orientation of amino acid residues at the lipid-protein interface have also been estimated.[22] Here, we aim to provide a comprehensive analysis of packing and interhelical interactions of membrane proteins based on known structures. We seek to answer the following questions: how are voids and pockets in the TM regions distributed? What are the coordination numbers for residues in the TM regions? What are the specific interhelical interactions? Do helical membrane proteins pack differently from helical soluble proteins?

A detailed structural characterization of the TM helices is a prerequisite for answering these questions. However, manual inspection of interhelical interactions in the TM region is not feasible because it is difficult to clearly identify the interacting atoms from different helices, and this approach does not scale up. A common strategy is to define a cut-off distance and search for all atoms and residues within this distance and count them as contact partners. This has been successfully applied in the analysis of heptad repeat pattern of membrane proteins.[23] The problem with this strategy is that in order to identify all contacting nearest neighbor atoms, a large enough distance will have to be chosen, which may lead to the inclusion of many non-contacting neighbors (Figure 1).

Here, we use a new method based on computational geometry to characterize interhelical interactions of membrane proteins. Our goal is to accurately measure the nearest atomic neighbor in physical contact, without the use of a distance cut-off. Analysis of 14 alpha-helical membrane protein structures indicates that bacteriorhodopsin, halorhodopsin, and rhodopsin are tightly packed, but mechanosensitive channel has extensive voids and pockets in the TM region. Each TM helix, on average, interacts with three to four other helices, and the coordination number of non-bonded interactions for residues in the TM regions is correlated with the size of the side-chain. We found that there are many specific interhelical pairwise interactions, which often involve polar atoms and/or a hydrogen bond. Using a simple probabilistic model, we estimate the single residue propensities for the 20 amino acid residue types to be located in a void or a pocket in TM region, and the single residue propensities to be in interhelical contact with another residue from a different helix. The propensities of specific residue pairs for interhelical interactions are then estimated using an atom-based probabilistic model, and are summarized by a membrane helical interfacial pairwise (MHIP) contact propensity matrix. The algorithms for calculating membrane interhelical interactions, and the probabilistic models used for calculating the single residue propensities and the MHIP propensities are described in Materials and Methods. After we describe the results of the analysis of 14 membrane protein structures, including details of the high-propensity residue pairs that are preferred for interhelical contacts, we show that interhelical packing is different for membrane proteins and soluble proteins, and we discuss the implications of our findings for understanding the mechanism of membrane protein folding.

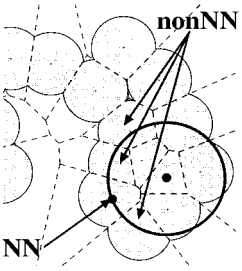

Figure 1. Distance cut-off approach. To include all atoms that are contacting nearest neighbors (NN) to the atom in the center of the circle, the radius of the circle, which is the cut-off distance, must be large enough to include the atom labeled NN. However, several atoms that are not contacting nearest neighbors (nonNN) will also have to be included.

Results

Helix-helix interactions

In bacteriorhodopsin, a TM helix packs only with two other TM helices that are consecutive in primary sequence, except the first and the last TM helices. In more complex membrane proteins such as calcium transporting ATPase, a TM helix often packs with three or more TM helices, frequently involving non-sequentially neighboring TM helices (Figure 2a and (b)). In general, each TM helix typically interacts with three helical partners from the same subunit (Figure 2(c)), although the number of interhelical partners can be as high as five or more. Alpha helices have dipoles, but we do not observe strong preference for either parallel or anti-parallel orientation. This is consistent with an earlier computational study where it was found that non-specific electrostatic interactions play minimal roles in membrane protein packing.[24]

Voids and pockets in TM regions

We examine the overall distribution of voids and pockets in the TM regions of membrane proteins. Some of the voids and pockets are occupied by ligands or prosthetic groups such as heme and retinal, or by water molecules. These non-protein mol-

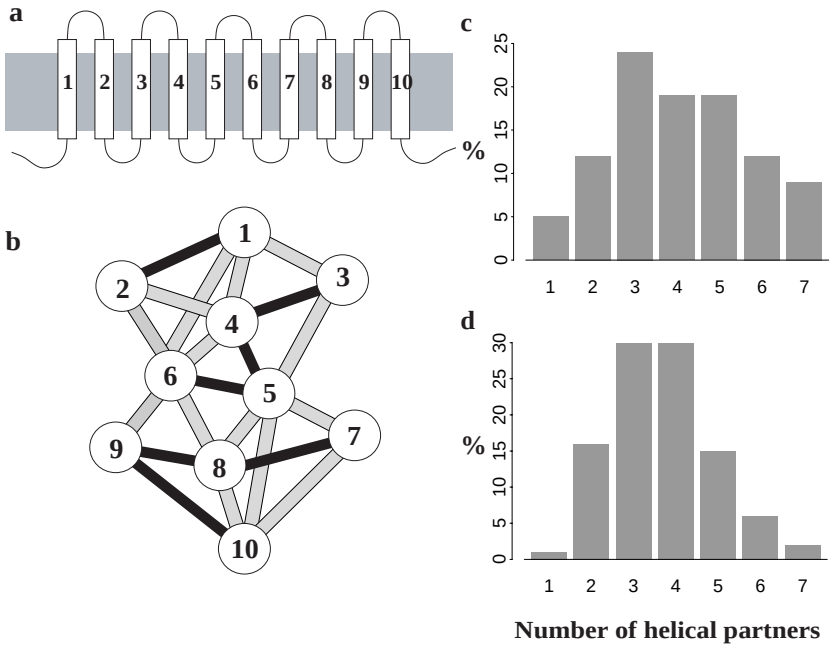

Figure 2. Helix-helix interactions in membrane proteins and soluble proteins. (a) The topology of calcium-tranporting ATPase. (b) The interaction graph of TM helices in calcium-transporting ATPase, where each vertex represents a helix, and each edge indicates a pairwise helix-helix interactions. Note that interactions between non-neighboring TM helices in primary sequences occur frequently, and are shown in gray. Edges representing interacions between sequential neighboring TM helices are colored in black. There are ten parallel helical interactions and nine anti-parallel interactions. Edges connecting an odd-numbered helix with an even-numbered helix represent anti-parallel interactions, and edges connecting two odd-numbered helices or two even-numbered helices represent parallel interactions. (c) The distribution of the number of interhelical contact partners from the same subunit for helices in membrane protein, and (d) for helices in soluble protein.

ecules play important roles for the functions of membrane protein. How residues in the TM regions pack against ligands and water molecules is an important issue that requires detailed analysis, which is beyond the scope of this paper. Here, we are interested in the packing of peptide chains, therefore all ligands and water molecules are removed before computation. Voids and pockets in the TM regions are identified and measured with a probe radius of 1.4 Å using the CAST method.[25,26] We find that the percentage of residues in the TM regions that is in a void or a pocket ranges from 26 % (1c3w, bacteriorhodopsin) to 98 % (1msl, mechanosensitive channel) (Table 1A). Bacteriorhodopsin and halorhodopsin are the most tightly packed membrane proteins that have the smallest number of amino acid residues found in voids and pockets. Conversely, mechanosensitive channel (1msl) has almost all of its residues in the TM region (98 %) found in voids or pockets. It is possible that large lateral motions are necessary for the functional role of mechanosensitive ion channel. The abundance of voids and pockets, which are deformable under mechanical forces in the TM regions, is consistent with this possibility. For the majority of the membrane proteins (11 out of 14), about 57 % of the residues in the helical bundles of the TM regions participate in the formation of voids and pockets. For comparison, we also examine voids and pockets in a set of 26 soluble proteins. In soluble proteins, the percentage of residues participating in voids and pockets ranges from ~11 % in Rop protein (62 residues, 1gto) to ~63 % in monooxygenase (1070 residues in three chains, 1mty). On average, 49 % of the residues in soluble proteins participate in forming voids and pockets.

Which type of amino acid residue is most likely to be located in a void or pocket? We estimate the single residue propensity to be in a pocket or a void for each amino acid residue type (Table 1B). Residues F, W, and H are more likely to be found in a TM pocket or a void. Aromatic residues, including F and W, form an ''aromatic belt'', and are frequently located near the ends of the TM helices, regions not tightly packed.[27] Residue H is frequently found in the active sites and often play important functional role. Residues S, G, A and T are least likely to be found in voids or pockets. These are small, hydrophobic or non-ionizable polar residues and tend to be away from voids or pockets. They are located mostly in well-packed regions of the TM helices. The overall pattern of propensities for being located in a void or a pocket in soluble proteins is similar to that of TM helices. The exceptions are residues R and Y, which have higher propensities in soluble proteins to be in a void or a pocket, and residues E and G, which have lower propensities.

Table 1. Amino acid residues in the transmembrane regions

A. *The percentage of the residues in the TM regions that participate in the formation of the voids and pockets*

PDB ID	Total number of residues in TM helices	Total number of residues in pockets	Percentage of residues in the pockets
1c3w	168	44	26.2
1e12	165	47	28.5
1f88	193	89	46.1
1fx8	158	74	46.8
1eul	226	118	52.2
1qla	258	142	55.0
1ehk	480	268	55.8
1occ	685	388	56.6
1prc	267	153	57.3
1fum	173	103	59.5
1ar1	389	247	63.5
1be3	300	191	63.7
1bl8	288	189	65.6
1msl	220	216	98.1

B. *The estimated propensities for each of the 20 amino acid residues to be in a pocket or a void in the TM region of membrane proteins (MB) and in soluble proteins (SP)*

	F	W	H	I	R	E	Y	V	N	L	P	D	Q	M	C	K	T	A	G	S
MP	1.28	1.24	1.16	1.12	1.07	1.06	1.05	1.05	1.05	1.05	1.01	1.01	1.00	0.99	0.96	0.94	0.87	0.85	0.84	0.78
SP	1.29	1.32	1.21	1.15	1.28	0.84	1.46	1.07	0.99	1.16	1.05	0.91	0.90	1.11	0.98	0.97	0.94	0.81	0.66	0.91

Coordination numbers of non-bonded interactions

The non-bonded atomic contacts among residues in the TM regions reflect van der Waals and hydrogen bonding interactions. Here, we measure the number of atomic contacts for each residue in the TM region of membrane proteins and in soluble proteins. Because of potential uncertainty in the original atomic coordinates, a number count of atomic contacts may provide a more robust measure of interactions than a numerical value such as volume overlaps, although an algorithm exists that calculates such parameters.[28] We term this contact number ''coordination number of non-bonded interactions''.

Figure 3(a) and (b) show the non-bonded, inter-residue coordination numbers per residue for each residue type in TM regions of membrane proteins and in soluble proteins. The residues are arranged on the *x*-axis in ascending order of the number of atoms in the side-chain. The coordination number increases with the size of the residue in both types of proteins. For both membrane and soluble proteins, small residues such as G, A, S and T, have lower median coordination numbers, whereas large residues, such as W, Y and F, have the highest median coordination numbers. This suggests that bulkier residues have more total non-bonded contacts. However, when normalized by the number of atoms in the side-chains, the coordination number decreases with residue size (Figure 3(c) and (d)). Smaller residues, such as G and A, have the highest median coordination numbers on a per atom basis in both membrane and soluble proteins. The overall patterns of coordination numbers for both TM helices and soluble proteins are rather similar, except for residues P and S. Proline seems to have a higher per atom median coordination number in membrane proteins than in soluble proteins. This is expected, because in soluble proteins proline is often found in turns. Serine has higher median coordination number per atom in TM helices than in soluble proteins. It is packed more tightly than abundantly found hydrophobic residues V, I, L and M. In soluble proteins, residues V, I, L and M all have more per atom contacts than S. These data indicate that although larger residues have more non-bonded total contacts with other residues, the smaller residues are often located in a tightly packed regions in both membrane and soluble proteins, and have on average more non-bonded interactions for each atom.

Single residue propensities for interhelical interaction

Which residue type is most likely to participate in interhelical contacts? For membrane proteins, the single residue propensity values of each residue to participate in interhelical contacts are close to 1.0, where residue M has the highest propensity (1.17) and residue R the lowest propensity (0.78) (Figure 4). In contrast, the propensity values for interhelical contact in soluble proteins have a wider range, with residue W of the highest value (1.32) and residue P of the lowest value (0.62). For soluble proteins, the hydrophobic and aromatic residues (W, C, L, F, Y, M, I, and V) commonly found in protein cores have higher propensity for interhelical contacts. Polar and ionizable residues (D, E, Q, S, K, and T), as well as helix breakers (G

and P), are less likely to participate in interhelical contacts. The lack of simple correlation between propensity values for residues in membrane proteins and in soluble proteins, and the different patterns in the range of propensity values perhaps reflects the constraints from different environments of solvent and lipid membranes.

Pairwise interhelical contact

Any contact point between two helices involves at least two residues, one from each helix. The single residue propensity for interhelical interactions cannot capture any packing interaction that depends on the partner residue on the other helix. Here, we consider the pairwise interhelical interactions. Since we can choose two residues from 20 amino acid types with replacement allowed, there are 210 possible pairs of residues. Only 203 of them are observed from a total of 4,144 contacting residue pairs, with 18,991 atomic interhelical contacts, or ~4.6 atomic contacts per interacting pair in the small set of 14 membrane proteins. Residue pairs D-D, D-C, D-Q, C-Q, C-K, K-K, and Q-Q are never observed to be in interhelical contacts. The data set of soluble proteins was constructed for comparison, and is also of limited size. Here, we observed all 210 residue pairs from a total of 32,659 atomic contacts.

Residue pair C-C is an example of rarely observed interacting residue pairs in a membrane protein. In soluble proteins, disulfide bonds formed between C-C residue pairs play important roles in maintaining protein stability. In contrast, the C-C pair is observed only once in all 14 membrane proteins, where there are in total 341 atomic contact pairs involving a C residue. It seems that disulfide bonds are less prevalent in membrane proteins and do not play important roles in maintaining the stability of the TM helical bundle.

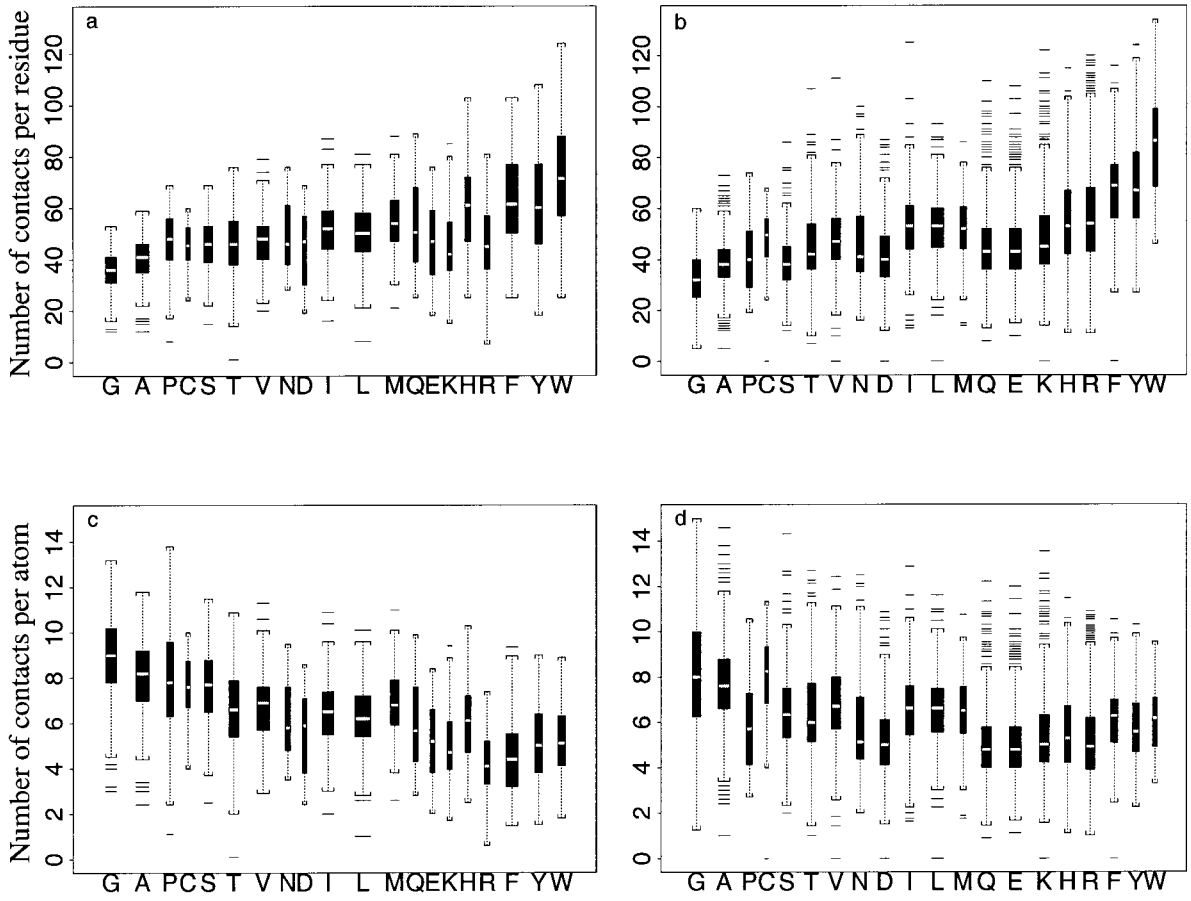

Figure 3. Non-bonded inter-residue coordination numbers of residues in the TM regions of membrane proteins and in soluble proteins. (a) and (b) Coordination number per residue, membrane and soluble proteins, respectively. (c) and (d) Coordination number per atom, membrane and soluble proteins, respectively. The amino acid residues on the *x*-axis are ordered by the number of atoms in their side-chains from left to right. Here the set of coordination numbers of a residue type is plotted as a boxplot, where the central box shows the data between the quartiles, and the median value is represented by a line. Whiskers represent the extremes of the data, and the width of each box reflects the frequency of occurrence of the corresponding amino acid residue in the TM regions. Outliers are drawn as individual extra whiskers in the boxplot.

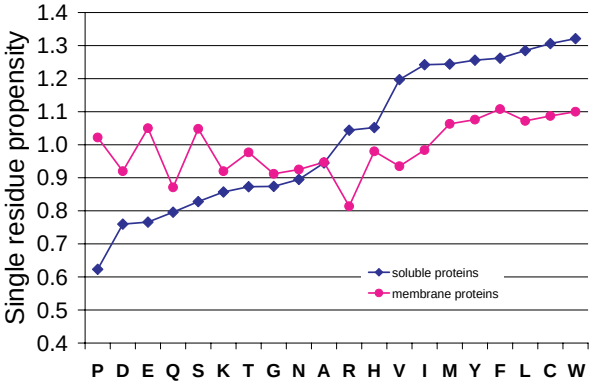

Figure 4. Single residue propensities of the 20 amino acid residues for interhelical contacts. The propensity values for soluble proteins have a wider range, whereas the propensity values for membrane proteins are close to 1.

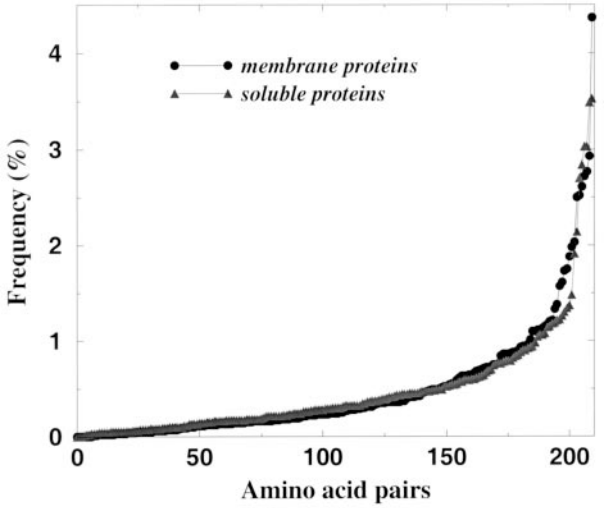

Figure 5. Relative frequencies of pairs of amino acid residues in interhelical contact. The residue pairs are arranged along the *x*-axis in ascending order of frequency. The top 20 pairs account for about 40 % of all residue pairwise contacts in both membrane and soluble proteins.

The overall distributions of the relative frequencies of the interacting residue pairs, i.e. the number of observed occurrences of a specific pair of interacting residues divided by the total number of all pairs of residues, is shown in Figure 5 and Table 2. The residue pairs are arranged along the *x*-axis in ascending order of frequency. The overall patterns of the frequency of pairwise interhelical contacts are similar for membrane proteins and soluble proteins. The top 20 residue pairs account for about 40 % of all residue pairwise contacts in both membrane and soluble proteins. Eleven out of these 20 pairs are identical for both data sets: F-L, L-L, L-V, I-F, L-W, A-L, L-Y, F-V, A-F, L-M, and G-L, seven of which contain L. For soluble proteins, the top 20 residue pairs also include six residue pairs containing one or two polar or ionizable residues: L-R, K-L, E-R, and R-Y. These residue pairs, however, are not frequently seen in membrane proteins. In summary, 20 or so residue pairs dominate interhelical interactions for both membrane and soluble proteins, with roughly half of them common to both membrane and soluble proteins.

MHIP propersity values

For a given residue in the TM region, which residue types are mostly likely to have interhelical contacts with it? We estimate the pairwise propensities for interhelical interactions for all residue pairs using a simple probabilistic model. The estimated values (Table 3) are odds ratio, namely, the ratio of the observed frequency of interhelical atomic contacts between a specific type of residue pair against the frequency expected if the pairs of interacting atoms are drawn independently and randomly from the same set of residues. As an example, the G-H pair has an odds ratio of 3.1. This means that the observed frequency of atomic contacts from a G-H residue pair on two helices is 3.1 times what would be expected if the two contacting atoms happen to be from a Gly residue and a His residue when picked randomly and independently. The residue pair L-T has a propensity of 0.6, indicating that it is less likely to find contacting atoms from L-T residue pairs on two helices than would be expected from random sampling. For comparison, helical interfacial pairwise propensity values for soluble proteins are also estimated (Table 4). The list of number count for each type of residue pair is provided as Supplementary Material.

Residue pairs with high propensity

For membrane proteins, there are five frequently observed residue pairs (F-F, 1.7, F-W, 1.4, F-M, 1.4,

Table 2. Observed frequence (%) for the top 20 interacting pairs

A. *In interacting helices from TM regions of membrane proteins*

L-F	L-L	L-V	I-L	F-W	L-W	F-F	A-L	L-Y	F-V	F-I	A-F	M-F	L-M	L-S	I-W	I-V	G-L	W-V	G-F
4.37	2.93	2.78	2.72	2.61	2.52	2.5	2.03	1.98	1.88	1.75	1.73	1.61	1.57	1.38	1.33	1.21	1.2	1.17	1.15

B. *In interacting helices from soluble proteins*

L-L	I-L	F-L	L-V	A-L	L-Y	L-R	L-M	G-L	A-V	I-F	L-W	F-Y	L-T	A-I	K-L	A-F	E-R	F-V	R-Y
3.41	3.09	2.99	2.91	2.88	2.65	1.94	1.66	1.38	1.33	1.30	1.24	1.24	1.22	1.22	1.17	1.13	1.13	1.12	1.11

Table 3. Membrane helical interfacial pairwise contact propensity

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	1.3	0.5	1.1	1.2	1.7	1.0	0.8	1.1	1.3	1.0	0.9	0.7	1.7	1.1	2.1	0.9	1.0	1.1	0.8	0.8
ARG	0.5	1.3	0.7	3.7	0.4	2.8	1.8	0.6	0.3	0.2	0.6	0.8	0.6	0.4	0.3	1.4	0.7	2.5	1.4	0.7
ASN	1.1	0.7	6.0	4.8	0.8	2.4	1.3	1.6	0.2	1.1	0.9	2.8	0.8	0.5	1.2	1.4	0.8	0.7	1.5	0.9
ASP	1.2	3.7	4.8	0.0	0.0	0.0	0.4	0.1	0.6	1.0	0.7	2.2	0.5	0.3	1.8	1.2	1.0	0.4	2.8	0.1
CYS	1.7	0.4	0.8	0.0	0.8	0.0	0.8	2.4	1.3	1.1	1.0	0.0	1.5	1.8	0.6	2.2	1.2	0.4	0.3	0.3
GLN	1.0	2.8	2.4	0.0	0.0	0.0	0.4	1.4	2.5	0.2	0.8	1.6	1.3	0.4	0.9	2.5	1.4	1.3	2.3	0.7
GLU	0.8	1.8	1.3	0.4	0.8	0.4	1.6	0.7	1.0	0.4	0.5	1.7	1.0	0.5	1.9	1.2	1.0	0.1	0.7	0.7
GLY	1.1	0.6	1.6	0.1	2.4	1.4	0.7	3.0	3.1	0.6	1.0	0.4	1.3	1.3	0.6	1.0	0.6	1.4	1.6	1.0
HIS	1.3	0.3	0.2	0.6	1.3	2.5	1.0	3.1	3.9	0.7	0.7	0.5	1.0	1.1	0.3	1.3	2.3	1.9	1.2	0.5
ILE	1.0	0.2	1.1	1.0	1.1	0.2	0.4	0.6	0.7	1.3	1.0	0.5	1.1	0.8	1.2	0.6	0.9	1.0	0.5	0.8
LEU	0.9	0.6	0.9	0.7	1.0	0.8	0.5	1.0	0.7	1.0	1.1	0.7	1.0	1.1	0.7	1.1	0.6	1.0	1.0	1.0
LYS	0.7	0.8	2.8	2.2	0.0	1.6	1.7	0.4	0.5	0.5	0.7	0.0	2.2	0.4	1.0	1.1	0.2	0.6	2.5	0.4
MET	1.7	0.6	0.8	0.5	1.5	1.3	1.0	1.3	1.0	1.1	1.0	2.2	1.5	1.4	1.4	1.9	0.7	1.2	0.6	0.9
PHE	1.1	0.4	0.5	0.3	1.8	0.4	0.5	1.3	1.1	0.8	1.1	0.4	1.4	1.7	0.6	1.0	0.7	1.4	0.8	0.9
PRO	2.1	0.3	1.2	1.8	0.6	0.9	1.9	0.6	0.3	1.2	0.7	1.0	1.4	0.6	1.8	1.2	1.3	1.2	2.0	0.6
SER	0.9	1.4	1.4	1.2	2.2	2.5	1.2	1.0	1.3	0.6	1.1	1.1	1.9	1.0	1.2	4.4	1.5	1.1	1.0	0.8
THR	1.0	0.7	0.8	1.0	1.2	1.4	1.0	0.6	2.3	0.9	0.6	0.2	0.7	0.7	1.3	1.5	1.1	1.1	1.2	1.1
TRP	1.1	2.5	0.7	0.4	0.4	1.3	0.1	1.4	1.9	1.0	1.0	0.6	1.2	1.4	1.2	1.1	1.1	0.8	0.9	0.9
TYR	0.8	1.4	1.5	2.8	0.3	2.3	0.7	1.6	1.2	0.5	1.0	2.5	0.6	0.8	2.0	1.0	1.2	0.9	0.6	0.5
VAL	0.8	0.7	0.9	0.1	0.3	0.7	0.7	1.0	0.5	0.8	1.0	0.4	0.9	0.9	0.6	0.8	1.1	0.9	0.5	1.0

F-G, 1.3, and A-M, 1.7) with high MHIP propensity values. Each pair accounts for more than 1.1 % of all interhelical contacts, far more than would be expected ($1/210 = 0.48\,\%$) if interhelical contacts were distributed evenly among the possible 210 residue pairs. Four of these contain the F residue, which is abundantly found in helices. The hydrophobic and bulky side-chain of an F residue efficiently packs with non-polar and aromatic residues in membranes, often producing multiple interhelical atomic contacts. In addition, as residues commonly found at the membrane-solution interfaces,[29–31] W and Y residues have higher propensity to interact with residues similarly enriched in the membrane-solution interfaces:[30] W with R and H, Y with K. Polar residue S has a high propensity to form self-pairs (S-S, 4.4), and to interact with M residues (S-M, 1.9). Another polar residue, T, has a high propensity to interact with H residues. For soluble proteins, the top 14 high-propensity residue pairs that each accounts for 1.1 % or more interhelical contacts are: L-L (1.8), I-L (1.7), F-L (1.9), L-V (1.5), A-L (1.6), L-Y (1.5), L-M (1.6), A-V (1.5), F-I (1.8), F-Y (1.7), L-W (1.5), A-I (1.5), A-F (1.6), F-V (1.5). These are rich in hydrophobic residues (L, A, I, V, and F) commonly found in protein cores.

Some residue pairs occur less frequently, yet their pair-propensity values are higher than what would be expected from random sampling. For those high-propensity pairs that each account for less than 0.3 % of all pairwise contacts, there are

Table 4. Soluble proteins helical interfacial pairwise contact propensity

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	2.1	0.6	1.2	0.3	1.4	0.6	0.3	1.4	0.6	1.5	1.6	0.6	1.2	1.6	1.0	1.0	1.2	1.5	1.3	1.5
ARG	0.6	0.4	0.7	1.6	0.5	0.7	0.9	0.9	0.9	0.5	0.8	0.4	0.7	0.6	0.7	0.9	0.7	1.2	1.0	0.5
ASN	1.2	0.7	1.6	0.6	0.6	0.8	0.6	1.2	0.3	0.8	0.7	0.7	0.8	0.3	0.9	1.2	0.5	0.9	1.4	0.4
ASP	0.3	1.6	0.6	0.4	0.5	0.3	0.2	0.1	0.7	0.2	0.4	1.8	0.6	0.1	0.8	0.7	0.2	0.8	1.3	0.7
CYS	1.4	0.5	0.6	0.5	6.3	0.5	0.6	1.3	0.3	1.5	2.4	1.3	2.1	2.3	0.5	1.7	0.7	1.2	1.3	0.5
GLN	0.6	0.7	0.8	0.3	0.5	0.8	0.3	0.7	1.0	0.5	0.7	0.7	0.6	1.3	0.7	0.5	0.9	0.8	0.6	0.9
GLU	0.3	0.9	0.6	0.2	0.6	0.3	0.4	0.3	1.2	0.5	0.7	1.2	0.4	0.7	0.4	0.7	0.5	1.0	0.9	0.7
GLY	1.4	0.9	1.2	0.1	1.3	0.7	0.3	2.3	1.2	0.6	1.0	0.8	1.8	1.1	0.3	1.6	1.0	2.0	1.3	0.7
HIS	0.6	0.9	0.3	0.7	0.3	1.0	1.2	1.2	2.8	1.1	1.0	0.5	1.3	0.9	0.4	0.6	0.8	2.2	2.1	0.4
ILE	1.5	0.5	0.8	0.2	1.5	0.5	0.5	0.6	1.1	2.0	1.7	0.6	1.2	1.8	0.8	0.7	0.9	1.1	1.3	1.1
LEU	1.6	0.8	0.7	0.4	2.4	0.7	0.7	1.0	1.0	1.7	1.8	0.7	1.6	1.9	0.6	0.9	1.1	1.5	1.5	1.5
LYS	0.6	0.4	0.7	1.8	1.3	0.7	1.2	0.8	0.5	0.6	0.7	0.3	0.5	1.1	0.3	0.4	0.4	0.9	1.2	0.5
MET	1.2	0.7	0.8	0.6	2.1	0.6	0.4	1.8	1.3	1.2	1.6	0.5	2.2	2.1	0.3	0.7	1.2	1.7	1.7	1.3
PHE	1.6	0.6	0.3	0.1	2.3	1.3	0.7	1.1	0.9	1.8	1.9	1.1	2.1	3.1	1.4	1.1	1.1	2.4	1.7	1.5
PRO	1.0	0.7	0.9	0.8	0.5	0.7	0.4	0.3	0.4	0.8	0.6	0.3	0.3	1.4	0.6	1.0	1.6	1.7	1.3	0.4
SER	1.0	0.9	1.2	0.7	1.7	0.5	0.7	1.6	0.6	0.7	0.9	0.4	0.7	1.1	1.0	0.8	1.1	0.8	0.6	1.1
THR	1.2	0.7	0.5	0.2	0.7	0.9	0.5	1.0	0.8	0.9	1.1	0.4	1.2	1.1	1.6	1.1	1.5	1.0	0.9	0.8
TRP	1.5	1.2	0.9	0.8	1.2	0.8	1.0	2.0	2.2	1.1	1.5	0.9	1.7	2.4	1.7	0.8	1.0	2.0	1.2	1.4
TYR	1.3	1.0	1.4	1.3	1.3	0.6	0.9	1.3	2.1	1.3	1.5	1.2	1.7	1.7	1.3	0.6	0.9	1.2	1.5	0.8
VAL	1.5	0.5	0.4	0.7	0.5	0.9	0.7	0.7	0.4	1.1	1.5	0.5	1.3	1.5	0.4	1.1	0.8	1.4	0.8	1.1

several well-known examples, such as those pairs forming salt bridges (D-R, propensity value 3.7, D-K, 2.2, and E-R, 1.8) in membrane proteins and disulfide bonds in soluble proteins. Other such residue pairs found in membrane proteins often contain an N residue or a Q residue: D-N (4.8), N-N (6.0), K-N (2.8), Q-R (2.8), H-Q (2.5), Q-S (2.5), and N-Q (2.4). Among these, residue pairs N-N and D-N are often found in functional sites. For example, residue N86 from the conserved NPA motif[32] forms N-N contact with N203 in glycerol-conducting channel (1fx8). The side-chain of each of the N residues is constrained by two hydrogen bonds, and the amide group of the side-chain is oriented towards the acceptors on the permeant substrate. Another N-N pair appears in the functional region of the calcium pump of sarcoplasmic reticulum (1eul).[33] Here, hydrogen bonding between OD1 of residue N101 and ND2 of residue N796 helps to correctly orient the side-chain of residue N796, which is a key residue in the Ca^{2+} binding site. Point mutations of N101 of the calcium pump resulted in partial loss of the function of this channel protein[34], indicating the structural importance of the N-N residue pair. For these infrequently observed residue pairs that have high interhelical contact propensity, although no statistical inference can be drawn with strong confidence because of the limitations of the sample size of the data set, they point to specific interhelical interactions that may be functionally important. In soluble proteins, there are also infrequently observed residue pairs (<0.3 %) that have high pairwise contact propensities (>1.7): C-C (6.3), H-H (2.8), G-G (2.3), C-M (2.1), W-W (2.0), G-W (2.0), G-M (1.8). These residue pairs are all different from those observed in the membrane proteins.

Residue pairs with low propensity

Several residue pairs observed frequently (>0.5 %) in the TM regions of membrane proteins have lower than expected propensity for interhelical contacts. These include: (1) residue pairs between two bulky and branched residues, W-W (0.8), I-Y (0.5), H-L (0.7), F-Y (0.8); (2) pairs between a hydrophobic residue and a polar residue, L-T (0.6) and Y-V (0.5); and (3) a pair between two small residues, A-V (0.8). Some of these residue pairs are of high propensity for interhelical interactions in soluble proteins (e.g. I-Y, 1.3, and W-W, 2.0).

Polar-polar interactions

To examine the physicochemical nature of helical interactions, we analyze the details of interhelical contacts. We first examine interactions between polar-polar atoms, which are defined here conveniently as N and O atoms, and S atom in the thiol groups of C residues. For all 210 possible types of residue pairs, polar-polar atomic contacts are observed for 152 types of residue pairs in membrane proteins and 160 types in soluble proteins. There are 745 and 1617 pairs of polar-polar atomic contacts in membrane and soluble proteins, respectively. These polar contacts account for roughly the same fraction (4 %) of all atomic contact pairs in both soluble and membrane proteins. There are 572 interhelical contacting residue pairs with polar-polar atomic contacts in membrane proteins, 124 of them belong to the 22 types of residue pairs that have high propensity for polar-polar interhelical interactions. The total number of contacting residue pairs is 4,144, including polar-polar, polar-non-polar, and non-polar-non-polar contacts. Therefore, there is roughly one polar-polar atomic contact for every six contacting residue pairs in

Table 5. Contact propensities for interactions between polar atoms in membrane proteins

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.0	0.7	1.0	0.6	1.0	0.5	0.7	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.3	0.4	0.2	0.0
ARG	0.7	2.3	1.6	9.1	0.6	3.3	2.4	0.4	1.2	0.4	0.5	0.6	0.4	0.1	0.0	1.7	0.8	1.4	1.6	0.1
ASN	1.0	1.6	5.6	5.8	0.8	2.5	2.2	0.9	0.3	0.3	0.5	3.4	0.8	0.1	0.0	2.0	0.7	1.5	1.7	0.7
ASP	0.6	9.1	5.8	0.0	0.0	0.0	0.6	0.0	0.9	0.2	0.3	6.6	0.0	0.2	0.7	1.6	1.4	0.9	5.6	0.0
CYS	1.0	0.6	0.8	0.0	3.1	0.0	0.0	1.4	0.0	0.6	0.2	0.0	0.5	0.6	1.6	1.6	1.1	0.7	0.0	0.0
GLN	0.5	3.3	2.5	0.0	0.0	0.0	0.5	1.3	2.8	0.0	0.6	5.1	0.0	0.0	0.5	2.3	1.4	0.7	2.2	0.2
GLU	0.7	2.4	2.2	0.6	0.0	0.5	4.0	0.3	1.6	0.0	0.3	5.0	1.0	0.3	0.8	0.9	1.3	0.0	1.0	0.4
GLY	0.0	0.4	0.9	0.0	1.4	1.3	0.3	0.5	2.2	0.0	0.0	0.0	0.0	0.0	0.2	0.5	0.2	0.3	0.4	0.3
HIS	0.7	1.2	0.3	0.9	0.0	2.8	1.6	2.2	5.4	0.1	0.1	0.0	0.2	0.1	0.3	1.1	1.9	1.0	1.1	0.0
ILE	0.0	0.4	0.3	0.2	0.6	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.1	0.1	0.1	0.0
LEU	0.0	0.5	0.5	0.3	0.2	0.6	0.3	0.0	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.1	0.3	0.3	0.0
LYS	0.0	0.6	3.4	6.6	0.0	5.1	5.0	0.0	0.0	0.0	0.1	0.0	0.5	0.0	0.0	0.6	0.2	0.3	0.7	0.0
MET	0.0	0.4	0.8	0.0	0.5	0.0	1.0	0.0	0.2	0.0	0.0	0.5	0.0	0.0	0.3	0.6	0.0	0.4	0.4	0.0
PHE	0.0	0.1	0.1	0.2	0.6	0.0	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.0	0.2	0.2	0.2	0.7	0.0
PRO	0.0	0.0	0.0	0.7	1.6	0.5	0.8	0.2	0.3	0.0	0.0	0.0	0.3	0.0	0.9	0.2	0.3	0.2	0.2	0.0
SER	0.5	1.7	2.0	1.6	1.6	2.3	0.9	0.5	1.1	0.3	0.1	0.6	0.6	0.2	0.2	2.8	1.0	1.3	1.3	0.3
THR	0.3	0.8	0.7	1.4	1.1	1.4	1.3	0.2	1.9	0.1	0.1	0.2	0.0	0.2	0.3	1.0	0.8	0.5	1.2	0.2
TRP	0.4	1.4	1.5	0.9	0.7	0.7	0.0	0.3	1.0	0.1	0.3	0.3	0.4	0.2	0.2	1.3	0.5	0.0	1.0	0.0
TYR	0.2	1.6	1.7	5.6	0.0	2.2	1.0	0.4	1.1	0.1	0.3	0.7	0.4	0.7	0.2	1.3	1.2	1.0	0.2	0.1
VAL	0.0	0.1	0.7	0.0	0.0	0.2	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.2	0.0	0.1	0.2

membrane proteins (4,144/745 = 5.6) and one polar-polar atomic contact for every 4.5 contacting residue pairs in soluble proteins (7,284/1.617 = 4.5). Because there are N and O atoms in the peptide bond of hydrophobic residues, it is possible to observe atomic polar contacts between hydrophobic residues such as G-V, although they appear very rarely. The majority of atomic polar-polar contacts involve polar and ionizable residues.

Because there are only a few polar atoms for each residue, atomic polar-polar contacts are observed far less frequently than non-polar-non-polar interactions, and the estimation of interhelical contact propensity of polar-polar atoms is subject to significant sampling errors due to the small sample size. Nevertheless, the propensity values (Table 5) seem to suggest that atomic polar interactions in membrane proteins are different from those in soluble proteins. For example, polar-polar interaction residue pairs of high propensity ($\geqslant 1.9$) are more diverse in membrane proteins (22 types of residue pairs with 219 observed atomic contacts: D-R (9.1), D-K (6.6), D-N (5.8), N-N (5.6), D-Y (5.6), H-H (5.4), K-Q (5.1), E-K (5.0), E-E (4.0), K-N (3.4), Q-R (3.3), S-S (2.8), H-Q (2.8), N-Q (2.5), E-R (2.4), Q-S (2.3), R-R (2.3), G-H (2.2), Q-Y (2.2), E-N (2.2), N-S (2.0), and H-T (1.9)) than in soluble proteins (three residue pairs with 189 observed atomic contacts: D-R, 2.7, H-H, 2.5, D-K, 2.3). Polar-polar interactions in soluble proteins are mostly reflections of salt-bridge interactions between ionizable amino acid residues. In membrane proteins, salt-bridge interactions also exist, but there are many high-propensity pairs that contain only one ionizable residue. Polar residues such as S, T, Y, N, and Q contribute most to interhelical polar-polar atomic contacts. Many such polar-polar atomic contacts represent interhelical H-bondings between residues buried in a hydrophobic membrane environment. These polar interactions may play important roles for the interhelical recognition between TM helices.

Non-polar-non-polar interactions

Because of the abundance of hydrophobic residues in the TM regions, non-polar-non-polar atomic interhelical contacts are observed frequently, and they account for 69 % of all atomic pairwise contacts. This is similar to soluble proteins, where non-polar-non-polar contacts account for 65 % of all interhelical atomic interactions. Frequently observed residue pairs with high propensity for non-polar-non-polar interactions are: A-P (3.7), A-M (2.8), H-W (2.7), G-W (2.4), G-F (2.4), F-F (2.2), I-I (2.0), T-V (1.9), F-M (1.9), F-W (1.8), G-L (1.8), A-F (1.8), A-I (1.8), each accounts for > 0.5 % of total interhelical atomic contacts. Residue G has no side-chain, and tends to form interhelical non-polar contacts with bulky aromatic residues (W, F and Y). Sterically, they fit well between two helices. For the G-W residue pair, the C and C^{α} atoms of residue G are often found in contact with the aromatic ring of residue W. In fumarate reductase from *E. coli* (1fum), the aromatic ring of W86 on helix II is positioned on top of two glycine residues from helix IV (G18 and G21) (Figure 6(a)). In subunit III of bovine cytochrome *c* oxidase (1occ), W242 from helix VII interacts with two G residues (G170, helix V and G211, helix VI, Figure 6(b)). Another example of G-W interactions is found in rhodopsin (PDB 1f88, Figure 6(c)), where G121 from helix III interacts with aromatic ring of W265 from helix VI. The pattern of non-polar-non-polar interactions between residue G and aromatic residues is common in membrane proteins, but is rarely seen in soluble proteins. Soluble proteins have a different set of frequently observed residue pairs with high interhelical contact propensity values for non-polar-non-polar interactions. Only F-M, A-F, and A-I are high propensity pairs frequently seen in both membrane proteins and soluble proteins.

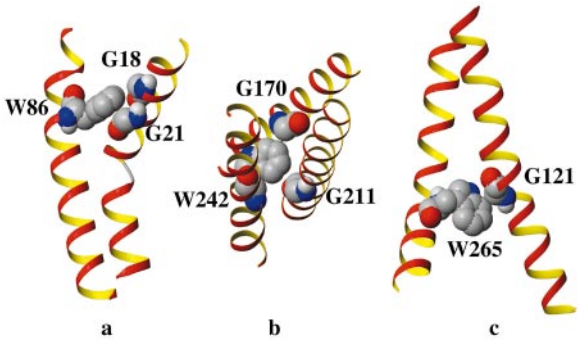

Figure 6. Interhelical packing of W-G pairs. (a) In fumarate reductase (1fum), W86 from helix II is positioned on top of two glycines from helix IV (G18 and G21). (b) In bovine cytochrome c oxidase (1occ), W242 from helix VII interacts with G170, helix V and G211, helix VI (top view). (c) W265-G121 pair in rhodopsin (1f88). All of the molecular structure representations in this Figure were drawn with the program MOLMOL.[77]

Backbone contacts

Interactions between backbone atoms (i.e. O, N, C and C^{α} atoms) are rare. Compared with soluble proteins, membrane proteins may have more backbone-backbone interactions (2.8 % of all atomic contacts *versus* 1.1 % in soluble proteins). Such interactions usually involve small amino acid residues. The only residue pair that shows a high MHIP propensity value (3.0) with a non-trivial number of observations is the G-G pair (76 contacts observed). This confirms results from a previous study where glycine was shown to play important role in providing the closest contact point for helix-helix interactions.[16] In soluble proteins, G-G is also the only pair that has high propensity (2.3) for interhelical interactions through backbone with total 20 backbone-backbone atomic contacts.

Discussion

Statistical analysis of residue contacts

Empirical statistical analysis of residue-residue interaction have long been usefully applied to study a variety of problems, including protein folding, protein threading, and protein-protein interactions.[36–42] Derived from databases of protein structure, they can capture broad information about the specific protein environment of amino acid residues. Similar to these empirical potentials, the MHIP reported here is also empirically derived from a database of membrane protein structures. However, there are two characteristics that distinguish MHIP from other residue-residue potentials. First, MHIP is strictly about immediate nearest neighbors that are in physical contact. Although, strictly, contact potentials have been used widely in lattice simulations,[40] they are difficult to derive from real protein structures. Unlike residue-residue potential, any atoms that are one layer of atoms away are not considered, and therefore in MHIP there is no distance dependency, and no atoms are considered beyond the first contact layer in residue-residue correlation. Two atoms are in contact only if their Voronoi cells intersect and only if at least part of the intersecting Voronoi interface plane is contained within the two atoms. These strict criteria dictate that there will not be a third atom in the way of the first two atoms that are recorded to be in contact. Second, the pairwise propensity is estimated from an atomic model instead of a residue model. Each residue is not represented by a single point, and the size information of each residue is encoded explicitly in the random model. In addition, the random model used for odds-ratio calculation is constructed combinatorically, and the probability of atomic contact in the random model is given analytically. This eliminates the need for extensive numerical randomization tests. Like any other empirical potentials derived from finite size systems, MHIP values between two different residue pairs may not be fully independent.[40] However, we expect MHIP does not suffer significantly from chain-length dependence, since we are examining interfacial contacts between different TM helices, which are all of similar lengths. In addition, MHIP does not suffer from composition bias because the random model used for odds-ratio calculation is of the same set of amino acid residues in the TM helices. It is hoped that MHIP estimates can provide more discriminating information about helix packing.

Comparison with other studies of membrane protein packing

The packing of helical membrane proteins has been analyzed in several previous studies. Using the method of occluded surface (OS),[43] Eilers *et al.*[17] analyzed internal packing of helical membrane proteins. Despite of the differences in the methodology and the data sets, in several cases our data confirm their earlier results. For example, mechanosensitive protein is found to be the least well-packed membrane protein in both studies, and both studies find residue P to be well packed in membrane proteins. Based on analysis of four polytopic membrane proteins, Javadpour *et al.*[16] found that residue G plays an important role in facilitating helical packing. It has high propensity for helix-helix interactions and is rarely found in voids. We find that residue G has a high coordination number on a per atom basis (Figure 3(b)), and has one of the smallest values of the single residue propensity to be located in a void or a pocket of the TM region (Table 1B). These are consistent with results reported by Javadpour *et al.*[16] In addition, the coordination number per atom for the 20 amino acid residues reported in Figure 3(b) closely resembles the normalized occurrence at the helix interface.[16]

Correlating MHIP profiles

The behavior of interhelical interactions for a specific type of amino acid residue is determined by the 20 values of its MHIP propensity, one for each type of amino acid residue located on a neighboring helix. These values represent the residue-specific profile for interhelical contacts, and can be represented as a 20-dimensional vector X. Let μ_1, μ_2 and σ_1, σ_2 be the means and standard deviations of the MHIP vectors X_1 and X_2 of residue type 1 and type 2, respectively. The correlation coefficient:

$$\rho = \frac{E(X_1 - \mu_1)(X_2 - \mu_2)}{\sigma_1 \sigma_2}$$

measures how well residue type 1 and residue type 2 are correlated globally in participating interhelical interactions across the board with all 20 types of residues. The strongest correlation ($\rho = 0.74$) is found between residues N and Y (Figure 7(a)), and the strongest anti-correlation ($\rho = -0.61$) is found between residues F and Y (Figure 7(b)).

To further examine the profiles of interhelical pairwise interactions of residues, we plot the correlation coefficient values of residue N and Y to each of the 20 amino acid residue types in Figure 7(c), respectively. Here the ρ values on the y-axis are plotted against the 20 residue types, which are arranged along the x-axis in ascending order of their side-chain volumes. The ρ values of N and Y closely track one another, indicating that their packing behavior is similar in the TM regions. Figure 7(d) shows the profiles of correlation coefficients of residues F and Y. Except for a few residues, the profiles of correlation coefficients of residue F and residue Y are strongly anticorrelated.

We can further group the 20 types of amino acid residues by their ρ values. Figure 8 shows the clustering of the 20 amino acid residues using the cri-

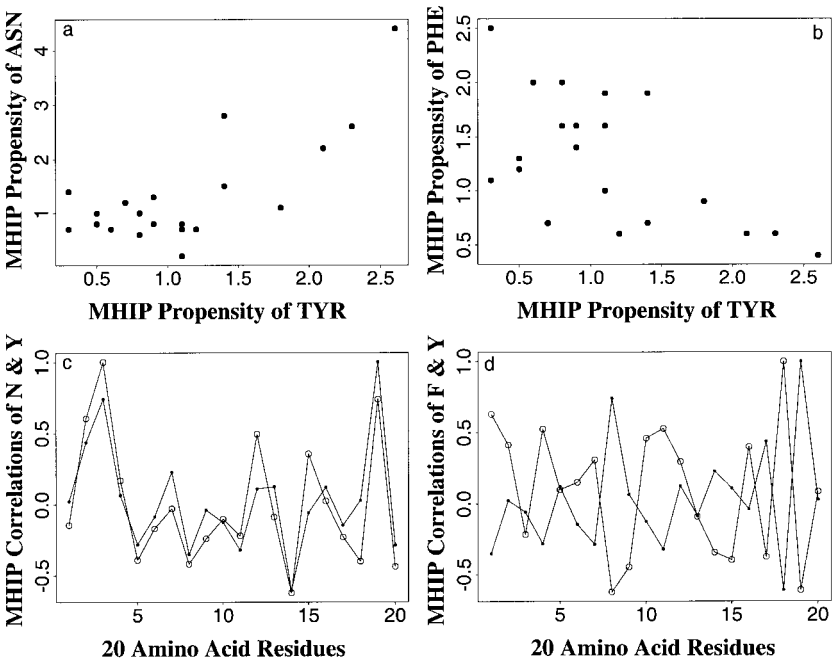

Figure 7. The behavior of interhelical interactions for a specific residue type is determined by the 20 MHIP propensity values. The correlation coefficient of the twenty MHIP values for two residue types indicates how well they are correlated globally in participating interhelical interactions. (a) Residues N and Y are strongly correlated ($\rho = 0.74$). (b) Residues F and Y are strongly anticorrelated ($\rho = -0.61$). (c) The profiles of MHIP correlation of residue N and Y closely track one another. (d) The profiles of MHIP correlation of residue F and Y are strongly anticorrelated.

terion of correlation by hierarchical clustering. As an exploratory tool for data analysis, hierarchical clustering can reveal interesting and informative grouping patterns of the data,[44] although an accurate and robust analysis of the data requires more detailed statistical modeling (see Zhang *et al*.[45] for an example where statistical resampling is used to assess the reliability of clusters identified from hierarchical clustering). In this Figure, residues that are correlated in their MHIP propensity values to the 20 residue types are grouped together. Among residues clustered together, K and M both have a long unbranched side-chain that differs in size by only one atom. C and S are grouped together, and both have unbranched residues of the same size, the difference is an S atom *versus* an O atom. T and V both have branched side-chains and are of the same size. Similarly, both E and L have branched side-chains, P and W both have ring structures. In summary, the pattern of clustering by correlation suggests that residues similar in side-chain branchness, cyclic structures, and size tend to be correlated in participating interhelical interactions.

Glycophorin A and engineered GCN4 leucine zipper

Glycophorin A (GpA) and engineered GCN4 leucine zipper peptide (GCN4-LZ) are two well-studied model systems that have provided much insight about the association of the TM helices of membrane proteins.[46–49] The modes of helical packing found in these two systems may be representative of other membrane proteins. For example, the packing mode of GpA may be shared by synaptobrevin II,[50] the mode of GCN4-LZ may be representative of phospholamban,[51,52] M2 proton channel,[52] as well as bacteriorhodopsin,[54] photosynthetic reaction center,[55] and cytochrome c oxidase.[56] It is therefore important to characterize helical packing in these two model protein systems.

Do helices in GCN4-LZ pack differently from helices in GpA? In engineered GCN4 leucine zipper proteins,[47,48] the two helices are in coiled-coil association as seen in the soluble GCN4 leucine zipper, and are tightly wrapped around each other. In glycophorin A, the TM helices associate to form a stable dimer. Their packing appears to

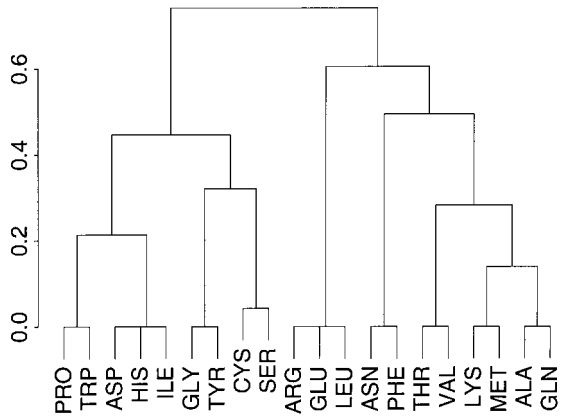

Figure 8. Grouping by hierarchical clustering of the 20 types of amino acid residues by their correlation coefficients of the MHIP values. Clusters are often determined by the side-chain branchness, cyclic structure, and size of the residues. Clustering by compact linkage, average linkage, and complete linkage gives the same result.

involve a less extensive contact surface, because the two helices adopt a splayed configuration.[49] Based on the configuration of the helical backbones, it seems that TM helices in GCN4-LZ pack more tightly than helices in GpA.

We assess the packing efficiency between the TM helices of GCN4-LZ and GpA by measuring the numbers of interhelical contacts per residue and per atom in both proteins. For the 16 residues of the TM region of the GpA dimer and the 16 central residues of GCN4 leucine zipper, the number of interhelical contacts per residue for GpA and GCN4-LZ is 5.3 and 6.0, respectively. Because leucine zipper uses bulkier amino acid residues for interhelical interactions, the average number of atoms per atom participating in the interhelical contact is 1.6 for GCN4 and 1.5 for glycophorin A. This suggests that packing efficiency is very similar for the two motifs represented by GpA and GCN4-LZ. Therefore, specific residue packing may play more important roles in these two proteins than the generic effects of packing efficiency.

Do membrane proteins pack and fold differently from soluble proteins?

Helical membrane proteins share the same alpha helical up-down bundle topology with many soluble proteins. As shown in previous studies, they also often share the same packing motifs such as the *abcdefg* heptads found first in soluble coiled-coil proteins.[23] However, the distributions of packing density for individual amino acid residue type are different for membrane proteins and soluble proteins.[16] Several pieces of evidence gathered in this study further suggest that interhelical pairwise interactions in membrane proteins are different from interactions in soluble proteins. First, disulfide bonds formed by two contacting cysteine residues play important roles in maintaining the stability of soluble proteins and are frequently observed, but they rarely occur in the TM regions of membrane proteins. Second, after adjustment of background composition by comparing log odds-ratio, helices in membrane proteins and soluble proteins do not share common residue pairs with a high degree of helical interfacial pairwise propensity. Third, the pattern of interhelical atomic contacts between polar atoms is very diverse in membrane proteins, including pairs between ionizable residues (salt bridges), ionizable residue and polar residue, as well as polar-polar residues. In contrast, interhelical atomic contacts of polar atoms are exclusively found in residue pairs of two ionizable residues (salt bridges) in soluble proteins. Fourth, packing between backbone atoms, e.g. between residues G-G seems to be more common in membrane proteins than in soluble proteins. Finally, a chi-square test rejects the null hypothesis with a *P*-value of 0.00002 that the two matrices of helical interfacial propensity values for membrane proteins (Table 3) and soluble proteins (Table 4) are drawn from the same underlying probability distribution (chi square = 302.2, $df = 207$).

The finding that membrane proteins pack differently from soluble proteins may have important implications in understanding the mechanism of membrane protein folding. According to the two-stage model of membrane protein folding,[8] hydrophobic helices are first inserted into the membrane to form independently stable TM helices. These TM helices then associate laterally through specific interactions between TM helices to form the tertiary structure. Recently, a detailed Monte Carlo simulation[57] of the folding kinetics of a C-alpha-based two-helix bundle fragment of bacteriorhodopsin in membrane[9] showed that TM helices do not pre-assemble in the solution phase. Rather, they first became completely embedded in the membrane phase horizontally without interhelical contacts, then became vertically oriented and gain in packed tertiary structure. A critical assumption in this computational study is that the modified Lennard-Jones 10-12 potential has different values in the membrane region and in the solution region. This assumption is crucial and determines the mechanism of protein insertion across the membrane, and the difference between the potentials in the two regions controls the amount of tertiary structure formation outside the membrane. The difference in structure-derived empirical potential such as the helical interfacial propensity values between membrane proteins and soluble proteins lend support to this assumption.

Summary

We have developed a novel computational approach to analyze contacting nearest neighbor atoms. On the basis of an atom-based probabilistic model, the pairwise propensity values for interhelical interactions in the TM region are estimated for residue pairs. Our results indicate that there are many specific pairwise interactions in the TM helices. These often involve a diverse pattern of polar-polar atomic interactions. Our results suggest that membrane proteins and soluble proteins have different interhelical interactions, and this observation may help us to understand the folding mechanism of membrane proteins.

Materials and Methods

Membrane and soluble protein data

The 14 membrane proteins used in this study are: cytochrome *c* oxidase from *Paracoccus dentrificans* (PDB accession number: 1ar1),[58] *Thermus thermophilus* (1ehk)[59] and *Bos taurus* (1occ),[60] cytochrome *bc*1 complex from *B. taurus* (1be3),[61] photosynthetic reaction center from *Rhodopseudomonas viridis* (1prc),[53] bacteriorhodopsin (1c3w),[54] halorhodopsin (1e12),[62] rhodopsin (1f88),[63] fumarate reductase flavoprotein from *Escherichia coli* (1fum),[64] and *Wolinella succinogenes* (1qla),[65] glycerol-conducting channel (1fx8),[32] potassium channel (1bl8),[64] calcium-transporting ATPase (1eul),[33] and mechanosensi-

tive ion channel (1msl).[65] The two structures of prokaryotic cytochrome *c* oxidase from *P. dentrificans* and *T. thermophilus* have low-sequence identity (~21 %), and are both included. Two structures of fumarate reductase flavoprotein subunits (PDB accession numbers: 1fum and 1qla) with low sequence identity are also included. All loops in the soluble regions are manually removed, leaving only the alpha helices that span the TM regions.

Determining the exact boundaries of the TM regions is a difficult task even when structures are available.[68] Javadpour *et al*.[16] assigned the TM regions based on the position of basic and acidic residues. Senes *et al*.[68] used short 18-residue windows for the analysis of sequences of TM helices. None of these approaches is error-free under all circumstances. Here, we are interested in assessing the interhelical interactions and the packing of TM helices as a whole in integral membrane proteins, and we use the simple definition of the TM helices from the secondary-structure assignment.

A set of soluble alpha-helical proteins was also constructed for comparison with the membrane proteins. It consists of 31 X-ray structures (PBD accession numbers: 1a0b, 1a17, 1aue, 1b3u, 1cun, 1dkx, 1dow, 1e2a, 1evs, 1ez3, 1ezf, 1few, 1fio, 1gnw, 1gto, 1gux, 1he1, 1le4, 1mty, 1pbv, 1qgh, 1qgr, 1qjb, 1qkr, 1qsa, 1qsd, 1qu7, 1quu, 1vlt, 256b, 2mhr). These proteins all consist of 50 % or more alpha helix and have negligible amount of beta strands. After manually removing the connecting loops, there are a total of 288 unique helices in the data set. We exclude five soluble proteins (PDB accession numbers: 1a17, 1b3u, 1qgr, 1qkr, 1qsa) from void analysis because they contain interacting helices that form voids and pockets of large global length scale, preventing a meaningful comparison with pockets found in membrane proteins.

Computation of voids, pockets and interhelical contacts

There are two aspects in studying the packing of membrane proteins.[69] First, there are unfilled spaces in the TM regions, namely the voids and the pockets that are not occupied. There are well-developed computational methods for the identification and measurement of voids and pockets, including the CAST program[25,70] (http://cast.engr.uic.edu), which are used in this study. Second, there are non-bonded atomic contacts or volume overlaps between residues. We describe below the computational approach used for characterizing such atomic contacts.

The main components of our approach are geometric constructs derived from the coordinates of the protein, namely, the Voronoi diagram, the Delaunay triangulation and the alpha complex (Figure 9(a), (b) and (c)).[25,71–73] A similar approach has been applied to the study of packing in soluble proteins.[69] To illustrate, Figure 9(a) shows a two-dimensional molecule formed by a collection of disks of uniform size. The Voronoi diagram is also shown in Figure 9(a). Each Voronoi cell is defined by its boundaries, shown as broken lines. Every Voronoi edge is a perpendicular bisector of the line between two atom centers. Each Voronoi cell contains one atom, and every point inside a Voronoi cell is closer to this atom than to any other atom. Three connected Voronoi edges meet at a Voronoi vertex. Another geometric construct, the Delaunay triangulation (Figure 9(b)) is mathematically dual to the Voronoi diagram, and can be explained by the following procedure. For each Voronoi edge, connect the corresponding two atom centers with a line segment, and for each Voronoi vertex, place a triangle spanning the three atom centers of the three Voronoi cells. Completing this for all Voronoi edges and Voronoi vertices gives a collection of line segments and triangles. Together with the vertices representing atom centers, they form the ''Delaunay complex'', which is the underlying structure of Delaunay triangulation.

Now we remove all Delaunay edges (or line segments) where the two atoms have no two-body volume overlaps (Figure 9(c)). When two atoms are spatially very close, the balls representing the two atoms intersect, and these two atoms have non-zero, two-body volume overlap. When three atoms are spatially very close, they intersect and have non-zero, three-body volume overlap. We

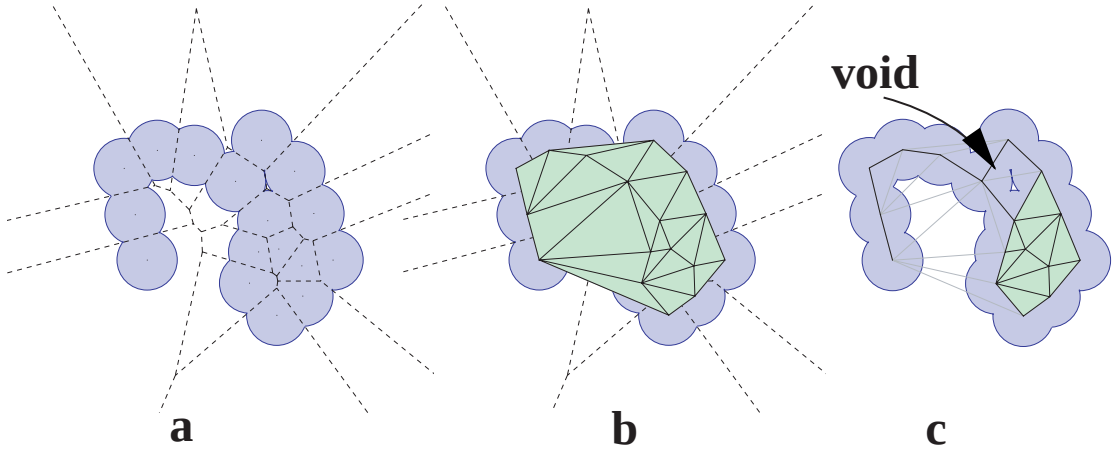

Figure 9. Geometry of protein. (a). The molecule formed by the union of atom disks of uniform size. Voronoi diagram is shown in dashed lines. (b) The shape enclosed by the boundary polygon is the *convex hull*. It is tessellated by the *Delaunay triangulation*. (c) Removing those Delaunay edges and triangles that are not completely contained within the molecule forms the alpha shape of the molecule. A molecular void is represented in the alpha shape by two empty triangles.

further remove all Delaunay triangles where the three corresponding atoms do not have three-body volume overlaps. The subset of the Delaunay complex formed by the remaining triangles, edges and vertices (atom centers) is called the ''alpha complex''. We are interested in identifying only contacting atoms that are spatial nearest neighbors. These are precisely atoms with two-body volume overlaps whose Voronoi cells intersect. By following the mathematical dual structure, i.e. the edges in the alpha complex, we can accurately identify all contacting nearest neighboring atom pairs. Restricting ourselves only to edges connecting atoms from two helices, we can identify all interhelical atomic nearest neighbor contacts. All such contacts are within a distance that depends on the identities of the two atoms. This distance is equal to the sum of the van der Waals radii of the two atoms, plus 2×0.5 Å. We follow Singh & Thornton[74] and choose the increment of van der Waals radii to be 0.5 Å. This increment is small and comparable with the resolution of the structure, and it enables the modeling of imprecisely determined atomic coordinates without introducing many spurious two-body and three-body volume overlaps. Unlike cut-off-based methods, this distance is not a single fixed constant but depends on the atom types. Another important uniqueness of our approach is that only a subset of atoms satisfying this distance criterion will be counted as a physical nearest neighbor. Our approach has an additional criterion that contacting atoms must have intersecting Voronoi cells. Using the alpha shape API kindly provided by the Edelsbrunner group, a program INTERFACE has been implemented to compute interhelical contact atoms, using precomputed Delaunay triangulation and precomputed alpha shape. The Delaunay triangulation of membrane proteins is computed using the DELCX program,[71,75] and the alpha shapes computed using the MKALF program[70,75]. Both can be downloaded from the website of NCSA (http://www.ncsa.uiuc.edu). The van der Waals radii of protein atoms are taken from Tsai *et al.*[77]

Probabilistic model for single residue propensity

To assess how likely a residue of type i is to participate in interhelical contact, we calculate the single residue propensity $p(i)$ for interhelical interaction:

$$p(i) = \frac{\sum_j c(i,j) / \sum_{i',j} c(i',j)}{n(i) / \sum_{i'} n(i')}$$

Here, $\Sigma_j c(i,j)$ is the number count of interhelical residue pairs containing residue type i, $\Sigma_{i',j} c(i',j)$ is the total number of interhelical residue pairs, $n(i)$ is the number count of residue type i in the helices, and $\Sigma_{i'} n(i')$ is the total number of residues of all types in the helices. $n(i)/\Sigma_{i'} n(i')$ is the fraction of type i residue in the TM helices. This denominator is the estimated probability that a randomly picked residue happens to be of residue type i. The single residue propensity $p(i)$ is therefore an odds ratio that corrects the bias due to different residue composition. To simplify the calculation, a pair of interacting residues was counted once, regardless of the actual number of atomic contacts from these residues.

The single residue propensity to be in a void or a pocket p (i) for residue type i can be calculated similarly:

$$p(i) = \frac{c(i) / \sum_{i'} c(i')}{n(i) / \sum_{i'} n(i')}$$

Here, $c(i)$ is the number count of residues of type i that are part of a void or a pocket in the TM region, $\Sigma_{i'} c(i')$ is the total number of residues of any type located in a pocket or a void, $n(i)$ is the number count of residue type i in the TM helices, and $\Sigma_{i'} n(i')$ is the total number of residues of all types in the TM helices.

Probabilistic model for MHIP propensity

To evaluate pairwise helical interaction propensity $P(i,j)$ of residue type i and type j, we first estimate the observed probability $q(i,j)$ of interhelical contacting atom pairs involving both residue type i and residue type j. Using maximum likelihood estimate, we have:

$$q(i,j) = a(i,j) / \sum_{i',j'} a(i',j')$$

Here, $a(i,j)$ is the number count of interhelical atomic contacts between residue type i and residue type j, $\Sigma_{i',j'} a(i',j')$ is the number of all atomic interhelical contacts. The observed probability is then compared against the random probability $p(i,j)$ that a pair of contacting atoms is picked from a residue of type i and a residue of type j, respectively, when chosen randomly and independently from the same set of interacting residues in the TM regions. We have:

$$p(i,j) = N_i N_j \left(\frac{n_i n_j}{n(n-n_i)} + \frac{n_i n_j}{n(n-n_j)} \right), \text{ when } i \neq j$$

Here, N_i is the number of interacting residues of type i in the TM region, n_i is the number of atoms a residue of type i has, and n is the total number of interacting atoms in the TM region. For calculating polar-polar or nonpolar-non-polar interhelical contact propensities, we replace n_i in the numerator with the number of polar or non-polar atoms for this type of amino acid residue. The formula is different if both residues are of the same type:

$$p(i,i) = N_i(N_i - 1) \frac{n_i n_i}{n(n-n_i)}$$

The MHIP propensity $P(i,j)$ is the odds ratio of the observed probability and the random probability:

$$P(i,j) = \frac{q(i,j)}{p(i,j)}$$

Acknowledgments

We thank Dr Connie Jeffrey for stimulating conversations and suggestions, Drs Nir Ben-Tal, Renhao Li, and Clare Woodward for helpful discussions, Drs Nir Ben-Tal and Stephen White for sharing their results prior to publication. We thank all structural biologists for depositing the coordinates of membrane proteins in the Protein Data Bank. We thank Dr Soulimane for providing us with coordinates of cytochrome *c* oxidase (PDB file 1ehk) before they were available from the PDB. Research

support from NSF (DBI-0078270, MCB998008) and ACS (Petroleum Research Fund #35616-G7) is gratefully acknowledged.

References

1. Boyd, D., Schierle, C. & Beckwith, J. (1998). How many membrane proteins are there? *Protein Sci.* **7**, 201-205.
2. Wallin, E. & von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7**, 1029-1038.
3. Sipos, L. & von Heijne, G. (1993). Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* **213**, 1333-1340.
4. Claros, M. G. & von Heijne, G. (1994). TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* **10**, 685-686.
5. Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5**, 1704-1718.
6. Rost, B., Casadio, R., Fariselli, P. & Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**, 521-533.
7. Cserzo, M., Wallin, E., Simon, I., von Heijne, G. & Elofsson, A. (1997). Prediction of transmembrane alpha-helices in procariotic membrane proteins: the Dense Alignment Surface method. *Protein Eng.* **10**, 673-676.
8. Popot, J. L. & Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry,* **29**, 4031-4037.
9. Kahn, T. W. & Engelman, D. M. (1992). Bacteriorhodopsin can be refolded from two independently stable transmembrane helices and the complementary five-helix fragment. *Biochemistry,* **31**, 6144-6151.
10. Treutlein, H. R., Lemmon, M. A., Engelman, D. M. & Brunger, A. T. (1992). The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. *Biochemistry,* **31**, 12726-12732.
11. Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J. & Engelman, D. M. (1992). Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry,* **31**, 12719-12725.
12. Lemmon, M. A., Treutlein, H. R., Adams, P. D., Brunger, A. T. & Engelman, D. M. (1994). A dimerization motif for transmembrane alpha-helices. *Nature Struct. Biol.* **1**, 157-163.
13. Langosch, D., Brosig, B., Kolmar, H. & Fritz, H. J. (1996). Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. *J. Mol. Biol.* **263**, 525-530.
14. Mingarro, I., Whitley, P., Lemmon, M. A. & von Heijne, G. (1996). Ala-insertion scanning mutagenesis of the glycophorin A transmembrane helix: a rapid way to map helix-helix interactions in integral membrane proteins. *Protein Sci.* **5**, 1339-1341.
15. Mingarro, I., Elofsson, A. & von Heijne, G. (1997). Helix-helix packing in a membrane-like environment. *J. Mol. Biol.* **272**, 633-641.
16. Javadpour, M. M., Eilers, M., Groesbeek, M. & Smith, S. O. (1999). Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophys. J.* **77**, 1609-1618.
17. Eilers, M., Shekar, S. C., Shieh, T., Smith, S. O. & Fleming, P. J. (2000). Internal packing of helical membrane proteins. *Proc. Natl Acad. Sci. USA,* **97**, 5796-5801.
18. Li, S. C. & Deber, C. M. (1992). Glycine and beta-branched residues support and modulate peptide helicity in membrane environments. *FEBS Letters,* **311**, 217-220.
19. Smith, S. O., Jonas, R., Braiman, M. & Bormann, B. J. (1994). Structure and orientation of the transmembrane domain of glycophorin A in lipid bilayers. *Biochemistry,* **33**, 6334-6341.
20. Jiang, S. & Vakser, I. A. (2000). Side-chains in transmembrane helices are shorter at helix-helix interfaces. *Proteins: Struct. Funct. Genet.* **40**, 429-435.
21. White, S. H. (2001). Tryptophan and the folding of proteins into membranes. *Biophys. J.* **80**, 5a.
22. Pilpel, Y., Ben-Tal, N. & Lancet, D. (1999). kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J. Mol. Biol.* **294**, 921-935.
23. Langosch, D. & Heringa, J. (1998). Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils. *Proteins: Struct. Funct. Genet.* **31**, 150-159.
24. Ben-Tal, N. & Honig, B. (1996). Helix-helix interactions in lipid bilayers. *Biophys. J.* **71**, 3046-3050.
25. Edelsbrunner, H., Facello, M. & Liang, J. (1998). On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math.* **88**, 83-102.
26. Liang, J., Edelsbrunner, H. & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7**, 1884-1897.
27. von Heijne, G. (1999). Recent advances in the understanding of membrane protein assembly and structure. *Quart. Rev. Biophys.* **32**, 285-307.
28. Edelsbrunner, H., Facello, M., Fu, P. & Liang, J. (1995). Measuring proteins and voids in proteins. *Proceedings of the 28th Annual Hawaii International Conference on System Science,* vol. 5, pp. 256-264.
29. Landolt-Marticorena, C., Williams, K. A., Deber, C. M. & Reithmeier, R. A. (1993). Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins. *J. Mol. Biol.* **229**, 602-608.
30. Arkin, I. T. & Brunger, A. T. (1998). Statistical analysis of predicted transmembrane α-helices. *Biochim. Biophys. Acta,* **1429**, 113-128.
31. Yau, W. M., Wimley, W. C., Gawrisch, K. & White, S. H. (1998). The preference of tryptophan for membrane interfaces. *Biochemistry,* **37**, 14713-14718.
32. Fu, D., *et al*. (2000). Structure of a glycerol-conducting channel and the basis for its selectivity. *Science,* **290**, 481-486.
33. Toyoshima, C., Nakasako, M., Nomura, O. H. & Ogawa, H. (2000). Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature,* **405**, 647-655.
34. MacLennan, D. H., Rice, W. J. & Green, M. N. (1997). The mechanism of Ca^{2+} transport by sarco(Endo)plasmic reticulum Ca^{2+}-ATPases. *J. Biol. Chem.* **272**, 28815-28818.
35. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51-55.
36. Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective inter-residue contact energies from protein

crystal structures: quasi-chemical approximation. *Macromolecules,* **18**, 534-552.
37. Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623-644.
38. Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229-235.
39. Zhang, L. & Skolnick, J. (1998). How do potentials derived from structural databases relate to ''true'' potentials? *Protein Sci.* **7**, 112-122.
40. Thomas, P. D. & Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **257**, 457-469.
41. Vijayakumar, M. & Zhou, H.-X. (2000). Prediction of residue-residue pair frequencies in proteins. *J. Phys. Chem. ser. B,* **104**, 9755-9764.
42. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Struct. Funct. Genet.* **43**, 89-102.
43. Pattabiraman, N., Ward, K. B. & Fleming, P. J. (1995). Occluded molecular surface: analysis of protein packing. *J. Mol. Recogn.* **8**, 334-344.
44. Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data*, John Wiley & Sons, New York.
45. Zhang, K. & Zhao, H. (2000). Assessing reliability of gene clusters from gene expression data. *Funct. Integr. Genom.* **1**, 156-173.
46. MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science,* **276**, 131-133.
47. Choma, C., Gratkowski, H., Lear, J. D. & DeGrado, W. F. (2000). Asparagine-mediated self-association of a model transmembrane helix. *Nature Struct. Biol.* **7**, 161-166.
48. Zhou, F. X., Cocco, M. J., Russ, W. P., Brunger, A. T. & Engelman, D. M. (2000). Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nature Struct. Biol.* **7**, 154-160.
49. Bowie, J. U. (2000). Understanding membrane protein structure by design. *Nature Struct. Biol.* **7**, 91-94.
50. Laage, R. & Langosch, D. (1997). Dimerization of the synaptic vesicle protein synaptobrevin (vesicle-associated membrane protein) II depends on specific residues within the transmembrane segment. *Eur. J. Biochem.* **249**, 540-546.
51. Arkin, I. T. *et al*. (1994). Structural organization of the pentameric transmembrane alpha-helices of phospholamban, a cardiac ion channel. *EMBO J.* **13**, 4757-4764.
52. Simmerman, H. K., Kobayashi, Y. M., Autry, J. M. & Jones, L. R. (1996). A leucine zipper stabilizes the pentameric membrane domain of phospholamban and forms a coiled-coil pore structure. *J. Biol. Chem.* **271**, 5941-5946.
53. Pinto, L. H. *et al*. (1997). A functionally defined model for the M2 proton channel of influenza A virus suggests a mechanism for its ion selectivity. *Proc. Natl Acad. Sci. USA,* **94**, 11301-11306.
54. Luecke, H., Schobert, B., Richter, H.-T., Cartailler, J.-P. & Lanyi, J. K. (1999). Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.* **291**, 899-911.
55. Deisenhofer, J., Epp, O., Sinning, I. & Nichell, H. (1995). Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction center from *Rhodopseudomonas viridis*. *J. Mol. Biol.* **246**, 429-457.
56. Gurezka, R., Laage, R., Brosig, B. & Langosch, D. (1999). A heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments. *J. Biol. Chem.* **274**, 9265-9270.
57. Orlandini, E., Seno, F., Banavar, J. R., Laio, A. & Maritan, A. (2000). Deciphering the folding kinetics of transmembrane helical proteins. *Proc. Natl Acad. Sci. USA,* **97**, 14229-14234.
58. Ostermeier, C., Harrenga, A., Ermler, U. & Michel, H. (1997). Structure at 2.7 Å resolution of the Paracoccus denitrificans two-subunit cytochrome *c* oxidase complexed with an antibody FV fragment. *Proc. Natl Acad. Sci. USA,* **94**, 10547-10553.
59. Soulimane, T. *et al*. (2000). Structure and mechanism of the aberrant ba(3)-cytochrome *c* oxidase from *Thermus thermophilus*. *EMBO J.* **19**, 1766-1776.
60. Tsukihara, T. *et al*. (1996). The whole structure of the 13-subunit oxidized cytochrome *c* oxidase at 2.8 Å. *Science,* **272**, 1136-1144.
61. Iwata, S. *et al*. (1998). Complete structure of the 11-subunit bovine mitochondrial cytochrome *bc*1 complex. *Science,* **281**, 64-71.
62. Kolbe, M., Besir, J., Essen, L. O. & Oesterhelt, D. (2000). Structure of light-driven chloride pump halorhodopsin at 1.8 Å resolution. *Science,* **288**, 1390-1396.
63. Palczewski, K. *et al*. (2000). Crystal structure of rhodopsin: a G protein-coupled receptor. *Science,* **289**, 739-745.
64. Iverson, T. M., Luna-Chavez, C., Cecchini, G. & Rees, D. C. (1999). Structure of the *E. coli* fumarate reductase respiratory complex. *Science,* **284**, 1961-1966.
65. Lancaster, C. R. D., Kroeger, A., Auer, M. & Michel, J. (1999). Structure of fumarate reductase from *Wolinella succinogenes* at 2.2 Å resolution. *Nature,* **402**, 377-385.
66. Doyle, D. A. *et al*. (1998). The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science,* **280**, 69-77 (1998).
67. Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. (1998). Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. *Science,* **282**, 2220-2226.
68. Senes, A., Gerstein, M. & Engelman, D. M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with β-branched residues at neighboring positions. *J. Mol. Biol.* **296**, 921-936.
69. Liang, J. & Dill, K. A. (2001). Are proteins well-packed? *Biophys. J.* **81**, 751-766.
70. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. & Subramaniam, S. (1998). Analytical shape computation of macromolecules. I. Molecular area and volume through alpha shape. *Proteins: Struct. Funct. Genet.* **33**, 1-17.
71. Edelsbrunner, H. & Mucke, E. P. (1994). 3-Dimensional alpha-shapes. *ACM Transact. Graph.* **13**, 43-72.
72. Edelsbrunner, H. (1995). The union of balls and its dual shape. *Discrete Comput. Geom.* **13**, 415-440.
73. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. & Subramaniam, S. (1998). Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins: Struct. Funct. Genet.* **33**, 18-29.
74. Singh, J. & Thornton, J. M. (1992). *Atlas of Protein Side-chain Interactions*, IRL Press, Oxford.

75. Edelsbrunner, H. & Shah, N. R. (1996). Incremental topological flipping works for regular triangulations. *Algorithmica,* **15**, 223-241.
76. Facello, M. A. (1995). Implementation of a randomized algorithm for Delaunay and regular triangulation in 3 dimensions. *Comput. Aided Geom. sect. D,* **12**, 349-370.
77. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **290**, 253-266.

Edited by G. von Heijne

(Received 17 April 2001; received in revised form 4 July 2001; accepted 4 July 2001)

http://www.academicpress.com/jmb

Supplementary Material comprising one Table is available on IDEAL