doi:10.1016/S0022-2836(03)00041-X J. Mol. Biol. (2003) 327, 251–272

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

# Higher-order Interhelical Spatial Interactions in Membrane Proteins

**Larisa Adamian, Ronald Jackups Jr, T. Andrew Binkowski and Jie Liang***

*Department of Bioengineering University of Illinois at Chicago, IL 60607, USA*

Higher-order interactions are important for protein folding and assembly. We introduce the concept of interhelical three-body interactions as derived from Delaunay triangulation and alpha shapes of protein structures. In addition to glycophorin A, where triplets are strongly correlated with protein stability, we found that tight interhelical triplet interactions exist extensively in other membrane proteins, where many types of triplets occur far more frequently than in soluble proteins. We developed a probabilistic model for estimating the value of membrane helical interaction triplet (MHIT) propensity. Because the number of known structures of membrane proteins is limited, we developed a bootstrap method for determining the 95% confidence intervals of estimated MHIT values. We identified triplets that have high propensity for interhelical interactions and are unique to membrane proteins, e.g. AGF, AGG, GLL, GFF and others. A significant fraction (32%) of triplet types contains triplets that may be involved in interhelical hydrogen bond interactions, suggesting the prevalent and important roles of H-bond in the assembly of TM helices. There are several well-defined spatial conformations for triplet interactions on helices with similar parallel or antiparallel orientations and with similar right-handed or left-handed crossing angles. Often, they contain small residues and correspond to the regions of the closest contact between helices. Sequence motifs such as GG4 and AG4 can be part of the three-body interactions that have similar conformations, which in turn can be part of a higher-order cooperative four residue spatial motif observed in helical pairs from different proteins. In many cases, spatial motifs such as serine zipper and polar clamp are part of triplet interactions. On the basis of the analysis of the archaeal rhodopsin family of proteins, tightly packed triplet interactions can be achieved with several different choices of amino acid residues.



*Keywords:* membrane protein; three-body interaction; hydrogen bond; structure clustering; alpha shape

*Corresponding author

## Introduction

Membrane proteins play essential cellular roles, including signal transduction, proton pumping, ion transport, and light harvesting. Although transmembrane (TM) helices can be predicted reliably from sequences,[1–2] only a limited number of structures of membrane proteins are known. Understanding how TM helices interact with each other in the lipid environment may help us to understand how helical membrane proteins fold and assemble, and how the assembled structures carry out biological functions. In addition, it will aid in developing computational methods for predicting membrane protein structures.

The TM regions of a large number of membrane proteins are likely to be helical bundles. The compositional and structural simplicity of TM helices suggests that a finite number of common sequential and spatial packing patterns may exist to

Abbreviations used: TM, transmembrane; GpA, glycophorin A; MHIT, membrane helical interfacial triplet; AR, archaeal rhodopsins; bR, bacteriorhodopsin; hR, halorhodopsin; sR, sensory rhodopsin I; pR, sensory rhodopsin II.

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

mediate helix–helix interactions. Extensive experimental work on oligomerization of single-span glycophorin A in detergents revealed the importance of GxxxG (GG4) sequence motif, which facilitates the "knob-into-hole" packing and efficiently enhances van der Waals interactions.[3–4] Polytopic TM proteins are more complex, as they are assembled from multiple non-homologous TM helices. Examination of high-resolution structure of bacteriorhodopsin allowed Luecke *et al*.[5] to conclude that the transmembrane region of bacteriorhodopsin is more compact and rigid than the solvent-exposed region. Recent comparative study of membrane and soluble helical-bundle proteins showed that the average packing values for amino acid residues in membrane proteins are higher.[6] The compactness of the TM region facilitates interactions between amino acid residues on neighboring helices. It is likely that there are additional sequence and spatial motifs in polytopic membrane proteins that mediate helix–helix interactions. An indication comes from recent statistical analysis of TM sequences where several over-represented sequence motifs in addition to GG4, including II4, GA4, IG1 and others are identified.[7]

The pairwise propensity of residues for interhelical interactions in the TM regions has been examined in detail, where sets of empirical propensity scales have been developed,[6,8] and a number of spatial motifs, including the polar clamp and the serine zipper, have been discovered.[9] However, recent studies showed that pairwise contacts alone are, in general, inadequate for studying protein folding.[10–11] The need for introducing higher-order interactions in folding proteins has been highlighted in recent works.[12] Here, we analyze higher-order spatial interactions in membrane proteins. When helices are packed tightly, three atoms from three different amino acid residues may be in close contact with each other. We identify this type of packing interaction as three-body interaction or "triplet". All three-body interactions or triplets that contain amino acid residues A, B and C belong to "triplet type" ABC. We further focus on interhelical triplets; namely, interactions from three residues residing on at least two different helices. The computational methods we use are geometric algorithms (Voronoi diagram, Delaunay triangulation and alpha shape) that have been applied previously to study pairwise or two-body interhelical interactions in membrane and soluble proteins.[8] These methods provide accurate characterization of the nearest-neighbor interactions that are in physical contact. To identify higher-order interactions of amino acid residues within the TM region reliably from a relatively small data set (17 proteins), we further develop a bootstrap procedure[13–15] to assess the confidence intervals of estimated propensity values.

We first discuss three-body interactions in membrane proteins with the example of glycophorin A. We then discuss the descriptive statistics of triplet interactions found in membrane proteins and soluble proteins. Triplet types with high propensity for interhelical interactions are then identified in membrane proteins and are compared with those from soluble proteins. We further discuss the roles of H-bonds, side-chain size and chemical properties in triplet interactions. Clusters of triplets that closely resemble each other are then identified, and the relationships between spatial and sequence motifs are discussed. Finally, we discuss evolutionary conservation of residues in triplets found in the archaeal rhodopsin family.

## Results

### Examples of triplet interactions from glycophorin A (GpA)

Glycophorin A is an exemplary protein that provides a structural paradigm for studying interhelical interactions of residues in the TM region.[16] GpA consists of two identical TM helices (chain A and chain B) that form a dimer in the TM region, with a tightly packed middle region (residues 73–88).[16] Interacting amino acid residues in this region often form pairs with high propensity for interhelical interactions:[8] G-G (propensity 3.0), I-I (1.3) and T-V (1.1). These tight packing interactions are important for folding and stability of GpA. We use the structure of GpA (pdb 1AFO) to illustrate the triplets or three-body interactions in membrane proteins.

The INTERFACE-3 program detected six different types of triplets in the middle region of the dimer. They are GGV, GTV, GVV, IIL, IIT and ITT. Figure 1(a) shows an example of three-body interaction in a GVV triplet. This triplet is formed

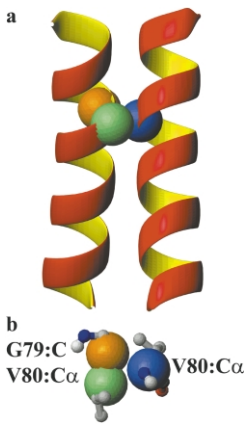


**Figure 1**. GVV triplet in GpA dimer (chains A and B, residues 73–91). (a) Tight three-body atomic cluster. Orange, C from G79, chain A; green, $C^{\alpha}$ from V80, chain A; blue, $C^{\alpha}$ from V80, chain B. (b) Top view of the same three-body atomic cluster shown in space-filling representation. Other atoms from residues G79A, V80A and V80B are shown in ball-and-stick representation. All of the molecular structure representations in Figures were drawn with the program MOLMOL.[36]

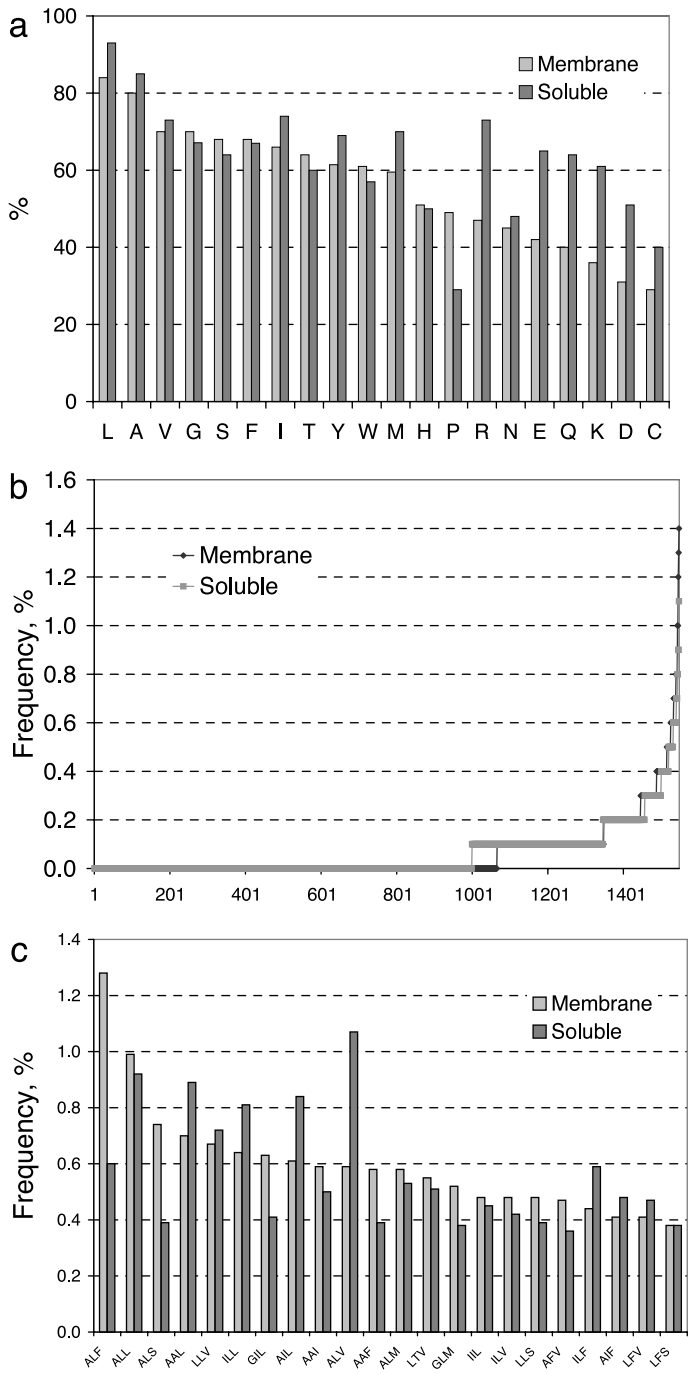


**Figure 2**. (a) The relative frequency of each amino acid residue to be found in different triplet types in comparison with the ideal frequency of 210 if a residue appears in all possible triplet types for helices in membrane and soluble proteins. (b) Relative frequencies of triplets of amino acid residues in interhelical contact, i.e. the number of observed occurrences of a specific triplet interactions divided by the total number of all triplets. Triplets are arranged along the *x*-axis in ascending order of frequency. (c) Frequencies of high propensity triplets that are common for membrane and soluble proteins.

by three carbon atoms shown in space-filling symbols. Figure 1(b) shows a top view of this interaction and demonstrates that atoms in this triplet are in close van der Waals contact. In other words, triplets can be considered as minimalistic elements of tight packing in membrane protein structure.

The introduction of clashes or the removal of favorable packing interactions is likely to disrupt or eliminate three-body interactions. Alanine scanning mutagenesis and sedimentation equilibrium analytical ultracentrifugation studies of GpA mutants identified several residues that are "hot spots" of interaction between the helices.[17] All hot spot residues (L75, I76, G79, V80, G83, V84, T87) were found in tightly packed triplets. At the same time, all triplets identified by computation contain hot spot residues. It seems that triplet interactions are strongly correlated with stability of GpA assembly.

## Triplets in membrane and soluble proteins

In principle, each of the 20 amino acid residues can be found in 210 different types of triplets, where it can appear once (e.g. residue M in triplet MTV), twice (MMV), or three times (MMM). In the set of 17 transmembrane proteins used in this study, the INTERFACE-3 program identified 846

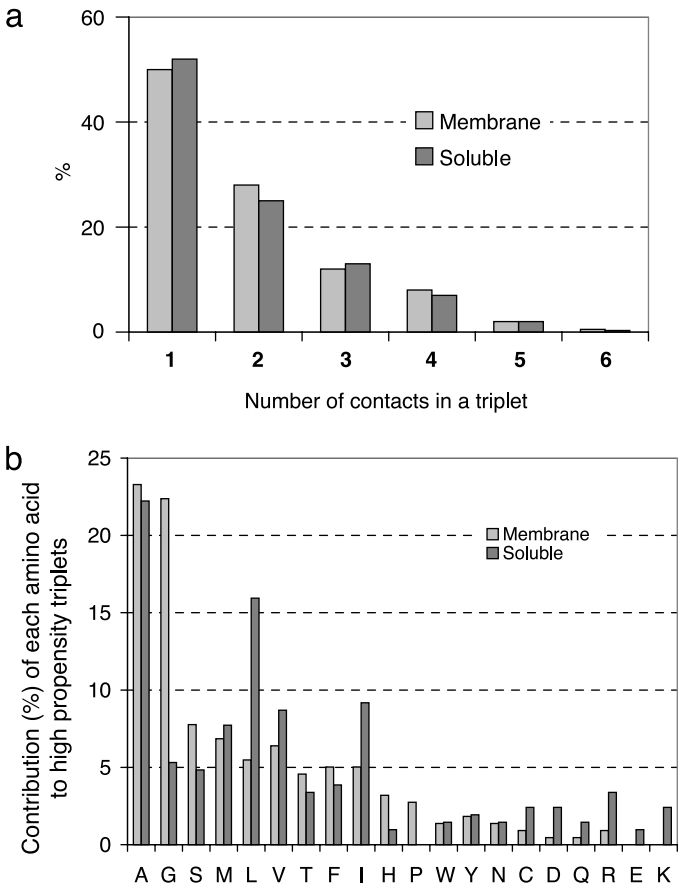


**Figure 3**. (a) Distributions of the number counts of three-body atomic contacts in a residue triplet observed in membrane and in soluble proteins. (b) Contributions (%) of amino acid residues in different types of triplet interactions of high propensity in membrane and proteins.

out of 1540 possible triplet types of amino acids†. In the set of soluble helical proteins of comparable size, INTERFACE-3 identified 951 out of 1540 possible triplet types.

Helices in the TM region of membrane proteins and helices in soluble proteins have different amino acid compositions. This is reflected in the compositional differences of triplets formed in TM and in soluble protein helices. Figure 2(a) shows the number of different triplet types in which each residue appears when compared with the number (210) of all possible triplet types containing a specific residue. No amino acid is found in all 210 possible triplet types. Leucine and alanine are found in the largest number of triplet types in both membrane and soluble proteins. Val, Gly, Ser, Phe, Thr, Ile, Tyr, Trp and Met have similar distributions of interactions with other types of amino acid residues in membrane and soluble proteins. Ionizable and polar residues (Arg, Glu, Gln, Lys, Asp and Cys) have a more diverse pattern of triplet types in soluble proteins than in membrane proteins, presumably due to ionic and polar interactions on the surfaces of soluble proteins. Proline has more diverse pattern of interactions in membrane than in soluble proteins.

† The coordinates of all triplets calculated in this study are available at http://gila.bioengr.uic.edu/triplets

The sorted distributions of relative frequencies of triplet types, i.e. the number of observed occurrences of a specific triplet of interactions divided by the total number of all triplets, is shown in Figure 2(b). The overall patterns are similar for membrane and soluble proteins. There is an exponential increase in frequencies for the top 100 triplet types. If all triplets were distributed uniformly among all possible triplet types, each triplet type would have a frequency of $1/1540 = 0.07\%$. The majority of triplet types (74% in membrane and 71% in soluble helices) are either missing or scarcely sampled with frequencies $<0.07\%$. The top 100 triplet types have frequencies higher than 0.2%, which is about three times the frequency of the uniform distribution. The high-frequency triplet types (frequency $>0.25\%$) are relatively scarce: there are 92 triplet types (6%) in membrane and 91 (6%) in soluble helices.

There are 22 common high-frequency triplet types among the 50 top high-frequency triplet types from both membrane and soluble proteins (Figure 2(c)). The tight packing between residues in ALF, ALS, GIL and AAF triplets is more preferred in membrane helices than in soluble helices, while contacts between residues AAL, ILL, AIL and ALV are more frequent in soluble proteins. Residues Ala and Leu pack more often with Phe in membrane than in soluble proteins, forming ALF triplets with frequencies of 1.3 and 0.6,

**Table 1.** Odds ratios of observed and expected frequencies together with Studentized confidence intervals for high-propensity triplets of amino acid residues in membrane proteins

| Triplet | Odds ratio | Bootstrap interval | Count | Triplet | Odds ratio | Bootstrap interval | Count |
|---|---|---|---|---|---|---|---|
| **AAA** | 4.1 | (2.2…13.8) | 12 | APT | 4.1 | (2.2…13.4) | 14 |
| **AAF** | 2.8 | (1.5…6.1) | 37 | APY | 2.6 | (1.1…12.5) | 11 |
| **AAG** | 9.7 | (6.0…23.5) | 50 | ARG | 3.4 | (2.2…19.8) | 18 |
| **AAI** | 4.3 | (2.7…14.4) | 38 | ASS | 8.2 | (3.2…41.1) | 18 |
| **AAL** | 2.3 | (1.5…4.9) | 45 | **ASV** | 2.2 | (1.2…7.1) | 19 |
| **AAM** | 5.1 | (2.2…14.3) | 26 | ATV | 2.5 | (1.4…6.9) | 26 |
| AAP | 3.4 | (1.3…25.4) | 10 | CGF | 8.5 | (4.3…26.9) | 13 |
| **AAS** | 2.9 | (1.3…7.9) | 13 | GFF | 3.3 | (2.0…9.0) | 37 |
| **AAV** | 2.2 | (1.4…5.3) | 19 | GFS | 2 | (1.3…4.8) | 15 |
| AAW | 2.2 | (1.2…6.1) | 19 | GGF | 6.8 | (4.3…16.5) | 30 |
| ACI | 8.6 | (4.5…35.8) | 15 | GGG | 17.4 | (8.2…121.1) | 10 |
| AFP | 2.1 | (1.2…5.5) | 18 | GGI | 5.8 | (3.7…16.2) | 17 |
| AGF | 4.9 | (3.8…9.0) | 75 | GGL | 3.5 | (1.7…32.7) | 23 |
| AGG | 21.7 | (15.1…49.8) | 65 | GGM | 9.9 | (6.8…28.2) | 17 |
| **AGI** | 4.3 | (2.9…11.2) | 44 | GGS | 10.7 | (6.0…36.5) | 16 |
| **AGL** | 3.9 | (2.8…6.6) | 89 | GGV | 5.7 | (2.1…28.4) | 17 |
| AGM | 3.9 | (2.4…8.4) | 23 | GHS | 5.1 | (2.2…16.3) | 11 |
| **AGS** | 5.8 | (3.8…14.4) | 30 | GHT | 6.2 | (2.3…19.6) | 16 |
| AGT | 4.4 | (2.4…20.3) | 27 | GIL | 1.8 | (1.2…3.5) | 40 |
| AGV | 4.3 | (3.0…7.3) | 44 | GLL | 2.2 | (1.5…4.7) | 54 |
| AGY | 2.9 | (1.4…16.1) | 22 | **GLM** | 2.5 | (1.5…5.4) | 33 |
| AHS | 5.6 | (2.9…41.1) | 21 | **GLV** | 2.4 | (1.7…4.0) | 53 |
| AHT | 3.6 | (2.0…11.0) | 16 | GMF | 3.2 | (2.0…6.7) | 28 |
| AIM | 2.5 | (1.6…6.0) | 25 | GMS | 6.8 | (3.1…21.0) | 20 |
| AIP | 2.6 | (1.5…9.9) | 15 | GMV | 4.8 | (3.2…9.9) | 28 |
| **AIT** | 2.2 | (1.2…5.0) | 23 | GMW | 2.8 | (1.2…13.2) | 16 |
| **ALL** | 1.5 | (1.1…2.2) | 63 | GST | 6.6 | (3.4…18.5) | 20 |
| ALM | 1.7 | (1.1…2.6) | 37 | GSV | 2 | (1.1…5.9) | 10 |
| ALS | 2.4 | (1.7…4.4) | 47 | GTV | 2.5 | (1.2…20.4) | 15 |
| AMF | 2.2 | (1.4…4.4) | 34 | GVV | 2.2 | (1.2…7.9) | 11 |
| AMV | 2.4 | (1.2…6.3) | 24 | IIM | 3.6 | (1.7…11.3) | 18 |
| ANG | 5.2 | (2.5…32.3) | 12 | NGS | 9.5 | (4.4…32.6) | 11 |
| ANY | 4.1 | (2.0…16.4) | 12 | STV | 2.9 | (1.5…8.3) | 15 |

All entries of triplets have ten or more three-body atomic contacts. Triplets that have high propensity in both membrane and soluble proteins are highlighted in bold face.

respectively. Residues Ala and Leu pack with Val almost twice as often in soluble than in membrane proteins, forming ALV triplets with frequencies of 1.1 and 0.6, respectively.

### Extent of atomic contacts in an individual triplet

There may be multiple three-body atomic contacts in a triplet interaction. The total number of three-body atomic contacts provides a relative measure of the extent of atomic contacts among three amino acid residues, although this measure is subject to the uncertainty in atomic coordinates. We found that three interacting amino acid residues may share several three-body atomic contacts in both membrane and soluble proteins. For example, three three-body atomic contacts are found in triplet G28-L32-F80 in 1EHK (cytochrome *c* oxidase from *Thermus thermophilus*). Each three-body interaction is composed of a different set of atoms: G28(O)-L32(CB)-F80(CD1), G28(O)-L32(N)-F80(CB) and G28(O)-L32(CB)-F80(CB). The number count of three-body atomic contacts in triplets shows approximately the same distributions for helices in membrane proteins and in soluble proteins (Figure 3(a)). The average number of contacts per triplet (total number of contacts divided by total number of triplets) is 1.87 for membrane proteins and 1.83 for soluble proteins.

### High and low-propensity triplets in polytopic membrane proteins

To remove compositional bias, we calculate the propensity values of membrane helical interfacial triplet (MHIT) interactions. MHIT values are obtained by comparing the observed frequency of a triplet against the expected frequency if the three atoms are drawn randomly from a pool of amino acid residues of the same composition (see Materials and Methods). There are 73 triplet types with high propensity in TM helices, and 71 triplet types with high propensity in helices from soluble proteins. These are the triplet types with confidence intervals for the MHIT values estimated from the bootstrap method. The MHIT values are listed in Table 1 for triplet types from membrane proteins. The propensity values for triplet types

**Table 2.** Odds ratios of observed and expected frequencies together with Studentized confidence intervals for high-propensity triplets of amino acid residues in soluble proteins

| Triplet | Odds ratio | Bootstrap interval | Count | Triplet | Odds ratio | Bootstrap interval | Count |
|---|---|---|---|---|---|---|---|
| **AAA** | 15.7 | (7.7…94.4) | 32 | CLL | 5.7 | (2.1…34.2) | 35 |
| **AAF** | 4.5 | (2.4…12.5) | 26 | CLM | 3.8 | (2.1…24.5) | 12 |
| **AAG** | 6.7 | (4.4…38.5) | 10 | DEK | 3.0 | (1.8…10.4) | 15 |
| **AAI** | 5.2 | (3.9…9.0) | 33 | DKY | 3.5 | (2.1…10.6) | 17 |
| AAK | 4.0 | (2.2…9.9) | 21 | EHY | 3.2 | (1.8…34.5) | 17 |
| **AAL** | 4.2 | (2.9…8.0) | 59 | **GIL** | 3.8 | (2.1…7.6) | 27 |
| **AAM** | 5.6 | (3.0…17.4) | 20 | GIM | 5.5 | (3.5…31.2) | 10 |
| AAR | 3.0 | (1.1…18.8) | 21 | **GLM** | 6.3 | (3.3…18.1) | 25 |
| **AAS** | 5.5 | (3.3…12.2) | 15 | **GLV** | 2.6 | (1.3…6.9) | 17 |
| AAT | 3.7 | (1.6…18.1) | 12 | GMT | 18.1 | (5.6…149.7) | 17 |
| **AAV** | 7.0 | (4.2…11.6) | 42 | III | 8.0 | (4.1…97.5) | 18 |
| ACF | 5.3 | (1.1…47.9) | 12 | IIL | 2.0 | (1.1…5.8) | 30 |
| ACL | 4.2 | (2.6…8.0) | 23 | ILL | 1.7 | (1.2…2.5) | 54 |
| AFV | 2.1 | (1.1…6.0) | 24 | ILT | 1.9 | (1.1…5.3) | 29 |
| **AGI** | 4.2 | (2.0…17.5) | 13 | IMV | 2.3 | (1.2…10.1) | 17 |
| **AGL** | 3.1 | (1.5…20.1) | 21 | KFY | 2.9 | (1.2…43.5) | 28 |
| AHL | 2.5 | (1.1…49.4) | 30 | LFS | 2.2 | (1.2…5.5) | 25 |
| AIF | 2.7 | (1.7…5.5) | 32 | LLL | 1.6 | (1.1…2.7) | 37 |
| AII | 3.2 | (1.8…26.5) | 21 | LLS | 1.9 | (1.1…3.6) | 26 |
| AIL | 1.9 | (1.2…3.6) | 56 | LLV | 1.6 | (1.1…3.1) | 48 |
| **AIT** | 3.5 | (1.9…13.9) | 24 | LMF | 2.6 | (1.8…4.4) | 40 |
| **ALL** | 1.9 | (1.3…3.0) | 61 | LMM | 2.8 | (1.3…7.8) | 13 |
| **ALM** | 2.2 | (1.3…4.0) | 35 | LST | 2.0 | (1.2…4.6) | 13 |
| ALT | 3.0 | (2.0…5.7) | 44 | LVV | 2.0 | (1.1…5.4) | 26 |
| ALV | 2.6 | (1.8…4.1) | 71 | LWV | 1.8 | (1.2…5.0) | 23 |
| ALW | 1.6 | (1.1…3.6) | 21 | MTV | 2.7 | (1.3…13.1) | 10 |
| **AMV** | 2.3 | (1.4…5.6) | 16 | QFV | 2.9 | (1.9…6.3) | 22 |
| ANI | 2.3 | (1.3…8.9) | 12 | QTV | 3.0 | (1.3…28.3) | 13 |
| ARV | 1.5 | (1.1…3.2) | 21 | RDS | 4.4 | (2.5…20.7) | 13 |
| **ASV** | 4.1 | (2.3…14.8) | 22 | RDV | 2.5 | (1.3…7.7) | 16 |
| AVV | 3.6 | (1.7…21.7) | 21 | RGS | 8.5 | (4.1c30.1) | 13 |
| CIL | 1.8 | (1.1…5.6) | 10 | RIS | 2.9 | (1.5…6.8) | 19 |

All entries of triplets have ten or more three-body atomic contacts. Triplets that have high propensity in both membrane and soluble proteins are highlighted in bold face.

from soluble proteins are listed in Table 2. Frequently observed high-propensity triplet types in TM helices are: AGL (1.4%), AGF (1.2%), AGG (1.0%), ALL (1.0%), GLL (0.8%), GLV (0.8%), AAG (0.8%), ALS (0.7%), AAL (0.7%), AGV (0.7%) and AGI (0.7%). Frequently observed high-propensity triplet types in helices from soluble proteins are: ALV (1.1%), ALL (0.9%), AAL (0.9%), AIL (0.8%), ILL (0.8%), LLV (0.7%), ALT (0.7%), AAV (0.6%), LMF (0.6%) and LLL (0.6%). Triplet types with high propensity observed in both membrane and soluble proteins are: AAA, AAF, AAG, AAI, AAL, AAM, AAS, AAV, AGI, AGL, AIT, ALL, ALM, AMV, ASV, GIL, GLM and GLV. These are highlighted in bold face in Tables 1 and 2. Overall, the amino acid composition of high-propensity triplets is different in membrane and soluble proteins (Figure 3(b)). Residue Gly occurs almost four times more and residue Ser occurs about 1.5 times more in different triplet types in membrane proteins than in soluble proteins. Residues Leu and Ile occur in about twice as many triplet types in soluble proteins than in membrane proteins. Ala is very versatile and participates in many high-propensity triplet types in both membrane and soluble proteins. His and Pro are found mainly in high-propensity triplet types formed in membrane proteins, while ionizable and polar residues Gln and Asn are more often found in high-propensity triplet types formed in soluble proteins.

There are triplet types with low propensity of MHIT values. For example, triplet LFW (0.4 with 95% confidence interval 0.2–0.8) is composed of three bulky hydrophobic amino acid residues, which are difficult to fit in the constrained environment of TM helices.

We find that, on average, the portion of amino acid residues that are involved in interhelical triplet and interhelical pairwise interactions is approximately the same in both TM and soluble helices. In these data sets, ~58% of residues in TM and ~60% of residues in soluble helices are found in triplet interactions and ~69% of residues in TM and ~67% of residues in soluble helices are found in pairwise interactions.

## Triplets with interhelical hydrogen bonds

In a previous study, we found that almost all TM helices have interhelical H-bonds. In addition, pairs of interacting helices are packed tighter when there is interhelical H-bond between them.[9] Analysis of triplets showed that 273 out of 846

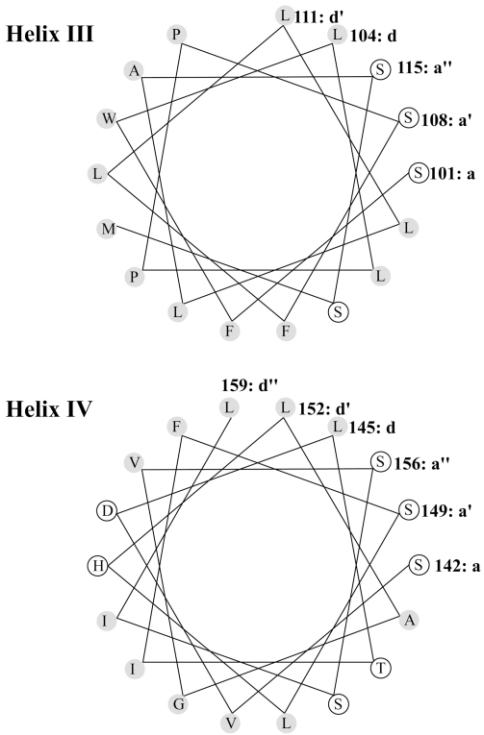


**Figure 4**. Helical wheels of helices III and IV from subunit I of bovine cytochrome *c* oxidase (1OCR). Leucine-serine zipper is formed by pairs of residues at positions "a" and "d": S101(a)-S156(a″) and L104(d)-L159(d″), S108(a′)-S149(a′) and L111(d′)-L152(d′), S115(a″)-S142(a) and L145(d). There are three Ser-Ser pairs and two Leu-Leu pairs. Helical wheels were generated using http://marqusee9.berkeley.edu/kael/helical.htm

(32%) observed triplet types contain at least one triplet with interhelical H-bond. Seventeen high-propensity triplet types have triplets with 1–4 interhelical H-bonds (AAS, AAW, AGS, AGT, AGY, AHT, ALS, AMS, ANG, ASV, GFS, GGS, GHS, GHT, GLW, GMS, GTW). However, the majority of interhelical H-bonds are found in low-count triplet types with uncertain confidence intervals for their MHIT values. These triplet types usually contain polar amino acid residues of low occurrences in TM helices. Frequently, all triplets of such triplet types contain H-bonds. For example, triplet types RQS and RDL each has two triplets, all contain interhelical H-bonds. Triplet types RNS and REV each has three triplets, and all are due to H-bond interactions.

The serine zipper is a spatial motif[9] studied previously, where three pairs of interhelical H-bonds are found between two tightly packed TM helices (e.g. TM helices III and IV from subunit I of bovine cytochrome *c* oxidase, 1OCR). In addition, the serine zipper is a part of a proposed channel that participates in proton transfer in bovine cytochrome *c* oxidase.[18] Triplet analysis revealed that each pair of H-bonded Ser residues (S101-S156, S108-S149 and S115-S142) is packed additionally with a Leu residue, forming an interacting LSS triplet. Altogether there are 13 LSS triplets in the 17 membrane protein structures, nine of which contain interhelical H-bonds and seven are part of the serine zipper spatial motif. Helical wheels of helices III and IV of subunit I of bovine cytochrome *c* oxidase show that if Ser residues are placed at "a" positions, then Leu residues are at "d" positions (Figure 4). Consequently, a mixed serine-leucine zipper interface is formed between the two interacting helices. The flat and polar "serine surface" is oriented towards the interior of subunit I, whereas the rougher and hydrophobic "leucine surface" is oriented towards the exterior.

Another previously studied spatial motif is the polar clamp.[9] It is formed by three amino acid residues on two different helices, with two interhelical H-bonds. In most cases, the side-chain of an amino acid residue capable of forming at least two hydrogen bonds (i.e. D, E, K, N, Q, R, S, T) is "clamped" twice by H-bonds, formed with either two side-chains, or a side-chain and a main-chain oxygen or nitrogen atom, or two main-chain oxygen (nitrogen) atoms from residues at positions $i$ and $i+1 \ldots i+4$. Therefore, the polar clamp requires three polar atoms to interact with each other. Frequently, polar clamps correspond to triplet interactions. They are found in 25 triplet types with a low number count of contacts: AEM, AEL, ARS, ART, DKF, DRY, DSW, EHL, ELS, NGS, NSS, NST, NTW, QEK, QST, RIV, RLM, RMV, RNS, RQS, RRD, RSK, RSS, SST and TWY. There are one to four instances of polar clamps in the 17 TM protein structures for each of the above triplet type. For eight rare triplet types, the only triplet is formed due to polar clamp interaction between interacting TM helices.

## Packing of residues of different size and chemical properties in the TM region

Different types of triplets can be grouped together by the size and the chemical properties of the side-chains of participating amino acid residues. We classify residues in triplets into small S residues (A, G, S and C), large aliphatic A residues (I, L, M, V), aromatic R residues (F, Y and W), polar P residues (R, K, H, Q, E, D, N, T) and proline O. Proline plays a special role in TM helices and is placed in its own group. Few packing preferences of proline-containing triplets are shown here due to insufficient sampling. Altogether we have 34 groups for all triplet types.

The majority of triplets in these groups are composed of amino acid residues residing on two adjacent helices, but there are triplets (2%–22%) in every group where all amino acid residues originate from three different helices. The majority of such cases is observed in triplet types composed of polar and aromatic residues: PPR (22%), PPP (14%), APR (13%) and RRR (12%). Triplet types containing two small residues have the smallest number of triplets formed by residues residing on three helices: SSS (2%), RSS (3%), PSS (4%) and ASS (4%).

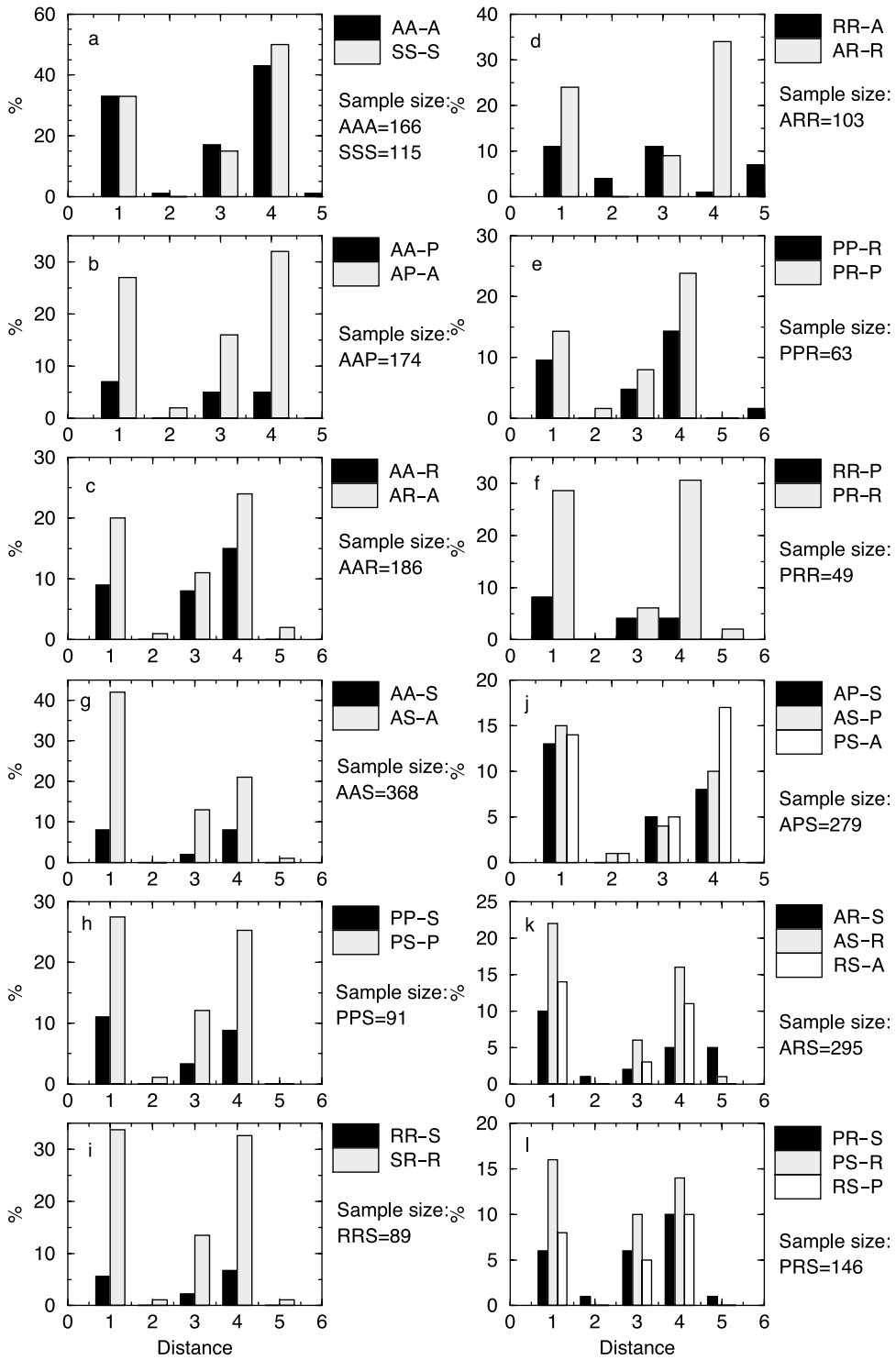


**Figure 5** (*legend opposite*)

For triplets formed by residues from two helices, the sequence distance between the two residues residing on the same helix can range from one to eight, with the highest preference for the positions $i$, $i+1$ and $i$, $i+4$ (Figure 5(a)–(r)). In the group composed of only small (SSS) or only aliphatic (AAA) residues, 45–50% of all triplets with two residues from the same helix have the two residues at a distance $i$, $i+4$ relative to each other and 33% of all triplets have the two residues at a distance $i$, $i+1$. Similar bias is observed in AP-A, AR-A, AR-R, PR-P and PR-R subgroups (Figure 5(b)–(f)) (here, a dash separates residues residing on different helices). Triplets of the same composition but

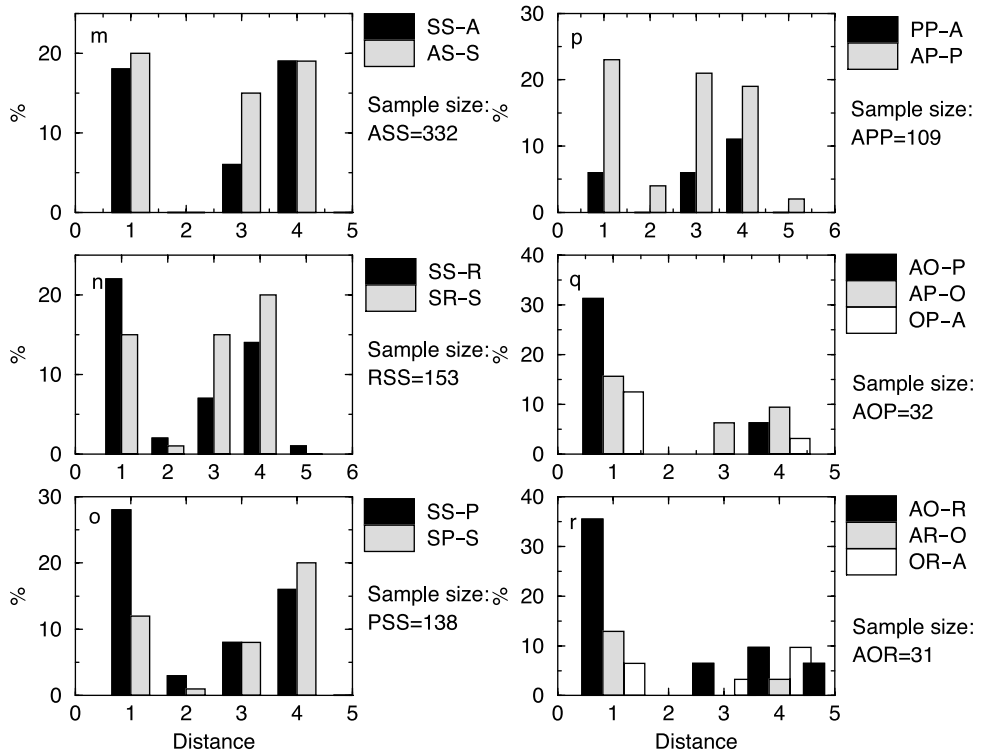


**Figure 5**. Frequencies of triplets with different sequence distance between two residues located on the same TM helix. Triplets are grouped by the size and chemical properties of side-chains of residues. The symbol – separates two residues located on one helix and a third residue located on another helix, e.g. SS–A represents a subgroup of triplets formed by two small residues residing on one helix, and an aliphatic residue on another helix. Symbols: S, small (A, G, S, C); A, aliphatic (I, L, M, V); R, aromatic (F, W, Y); P, polar (R, K, H, Q, E, D, N); O, proline.

with two amino acid residues of the same type on one helix (e.g. AA-R, PP-R, RR-A, RR-P, AA-P) are less populated in any of these groups (Figure 5(b)–(f), black bars). Triplets in these subgroups are composed of amino acid residues with larger side-chains. The introduction of one small amino acid residue shifts bias to triplets with $i$, $i+1$ intra-helical distances between residues. This is observed in subgroups AS-A, PS-P and SR-R (Figure 5(g)–(i)). Subgroups where a small residue is packed against two aliphatic (AA-S), two polar (PP-S) or two aromatic (RR-S) residues residing on the same helix have significantly smaller representation than those where small and other (A, P or R) residues originate from the same helix (Figure 5(g)–(i)). Similarly, in the groups APS, ARS and PRS (Figure 5(j)–(l), there is a preference for two amino acid residues with larger side-chains to pack on different helices and the subgroups AR-S, PR-S and AP-S have the lowest frequencies. This preference is very strong for some triplets. For triplet types GMW, STV and GHT, all bigger residues are on the opposite helices. For each of the triplet types GFF and ATV, only one triplet has both big residues on the same helix.

The packing preferences of two small and an aliphatic, aromatic or polar residue are rather similar (Figure 5(m)–(o)). There is significant preference in the SS-P type for the two small residues on the same helix to be located at $i$, $i+1$ positions. For some triplet types in SSA and SSR groups, there is a strong preference for two specific residues to be on the same helix, and the third residue on the other helix. These include AAI and AAW triplets, where the two Ala residues are more likely to be located on the same helix than on two helices (14 AA-I *versus* seven AI-A triplets, and five AA-W *versus* one AW-A triplets, respectively). For the AAL type, it is more likely that the Ala residues are located on different helices (14 AL-A *versus* nine AA-L triplets). Similarly, GGV triplets show a strong preference for Gly residues to be on different helices (nine GV-G *versus* one GG-V triplets).

Figure 5(q) and (r) show packing of proline residues with aliphatic and polar or aliphatic and aromatic residues. Although the sample size in both cases is rather small, there is a clear preference for proline and a larger alphatic residue at a distance $i$, $i+1$ to pack against polar or aromatic residue on the adjacent helix.

## Conformational library of cooperative three-body packing units

Triplets of a specific triplet type formed by the same three amino acid residues can have different conformations. There exist preferred cooperative spatial packing conformations for residues located on different helices in the TM region. The coordinates of 905 triplets from 73 different triplet types

**Table 3.** Well-defined conformations or tight clusters of TM triplet types with high propensity for interhelical interactions

| Triplet type | Cluster no. | Seq. mf.[a] | Average RMSD mean, (Å) | PDB | Amino acid residues in triplet[b] | TM helices[c] | Helix orientation and ω crossing angle (deg.) | $Dist_{C\alpha}$/ $Dist_{MIN}$ (Å)[d] |
|---|---|---|---|---|---|---|---|---|
| **AAA** | 1 | AA4 | 0.4 | 1L7V: | A121 + A125–A97 | TM3, TM4 | ↑ ↓ 138 | 4.1/3.8• |
| | | | | 1L7V: | A280 + A284–A68 | TM2, TM9 | ↑ ↓ 132 | **3.7/3.7** |
| **AAG** | 1 | AG3 | 0.4 | 1L7V: | A121 + G124–A97 | TM3, TM4 | ↑ ↓ 138 | **3.8/3.8** |
| | | | | 1L7V: | A280 + G283–A68 | TM2, TM9 | ↑ ↓ 132 | **3.7/3.7** |
| | 2 | AG4 | 0.6 | 1J4N: | A196 + G200–A107 | M4, M7 | ↑ ↑ −42 | 4.5/4.0• |
| | | | | 1J4N: | A80 + G84–A222 | M3, M8 | ↑ ↑ −33 | **5.0/5.0** |
| | 3 | AA4 | 0.4 | 1JB0: | A128 + A132–G55 | PsaL-I, PsaL-III | ↑ ↑ −33 | **3.9/3.9** |
| | | | | 1KPL: | A432 + A436–G263 | Q, J | ↑ ↑ −48 | **3.9/3.9** |
| | 4 | GA4 | 0.5 | 1DXR: | A215 + G211–A237 | E(L), D(M) | ↑ ↓ −127 | 5.0/4.1 |
| | | | | 1JB0: | A342 + G338–A389 | PsaB-e, PsaB-f | ↑ ↓ −162 | **3.9/3.9** |
| | | | | 1JB0: | A547 + G543–A578 | PsaB-h, PsaB-i | ↑ ↓ −162 | **4.9/4.9** |
| | 5 | GA1 | 0.5 | 1EHK: | G148 + A149–A120 | α3(I), α 4(I) | ↑ ↓ −159 | 4.5/4.0• |
| | | | | 1L7V: | G96 + A97–A125 | TM3, TM4 | ↑ ↓ 138 | 4.4/3.8 |
| | | | | 1L7V: | G67 + A68–A284 | TM2, TM9 | ↑ ↓ 132 | 4.3/3.7 |
| | 6 | GA1 | 0.3 | 1L7V: | G124 + A125–A97 | TM3, TM4 | ↑ ↓ 138 | **3.8/3.8** |
| | | | | 1L7V: | G283 + A284–A68 | TM2, TM9 | ↑ ↓ 132 | 4.3/3.7 |
| **AAI** | 1 | AA4 | 0.6 | 1L7V: | A310 + A314–I154 | TM5, TM10[e] | ↑ ↓ −144 | 6.2/4.7 |
| | | | | 1OCR: | A337 + A341–I257 | VI(I), IX(I) | ↑ ↓ −136 | 7.0/5.9 |
| | 2 | AI4 | 0.3 | 1L7V: | A151 + I155–A314 | TM5, TM10[e] | ↑ ↓ −144 | **4.7/4.7** |
| | | | | 1OCR: | A161 + I165–A192 | IV(I), V(I) | ↑ ↓ −150 | **4.4/4.4** |
| **AAL** | 1 | AL4 | 0.6 | 1JB0: | A169 + L173–A81 | PsaA-a, PsaA-b | ↑ ↓ −167 | **4.2/4.2** |
| | | | | 1JGJ: | A12 + L16–A48 | A, B | ↑ ↓ −164 | **4.3/4.3** |
| | 2 | AL4 | 0.6 | 1JB0: | A54 + L58–A145 | PsaB-a, PsaB-b | ↑ ↓ −164 | **4.1/4.1** |
| | | | | 1JB0: | A298 + L302–A213 | PsaA-c, PsaA-d | ↑ ↓ −161 | 4.9/3.8 |
| | | | | 1OCR: | A192 + L196–A161 | IV(I), V(I) | ↑ ↓ −150 | **4.4/4.4** |
| **AAM** | 1 | AM3 | 0.6 | 1JB0: | A169 + M172–A81 | PsaA-a, PsaA-b | ↑ ↓ −167 | **4.2/4.2** |
| | | | | 1JGJ: | A12 + M15–A48 | A, B | ↑ ↓ −164 | **4.3/4.3** |
| **AFS** | 1 | AF4 | 0.5 | 1DXR: | A38 + F34–S99 | A(L), B(L) | ↑ ↓ −157 | 5.0/4.7• |
| | | | | 1F88: | A41 + F37–S98 | I, II | ↑ ↓ −162 | 5.3/4.5 |
| **AGF** | 1 | AG4 | 0.6 | 1JB0: | A148 + G152–F50 | PsaB-a, PsaB-b | ↑ ↓ −164 | 5.0/4.1 |
| | | | | 1JB0: | A711 + G715–F669 | PsaB-j, PsaB-k | ↑ ↓ −154 | 4.7/3.9 |
| **AGG** | 1 | GG4 | 0.2 | 1FX8: | G243 + G247–A157 | M5, M8 | ↑ ↑ −42 | **3.7/3.7** |
| | | | | 1JB0: | G51 + G55–A128 | PsaL-I, PsaL-III | ↑ ↑ −33 | **3.9/3.9** |
| | | | | 1KPL: | G259 + G263–A432 | Q, J | ↑ ↑ −48 | **3.9/3.9** |
| | 2 | GA4 | 0.4 | 1FX8: | A53 + G49–G184 | M2, M6 | ↑ ↓ 150 | **3.9/3.9** |
| | | | | 1J4N: | A63 + G59–G175 | M2, M6 | ↑ ↓ 155 | **4.2/4.2** |
| | | | | 1L7V: | A71 + G67–G283 | TM2, TM9 | ↑ ↓ 132 | 4.6/3.7 |
| | 3 | GA4 | 0.4 | 1EHK: | G116 + A120–G148 | α3(I), α4(I) | ↑ ↓ −159 | **4.0/4.0** |
| | | | | 1FX8: | G184 + A188–G49 | M2, M6 | ↑ ↓ 150 | **3.9/3.9** |
| | | | | 1J4N: | G175 + A179–G59 | M2, M6 | ↑ ↓ 155 | **4.2/4.2** |
| | 4 | GA1 | 0.5 | 1KPL: | G181 + A182–G156 | F, G | ↑ ↓ 150 | 4.2/4.0• |
| | | | | 1L7V: | G67 + A68–G283 | TM2, TM9 | ↑ ↓ 132 | 3.7/4.3 |
| | | | | 1L7V: | G96 + A97–G124 | TM3, TM4 | ↑ ↓ 138 | **3.8/3.8** |
| **AGL** | 1 | GA1 | 0.5 | 1EZV: | G24 + A25–L65 | QCR9, RIP1 | ↑ ↑ −67 | 5.4/5.1• |
| | | | | 1JB0: | G388 + A389–L341 | PsaB-e, PsaB-f | ↑ ↓ −163 | 4.7/3.9 |
| | | | | 1JB0: | G137 + A138–L64 | PsaB-a, PsaB-b | ↑ ↓ −164 | 5.6/4.1 |
| | | | | 1KPL: | G140 + A141–L89 | E, C | ↑ ↓ −165 | 5.4/3.8 |
| | 2 | LG4 | 0.5 | 1EHK: | L395 + G399–A348 | α9(I), α10(I) | ↑ ↓ −162 | 5.1/3.8 |
| | | | | 1EHK: | L439 + G443–A379 | α10(I), α11(I) | ↑ ↓ −156 | 4.3/4.1• |
| | | | | 1JGJ: | L126 + G130–A111 | D, E | ↑ ↓ −157 | 5.2/3.8 |
| | 3 | AG4 | 0.5 | 1EHK: | A348 + G352–L395 | α9(I), α10(I) | ↑ ↓ −162 | 5.1/3.8 |
| | | | | 1EHK: | A31 + G35–L75 | α1(I), α2(I) | ↑ ↓ −148 | 5.2/4.2 |
| | | | | 1EHK: | A379 + G383–L439 | α10(I), α11(I) | ↑ ↓ −156 | 4.8/4.1 |
| | | | | 1JGJ: | A111 + G115–L126 | D, E | ↑ ↓ −157 | 5.2/3.8 |
| **AGV** | 1 | VA1 | 0.5 | 1EHK: | V119 + A120–G148 | α3(I), α4(I) | ↑ ↓ −159 | 4.5/4.0• |
| | | | | 1J4N: | V178 + A179–G59 | M2, M6 | ↑ ↓ 155 | **4.2/4.2** |
| **AHS** | 1 | SA4 | 0.3 | 1JB0: | A213 + S209–H301 | PsaA-c, PsaA-d | ↑ ↓ −157 | 5.4/3.8 |
| | | | | 1JB0: | A371 + S367–H397 | PsaA-e, PsaA-f | ↑ ↓ −162 | 5.4/3.6 |
| **AIM** | 1 | MI1 | 0.4 | 1C3W: | I117 + M118–A144 | D, E | ↑ ↓ −161 | 6.7/4.3 |
| | | | | 1EHK: | I434 + M435–A473 | α11(I), α12(I) | ↑ ↓ −156 | 5.0/4.3 |
| | | | | 1OCR: | I416 + M417–A464 | XI(I), XII(I) | ↑ ↓ −165 | 5.9/4.1 |
| **AIP** | 1 | AI4 | 0.3 | 1E12: | A40 + I44–P70 | A, B | ↑ ↓ −162 | **5.1/5.1** |
| | | | | 1OCR: | A276 + I280–P315 | VII(I), VIII(I) | ↑ ↓ −154 | **4.7/4.7** |
| **ALF** | 1 | FL1 | 0.6 | 1EHK: | F24 + L25–A87 | α1(I), α2(I) | ↑ ↓ −148 | 5.6/4.2 |
| | | | | 1JB0: | F57 + L58–A145 | PsaB-a, PsaB-b | ↑ ↓ −164 | 5.4/4.1 |
| **ALS** | 1 | AS1 | 0.5 | 1OCR: | A139 + S140–L169 | IV(III), V(III) | ↑ ↓ −161 | 7.8/4.4 |
| | | | | 1OCR: | A114 + S115–L145 | III(I), IV(I) | ↑ ↓ −160 | 7.0/4.3 |

(*continued*)

Table 3 Continued

| Triplet type | Cluster no. | Seq. mf.[a] | Average RMSD mean, (Å) | PDB | Amino acid residues in triplet[b] | TM helices[c] | Helix orientation and ω crossing angle (deg.) | $Dist_{C\alpha}$/ $Dist_{MIN}$ (Å)[d] |
|---|---|---|---|---|---|---|---|---|
| | 2 | SL3 | 0.5 | 1JB0: | S61 + L64–A138 | PsaB-a, PsaB-b | ↑ ↓ −164 | 4.3/4.1• |
| | | | | 1KPL: | S86 + L89–A141 | C, E | ↑ ↓ −165 | 3.9/3.8• |
| **GGF** | 1 | GF4 | 0.5 | 1EHK: | G148 + F152–G116 | α3(I), α4(I) | ↑ ↓ −159 | **4.0/4.0** |
| | | | | 1JB0: | G715 + F719–G666 | PsaB-j, PsaB-k | ↑ ↓ −154 | 4.1/3.9• |
| **GGL** | 1 | GL4 | 0.4 | 1JB0: | G666 + L670–G715 | PsaB-j, PsaB-k | ↑ ↓ −154 | 4.1/3.9• |
| | | | | 1OCR: | G16 + L20–G77 | I(I), II(I) | ↑ ↓ −150 | **4.3/4.3** |
| **GGV** | 1 | GV3 | 0.3 | 1EHK: | G116 + V119–G148 | α3(I), α4(I) | ↑ ↓ −159 | **4.0/4.0** |
| | | | | 1J4N: | G175 + V178–G59 | M2, M6 | ↑ ↓ 155 | 4.4/4.2• |
| | | | | 1JB0: | G659 + V662–G722 | PsaB-j, PsaB-k | ↑ ↓ −154 | 4.1/3.9• |
| **GHT** | 1 | TG1 | 0.4 | 1EZV: | T46 + G47–H82 | A, B | ↑ ↓ −124 | 6.8/5.6 |
| | | | | 1OCR: | T354 + G355–H376 | IX(I), X(I) | ↑ ↓ −152 | 5.9/4.2 |
| | | | | 1QLA: | T33 + G34–H93 | I, II | ↑ ↓ −136 | **5.4/5.4** |
| | | | | 1QLA: | T132 + G133–H182 | IV, V | ↑ ↓ −128 | 5.9/5.0 |
| **GLF** | 1 | FG4 | 0.5 | 1EHK: | F24 + G28–L83 | α1(I), α2(I) | ↑ ↓ −152 | 5.4/4.2 |
| | | | | 1JB0: | F308 + G312–L201 | PsaA-c, PsaA-d | ↑ ↓ −161 | 4.7/3.8 |
| **GLL** | 1 | LG1 | 0.6 | 1JB0: | L432 + G433–L531 | PsaB-g, PsaB-h | ↑ ↓ −151 | **5.5/5.5** |
| | | | | 1QLA: | L139 + G140–L175 | IV, V | ↑ ↓ −128 | 6.2/5.0 |
| | 2 | GL1 | 0.5 | 1EHK: | G352 + L353–L392 | α9(I), α10(I) | ↑ ↓ −152 | **3.8/3.8** |
| | | | | 1OCR: | G30 + L31–L43 | I(III), II(III) | ↑ ↓ −158 | **3.7/3.7** |
| | 3 | LG1 | 0.5 | 1EHK: | L351 + G352–L395 | α9(I), α10(I) | ↑ ↓ −152 | 6.0/3.8 |
| | | | | 1EHK: | L27 + G28–L83 | α1(I), α2(I) | ↑ ↓ −152 | 7.3/4.2 |
| | | | | 1JB0: | L344 + G345–L385 | PsaB-e, PsaB-f | ↑ ↓ −163 | 7.1/3.9 |
| **HHV** | 1 | VH3 | 0.4 | 1OCR: | H290 + V287–H240 | VI(I), VII(I) | ↑ ↓ 112 | 7.4/5.5 |
| | | | | 1EHK: | H282 + V279–H233 | α6(I), α7(I) | ↑ ↓ 111 | 7.2/4.8 |
| IIL | 1 | LI1 | 0.3 | 1JGJ: | I83 + L82–I43 | B, C | ↑ ↓ −169 | 6.1/5.0 |
| | | | | 1OCR: | I216 + L215–I167 | V(III), VI(III) | ↑ ↓ −159 | 7.0/6.3 |
| **LLS** | 1 | LL1 | 0.6 | 1JB0: | L141 + L142–S61 | PsaB-a, PsaB-b | ↑ ↓ −164 | 5.5/4.1 |
| | | | | 1OCR: | L111 + L112–S149 | III(I), IV(I) | ↑ ↓ −160 | 5.3/4.3 |
| **LSV** | 1 | VS1 | 0.4 | 1JB0: | S61 + V60–L141 | PsaB-a, PsaB-b | ↑ ↓ −164 | 5.5/4.1 |
| | | | | 1OCR: | S156 + V155–L104 | III(I), IV(I) | ↑ ↓ −160 | 5.9/4.3 |

[a] Sequence motif or intrahelical pair as defined by Senes *et al.*[7]
[b] Residues originating from the same helix are listed first in the alphabetical order with a plus (+) sign between them followed by the residue on the adjacent helix.
[c] The numbering of helices is taken from the original X-ray papers (for references, see Table 5).
[d] Minimal distance between interhelical $C^{\alpha}$ atoms in triplet *versus* minimal global interhelical $C^{\alpha}$–$C^{\alpha}$ distance. Cases where minimal interhelical $C^{\alpha}$–$C^{\alpha}$ distance in triplet is identical with a global minimal interhelical $C^{\alpha}$–$C^{\alpha}$ distance are in bold face. The bullet (•) denotes the cases where the difference between these two distances is less than or equal to 0.5 Å.
[e] TM5 and TM10 are from different transmembrane subunits of vitamin $B_{12}$ transporter.

with high propensity values and with estimated confidence intervals were used to build a library of conformations of triplets. Pairwise RMSD are first calculated for all triplets of the same triplet type. Distance-based hierarchical clustering is then applied to group them into three-body packing units that share structural similarity. Table 3 lists 40 packing units of a total of 97 triplets from 25 triplet types. Triplet structures within each packing unit listed in Table 3 have RMSD $\leq 0.6$ Å to the mean coordinates. These triplets originate either from different proteins, or from non-homologous helices of the same protein. We refer such packing units as "tight clusters". Tight clusters are found mainly in triplet types that are sampled frequently (e.g. AGF triplet type (31 triplets), AGL (44 triplets)). It is less frequent to find tight clusters in triplet types that do not have adequate sampling of triplets. These include: CGF (six triplets), GFS (nine), and ATV (15).

Triplets in tight clusters tend to come from helical pairs where helices have similar parallel or antiparallel orientation, similar right-handed or left-handed ω packing angles, and have the same sequential distance between the two amino acid residues located on the same helix. Notable exceptions are three tight clusters in Table 3 (AGG, cluster 2, AGL, cluster 1 and GGV, cluster 1). The majority of triplets in tight clusters (66) come from 33 anti-parallel left-handed helical pairs with ω crossing angles in the −136° to −167° range. The bias of helix–helix interactions towards this particular type of packing was described earlier by Bowie[20] and by Senes *et al.*[19] The triplets in these clusters are mainly of AB1-C (27 triplets) and AB4-C (33 triplets) types, where residues A and B reside on the same helix, the number shows the intrahelical distance between amino acids, and residue C resides on the interacting helix. Only four triplets (AB1-C, 3 triplets and AB4-C 1 triplet) are found in three left-handed anti-parallel helical pairs with ω crossing angle in the −120° to −130° range. Notably, one of the latter triplets represents a tight intersubunit interaction between chains L and M in photosynthetic reaction center from *Rhodopseudomonas viridis* (1DXR, cluster 4 in AAG

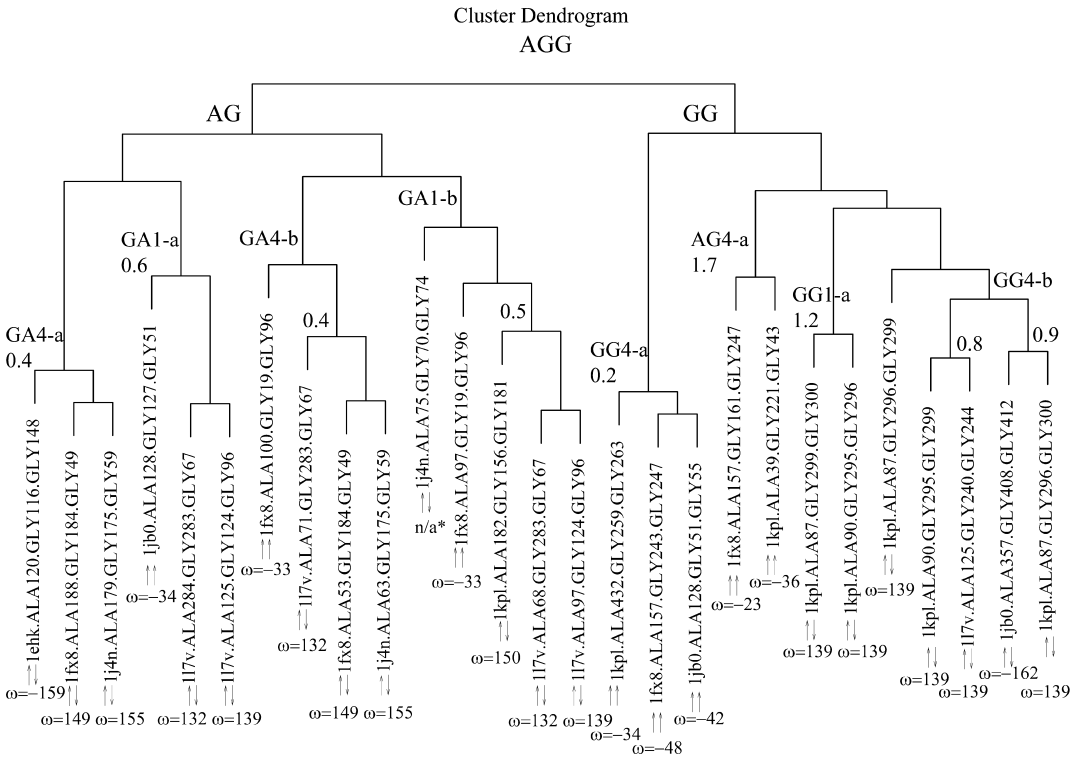


**Figure 6**. Cluster dendrogram for AGG triplets. The ↑ ↑ and ↑ ↓ arrows denote orientation of helices in the helical pairs. Average RMSD (Å) to the mean coordinates for the structures in a cluster is placed above each cluster, along with the name of the cluster.

triplet type) featuring an over-represented GA4 sequence motif. The remaining triplets in this group are from helical pairs containing residues interacting with heme molecules in fumarate reductase (1QLA, GHT triplet type, helices IV and V) and in cytochrome *bc*1 complex (1EZV, GHT triplet type, helices A and B). Right-handed antiparallel helical pairs (seven in total) with ω crossing angle in the 110°–155° range are represented almost equally by AB1-C (eight triplets), AB3-C (five) and AB4-C (seven) types. There are six right-handed parallel helical pairs with ω crossing angle in the −33°– −67° range. Triplets in tight clusters from these helical pairs have higher preference for AB4-C type of interactions (seven triplets) rather than for AB1-C type of interaction (one triplet). No triplets from tight clusters originated from parallel left-handed helices.

Small residues promote the formation of conformationally well-defined tight spatial clusters: there is at least one small residue (Ala, Gly or Ser) in 23 out of 25 triplet types listed in Table 3. There are 50 triplets listed in Table 3 where one of two interhelical $C^{\alpha}$–$C^{\alpha}$ distances between amino acid residues coincides with the global minimal $C^{\alpha}$–$C^{\alpha}$ interhelical distance of the helical pair (marked in bold face) or differs from it for not more than 0.5 Å (marked by bullets). Both of these distances are listed in the last column of Table 3. Triplets containing at least two small residues frequently (42 out of 50 triplets) correspond to the regions of the closest contact between helices. In these cases, the amino acid residues originating from the same helix are often separated by two or three residues, forming the following sequence motifs: AA4, AG3, AG4, AI4, AL4, AM3, GA1, GA4, GG4, GF4, GV3 and VA1. Among them, intrahelical pairs AG4, GA4 and GG4 are over-represented in TM helices.[7] The triplets corresponding to the regions of the closest contacts between two helices often incorporate high-propensity interhelical pairs such as A-A (propensity 1.3), A-G (1.1), A-M (1.7), A-F (1.1), G-G (3.0) and A-P (2.1). These triplet data point to the important roles that residues Ala and Gly play in helix–helix interactions in polytopic membrane proteins.

## Spatial and sequence motifs: clustering of AGG, AGL and GHT triplets

The work of Senes *et al*. provides an important resource of sequence patterns that are important for studying TM helix assembly.[7] The different preference of packing mode of residues with different size discussed earlier indicates that spatial arrangement of residues is related to sequence motifs. In this section, we further explore the relationship between sequence and spatial motifs, as well as the role of interhelical H-bond using the examples of tight clusters from three triplet types with high propensity for interhelical interaction (AGG, AGL, and GHT).

Figure 6 shows the hierarchical clustering dendrogram of 27 triplets of the AGG triplet type.

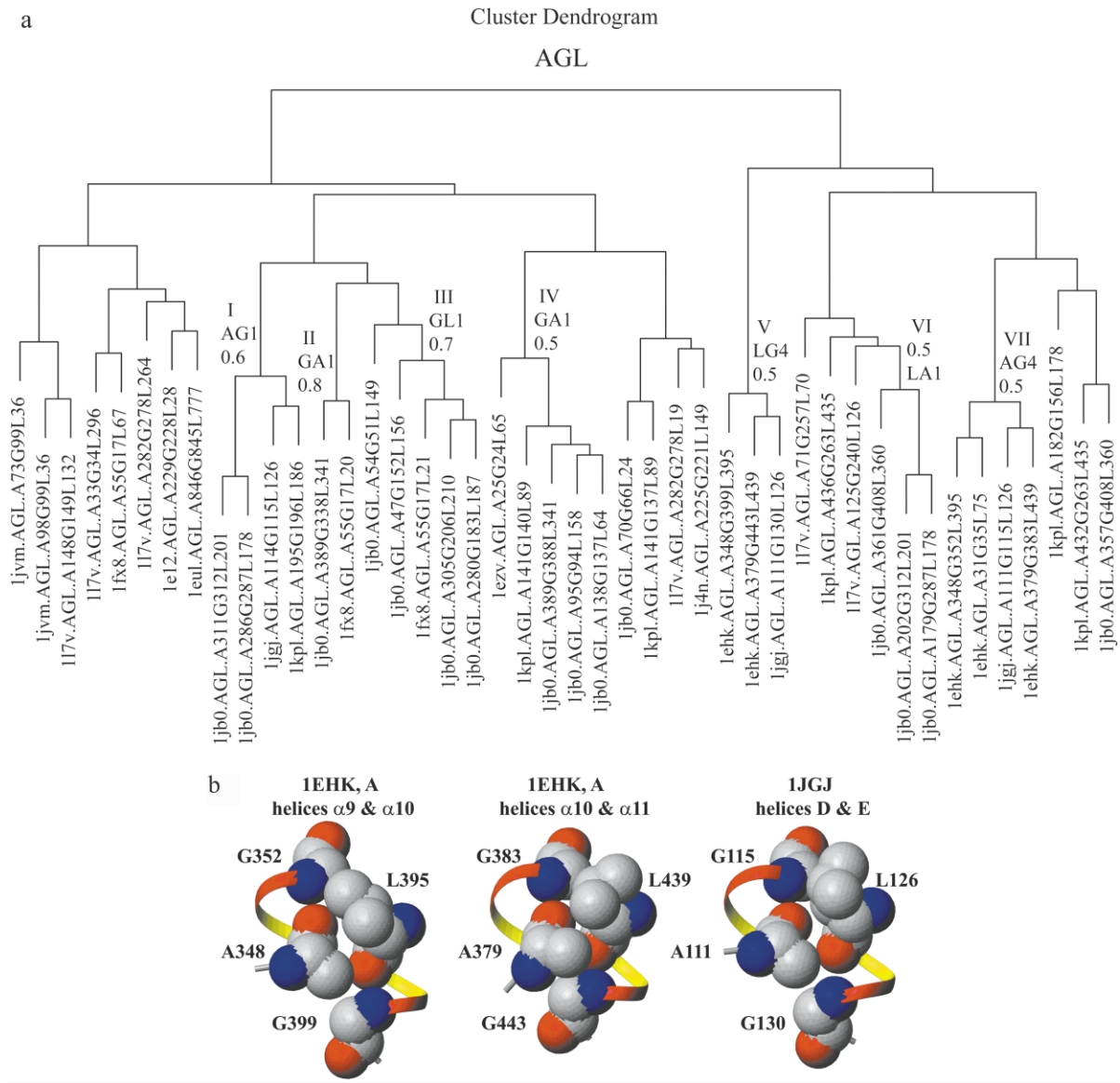


**Figure 7**. (a) Cluster dendrogram for AGL triplets. In this dendrogram, all clusters with RMSD to mean $<1.0$ Å are marked in roman numerals together with the corresponding sequential motifs for the residues residing on the same helix following Senes *et al*.[7] nomenclature. (b) Three helical pairs containing two AGL triplets from clusters V and VII on (a). AGL triplets from cluster V: 1EHK, chain A, helices α9 and α10 (L395–G399 + A348), 1EHK, chain A, helices α10 and α11 (L439–G443 + A379), 1JGJ, helices D and E (L126–G130 + A111). AGL triplets from cluster VII: 1EHK, chain A, helices α9 and α10 (A348–G352 + L395), 1EHK, chain A, helices α10 and α11 (A379–G383 + L439), 1JGJ, helices D and E (A111–G115 + L126). The dash denotes the residues on the same helix. Each helical fragment is seven residues long.

This triplet type is formed by small residues with little side-chain degrees of freedom. The conformational space is therefore determined by the relative orientations of residues. The dendrogram clearly separates the triplets into two major clusters, which we call AG cluster (15 triplets) and GG cluster (12 triplets). Triplets in the former all have a glycine and an alanine residue on the same helix, all triplets in the latter (with two exceptions) have two glycine residues on one helix. There are four smaller subclusters in the GG-cluster, GG4-a, GG4-b, GG1-a and AG4-b. The helical pairs that contain triplets from these subclusters have different orientation of helices and different ω crossing angles. All triplets in tight cluster GG4-a (RMSD to mean $<0.2$ Å) are formed by residues from parallel helices with right-handed crossing angle ω around $-40°$. All triplets in cluster GG4-b are formed by residues from antiparallel helices but with different handedness. The AG branch can be divided tentatively into four smaller clusters as well. A useful observation is that tight AGG clusters contain the sequence motifs GG4 and GA4,[7] which are among the most significantly over-represented pairs in the TM helices.[7]

Our second example is the hierarchical clustering of triplets in AGL triplet type (Figure 7(a)), which contains five tight clusters with RMSD to

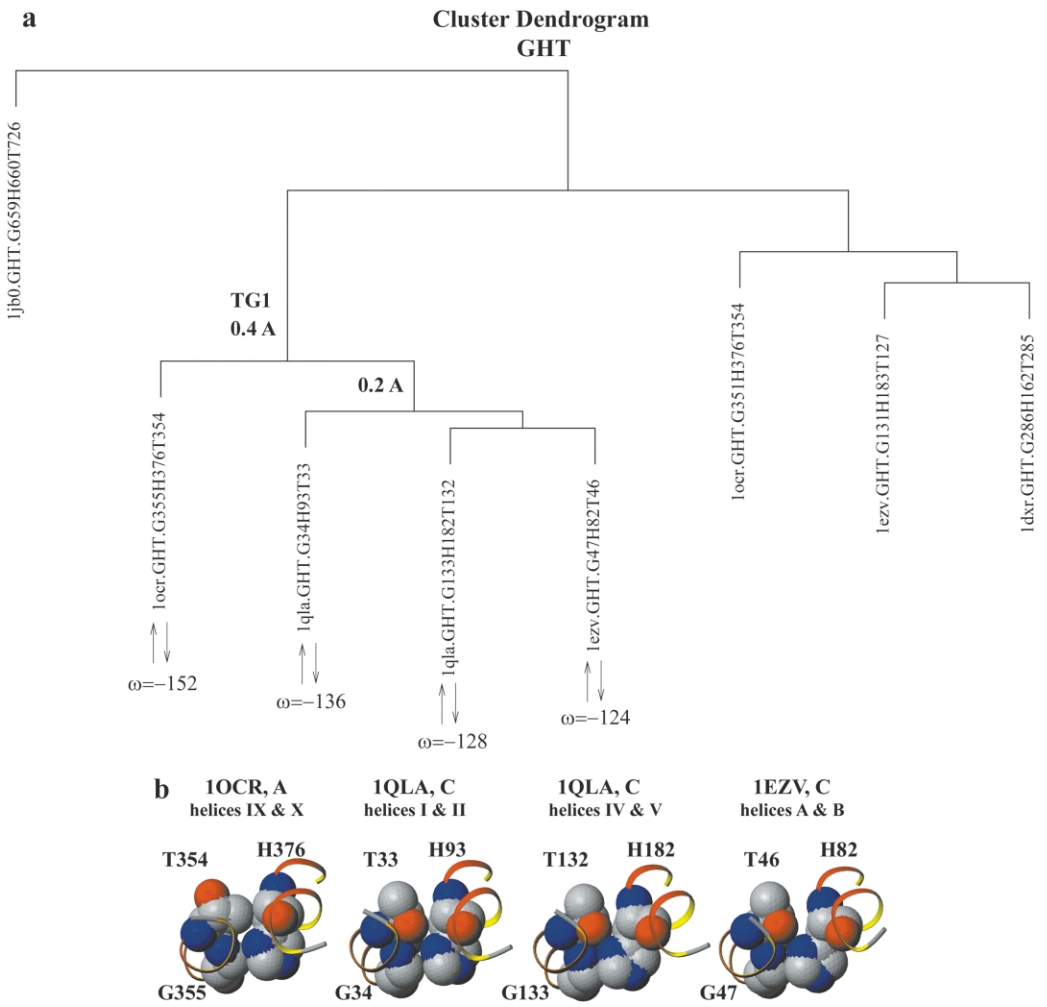


**Figure 8**. (a) Cluster dendrogram of GHT triplets. There is one tight cluster that contains consecutive Thr and Gly residues on the same helix and His on another helix. In this cluster, all helices are antiparallel with ω crossing angles in the range of −152° to −124°. The average RMSD to mean is 0.4 Å for all triplets. (b) H-bond stabilizes the conformations of GHT triplets in two triplets from two non-homologous helical pairs from fumarate reductase (1QLA) and one triplet from cytochrome *bc*1 complex (1EZV). Amino acid residues that form the triplet are shown in space-filling symbols. Interhelical H-bond forms between OG1 of Thr and ND1 of His.

mean <0.6 Å. More than two-thirds (13 out of 19) of all AGL triplets in tight clusters have residues Ala and Gly on the same helix (clusters I, IV and VII). The majority (nine) of which are consecutive residues AG (cluster III) or GA (cluster I) in primary sequence. These form two tight clusters with RMSD to mean 0.6 Å and 0.5 Å, respectively. Cluster VII contains four triplets with an AxxxG or AG4 sequence motif. Here, all of the helical pairs are antiparallel, left-handed, with the ω crossing angles of helical segments in the range of −154° to −157°. The Ala and Leu residues from three AGL triplets in cluster VII are part of the AGL triplets in cluster V, with the only difference that in cluster V Gly residue comes from the helix containing Leu. This results in a spatial four residue interacting motif involving two sequential pairs of residues: AG4 on one helix and LG4 on the other helix (Figure 7(b)). The structures of the seven residue fragments from helices containing two pairs of AGL triplets superimpose well with RMSD to mean 0.6 Å for all $C^{\alpha}$ atoms.

Our third example is the GHT triplet type. Four out of eight triplets form a tight cluster with RMSD to mean 0.4 Å (Figure 8(a)). Three triplets form a subcluster of structures with an H-bond. They form a bundle of almost identical structures with RMSD to the mean 0.2 Å (Figure 8(a)). Although residues His and Thr have rather large side-chains and high degrees of freedom in comparison with residues Ala or Gly, this particular triplet conformation is highly populated. The presence of an interhelical H-bond between OG1 of Thr residue and ND1 of His residue (Figure 8(b)), as well as the specific interaction of His residues with heme molecules play important roles. The H-bond fixes the positions of side-chains of residues Thr and His, and orients the imidazole ring to interact with the heme molecule. The sequential pair TG appears to be well conserved in the alignment of sequences of these helices. On the basis of these observations, we propose that the assignment of the –OH (OG1) group and the CG2 atom of Thr354 should be exchanged in the fourth GHT

**Table 4.** Conserved triplets in the family of archaeal rhodopsins (ARF)

| Triplet | protein | Positions I | Positions II | Positions III | RMSD (Å) | Helices | Frequencies (%) *f*(I) | Frequencies (%) *f*(II) | Frequencies (%) *f*(III) |
|---|---|---|---|---|---|---|---|---|---|
| A. | | | | | | | | | |
| 1 | bR: | L97 | L152 | T178 | 0.2 | C-E-F | L:80 | L:100 | V:48 |
| | hR: | L123 | L179 | T203 | | | V:12 | | T:44 |
| | pR: | L87 | L141 | T167 | | | I:8 | | I:8 |
| 2 | bR: | Y185 | L211 | D212 | 0.1 | F-G | Y:100 | L:80 | D:100 |
| | hR: | Y210 | L237 | D238 | | | | I:12 | |
| | pR: | Y174 | L200 | D201 | | | | M:8 | |
| 3 | bR: | L174 | F219 | I222 | 0.5 | F-G | L:100 | F:84 | I:44 |
| | hR: | L199 | F245 | I248 | | | | Y:16 | L:28 |
| | pR: | L163 | F208 | I211 | | | | | F:20 |
| | | | | | | | | | V:8 |
| B. | | | | | | | | | |
| 4 | bR: | W86 | T90 | I119 | 0.2 | C-D | W:100 | T:100 | I:64 |
| | hR: | W112 | T116 | C145 | | | | | C:32 |
| | pR: | W76 | T80 | M109 | | | | | M:4 |
| 5 | bR: | T90 | L94 | W182 | 0.2 | C-F | T:100 | L:72 | W:100 |
| | hR: | T116 | L120 | W207 | | | | V:28 | |
| | pR: | T80 | V84 | W171 | | | | | |
| 6 | bR: | T90 | P91 | D115 | 0.4 | C-D | T:100 | P:100 | D:92 |
| | hR: | T116 | P117 | D141 | | | | | N:4 |
| | pR: | T80 | P81 | N105 | | | | | Q:4 |
| 7 | bR: | T90 | D115 | W182 | 0.3 | C-D-F | T:100 | D:92 | W:100 |
| | hR: | T116 | D141 | W207 | | | | N:4 | |
| | pR: | T80 | N105 | W171 | | | | Q:4 | |
| 8 | bR: | L100 | T170 | F171 | 0.2 | C-F | L:84 | T:32 | F:92 |
| | hR: | L126 | I195 | F196 | | | A:8 | L:32 | Y:8 |
| | pR: | L90 | L159 | Y160 | | | V:4 | I:28 | |
| | | | | | | | N:4 | K:4 | |
| | | | | | | | | M:4 | |
| 9 | bR: | L94 | D115 | I148 | 0.4 | C-D-E | L:72% | D:92 | V:52 |
| | hR: | L120 | D141 | V175 | | | V:28% | N:4 | L:44 |
| | pR: | V84 | N105 | L137 | | | | Q:4 | I:4 |
| 10 | bR: | L94 | D115 | W182 | 0.3 | C-D-F | L:72 | D:92 | W:100 |
| | hR: | L120 | D141 | W207 | | | V:28 | N:4 | |
| | pR: | V84 | N105 | W171 | | | | Q:4 | |
| 11 | bR: | G125 | A126 | W189 | 0.2 | D-F | G:72 | A:68 | W:100 |
| | hR: | A151 | A152 | W214 | | | A:28 | T:16 | |
| | pR: | G115 | A116 | W178 | | | | V:16 | |
| 12 | bR: | A184 | Y185 | L211 | 0.2 | F-G | A:36 | Y:100 | L:80 |
| | hR: | G209 | Y210 | L237 | | | G:32 | | I:12 |
| | pR: | I173 | Y174 | L200 | | | V:20 | | M:8 |
| | | | | | | | L:8 | | |
| | | | | | | | I:4 | | |
| 13 | bR: | L174 | F219 | L223 | 0.5 | F-G | L:100 | F:84 | L:72 |
| | hR: | L199 | F245 | L249 | | | | Y:16 | F:16 |
| | pR: | L163 | F208 | A212 | | | | | A:12 |

triplet conformation from bovine cytochrome *c* oxidase (1OCR) (Figure 8(b)). This triplet would then have an interhelical H-bond, similar to the other members of this tight cluster.

## Cooperative four residue spatial motifs

The example of overlapping AGL clusters V and VII indicates that over-represented sequence motifs can be part of the three-body interactions that have similar conformations, which in turn can be a part of a cooperative four residue spatial motif. The examination of Table 3 revealed five more such four residue spatial motifs that contained both over-represented and regular sequence pairs. The first four residue motif occurs in the vitamin B12 transporter when triplets AAA (cluster 1) and AAG (cluster 1) share residues A97 and A121 for the TM3–TM4 interacting helical pair, and residues A68 and A280 for the TM2–TM9 helical pair. This interaction falls into the region of the closest interhelical contact between these two antiparallel right-handed helices and can be represented as an AG3A1–A motif, following the notation used by Senes *et al.*[7] The second four residue motif can be represented as AA4-GG4. It is formed at the region of the closest interhelical contact but between the parallel right-handed helices PsaL-I–PsaL-III of photosystem I (1JB0) and helices Q and J of ClC chloride channel (1KPL). They form the common subset of residues shared by AAG triplets from cluster 3 and AGG triplets from cluster 1 (Table 3). The third four residue motif is formed by the residues from AAL (cluster 1) and AAM (cluster 1) triplets from photosystem I (helices PsaA-a and

PsaA-b) and from sensory rhodopsin II (helices A and B). This interaction can be represented as AM3L1-A. Again, this four residue motif falls into the region of the closest approach for the left-handed antiparallel helical pair. The next motif is formed between antiparallel left-handed helices PsaA-a and PsaA-b of photosystem I (AGL and ALS clusters, S61 + L64–G137 + A138) and helices C and E of ClC chloride channel (AGL and ALS clusters, S86 + L89–G140 + A141). Although the last four residue motif (AGG, cluster 3 and AGV, cluster 1, G116 + V119 + A120–G148 in cytochrome *c* oxidase (1EHK) and G175 + B178 + A179–G59 in aquaporin (1J4N)) falls into the region of the closest interhelical contact, it is formed between antiparallel helices that have different handedness: the helical pair from cytochrome *c* oxidase is left-handed ($\omega = -159°$), while the helical pair from aquaporin is right-handed ($\omega = 155°$).

## Conserved three-body interactions in archaeal rhodopsin family (ARF)

Amino acid residues in protein families are often conserved for biological function, for maintaining stability, and for kinetic folding accessibility.[21] Triplets are clusters of tightly packed amino acid residues. Are they more likely to be conserved? We explore this issue by examining the protein family of archaeal rhodopsins (AR), where three high-quality structures from two different organisms are available. There are currently four members in the bacterial retinal protein family: bacteriorhodopsin (bR), halorhodopsin (hR), sensory rhodopsin I (sR) and sensory rhodopsin II (pR). Their functional roles are proton pumping, $Cl^-$ transport, and phototactic behavior, respectively. The sequence identity between any two of the four proteins range from 20% to 35%. Sixteen amino acid residues are fully conserved among all sequences, ten of which are located in the retinal binding pocket. Ihara *et al*.[22] studied the evolutionary relationship between 25 archaeal retinal proteins of 13 strains from five genera of halophilic archaea. They concluded that all four functionally differentiated proteins were probably derived from a single ancestral retinal protein by three gene duplication events.

We compare the packing of amino acid residues from triplets that are common in bR (1C3W, from *Halobacterium salinarium*, 131 triplet), hR (1E12, from *H. salinarium*, 120 triplets) and pR (1JGJ, *Natronobacterium pharaonis*, 107 triplets). Residues in these triplets are from aligned positions in the multiple sequence alignment presented by Ihara *et al*.[22] Table 4(A) lists three triplets that are conserved both in sequence and in structure. The average RMSD to the mean of these triplets does not exceed 0.5 Å when all atoms are superimposed. Conservative substitutions do occur at some positions. The frequencies of these substitutions are also listed in Table 4(A).

Table 4(B) extends this list to include triplets that are similar structurally but are not formed by identical residues. The majority of the 13 triplets listed in Table 4 come from helices C, D, E, and F, which surround the retinal molecule. No triplet is composed of only amino acid residues that are 100% identical throughout the 25 proteins sequences in the AR family. Five triplets contain two residues that are fully conserved, and six triplets contain one fully conserved amino acid residue. However, substitutions in these triplets are often isosteric. For example, residue Phe is likely to be replaced by residue Tyr (triplets 3, 8 and 13), both of which are aromatic. Residue Gly is often replaced by residue Ala, both of which have small size (triplet 11).

For triplet 1 listed in Table 4(A), the β-branchness of the amino acid residue seems to be important (Figure 9). Here, position III always contains a β-branched residue (I, T or V) in all the aligned sequences. Position II is occupied by residue L exclusively in all three triplets and is fully conserved in all 25 sequences. This conserved leucine packs with an L-V pair in bR and with an L-T pair in hR at positions I–III. Both L-V and L-T pairs are isosteric, with the only difference in the OG1 atom from Thr that forms an intrahelical H-bond with another fully conserved residue L at position (*i*-4). Position I in sensory rhodopsins I (sR) is also occupied by a β-branched residue (I or V). Residues at position I and III in sR sequences are interchangeable, but they always form an I–V pair. The requirement of β-branchness at position III, a bulky Leu at position II, and another residue of comparable size (L, V, I) at position I suggests that

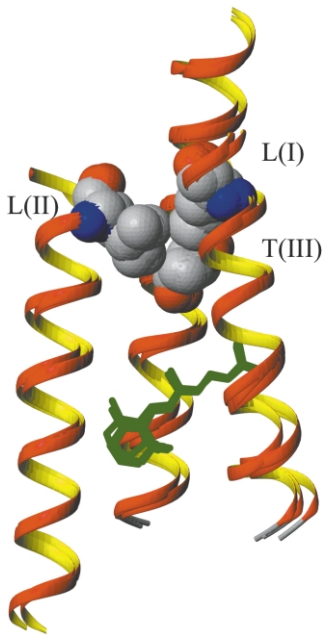


**Figure 9**. Superposition of helices C, E and F from bacteriorhodopsin, halorhodopsin and sensory rhodopsin II. Residues that form triplet 1 listed in Table 4(A) are shown in space-filling symbols. The bound retinal molecule is shown in green.

the tight packing of side-chains may be important for maintaining the correct assembly of these three helices, and this tight packing can be achieved with several different choices of residues at positions I and III.

There are cases where a wide variety of amino acid substitutions is observed at a single position. It often occurs in triplet elements that are part of the "knob-into-hole" packing motif. The residue with the most number of allowed substitutions in the AR family is usually facing the lipid environment and forms part of a "hole", e.g. R1 in triplet 12, R3 in triplets 3, 4 and 13, and R2 in triplets 8 and 11. The residue corresponding to "the knob" is frequently well conserved.

## Discussion

Membrane proteins are packed tightly in the environment of lipid bilayer and packing interactions are thought to play important roles in membrane protein folding.[16,23] Pairwise interhelical propensity provides rich information about the types of interactions and pairing preferences of amino acid residues in TM helices,[6,8] but it is intrinsically incapable of taking additional spatial context into consideration. In this work, we develop a novel approach to systematically study higher-order three-body interactions of amino acid residues in transmembrane helices. We decompose tightly packed regions of protein structure into interacting triplets of amino acid residues. We show with the example of glycophorin A that each triplet can be considered as a minimalistic element of packing. This method is applied to the set of 17 membrane proteins to identify the frequent and statistically significant tight three-body interactions in polytopic proteins. Although each three-body interaction of amino acid residues has a much smaller number of occurrences compared with pairwise interactions, the application of the bootstrap method allows the evaluation of the confidence intervals of estimated propensity for three-body interactions. This helps to guard against erroneous propensity values and interpretations.

The comparison of amino acid compositions of triplet types with a high propensity for interhelical interactions from membrane and soluble helices showed a preference of triplets from membrane proteins for Ala and Gly residues. These findings are in agreement with the recent experimental and computational studies, where small residues are shown to be important for oligomerization of monotopic TM helices.[24–25] These residues are often found at the places of tight helix–helix interactions.[6] A persistent feature of high-propensity triplets from the TM region (Table 1) is that they are composed mostly of a mixture of small and large residues. High-propensity triplets are rarely composed entirely of larger and branched residues (e.g. IIM). High-propensity triplets containing two residues with large aliphatic or aromatic side-chains show strong preference for these residues to be at the opposite positions on neighboring helices. This arrangement optimizes the side-chain packing and is a consequence of the restrictions imposed by the organization of membrane proteins as helical bundles.

An unexpected compositional feature of high-propensity triplet types in membrane proteins is the high frequency of Met residue (Figure 3(b)). Although the overall occurrence of this residue in membrane helices is only ~5%, which is much smaller than that of Leu (~15%),[6] Met is found with a large variety of residue pairs in high-propensity triplet types. Statistical analysis showed that there are only two (IM9 and PM4) significantly over-represented intrahelical sequence pairs of residues involving Met,[7] suggesting that this residue is distributed across TM helices rather randomly. On the other hand, several frequent high-propensity interhelical pairs with Met have been detected in TM helices: A-M (1.7), I-M (1.1), F-M (1.4), M-S (1.9).[8] Met has a large, flexible side-chain and can sample many different conformations in helical structure, depending on the local context. Consequently, Met may be favored for its flexible space-filling properties with a wide variety of pairs of amino acid residues to achieve tighter packing that enhances van der Waals interactions.

Although residues forming triplet in one protein may be all conserved in another aligned sequence, they may not always form a tight-packing triplet. For example, amino acid residues in triplet L97-L174-F219 in bR have identical counterparts in the other two proteins of the ARF family (L123-L199-F245 in hR, L87-L163-F208 in pR). The alignment of 25 proteins shows that Leu97 can be substituted conservatively with valine. Leu174 is fully conserved, and Phe219 is substituted to tyrosine in sensory rhodopsins I. However, spatial comparison (Figure 10) shows that although structurally conserved in bR and hR (with average RMSD to mean 0.5 Å), this triplet does not exist in pR, and the three residues do not pack tightly. The residues preceding the first Leu residue in this triplet are highly variable: Asp96 in bR (which is known to be functionally important), Ala122 in hR, and Tyr/Phe 86 in pR. The average RMSD to the mean for all atoms from the three amino acid residues increases to 1.0 Å when pR is included. One spatial feature preserved in all three structures is the interaction between side-chains of leucine and phenylalanine, which stack on top of each other. This type of packing for Leu and Phe or Leu and Tyr (e.g. L211-Y185 in bR) residues is often observed in membrane proteins. In fact, the L-F pair is the most abundant interhelical pairwise interaction in TM regions of membrane proteins.[8] This example illustrates that the spatial arrangement of higher-order packing interactions is still context-dependent. Residues above or below the triplet influence the conformation of a triplet.

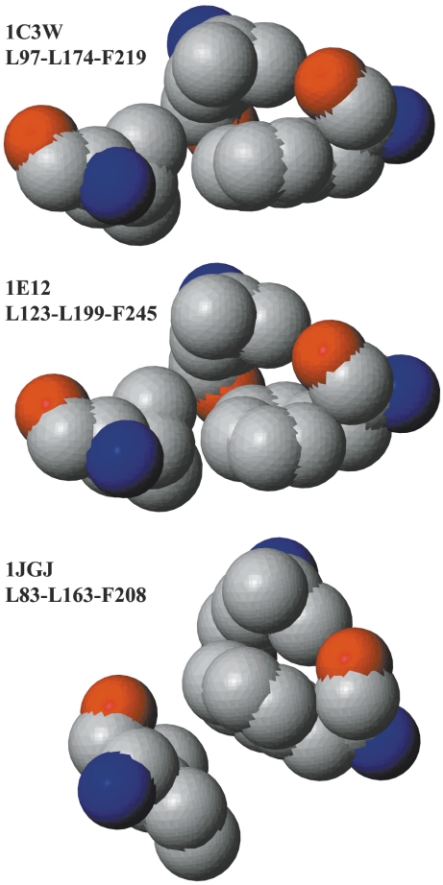


**Figure 10**. The spatial arrangement of three residues that are identical in sequence alignment of bR (1C3W), hR (1E12) and pR (1JGJ). These residues are packed tightly in bR and hR. They form triplets of very similar structure that can be superimposed with RMSD to mean 0.5 Å. The same three residues do not form a triplet in pR. The formation of triplet interactions depends on the context of additional residue(s).

We explore the issue of evolutionary conservation of triplet interactions by comparing structurally conserved triplets in three protein structures from the ARF family of archaea. The overall structures of bacteriorhodopsin, halorhodopsin and sensory rhodopsin II are very similar: the superposition of $C^{\alpha}$ atoms from helices C–G yields an RMSD to the mean around 0.8 Å for each pair of structures.[26] However, we found that there are very few common triplets between these proteins and there is no triplet that is composed of three amino acid residues that are 100% conserved. Similar packing of the TM helices in proteins from the ARF family is achieved by different combinations of amino acid residues. We conclude that the space-filling necessary to maintain the orientation of helices may be achieved in many different ways.

The organization of membrane proteins as helical bundles affects the size of residues that pack with each other, and limits the number of possible conformations of amino acid residues in triplets. The clustering of triplets of high-propensity triplet types indicates that there are regions in the conformational space that are strongly preferred by three-body packing. These triplet conformations are sampled more frequently than the other triplet conformations. We refer to the bundles of triplet structures with similar conformations as "tight clusters". Triplets in tight clusters often correspond to the regions of the closest contact between helices. As a general observation, the corresponding helical pairs have very similar geometry of helix–helix crossing. We found triplets forming tight clusters between both parallel and antiparallel helices (see Table 3). The majority of triplets in tight clusters are from anti-parallel left-handed and right-handed helical pairs. There are only six parallel helical pairs and all of them are right-handed with ω crossing angles in the −33° to −67° range. We did not find tight clusters of triplets from left-handed parallel coiled-coils with ω crossing angles about 20° in this data set.

In addition, we showed that some sequence motifs that are significantly over-represented in TM helices (GG4, AG4, GA4) have strong correlations with high-propensity triplets as well as with high-propensity interhelical pairs of residues (A-A, A-G, A-M, A-F and others). Interhelical H-bonds are among the important factors that stabilize such conformations. The results shown here indicate that it is possible to extract and identify these preferred conformations and sequence motifs, and to link them to the global structural parameters such as helix–helix crossing angle and helix–helix orientation, despite a limited sampling due to the small data set of available membrane protein structures.

An important advantage of three-body potentials over the pairwise potentials is that triplets contain additional information about interacting neighboring residues, which is not easy to obtain with other methods. For example, the triplet analysis showed that Ser-Ser pairs from the serine zipper motif[9] are packed predominantly with Leu residues forming a mixed serine-leucine packing interface between two interacting TM helices. In addition, triplet analysis also reveals "preferred functional packing", i.e. the preferred residues that provide necessary packing interactions for functional residues. For example, our analysis identified GHT as a high-propensity triplet type, as exemplified by four triplets of very similar conformation. All amino acid residues in this conformation are highly conserved. Discussion in the literature is usually limited to the role of the His residue, which interacts with a heme molecule, but there may be additional residues and interactions (i.e. interhelical H-bond) that play important roles in determining the correct orientation of imidazole ring.

## Summary

We have developed a novel computational approach to study higher-order three-body packing

**Table 5.** The set of 17 membrane proteins used in this study

| PDB ID | Protein name (organism) | Resolution (Å) | Reference |
|---|---|---|---|
| 1C3W | Bacteriorhodopsin (*H. salinarum*) | 1.6 | Luecke *et al.*[5] |
| 1DXR | Photosynthetic reaction center (*Rh. viridis*) | 2.0 | Lancaster *et al.*[37] |
| 1E12 | Halorhodopsin (*H. salinarum*) | 1.8 | Kolbe *et al.*[38] |
| 1EHK | Cytochrome *c* oxidase (*T. thermophilus*) | 2.4 | Soulimane *et al.*[39] |
| 1EUL | $Ca^{2+}$-transporting ATPase (*O. cuniculus*) | 2.6 | Toyoshima *et al.*[40] |
| 1EZV | Cytochrome *bc*1 complex (*S. cerevisiae*) | 2.3 | Hunte *et al.*[41] |
| 1F88 | Rhodopsin (*B. taurus*) | 2.8 | Palczewski *et al.*[42] |
| 1FUM | Fumarate reductase flavoprotein subunit (*E. coli*) | 3.3 | Iverson *et al.*[43] |
| 1FX8 | Glycerol conducting channel (*E. coli*) | 2.2 | Fu *et al.*[44] |
| 1J4N | Aquaporin 1 (*B. taurus*) | 2.2 | Sui *et al.*[45] |
| 1JB0 | Photosystem I (*S. elongatus*) | 2.5 | Jordan *et al.*[46] |
| 1JGJ | Sensory rhodopsin II (*N. pharaonis*) | 2.4 | Luecke *et al.*[47] |
| 1JVM | Kcsa potassium channel (*S. Lividans*) | 2.8 | Morais-Cabral *et al*[48] |
| 1KPL | Clc chloride channel (*S. typhimurium*) | 3.0 | Dutzler *et al.*[49] |
| 1L7V | Vitamin B12 transporter (*E. coli*) | 3.2 | Locher *et al.*[50] |
| 1OCR | Cytochrome *c* oxidase (*B. taurus*) | 2.4 | Yoshikawa *et al.*[51] |
| 1QLA | Fumarate reductase flavoprotein (*W. succinogenes*) | 2.2 | Lankaster *et al.*[52] |

interactions in protein structures. This approach provides additional information about the neighboring context, packing and conformational preferences of three tightly packed amino acid residues in membrane proteins. The results shown here indicate that the preferred conformations and sequence patterns can be linked to global structural parameters such as helix–helix crossing angle and helix–helix orientation. We expect that the utility of this approach will increase as the number of available X-ray structures of TM proteins increase.

## Materials and Methods

### Membrane and soluble protein data

The 17 membrane proteins used in this study are listed in Table 5. All loops in the soluble regions are removed manually, leaving only the α helices in the TM regions. As a result, each protein is represented by a bundle of TM helices. Determining the exact boundaries of the TM regions is a difficult task even when structures are available.[7] Javadpour *et al.*[27] assigned the TM regions on the basis of the positions of basic and acidic residues. Senes *et al.*[7] used a 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. Altogether, there are 192 unique helices in the data set. Here, we analyze only interhelical three-body interactions formed by atoms from three different amino acid residues residing on at least two TM helices.

A set of soluble α-helical proteins was constructed for comparison with the membrane proteins. It consists of 31 structures obtained from diffractions (pdb names: 1A0B, 1A17, 1AUE, 1B3U, 1CUN, 1DKX, 1DOW, 1E2A, 1EVS, 1EZ3, 1EZF, 1FEW, 1FIO, 1GNW, 1GTO, 1GUX, 1HE1, 1LE4, 1MTY, 1PBW, 1QGH, 1QGR, 1QJB, 1QKR, 1QSA, 1QSD, 1QU7, 1QUU, 1VLT, 256B, 2MHR). These proteins all have 50% or more α-helical content and have negligible amount of β-strands. After manually removing the connecting loops, there are a total of 288 unique helices in the data set.

### Computation of three body interhelical contacts

Using the alpha shape application program interface kindly provided by Professor Edelsbrunner and colleagues, a program INTERFACE-3 has been implemented to compute interhelical atomic triplets. INTERFACE-3 uses precomputed Delaunay triangulation and alpha shape. The Delaunay triangulation of membrane proteins is computed using the DELCX program,[28,29] and the alpha shape is computed using the MKALF program.[28,30] Both can be downloaded from the website of NCSA†. The advantage of using INTERFACE-3 compared to methods using distance cut-off is that only nearest-neighbor atoms in physical contacts are counted.[8] The van der Waals radii of protein atoms are taken from Tsai *et al.*[31] To account for uncertainty in the precision of atomic coordinates, the van der Waals radii are incremented by 0.5 Å, following Singh & Thornton.[32] Larger increments (e.g. 1 Å) introduce spurious triplets of amino acid residues that are not packed tightly.

### Probabilistic model for membrane helical interface triplet (MHIT) propensity

To evaluate three-body MHIT propensity $P(i,j,k)$ of residue type $i$, type $j$, and type $k$, we first estimate the observed probability $q(i,j,k)$ of interhelical atomic triplets involving residue types $i$, $j$, and $k$. We have:

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

Here, $a(i,j,k)$ is the number count of triple interhelical atomic contacts between residue types $i$, $j$, and $k$, and $\sum_{i',j',k'} a(i',j',k')$ is the number of all triple interhelical atomic contacts. The observed probability $q(i,j,k)$ is then compared against the random probability $p(i,j,k)$ that a triplet of contacting atoms is picked from a residue of type $i$, a residue of type $j$, and a residue of type $k$, respectively, when chosen randomly and independently

† http://www.ncsa.uiuc.edu

from the same set of interacting residues in the TM regions. The formula for $p(i,j,k)$ depends on how many of the three residues are drawn from the same residue type. When all three residues are of the same type (i.e. $i = j = k$), we have:

$$p(i,j,k) = N_i(N_i - 1)(N_i - 2)\frac{n_i n_i n_i}{n(n - n_i)(n - 2n_i)}$$

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. When exactly two of the three interacting residues are of the same type (e.g. $i = j \neq k$), we have:

$$p(i,j,k) = N_i(N_i - 1)N_k\left(\frac{n_i n_i n_k}{n(n - n_i)(n - 2n_i)} + \frac{n_i n_i n_k}{n(n - n_i)(n - n_i - n_k)} + \frac{n_i n_i n_k}{n(n - n_k)(n - n_k - n_i)}\right)$$

When all three interacting residues are of different types (i.e. $i \neq j \neq k$), we have:

$$=p(i,j,k) = N_i N_j N_k\left(\frac{n_i n_j n_k}{n(n - n_i)(n - n_i - n_j)} + \frac{n_i n_j n_k}{n(n - n_i)(n - n_i - n_k)} + \frac{n_i n_j n_k}{n(n - n_j)(n - n_j - n_i)} + \frac{n_i n_j n_k}{n(n - n_j)(n - n_j - n_k)} + \frac{n_i n_j n_k}{n(n - n_k)(n - n_k - n_i)} + \frac{n_i n_j n_k}{n(n - n_k)(n - n_k - n_j)}\right)$$

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

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

### Estimating confidence intervals of propensity values

Because the sample size of 17 membrane proteins is small, statistical modeling with approximations is prone to errors. Here, we apply bootstrap techniques to estimate the confidence intervals of the estimated propensity values from simulated data sets.[13,14] Let the true value of the MHIT propensity value of a triplet be θ. Our estimator $T$ takes the value $t$, which is the estimated value for θ. Our goal is to calculate a 95% confidence interval for θ. If we sample independently $R$ times from the 17 proteins with replication, we have a simulated data set of $Y_1^*,\ldots,Y_R^*$, each contains 17 structures. Some structures in the original set appear multiple times, some appear once, and some never appear. We estimate the propensity value for the triplet from each of the $R$ samples, and obtain $t_1^*,\ldots,t_R^*$. For an equitailed 95% confidence interval (95% = 1−2α, α = 2.5%), we have the basic bootstrap confidence intervals:

$$(t^*_{(R+1)(1-\alpha)},\ t^*_{(R+1)\alpha})$$

In our calculation, $R$ is chosen to be 30,000. The accuracy of these limits depends on $R$, and how well the distribution $T^* - t$ agrees with that of $T - \theta$. Perfect agreement occurs only when the distribution of $T - \theta$ does not depend on any unknown variables.

To reduce possible errors due to unknown variables, we use Studentized bootstrap. For the $r$ the bootstrapped sample, we calculate:

$$z_r^* = \frac{t_r^* - t}{v_r^{*1/2}}$$

To obtain a value for $v_r^*$ when calculating $z_r^*$, we bootstrap with replacement again $M$ times the $r$th sample of the original bootstrap. We have:

$$v_r^* = \frac{1}{M-1}\sum_{m=1}^{M}(t_m^* - \bar{t}^*)^2$$

where $t_1^*,\ldots,t_M^*$ are calculated from the second bootstrap sampling for $M = 100$. We use the $(R+1)$ αth order statistic of the simulated values $z_1^*,\ldots,z_R^*$, or $z^*_{((R+1)\alpha)}$ to estimate $z_\alpha$. The Studentized bootstrap confidence interval for θ has limits:

$$(t - v^{1/2}z^*_{(R+1)(1-\alpha)},\ \ t - v^{1/2}z^*_{(R+1)\alpha})$$

Since $M$ bootstrap samples from the $r$th sample are needed to obtain $v_r^*$, the required total number of bootstrap samples is: $R \times M = 30{,}000 \times 100 = 3{,}000{,}000$.

### RMSD, hydrogen bond and ω crossing angle calculations

To compare the spatial arrangement of amino acid residues in triplets, the root-mean-square distance (RMSD) was calculated between each pair of triplets of the same amino acid composition after implementing the method of Umeyama,[33] which calculates the least-squares estimation of transformation parameters through singular value decomposition.[33] We first identify individual structures for each occurrence of a given triplet type in the protein data set from all atomic coordinates of three amino acid residues forming the triplet. For example, there are 24 GLF triplets across the 17 membrane proteins and the structure of each triplet is defined by 46 atomic coordinates. The results of pairwise RMSD calculations for a triplet type were used to create a matrix of distances, which was processed by Gnu R, a statisical software for hierarchical clustering with complete linkage.

H-bonds are identified by HBPLUS program[34] using default parameters and allowing exchange of the nearly symmetrical side-chains of residues H, Q and N, since nitrogen, oxygen and carbon atoms are indistinguishable in electron density maps. Potential H-bonds that would be formed if histidine CD2 was actually ND1, CE1 was NE2 and the oxygen and nitrogen atoms in asparagine and glutamine residues were the other way around, were counted. We use the PROMOTIF suite of programs[35] to calculate ω crossing angles between interacting helices.

## Acknowledgements

We thank Drs William DeGrado and Vikas Nanda for valuable discussions. We thank two anonymous referees for insightful and valuable suggestions. We thank Patrick Freeman for design and technical support of data base of triplet

structures. We thank all structural biologists for depositing the coordinates of membrane proteins in the Protein Data Bank. This work is supported by National Science Foundation (DBI-0078270 and CAREER DBI-0133856) and American Chemical Society (Petroleum Research Fund, 35616-G7).

## References

1. Jayasinghe, S., Hristova, K. & White, S. H. (2001). Energetics, stability and prediction of transmembrane helices. *J. Mol. Biol.* **312**, 927–934.
2. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.
3. Lemmon, M. A., Flanagan, J. M., Hunt, J. F., Adair, B. D., Bormann, B. J., Dempsey, C. E. & Engelman, D. M. (1992). Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. *J. Biol. Chem.* **267**, 7683–7689.
4. Fleming, K. G., Ackerman, A. L. & Engelman, D. M. (1997). The effect of point mutations on the free energy of transmembrane alpha-helix dimerization. *J. Mol. Biol.* **272**, 266–275.
5. 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.
6. Eilers, M., Patel, A. B., Liu, W. & Smith, S. O. (2002). Comparison of helix interaction in membrane and soluble α-bundle proteins. *Biophys. J.* **82**, 2720–2736.
7. 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.
8. Adamian, L. & Liang, J. (2001). Helix–helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.* **311**, 891–907.
9. Adamian, L. & Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins*, **47**, 209–218.
10. Vendruscolo, M. & Domany, E. (1998). Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.* **109**, 11101–11108.
11. Tobi, D., Shafran, G., Linial, N. & Elber, R. (2000). On the design and analysis of protein folding potentials. *Proteins: Struct. Funct. Genet.* **40**, 71–85.
12. Rossi, A., Micheletti, C., Seno, F. & Maritan, A. (2001). A self-consistent knowledge-based approach to protein design. *Biophys. J.* **80**, 480–490.
13. Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, **57**, Chapman and Hall/CRC, Boca Raton, FL.
14. Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics, **1**, Cambridge University Press, London.
15. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. (2001). Residue frequences and pairing preferences at protein–protein interfaces. *Proteins: Struct. Funct. Genet.* **43**, 89–102.
16. MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science*, **276**, 131–133.
17. Fleming, K. G. & Engelman, D. M. (2001). Specificity in transmembrane helix–helix interactions can define a hierarchy of stability for sequence variants. *Proc. Natl Acad. Sci. USA*, **98**, 14340–14344.
18. Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. *Science*, **272**, 1136–1144.
19. Senes, A., Ubarretxena-Belandia, I. & Engelman, D. M. (2001). The Cα-H···O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc. Natl Acad. Sci. USA*, **98**, 9056–9061.
20. Bowie, J. U. (1997). Helix packing in membrane proteins. *J. Mol. Biol.* **272**, 780–789.
21. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123–129.
22. Ihara, K., Umemura, T., Katagiri, I., Kitajima-Ihara, T., Sugiyama, Y., Kimura, Y. & Mukohata, Y. (1999). Evolution of the archaeal rhodopsins: evolution rate changes by gene duplication and functional differentiation. *J. Mol. Biol.* **285**, 163–174.
23. White, S. H. & Wimley, W. C. (1999). Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 319–365.
24. Dawson, J. P., Weiner, J. S. & Engelman, D. M. (2002). Motifs of serine and threonine can drive association of transmembrane helices. *J. Mol. Biol.* **316**, 799–805.
25. Russ, W. P. & Engelman, D. M. (2000). The GxxxG motif: a framework for transmembrane helix–helix association. *J. Mol. Biol.* **296**, 911–919.
26. Royant, A., Nollert, P., Edman, K., Neutze, R., Landau, E. M., Pebay-Peyroula, E. & Navarro, J. (2001). X-Ray structure of sensory rhodopsin II at 2.1 Å resolution. *Proc. Natl Acad. Sci. USA*, **98**, 10131–10136.
27. 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.
28. Edelsbrunner, H. & Mucke, E. P. (1994). Three-dimensional alpha-shapes. *ACM Trans. Graph.* **13**, 43–72.
29. Edelsbrunner, H. & Shah, N. R. (1996). Incremental topological flipping works for regular triangulations. *Algorithmica*, **15**, 223–241.
30. Facello, M. A. (1995). Implementation of a randomized algorithm for Delaunay and regular triangulation in three dimensions. *Comput. Aided Geom. Des.* **12**, 349–370.
31. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **290**, 253–266.
32. Singh, J. & Thornton, J. M. (1992). *Atlas of Protein Side-chain Interactions*, vols 1 and 2, IRL press, Oxford.
33. Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 376–380.
34. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.

35. Hutchinson, E. G. & Thornton, J. M. (1996). A program to identify and analyze structural motifs in proteins. *Protein Sci.* **5**, 212–220.
36. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.
37. Lancaster, C. R. D., Bibikova, M. V., Sabatino, P., Oesterhelt, D. & Michel, H. (2000). Structural basis of the drastically increased initial electron transfer rate in the reaction center from a *Rhodopseudomonas viridis* mutant described at 2.0 Å resolution. *J. Biol. Chem.* **275**, 39364–39368.
38. 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.
39. Soulimane, T., Buse, G., Bourenkov, G. P., Bartunik, H. D., Hubert, R. & Than, M. E. (2000). Structure and mechanism of the aberrant ba(3)-cytochrome *c* oxidase from *Thermus thermophilus*. *EMBO J.* **19**, 1766–1776.
40. Toyoshima, C., Nakasako, M. & Nomura, O. H. (2000). Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature*, **405**, 647–655.
41. Hunte, C., Koepke, J., Lange, C., Rossmanith, T. & Michel, H. (2000). Structure at 2.3 Å resolution of the cytochrome *bc*(1) complex from the yeast *Saccharomyces cerevisiae* co-crystallized with an antibody Fv fragment. *Structure*, **8**, 669–684.
42. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A. *et al.* (2000). Crystal structure of rhodopsin: a G protein-coupled receptor. *Science*, **289**, 739–745.
43. 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.
44. Fu, D., Libson, A., Miercke, L. J. W., Weitzman, C., Nollert, P., Kucinski, J. & Stroud, R. M. (2000). Structure of a glycerol-conducting channel and the basis for its selectivity. *Science*, **290**, 481–486.
45. Sui, H., Han, B. G., Lee, J. K., Walian, P. & Jap, B. K. (2001). Structural basis of water-specific transport through the AQP1 water channel. *Nature*, **414**, 872–878.
46. Jordan, P., Fromme, P., Witt, H. T., Klukas, O. & Saenger, W. (2001). Three-dimensional structure of cyanobacterial photosystem I at 2.5 Å resolution. *Nature*, **411**, 909–917.
47. Luecke, H., Schobert, B., Lanyi, J. K. & Spudich, E. N. (2001). Crystal structure of sensory rhodopsin II at 2.4 angstoms: insights into color tuning and transducer interaction. *Science*, **293**, 1499–1503.
48. Morais-Cabral, J. H., Zhou, Y. & MacKinnon, R. (2001). Energetic optimization of ion conduction rate by the K + selectivity filter. *Nature*, **414**, 37–42.
49. Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. & MacKinnon, R. (2002). X-Ray structure of a ClC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature*, **415**, 287–294.
50. Locher, K. P., Lee, A. T. & Rees, D. C. (2002). The *E. coli* BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science*, **296**, 1091–1098.
51. Yoshikawa, S., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., Yamashita, E., Inoue, N. *et al.* (1998). Redox-coupled crystal structural changes in bovine heart cytochrome *c* oxidase. *Science*, **280**, 1712–1713.
52. 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.

***Edited by G. von Heijne***

*(Received 22 August 2002; received in revised form 26 November 2002; accepted 20 December 2002)*