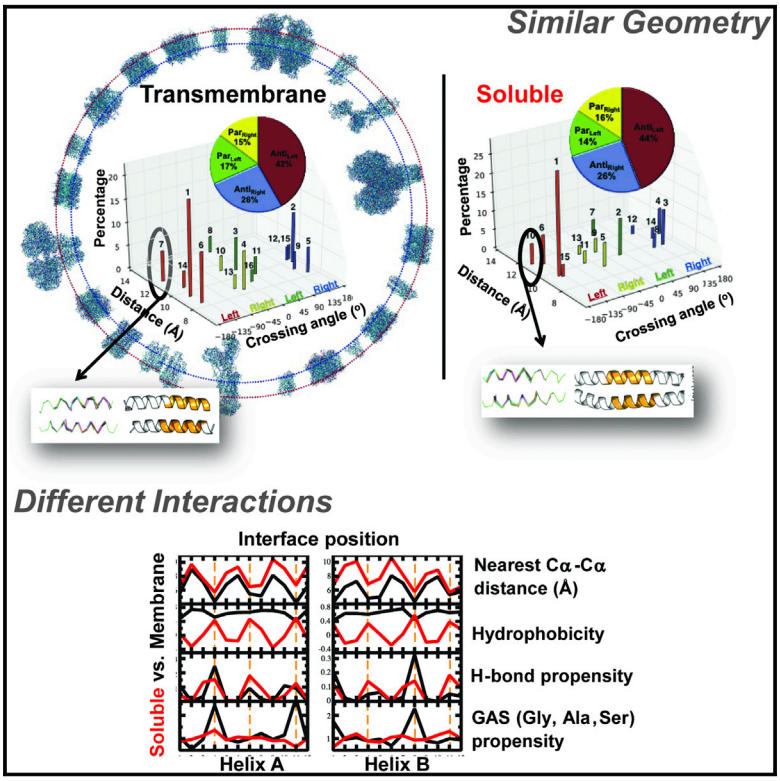
Article

Structure

# The Membrane- and Soluble-Protein Helix-Helix Interactome: Similar Geometry via Different Interactions

## Graphical Abstract




## Authors

Shao-Qing Zhang, Daniel W. Kulp, ..., Ilan Samish, William F. DeGrado

## Correspondence

william.degrado@ucsf.edu (W.F.D.), ilan.samish@weizmann.ac.il (I.S.)

## In Brief

Zhang et al. compare helix-helix geometries of soluble and membrane proteins and observe few common helix-helix geometries. But soluble dimer interfaces are enriched by hydrophobic residues, whereas the transmembrane class engages interfacial small residues and interhelical hydrogen bonds. The results should aid protein engineering, prediction, and design.

## Highlights

- Helix-helix interactions display just a few geometric clusters in all proteins
- Helix-helix interactome interfaces in soluble proteins pack via a hydrophobic core
- Small residues and H-bonds are engaged in helix-helix packing in membrane proteins
- The helix-helix interactome description should aid protein analysis and design

Zhang et al., 2015, Structure *23*, 527–541
March 3, 2015 
http://dx.doi.org/10.1016/j.str.2015.01.009

Structure
Article

# The Membrane- and Soluble-Protein Helix-Helix Interactome: Similar Geometry via Different Interactions

Shao-Qing Zhang,[1,2,10] Daniel W. Kulp,[3,6,10] Chaim A. Schramm,[3,7,10] Marco Mravic,[5] Ilan Samish,[4,8,9,*] and William F. DeGrado[2,*]
[1]Department of Physics and Astronomy, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, San Francisco, CA 94158, USA
[3]Graduate Group in Biochemistry and Molecular Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[5]Graduate Program in Biophysics, University of California San Francisco, San Francisco, CA 94158, USA
[6]Present address: Department of Immunology and Microbial Science, Scripps Research Institute, La Jolla, CA 92037, USA
[7]Present address: Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA
[8]Present address: Department of Plant Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel
[9]Present address: Department of Biotechnology Engineering, Braude College of Engineering, Karmiel 2161002, Israel
[10]Co-first author
*Correspondence: william.degrado@ucsf.edu (W.F.D.), ilan.samish@weizmann.ac.il (I.S.)
http://dx.doi.org/10.1016/j.str.2015.01.009

## SUMMARY

**α Helices are a basic unit of protein secondary structure and therefore the interaction between helices is crucial to understanding tertiary and higher-order folds. Comparing subtle variations in the structural and sequence motifs between membrane and soluble proteins sheds light on the different constraints faced by each environment and elucidates the complex puzzle of membrane protein folding. Here, we demonstrate that membrane and water-soluble helix pairs share a small number of similar folds with various interhelical distances. The composition of the residues that pack at the interface between corresponding motifs shows that hydrophobic residues tend to be more enriched in the water-soluble class of structures and small residues in the transmembrane class. The latter group facilitates packing via sidechain- and backbone-mediated hydrogen bonds within the low-dielectric membrane milieu. The helix-helix interactome space, with its associated sequence preferences and accompanying hydrogen-bonding patterns, should be useful for engineering, prediction, and design of protein structure.**

## INTRODUCTION

The α helix is by far the most common regular secondary structure element. In water-soluble proteins approximately 35% of all protein residues are in the α-helical conformation (Martin et al., 2005). Moreover, membrane proteins are almost exclusively α-helical bundles, with the exception of the β barrels found in the outer membrane of Gram-negative bacteria and mitochondria. More than 30% of the homologous superfamilies described in CATH are composed mainly or entirely of α helices (Greene et al., 2007). These domains are found in both soluble (SOL) and transmembrane (TM) proteins, and carry out a wide range of biological functions.

While SOL domains are well studied, TM domains have only recently begun to be elucidated. Since the first TM protein structure was solved in 1984 (Deisenhofer et al., 1984), the folding mechanism of these proteins has gradually become clearer (Bowie, 2005), yet much remains to be discovered. These proteins are estimated to make up 20%–30% of open reading frames in known genomes (Wallin and von Heijne, 1998), and are overwhelmingly α-helical, containing one or multiple membrane-spanning helices. Specific interaction patterns between helices play a critical role in the function, assembly, and oligomerization of these proteins (Langosch et al., 2010; Shai, 2001). Likewise, membrane protein misassembly can contribute to a myriad of disease states (Ng et al., 2012). However, due to experimental challenges in crystallization, TM proteins represent only 2% of deposited structures (White, 2009). Despite this shortage, deep computational and bioinformatics-based analyses of helix-helix interactions will accelerate our understanding the folding behavior of helical TM proteins (Nugent and Jones, 2012) and facilitate their design (Ghirlanda, 2009; Perez-Aguilar and Saven, 2012).

Consequently, the study of basic principles underlying the fold space of the helix-helix interactome, namely understanding the packing of helices, is intrinsic to understanding proteins. For example, in 1977 Chothia, Levitt, and Richardson presented simple helix-helix packing rules as determinants of protein structure (Chothia et al., 1977). An open question is whether helices from TM and SOL proteins are similar in the way they interact with each other and contribute to the overall protein structure. A small subset of SOL helix-helix pairs were

shown to be structurally homologous to TM pairs presenting similar properties, even though the overall distributions for SOL dimers are quite different from those of TM dimers (Gimpelev et al., 2004). Here, we investigate the range of currently known SOL helix-helix interactions and compare them with those found in TM proteins, focusing on the interplay between sequence and structure. To do this, we extend the approach used previously for characterizing TM dimers (Walters and DeGrado, 2006) to a larger database of TM dimers with stricter criteria, and compare the results with dimers from water-soluble proteins.

Analysis of sequences derived from helix-helix dimers propels our understanding of helix-helix interactions. The most extensively studied TM helix dimer is glycophorin A (GpA), a common model system (Lemmon et al., 1992; MacKenzie et al., 1997). Each helix of GpA contains two Gly separated by three amino acids, known as the GxxxG motif (Lemmon et al., 1994), which plays a key role in dimerization. The GxxxG motif is highly overrepresented in the sequences of TM proteins (Senes et al., 2000), and has been well-characterized structurally. GxxxG-containing dimers tend to have a parallel, right-handed geometry, compact helix-helix packing, and stabilizing interhelical backbone-mediated hydrogen bonds (MacKenzie et al., 1997; Mueller et al., 2014; Senes et al., 2001).

Comprehensive characterization via a variety of biophysical and biochemical methods has established the GxxxG motif as an important framework of TM helix-helix interaction (Russ and Engelman, 2000). Gly can be commonly replaced by another small residue, such as Ala or Ser in this motif (Mueller et al., 2014; Russ and Engelman, 2000; Senes et al., 2000). The Ala coil (Gernert et al., 1995) and GxxxxxxG motif are other prevalent sequence motifs found in membrane protein families (Liu et al., 2002). Additional sequence motifs have been identified, which depend on hydrogen bonds or weak polar interactions, and include derivatives of the small-residue motifs mentioned above (Adamian and Liang, 2002; Bowie, 2005; Gratkowski et al., 2002; Han et al., 2011; Hedin et al., 2011; Herrmann et al., 2009; Langosch and Arkin, 2009; Lawrie et al., 2010; Liang, 2002; Sal-Man et al., 2007; Unterreitmeier et al., 2007; Varriale et al., 2010; Wei et al., 2011; Zhou et al., 2001)

However, a systematic study of sequence-structure relationships on the scale of the whole protein structure database using structural bioinformatics is still lacking. Here we extract helix-helix pairs from high-resolution, non-homologous TM and SOL proteins from the PDB, and cluster them based on sequence-independent geometric similarity. We contrast the relative frequencies of each cluster in both environments and identify specific conformations that are unique to one or the other. Notably, sequence profiles can differ between the TM and SOL data sets, even for geometrically identical clusters. We also analyze the sidechain- and backbone-level interhelical hydrogen-bonding interactions of residues in seven clusters of TM helix dimers and in their structural counterparts, namely, SOL dimers, extending an early analysis of Adamian and Liang (2002). Characterization of these sequence, structural, and interaction motifs contribute to our understanding of the folding of helical proteins and aid both in structure prediction (Barth et al., 2009) and de novo design (Samish et al., 2011).

## RESULTS

### Clustering of TM Helical Pairs

Previously, Walters and DeGrado (2006) clustered the helical pairs culled from the existing crystal structures of membrane proteins to define distinct geometries for TM helical pairs, designated here as the WD analysis. Since then, the database has increased roughly 4-fold, allowing us to use more stringent criteria for clustering and resolve additional clusters. In our earlier work, we clustered a library of 455 pairs using a greedy clustering algorithm and a 1.5 Å cutoff, and found that 90% fell within geometric clusters. Here, as we hoped to find additional geometries, we used a more generous criterion for inclusion of helical pairs in the database but a more stringent cutoff of 1.25 Å as the clustering criterion. We again used greedy clustering and examined clusters with at least 25 members (representing 1.4% of the pairs, 16 in total). Clusters with fewer members are not considered here. Now we find 16 clusters (1,290 pairs), which comprise 48% of the pair library of 2,694 dimers (Figure 1). This coverage is smaller than the 90% seen previously (455 pairs in the library) for several reasons. We increased the minimal size of clusters to 25 members, so rare clusters are now excluded from the analysis. Secondly, the increased geometric stringency (root-mean-square deviation [RMSD] $\leq$ 1.25 Å) caused some of the WD clusters (RMSD $\leq$ 1.5 Å) to split into two clusters that did not separately meet the size threshold for inclusion in the analysis. Finally, and most importantly, we used different geometric criteria to define pairs, allowing large interhelical distances (up to 14 Å), whereas the previous study required that pairs should have an interhelical distance $\leq$12 Å. In the present study, most of these pairs with large interhelical distances did not fall within well-defined clusters, presumably because their geometries are determined by interactions with other portions of the protein. When we use a cutoff of 0.065 $Å^{-1}$ for the dimer mean inverse distance (see Experimental Procedures) we find that 67% of these more stringently defined pairs are in the 16 clusters. Moreover, 70% of the clustered dimers lie in the first seven clusters, each of which has more than 70 members. In summary, the geometries of most tightly interacting helices are well represented by the centroids of clusters 1–7 (Figure 2), which we discuss in detail below. Interestingly, Joo et al. (2012) mined the data sets of residues that contact each other and computed the crossing angles of the corresponding helices. Plotting the histogram distribution of these angles results in discrete peaks corresponding to the packing states described here (Figure 2). Similar crossing angle distributions have also recently been computed for membrane proteins (Lo et al., 2011).

Highly populated clusters of 70 members or more have been defined in the present analysis, even though the increased stringency split some of the previously defined clusters into two. The overall division between antiparallel and parallel and left- and right-handed clusters, i.e. the percentages of members in each class of cluster, is strikingly similar between the water-soluble and TM helix-helix interactome clusters (Figures 1A and 1B, inset). Yet the relative weight of helix-helix distances among these clusters displays differences (Figure 1). For example, as seen in Figure 1C, the largest cluster in the previous WD analysis (Walters and DeGrado, 2006) now splits into two

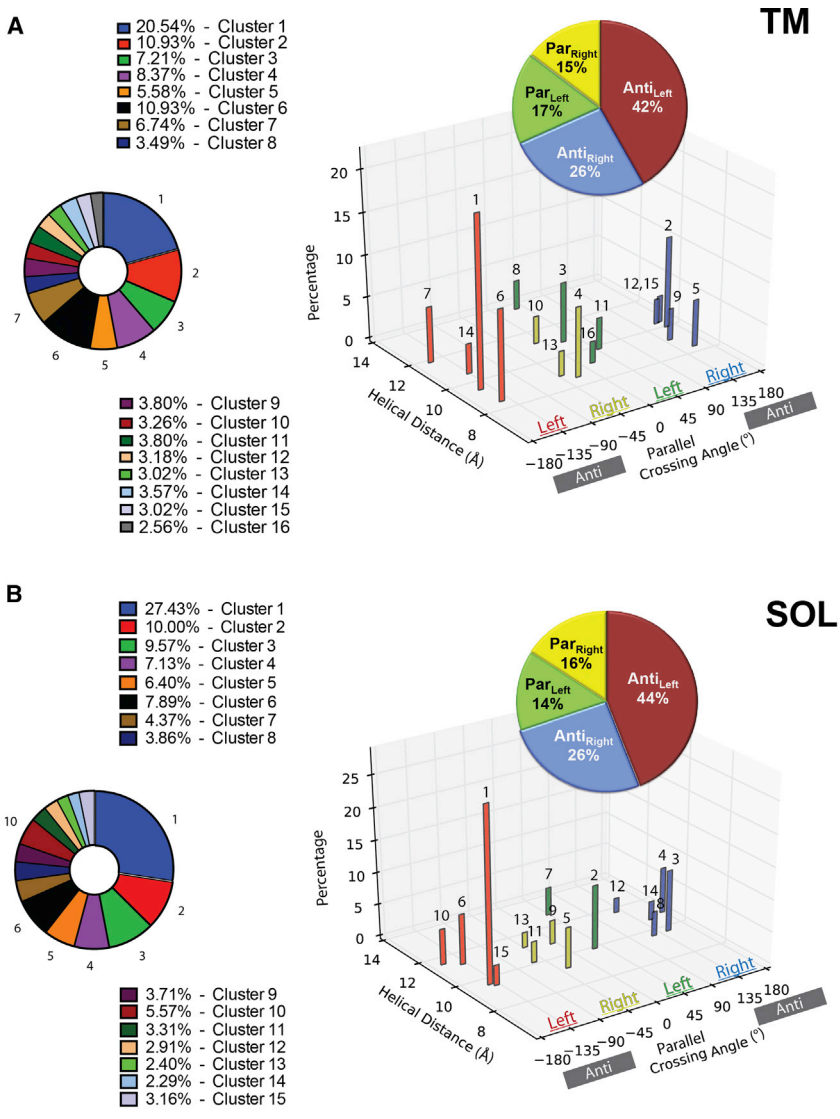



**Figure 1. Similarities between the TM and SOL Helix-Helix Clusters**

(A and B) Description of the 16 TM (A) and 15 SOL (B) clusters in respect of their crossing angle and interhelical distance. Helix-helix crossing angle is color coded by 90° segments as in the WD study (Walters and DeGrado, 2006) to $Anti_{left}$ (red), $Par_{right}$ (yellow), $Par_{left}$ (green), and $Anti_{right}$ (blue) with the percentage of each group (inset pie graph) and each cluster (pie graph on left) shown.

(C) The RMSD similarity of the top seven TM clusters relative to their SOL structural counterparts are measured on the 12-residue windows on the centroids with the smallest RMSDs along the most populated 15-residue regions. The corresponding cluster number from the WD study is depicted.

C

| Designation | TM Clusters | Cluster # | WD Clusters | SOL Counterparts | Cluster # | RMSD (Å) |
|---|---|---|---|---|---|---|
| **$Antiparallel_{left}$** | int | 1 | 1 | int | 1 | 0.59 |
| | close | 6 | 1 | int | 1 | 1.20 |
| | far | 7 | | far1 | 6 | 1.30 |
| | far | 7 | | far2 | 10 | 1.24 |
| **$Parallel_{left}$** | int | 3 | 4 | int | 2 | 0.91 |
| **$Antiparallel_{right}$** | close | 5 | 2 | close | 3 | 1.83 |
| | int | 2 | 5 | int | 4 | 0.53 |
| **$Parallel_{right}$** | close | 4 | 3 | close | 5 | 0.65 |

clusters (clusters 1 and 6), which we define as $Anti_{left}$(int) and $Anti_{left}$(close), respectively (Figure 1A). In this nomenclature, $Anti_{left}$(int) refers to an antiparallel dimer with a left-handed crossing angle and an interhelical distance that is intermediate between the other two major antiparallel left-handed clusters with close and far interhelical distances. Other than $Anti_{left}$, major clusters include $Par_{left}$(int), $Anti_{right}$(close), $Anti_{right}$(int) and $Par_{right}$(close). There are other less populated clusters that have, for example, closer and greater interhelical distances than $Par_{left}$(int), but they did not reach the criterion of 70 members that we have set for more in-depth structural analysis (Table S1).

## The Most Prevalent Water-Soluble Helical Pairs Have Geometries Closely Related to Their Membrane Counterparts

A total of 5,085 water-soluble helical pairs were extracted from a database of predominantly helical proteins, and clustered using the same methods as for the TM pairs, yielding a total of 15 clusters, ranging in size from 754 to 55 members (Figure 1A; Tables S1 and S2). Together this set comprises 52% of the total pairs. The TM and SOL helix pair clusters are geometrically highly similar with most being antiparallel (70% and 68% for the SOL and TM data sets, respectively) and left-handed pairs (58.8% and 59.0% for the SOL and TM data sets, respectively). Although these cluster groups also share similar interhelical distances (Table S1), they differ in the relative abundance of interhelical distances within each cluster (Figure 1).

The top seven SOL clusters (Figure 2B) include 74.0% of the clustered helix pairs. With the exception of the $Anti_{right}$(close) motif, these are highly similar to the top seven TM helical clusters (Cα RMSD ≤ 1.3 Å, Figure 1B). Thus, the differences between SOL and TM centroids are generally within the same range as the RMSD between members of a given cluster (up to 1.25 Å). Often there is a one-to-one relationship between the clusters, although this is not always the case. Three notable exceptions from this rule are: (1) the $Anti_{left}$(close) motif found in TM pairs

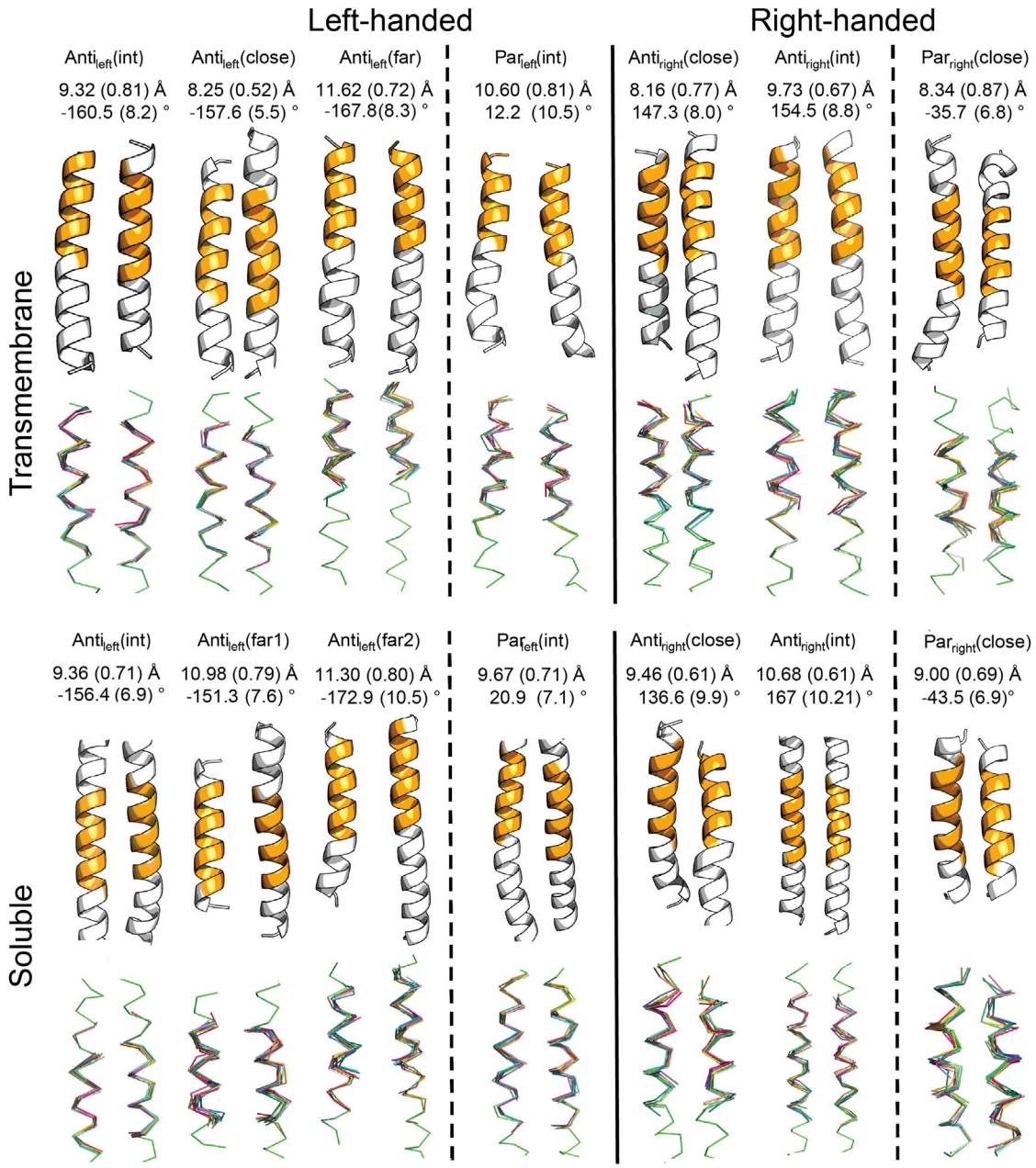



**Figure 2. Description of the Seven Frequent TM and SOL Clusters**

Average values of interhelical distance and crossing angle for the clusters are measured on the most populated 12-residue windows of the clusters colored in orange in the centroids, and SDs are shown in parentheses. The top ten members in the clusters with the closest RMSD to the centroid are overlapped on the lower panels.

is not among the top seven SOL clusters and is rare in the water-soluble database (cluster 15, see Table S2); (2) a motif in the soluble data set that is relatively close in geometry to the Anti$_{left}$(int) motif (RMSD = 0.6 Å for the centroids), and somewhat more distant from the Anti$_{left}$(close) motif (RMSD = 1.2 Å); (3) the Anti$_{left}$(far) motif shows high similarity to two different clusters of related geometry in the water-soluble database (Table S2).

Helices tend to pack more tightly and have shorter interhelical distances in membrane proteins compared with water-soluble proteins (Eilers et al., 2002; Oberai et al., 2009; Senes et al., 2004; Zhang et al., 2009). For example, the TM Anti$_{right}$(close) and Par$_{right}$(close) motifs have a closer interhelical distance than the corresponding water-soluble motifs by 0.9 and 0.5 Å, respectively. This tightening of the interhelical distance is well documented in previous studies of helix-helix packing of membrane proteins (Cross et al., 2013; Javadpour et al., 1999). Indeed, while the packing energetics of TM and SOL proteins is similar (Joh et al., 2009), TM proteins bury more residues,

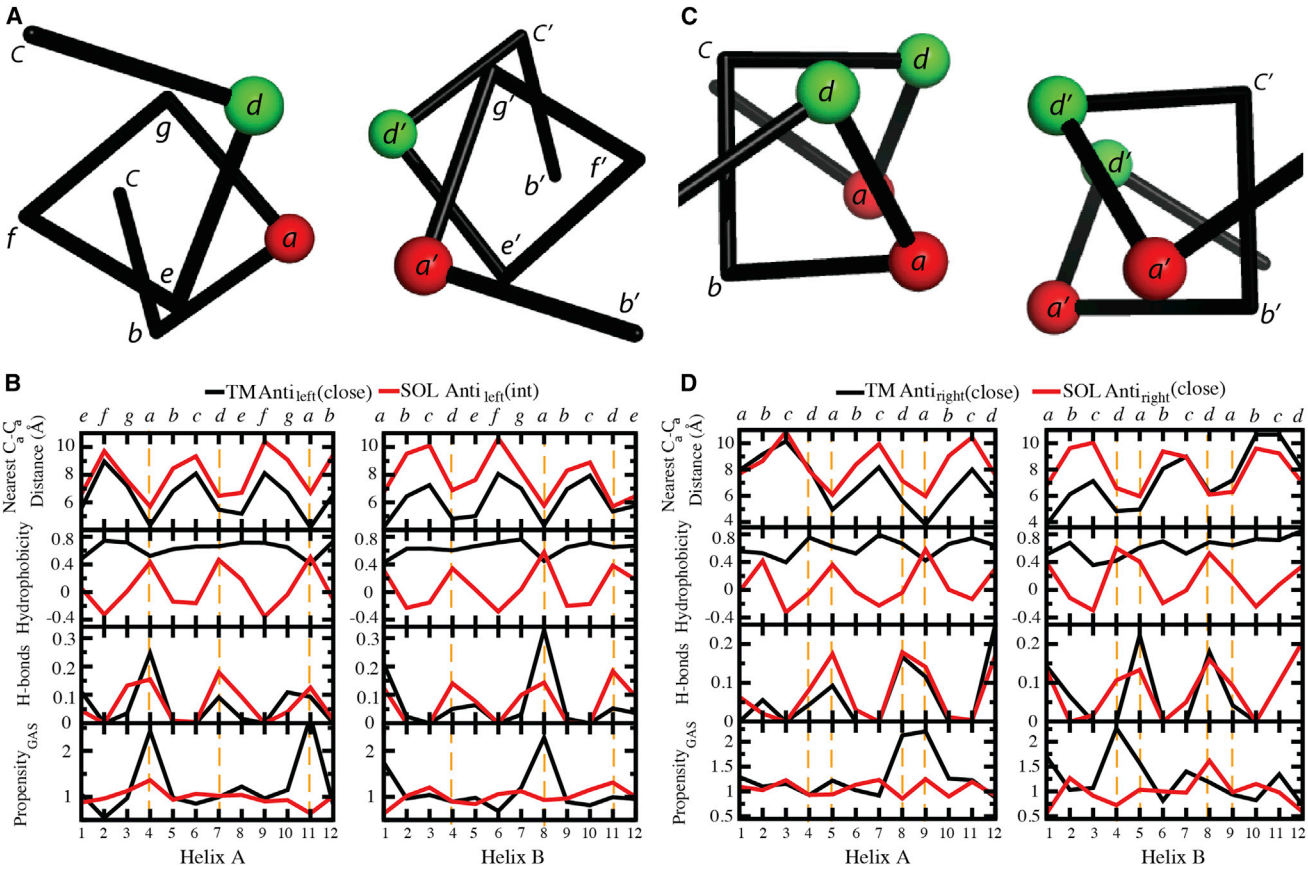



**Figure 3. Profiles of the Nearest Cα-Cα Distance, Average Hydrophobicity, Hydrogen-Bonding Fractions, and Propensity of Small-Residue GAS on Structurally Matched Windows between TM and SOL Clusters**

Residues at the interhelical interface are highlighted by orange dashed lines (B and D). The designation of positions in the heptad and tetrad repeats is shown at the top (A and C).

which are smaller on average, compared with SOL residues (Oberai et al., 2009), thus facilitating this phenomenon. In summary, the SOL and TM helix-helix interactomes display similar structural fold space with a small bias towards tighter helix-helix distances in the TM motifs.

## Correlations between Interhelical Distance, Hydrophobicity, Interhelical Hydrogen Bonding, and Residue Preferences in Aligned Sequences of TM and SOL Helical Pairs

We investigated possible similarities between the nearest Cα-Cα distances, the average hydrophobicity, the hydrogen-bonding fraction, and the sequence propensities for each position along the aligned windows of the top seven TM and SOL clusters (Figures 3 and 4; Figure S1). The structural resemblance of TM and SOL clusters is manifested in the highly similar patterns of the nearest Cα-Cα distance of their centroids. The periodicity of the nearest Cα-Cα distance tends to display the heptad and tetrad repeats for left- and right-handed helix dimers, respectively (Figure 3), confirmed by least squares fitting of a sinusoidal function to the data (Table S3). When helices cross with a left-handed crossing angle, the interaction pattern resembles that seen in classically left-handed coiled coils over a limited length of the chain (10–15 residues). We therefore denoted these positions using the classical coiled-coil heptad nomenclature, *abcdefg* (Crick, 1953a, 1953b; Sodek et al., 1972; Talbot and Hodges, 1982). By contrast, the interaction pattern between right-handed helix crossing approximately repeats each four residues, denoted *abcd*. In both cases, the positions *a* and *d* are at the interhelical interface.

Sequence profiles of the interhelical distance, hydrophobicity, interhelical hydrogen bond frequency, and the propensity for a position to be occupied by a small residue, Gly, Ala, or Ser (termed herein as GAS) provide information concerning the driving force for the assembly of helical pairs in different environments. Figure 3C presents data for the two helices in the TM $Anti_{left}$(close) motif, and its closest counterpart in the SOL database, the $Anti_{left}$(int) motif; the profiles for the TM and water-soluble helices are colored black and red, respectively. Focusing first on the interhelical distance profile, one can see that the water-soluble distances tend to be very similar to that of the TM at one end of the bundle, but diverge by about 2–3 Å at the C terminus of helix A and the N terminus of helix B in the antiparallel motif. We also see a clear 180° phase shift between the interhelical distance and the mean hydrophobicity at the corresponding position in water-soluble proteins. This relationship

reflects the tendency of water-soluble proteins to have apolar residues in buried positions and polar residues at water-accessible positions. This tendency to place hydrophobic residues at the *a* and *d* positions is reflected by different degrees of sinusoidal hydrophobicity propensities in practically all SOL clusters (Figure 4) and in propensities of the individual amino acids (Figure 5; Figure S2). By contrast, the hydrophobicity profile of the TM is uniformly high, reflecting the overall hydrophobic nature of TM helices. Hydrogen bonds are frequently observed along the interfacial *a* and *d* positions of the water-soluble $Anti_{left}$(int) pair, but are highly restricted to the *a* positions in the corresponding TM $Anti_{left}$(close) motif. The difference reflects the closer approach of the helices in the TM motif resulting in shorter interhelical distances at the *a* position. Finally, the TM $Anti_{left}$(close) motif has a very high propensity for GAS residues at only position *a* of the motif, a tendency that is not present in the water-soluble counterpart. The notable exception is of the significant preference for His at $Anti_{left}$(close) at *a* and *d* positions ($p > 0.01$, Figure 4). Upon further investigation, we found this to be due to 26 helical pairs (18.4% of $Anti_{left}$(int)) derived from chlorophyll binding proteins, which use His to coordinate metals (Braun et al., 2011). Meanwhile, a similar TM motif $Anti_{left}$(int) contains only 3% of pairs from such proteins. Otherwise, we observed a strong tendency to place small residues (Gly, Ala, or Ser) at these positions (Figure 5A), a phenomenon seen also for the TM $Anti_{right}$(close) (Figure 5D; Figure S2F) and $Par_{right}$(close) (Figure 5F; Figure S2H).

In parallel, bulky and β-branched amino acids are underrepresented in these close TM motifs yet are more abundant in their water-soluble counterparts, especially with increasing interhelical distance (Figure 5). Thus, the presence of small residues facilitates close helix-helix packing reflected by closer interhelical distances. In summary, the most striking difference in the profiles lies in the strong hydrophobic periodicity seen for the water-soluble pair, reflecting the hydrophobic driving force for assembly in water. In contrast, the TM (close) motifs show a strong periodicity in the GAS propensity, reflecting the strong driving force for folding in membranes associated with the packing of small residues along one face of a TM helix (Eilers et al., 2002; Oberai et al., 2009; Senes et al., 2004; Zhang et al., 2009).

The interhelical distance of related helical pairs is affected by the composition of the residues at the interface, as reflected in the profiles for the $Anti_{left}$(close), $Anti_{left}$(int), and $Anti_{left}$(far) motifs (Figures 4A–4C). A comparison of the interhelical distance profiles for these three left-handed antiparallel motifs shows that the TM and water-soluble motifs are essentially superimposable for the intermediate and far motifs (correlations, all $R^2 > 0.71$; periods shown in Table S3). The repeated pattern of hydrophobicity remains strong for all three SOL motifs, while the TM pairs remain uniformly hydrophobic. Conversely, the hydrogen-bonding profiles are only similar between the water-soluble and TM motifs for the $Anti_{left}$(int) and $Anti_{left}$(far) motifs ($R^2 = 0.55$ for $Anti_{left}$(far) helix A, but $R^2 > 0.67$ for the others; Figures 4B and 4C). However, for the $Anti_{left}$(close) motif, the frequency of interhelical hydrogen bonds at interfacial positions is 2- to 3-fold higher for water-soluble helices than for TM helices. This finding may reflect the relative paucity of polar residues to form hydrogen bonds in TM helices (Figure 4), rather than the favorability of their formation in an apolar environment (Senes et al., 2004). As the helices become increasingly distant in progressing from the $Anti_{left}$(close) to $Anti_{left}$(far) motifs, the propensity for GAS residues decreases, becoming unfavorable for $Anti_{left}$(far) for both water-soluble and TM motifs.

A comparison of the antiparallel right-handed motifs with the left-handed motifs (Figure 4 left versus right halves of the figure) shows precisely the same trends, although the periodicity of the profiles is shifted closer to 4-residues from the 3.5-residue period seen for the left-handed motifs. The water-soluble $Anti_{right}$(close) motif shows a systematic increase in the interhelical distance at one end of the pair while this divergence is not seen for the corresponding TM motif. The TM $Anti_{right}$(close) also shows a strong GAS propensity at the *a* and *d* positions where the helices make their closest contact. A strong GAS propensity is not seen in the corresponding SOL motifs, possibly reflecting the hydrophobic core (relative to hydrophilic surrounding) found only in the latter motifs (Figure 4). Also, as seen for the $Anti_{left}$ motifs, the geometry of the interacting helices became identical at intermediate interhelical distances for both the water-soluble and TM motifs.

GpA was an early example of a GxxxG motif. Geometrically, the $Par_{right}$(close) is similar to the GpA structure, and the RMSD between GpA and the centroid of the $Par_{right}$(close) cluster is 1.5 Å (by overlapping a window of 16 residues in the TM helix pairs). The GxxxG motif is rare in this analysis of multispan proteins, representing 11.9% of the top seven TM clusters. A possible explanation is that GpA is an anchor to a constitutively dimeric glycoprotein rather than a dynamically functioning protein, as is the case for most TM proteins. Interestingly, our sequence analysis shows the GAS propensity is stronger at one of the two helices. This finding matches recent results from mutagenesis analysis of the strengths of dimerization of integrin TM helices, which display an asymmetric GxxxG packing motif (Berger et al., 2010). Peaks in the GAS propensity are also seen in one of the two helices in the water-soluble $Par_{right}$(close) motifs (Figure 4F).

## The Clusters Have a Distinct Hydrogen-Bonding Connectivity Network

Antiparallel helices can form interhelical hydrogen bonds between residues from interacting helices. Depending on the sidechains and the interhelical geometry, a number of hydrogen-bonding patterns or ''connectivities'' are possible. For antiparallel left-handed helical motifs hydrogen bonding is geometrically feasible between *a* and *a′*, *d* and *d′*, *a* and *d′*, *d* and *g′*, or *a* and *e′*. However, these do not occur with equal frequencies. Classically, *a*-to-*d′* hydrogen bonding has been extensively studied and used in protein design (McClain et al., 2001, 2002; Oakley and Kim, 1998). However, this interaction pattern is the exception rather than the rule for antiparallel helices. For the motifs for which there are at least 25 observations of hydrogen bonds, *a*-to-*a′* and *d*-to-*d′* hydrogen bonding generally predominates over other hydrogen-bonding connectivities; this is particularly striking for the TM $Anti_{left}$(close) motif (Figure 4A), in which the proportion of *a*-*a′*, *a*-*d′*, and *a*-*e′* is 87:11:1 (Figure 6A). As the interhelical distance increases within a motif, the preference for *a*-*a′* and *d*-*d′* becomes less striking (Figure 6A), presumably because the greater interhelical distance provides greater flexibility for sidechain interactions. Interestingly, precisely the same preferences for *a*-*a′* and *d*-*d′* connectivities are seen in the antiparallel right-handed motifs (Figure 6C).

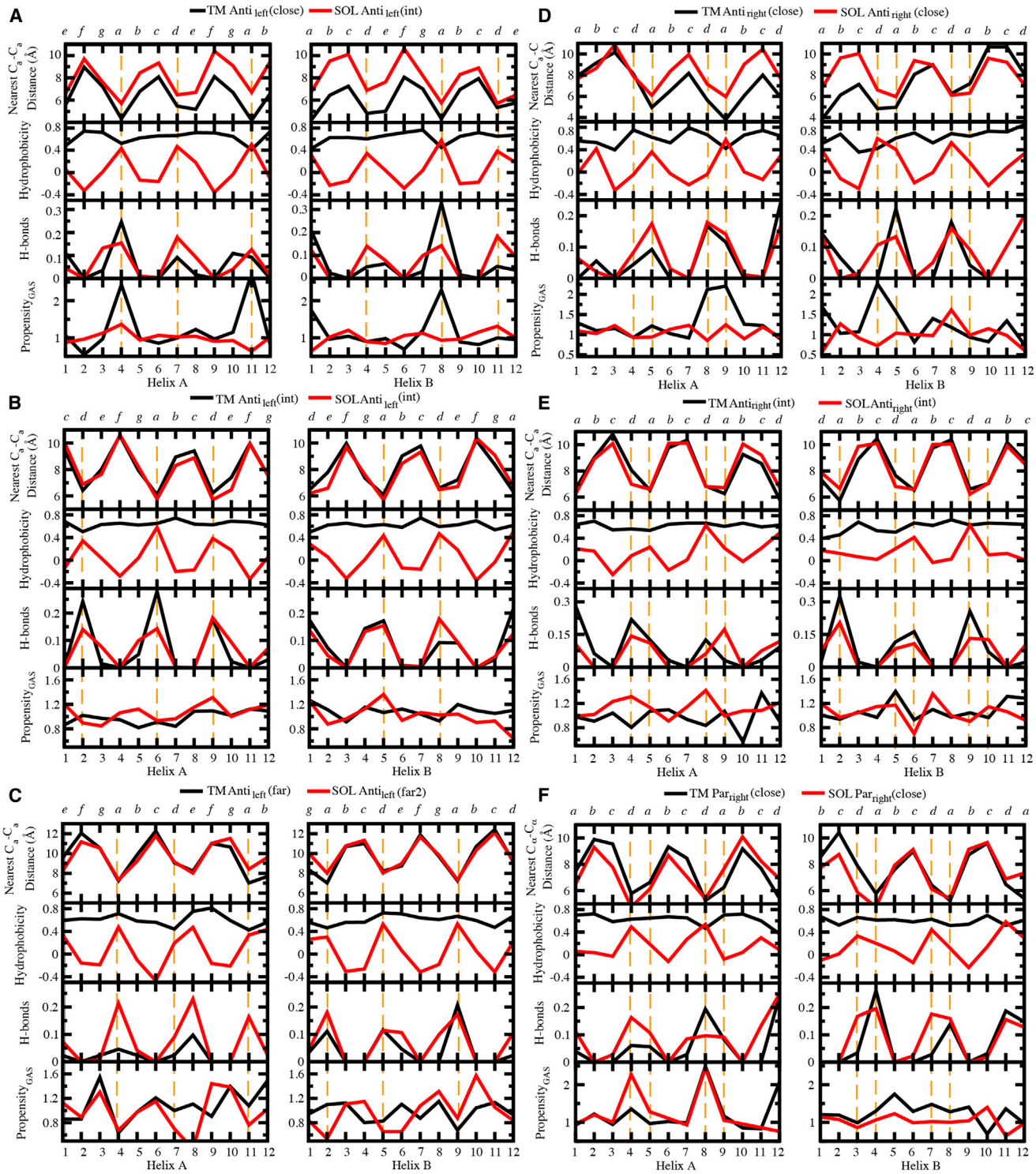



**Figure 4. Comparisons of Interhelical Distances, Average Hydrophobicity, Hydrogen-Bonding Fractions, and Propensity of Small-Residue GAS for Structurally Matched TM and SOL Motifs**

(A–F) The 12-residue window of each TM centroid that contains the most cluster members was chosen as a representative sample for analysis. These and the matching windows on each corresponding SOL cluster were analyzed together. Residues at the interhelical interface are highlighted by orange dashed lines. The interhelical distances refer to the closest distance at a given Cα for one helix to a Cα in the neighboring helix. This figure is continued for additional pairs in Figure S1.

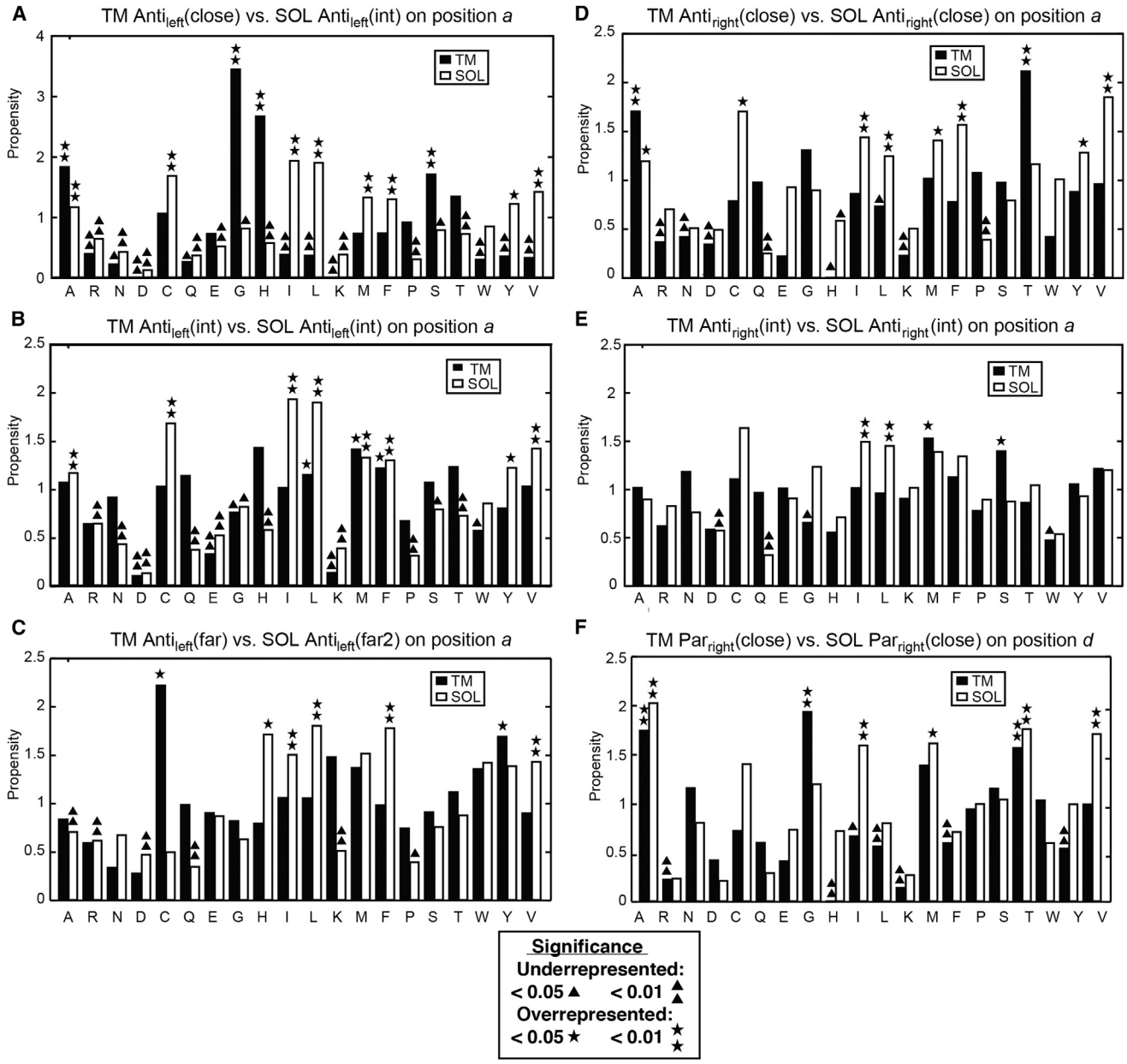


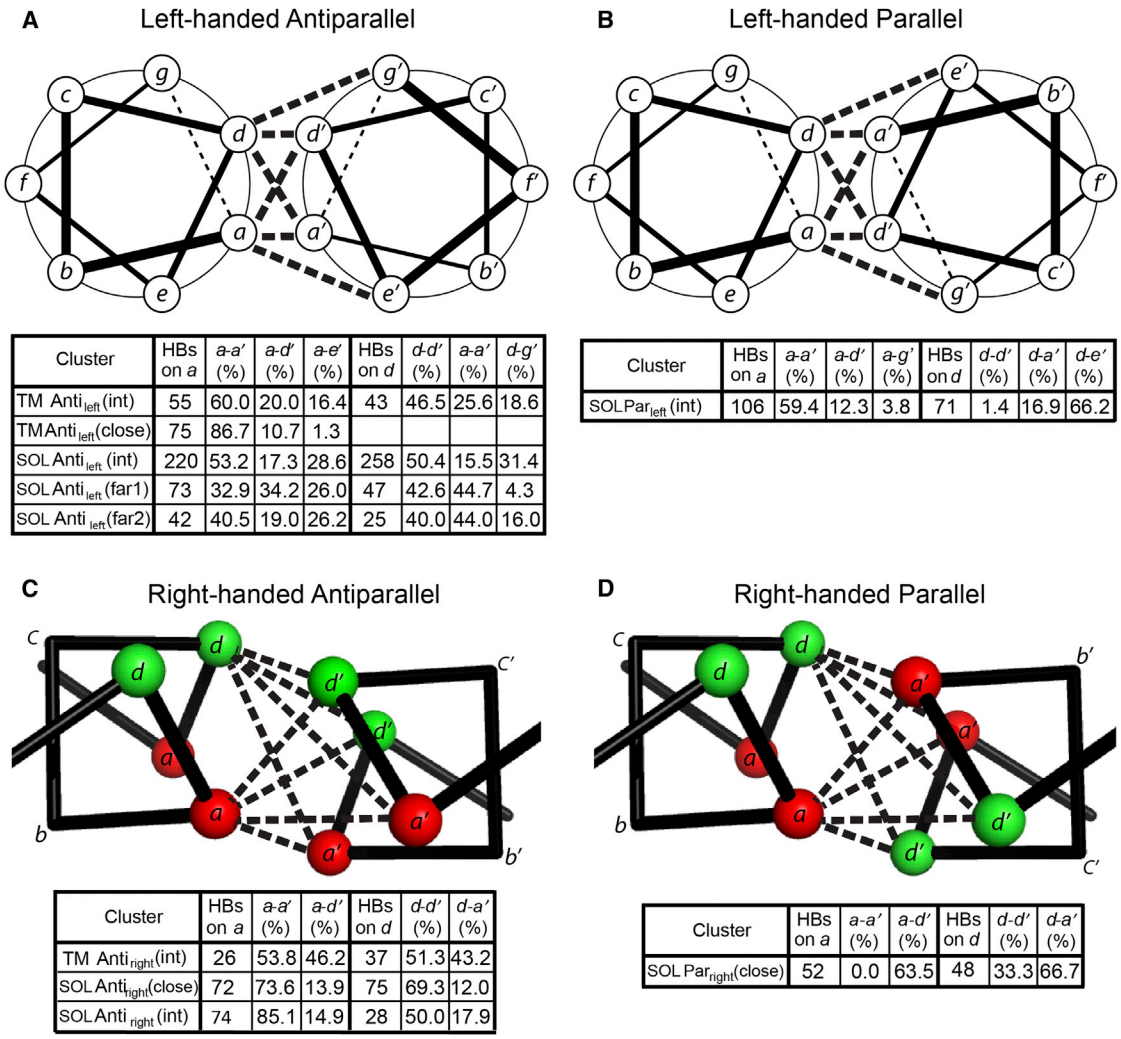

**Figure 5. Propensities of Amino Acids in Different Positions at the Interhelical Interface**

In A–F, residues labeled by asterisks or triangles are statistically overrepresented or underrepresented, respectively, as determined by the p value of a binomial test ($p < 0.05$ or $p < 0.01$), relative to the expected amino acid frequency as described in Experimental Procedures (Table S4). This figure is continued for additional pairs in Figure S2.

The hydrogen-bonding connectivities seen in parallel left-handed hydrogen-bonding patterns follow the familiar patterns expected from parallel coiled-coil motifs (Grigoryan and DeGrado, 2011). The preferred hydrogen bonding at *a* positions involves *a*-*a*′ connectivities. By contrast, *d*-*d*′ is rare, due to the geometry of the coiled coil. Instead, *d* residues tend to hydrogen bond to *e*′ of a neighboring helix (Figure 6B).

The only right-handed parallel cluster with sufficient numbers of interhelical hydrogen bonds to merit analysis was the water-soluble Par$_{right}$(close) motif (Figure 6D). In this case, *a*-*d*′ greatly outnumbered the *a*-*a*′ or *d*-*d*′ interactions. As mentioned above, the opposite was true for right-handed antiparallel motifs.

The hydrogen-bonding connectivity maps also shed light on the conformational specificity of TM and SOL helical bundles. Firstly, in prototypical parallel coiled coils, buried hydrogen bonds typically form between small polar residues in the same register of the heptad repeats (Woolfson, 2005). In antiparallel SOL coiled coils, strong *a*-*d*′ and *d*-*a*′ interactions are anticipated (Mason and Arndt, 2004), and this was observed in many cases.




| Cluster | HBs on *a* | *a-a′* (%) | *a-d′* (%) | *a-e′* (%) | HBs on *d* | *d-d′* (%) | *a-a′* (%) | *d-g′* (%) |
|---|---|---|---|---|---|---|---|---|
| TM $Anti_{left}$(int) | 55 | 60.0 | 20.0 | 16.4 | 43 | 46.5 | 25.6 | 18.6 |
| TM$Anti_{left}$(close) | 75 | 86.7 | 10.7 | 1.3 | | | | |
| SOL $Anti_{left}$ (int) | 220 | 53.2 | 17.3 | 28.6 | 258 | 50.4 | 15.5 | 31.4 |
| SOL $Anti_{left}$(far1) | 73 | 32.9 | 34.2 | 26.0 | 47 | 42.6 | 44.7 | 4.3 |
| SOL $Anti_{left}$(far2) | 42 | 40.5 | 19.0 | 26.2 | 25 | 40.0 | 44.0 | 16.0 |

| Cluster | HBs on *a* | *a-a′* (%) | *a-d′* (%) | *a-g′* (%) | HBs on *d* | *d-d′* (%) | *d-a′* (%) | *d-e′* (%) |
|---|---|---|---|---|---|---|---|---|
| SOL$Par_{left}$ (int) | 106 | 59.4 | 12.3 | 3.8 | 71 | 1.4 | 16.9 | 66.2 |

| Cluster | HBs on *a* | *a-a′* (%) | *a-d′* (%) | HBs on *d* | *d-d′* (%) | *d-a′* (%) |
|---|---|---|---|---|---|---|
| TM $Anti_{right}$(int) | 26 | 53.8 | 46.2 | 37 | 51.3 | 43.2 |
| SOL $Anti_{right}$(close) | 72 | 73.6 | 13.9 | 75 | 69.3 | 12.0 |
| SOL $Anti_{right}$ (int) | 74 | 85.1 | 14.9 | 28 | 50.0 | 17.9 |

| Cluster | HBs on *a* | *a-a′* (%) | *a-d′* (%) | HBs on *d* | *d-d′* (%) | *d-a′* (%) |
|---|---|---|---|---|---|---|
| SOL $Par_{right}$(close) | 52 | 0.0 | 63.5 | 48 | 33.3 | 66.7 |

**Figure 6. Hydrogen-Bonding Connectivity Networks for the Clusters with Different Geometry**

(A–D) The number of hydrogen bonds is the arithmetic summation of those on the most populated position *a* or *d* from both chains. The percentage of each contact type, e.g. *a-e′*, is the fraction of the sum on that position, i.e. sum on an *a* or *d*.

However, in $Anti_{left}$(int), there is a strong preference to form *a-a′* and *d-d′* hydrogen bonds, and a tendency to form *a-e′* and *d-g′* interactions. Hydrogen-bonding connectivity maps should help guide the design of complex SOL and TM helical bundles (Tatko et al., 2006).

## TM and SOL Clusters Utilize Different Residues for Hydrogen Bonding

Next, we examined differences and similarities between the interhelical sidechain-to-sidechain and sidechain-to-backbone interhelical hydrogen bonding in the TM versus the SOL helix dimers. In this nomenclature, e.g. sidechain-to-backbone, the first helix of the pair has a residue in which the sidechain participates in a hydrogen bond and the second helix of the pair has a backbone atom, which participates in the bond. Due to low number of counts for hydrogen bonds in the individual TM clusters, the hydrogen bonds of the top seven TM clusters are summed.

An expected major difference between TM and SOL clusters is the relative abundance of backbone-mediated interhelical hydrogen bonds expected. In the TM clusters, sidechain-to-sidechain and sidechain-to-backbone hydrogen bonds comprise 56% and 44% of the total, respectively, while in the SOL clusters sidechain-to-sidechain and sidechain-to-backbone hydrogen bonds have a population of 80% and 20%, respectively. Consistent with previous surveys of hydrogen bonding (Baker and Hubbard, 1984), the majority of sidechain-to-backbone hydrogen bonds is from sidechain donors to the backbone carbonyl hydrogen bond acceptors, with a portion of 93% and 94% in the TM and SOL clusters, respectively. Therefore, we analyze only sidechain-to-backbone carbonyl hydrogen bonds herein.

In the sidechain-to-sidechain hydrogen-bonding interactions among TM clusters (Figures 7 and 8), Ser is the largest contributor to hydrogen bonding, accounting for 25.4% of occurrences, and showing a significantly high propensity ($p < 0.001$) for these

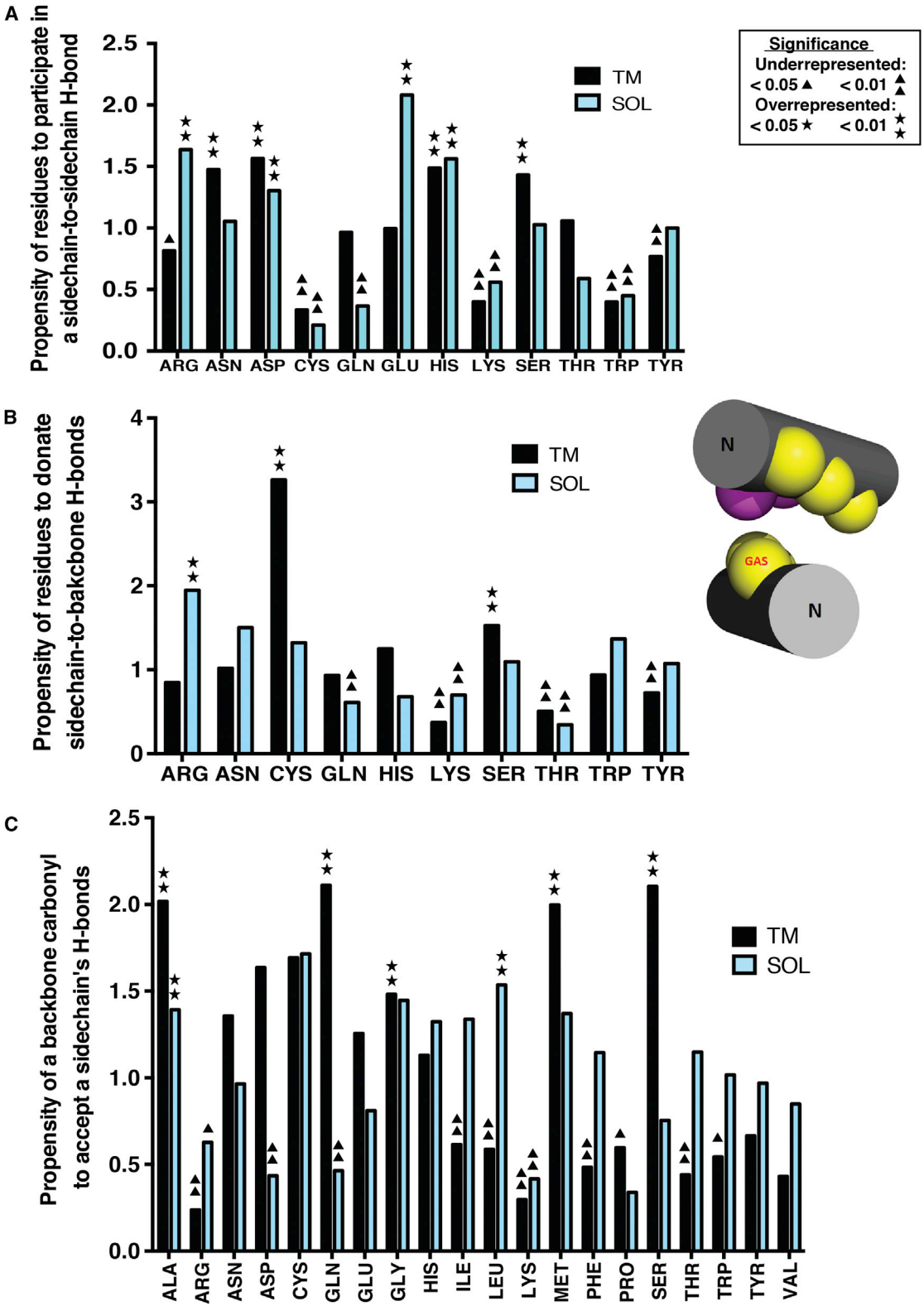



**Figure 7. Propensity of Residues in the Top Seven TM and SOL Clusters to Donate or Accept an Interhelical Hydrogen Bond of Different Types**

(A) Interhelical hydrogen-bonding propensity of residues participating in sidechain-to-sidechain hydrogen bonds.

(B) Interhelical hydrogen-bonding propensity of residues that donate a sidechain hydrogen bond to the backbone carbonyl on the helical pair.

*(legend continued on next page)*

interactions even relative to the high abundance of Ser in TM helices (Figure 7A). The other three residues with high propensity ($p < 0.01$) are Asn, His, and Asp, which have a much lower frequency in distribution (Table S4). Interestingly, Asn has a 4-fold preference to engage in hydrogen bonds in right-handed crossings, and His has a 4-fold preference in left-handed crossings (data not shown). Each of the other polar residues occurs in less than 12% of hydrogen bonds. The predominance of Ser among sidechain-to-sidechain interactions in the membrane environment is consistent with a previous report by Adamian and Liang (2002). Ser-Thr, Ser-Tyr, Ser-Ser, and Thr-Thr are the most common sidechain-to-sidechain hydrogen-bonding contributors, shown in Figure 8.

In the top seven SOL clusters, Arg displays a very high sidechain-to-sidechain hydrogen-bonding propensity (Figure 7A). The most frequent residues of this hydrogen-bonding class are Arg (19.8%), Glu (19.5%), and Asp (12.3%): Arg-Glu (19.0%), Arg-Asp (12.6%), and Lys-Glu (6.9%) are the three most common pairs of hydrogen-bonding partners (Figure 8B).

In the sidechain-to-backbone hydrogen bonds of the TM clusters (Figures 7B and 8C), Ser and Cys are overrepresented as hydrogen-bonding donors, with frequencies of 31.8% and 11.0%, respectively. Small residues Ala, Gly, and Ser are the major backbone carbonyl hydrogen-bonding acceptors, with 25.9%, 12.4% and 11.4% of the occurrences, respectively. The small residues may facilitate tight interactions, as found in the case of the $Par_{right}$(close) model protein GpA (Figure 7B, inset).

In contrast to the TM clusters, the SOL clusters have Arg as the main sidechain-to-backbone hydrogen-bonding donor (29.1%), with Gln (13.3%), Ser (11.5%), and Lys (10.9%) next (Figure 8D), but only Arg is overrepresented (Figure 7B). Aliphatic residues without β-branching, namely Leu (18.8%) and Ala (16.4%), are the two major backbone carbonyl hydrogen-bonding acceptors (Figure 7C). It is interesting to note the important role of Arg residues in forming both sidechain-to-sidechain and sidechain-to-backbone carbonyl interactions in water-soluble helical pairs. This finding agrees with experimental studies, which showed that this residue is unique among the polar residues in terms of its ability to contribute largely to conformational stability and specificity (Acharya et al., 2006; Borders et al., 1994).

## DISCUSSION

This work provides the most extensive analysis of TM and SOL helical interactions, providing a library of helical motifs and their corresponding sequence preferences. Moreover, the present study provides information concerning the pattern and positions of hydrogen-bonding residues and how they may provide specificity supporting different helical packing interaction motifs. This work also provides the first extensive comparison of geometrically similar TM and water-soluble helical pairs.

Comparing the helix-helix interactome of TM and water-soluble proteins leads to key differences, one of which lies in the greater abundance of tightly interacting helical pairs in TM compared with water-soluble proteins. Water-soluble structures tend to have more interhelical hydrogen bonds and utilize larger and more charged residues for this task. On one hand, the water-soluble helix-helix interactome generally displays a sinusoidal pattern of hydrophobicity. On the other hand, the TM helix-helix interactome displays a significantly more pronounced abundance of small residues at the helix-helix interface, which facilitate backbone-mediated interhelical hydrogen-bonding interactions. This contrasts with the old view that membrane proteins are inside-out versions of water-soluble proteins. Instead, the requirements to maintain membrane proteins within a low-dielectric transmembrane environment, or the requirements associated with helix insertion via the translocon, select for TM helices that are highly hydrophobic and do not necessarily use hydrogen bonds for stability as much as their soluble-protein counterparts. Nevertheless, small-residue sidechain- and backbone-mediated hydrogen bonds in the membrane milieu may guide helix-helix assembly and direct dynamic functionality (Bowie, 2011).

Helix-helix association is also affected by other factors, e.g. hydrophobic mismatch between a TM helix and the membrane (Benjamini and Smit, 2012). Investigation of the clusters will help greatly our understanding of the folding and structure of helical proteins, quantifying broad structural trends that will be useful in structure prediction and design.

## EXPERIMENTAL PROCEDURES

### Data Set Selection

The Orientation of Proteins in Membranes (OPM) database (Lomize et al., 2012) was used as the source for helical TM proteins. We obtained a list of all structures available as of September 26, 2014. To ensure accurate analysis, structures with X-ray resolution lower than 3.2 Å were removed from consideration. From the remaining structures, we used the PISCES server (Wang and Dunbrack, 2003) to cull at the PDB ID level for a maximum sequence homology of 30%. This resulted in a list of 139 representative structures, from which helix-helix pairs were derived. For the soluble database, a query was executed on the PDB as of February 9, 2012 for all structures classified in CATH (Greene et al., 2007) as "mainly α" and containing only protein. These were matched against the PDB-TM database (Tusnady et al., 2005), and any TM proteins were removed. This list was also culled using the PISCES server to a maximum of 30% sequence identity. To keep the size of the data set computationally tractable, only structures with a maximum resolution of 2.0 Å were kept, resulting in 765 proteins. For all soluble structures, the biological unit was downloaded from the PDB. The lists of TM and SOL structure covered for analysis are included in a spreadsheet file in the Supplemental Information.

We extracted the helical regions from the selected structures using the definitions of the TM segments in the OPM or the HELIX records in the PDB header information for soluble proteins. To ensure that these definitions were correct, the annotated regions were filtered to exclude helical breaks or sharp kinks (defined with a loose cutoff: $-130^\circ < \varphi < -20^\circ$ and $-90^\circ < \psi < 30^\circ$). They were also extended by up to four residues on both the N- and C-terminal sides if the positions met a stricter definition of helicity ($-90^\circ < \varphi < -35^\circ$; $-70^\circ < \psi < 0^\circ$). This helped to join soluble helices that otherwise might have been counted separately.

(C) Interhelical hydrogen-bonding propensity of residues that accept a hydrogen bond via the backbone carbonyl to the sidechains of their helical pair. As an example, the TM $Par_{right}$(close) motif adopts configuration shown in the inset. Positions *a* and *b* are represented by yellow and magenta spheres, respectively. The one-sided small-residue positions are labeled by GAS. The N termini of the helices are labeled.

Residues labeled by asterisks or triangles are statistically overrepresented or underrepresented as hydrogen bond participants, respectively, as determined by the p value of a binomial test ($p < 0.05$ or $p < 0.01$).

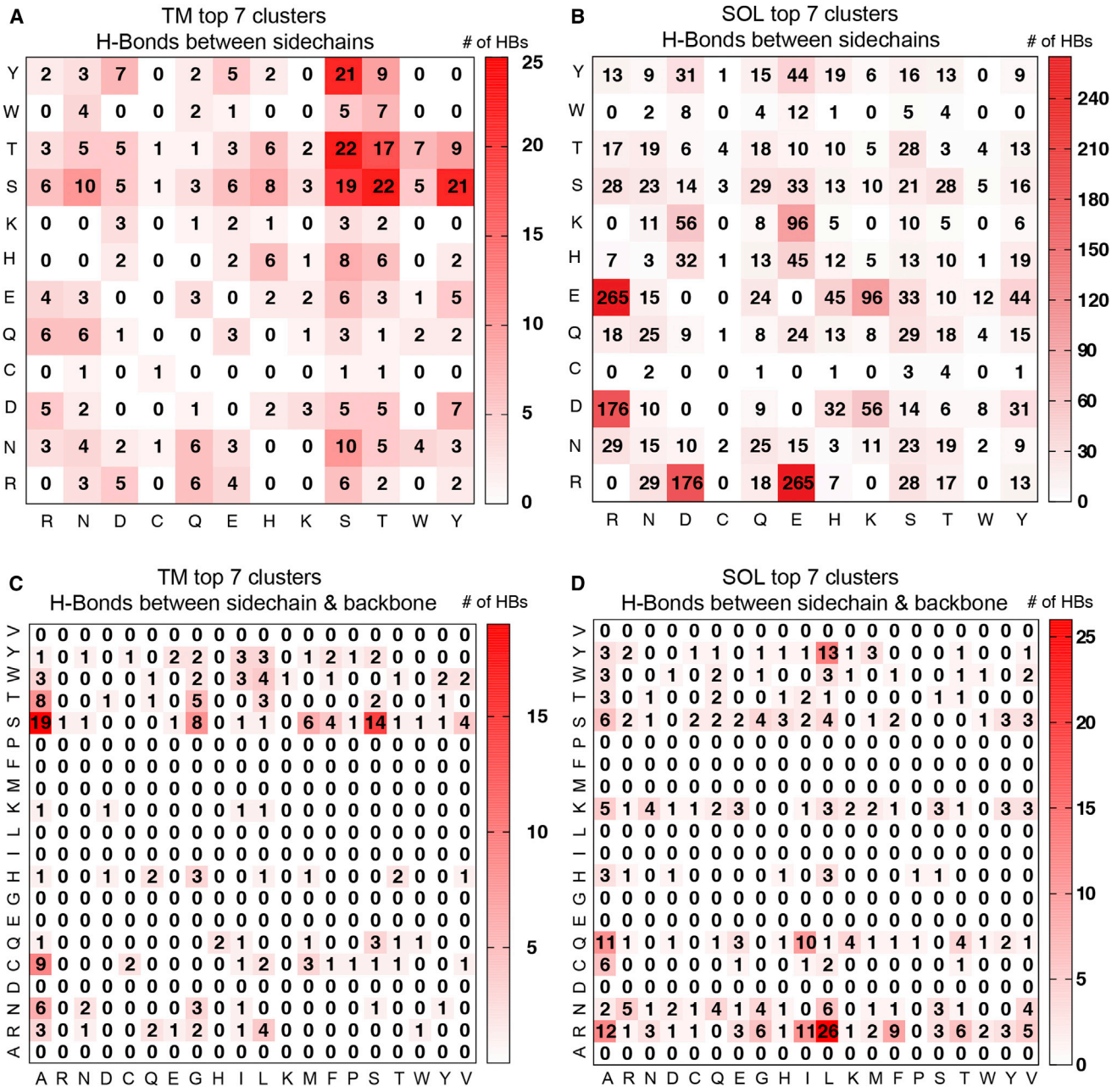



**Figure 8. Number of Interhelical Hydrogen Bonds between the Sidechains of Residues and between Sidechain and the Backbone Carbonyl in the Top Seven TM and SOL Clusters**

(A and B) Number of interhelical hydrogen bonds between the sidechains of residues. The numbers in the grids are the arithmetic summations of the numbers of specific sidechain-to-sidechain hydrogen bonds in the top seven clusters from each category (TM in A; SOL in B).

(C and D) Number of interhelical hydrogen bonds between sidechain and the backbone carbonyl (TM in C; SOL in D). The numbers of hydrogen bonds denote those from the sidechain of the residue on the column to the backbone carbonyl on the residue on the row.

### Creating the Pair Library

Two heuristic criteria were used to determine whether a given pair of helices was interacting. First, the minimum distance between the helical axes was required to be no more than 14 Å; second, the mean inverse distance was required to be at least 0.065 $Å^{-1}$ over a 12-residue window (see Window Selection and Alignment below for a definition of this quantity). Both of these were intended to be generous, as low specificity would merely result in a larger fraction of dimers which cannot be clustered, while low sensitivity would negatively affect our ability to detect and characterize real trends.

Although the overall structural libraries were filtered to reduce sequence homology, individual proteins often contain multiple copies of one or more subunits, resulting in several identical helix pairs. To remove this additional source of redundancy, polypeptide chains with identical sequences were assigned to a "chain group," which allowed us to identify and remove duplicate dimers. Two helices can come from the same chain, different chains, the same chain

group, or separate chains that also belong to disparate chain groups. The final helix pair library contains 2,694 TM dimers and 5,085 soluble dimers.

### Window Selection and Alignment

To be able to align pairs, we used a distance map representation of each dimer. In brief, the inverse distance between each Cα atom on one helix and every Cα atom on the other is stored in a matrix. (Residues more than 25 Å apart are given a value of 0.) We selected a 12-residue segment from each helix, chosen so that we captured the maximum amount of interaction for a given pair. Interaction strength was determined by averaging the interfacial distance map over a 12-residue window on each helix, as calculated using Equation 1:

$$M = \frac{1}{n^2} \sum_{i=a}^{a+n-1} \sum_{j=b}^{b+n-1} x_{ij}, \qquad \text{(Equation 1)}$$

where $M$ is the mean inverse distance, or interaction strength, $n$ is the window size (here 12 residues), $a$ and $b$ are the starting residues of the window on each helix, respectively, and $x_{ij}$ is the value of the distance map for residues $i$ and $j$, i.e. the inverse of the distance between the Cα atoms of residues $i$ and $j$ (in angstroms) or zero if they are more than 25 Å apart. $M$ was maximized by varying $a$ and $b$ over all possible values, from 1 to $L - n + 1$, where $L$ is the length of the particular helix. Since residues that are closer together in three dimensions have a larger entry in the distance map, this picks out the 12 residues on one helix that are closest to 12 residues on the other. Moreover, because of the inverse weighting, this emphasizes each residue's nearest neighbors, with the distances between the end of one helix and the far end of the other being less important.

We used MaDCaT (Zhang and Grigoryan, 2013) to conduct all-versus-all searches of the two dimer libraries. Interactions are not always symmetrical along the length of a helix, with six residues on either side of the point of closest approach: some are V-shaped rather than X-shaped. Thus had we merely compared the 12-residue windows with each other directly, we would have missed pairs that otherwise have the same geometry. We therefore searched each query window against the library of whole pairs, as extracted above. We limited the searches to a maximum of 10,000 hits each, which in practice exhausted all possible alignments within our clustering threshold.

### Structural Clustering

Examining the alignments calculated by MaDCaT, we chose a 1.25 Å RMSD cutoff for clustering as an appropriate balance between sensitivity and specificity. We used the same 12-residue windows described above; windows which overlapped by six residues or more on either helix were considered identical and clustered together, while windows with smaller overlaps were treated separately. This allows the total number of alignments to be greater than the number of unique pairs. To cluster the pairs, we computed all possible subthreshold alignments to each window. The window with the largest number of alignments from unique, previously unclustered pairs was selected as the next centroid. All matching windows were assigned to that cluster and removed from consideration for further rounds. This process was then repeated until none of the remaining windows matched at least ~1% of the associated database (25 pairs for TM and 55 pairs for SOL).

We found 16 TM clusters and 15 SOL clusters of helix pairs. Geometrical properties, including crossing angle and interhelical distance of the aligned windows in each cluster, were determined by HELANAL (Bansal et al., 2000) implemented by MSL (Kulp et al., 2012). Mean geometric properties (Figure 2; Tables S1 and S2) of each cluster were determined by the subset of pairs that fall within the most populated 12-residue window on the centroid. These same windows were those used to cluster, and are the subject of sequence, hydrophobicity, and hydrogen-bonding analysis (Figures 3, 4, 5, 6, and 7). The detailed information for TM and SOL clusters about the structural composition, RMSD to the centroids, interhelical distance, and crossing angle is provided as two spreadsheet files in the Supplemental Information.

### Comparing Clusters

For each centroid, we determined the 15-residue window that is most populated by members of that cluster. To compare clusters, we then used MaDCaT to find the best possible alignment of 12 residues between each pair of centroids approximate to those regions. This information allowed us to identify the most closely related clusters from different sets. The centroid of each cluster was fit to a sinusoidal curve using non-linear regression to estimate the cluster's periodicity. A two-tailed Student's t test assuming equal variances was performed to confirm that periods within the matching windows between TM and SOL were not significantly different.

### Sequence Analysis

We used the structural alignments generated by MaDCaT for each cluster to create sequence alignments. In brief, each centroid pair was renumbered so that the C-terminal residue of the centroid window would be residue 100. Each member of a cluster was then renumbered to match the centroid numbering, such that residues with the same number corresponded in the structural alignment. The numbers of observations for every amino acid type were computed for each position in each cluster and normalized to frequencies by dividing by the total number of observations at that position. The frequencies were compared with the expected frequencies of amino acids in helical regions of TM or SOL proteins that form interacting helical pairs using a binomial distribution. We derived the expected frequency of TM amino acids from the percent distribution of amino acids observed at helical, TM residues in the subset of our TM protein data set that formed interacting pairs. Likewise, only α-helical residues from the analogous SOL subset, determined by the DSSP Program (Kabsch and Sander, 1983), were observed in deriving the SOL amino acid distribution. These background frequencies are listed in Table S4. The propensity is defined as the ratio between the observed and expected (or background) frequencies. Significant overrepresentation or underrepresentation of an amino acid at a given position, relative to the expected frequency, was determined by the p value of respective one-tailed directional binomial tests. The counts of observation, frequency, and propensity for each amino acid on the positions with at least 25 and 55 total counts of observation for TM and SOL clusters, respectively, are provided as two spreadsheet files in the Supplemental Information. Hydrophobicity profiles were calculated based on the normalized consensus scale (Eisenberg et al., 1984).

### Hydrogen-Bonding Analysis

Hydrogen bonds were determined by the HBPLUS program (McDonald and Thornton, 1994) with default parameters. Weak Cα-H-O hydrogen bonds are not included. Two set of hydrogen bond data on positions *a* and *d* on the most populated region from each helix were used to calculate the hydrogen-bonding fraction, which is defined as the ratio between the numbers of residues forming interhelical hydrogen bonds and of the population accumulated on the four positions both for *a* and *d*. The hydrogen-bonding connectivity was calculated by assigning the interhelically hydrogen-bonded residues in the heptad or tetrad repeats from the most populated positions *a* and *d* from both chains. The sidechain-to-sidechain interhelical hydrogen-bonding propensity is calculated as the ratio between the fraction of Arg, Asn, Asp, Cys, Gln, Glu, His, Lys, Ser, Thr, Trp, and Tyr to make sidechain-to-sidechain hydrogen bonds and their fraction in the subset of background distribution (Table S4). Significant overrepresentation or underrepresentation of an amino acid to participate in a hydrogen bond was determined by the binomial test.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes four tables, two figures, and five supplemental spreadsheets and can be found with this article online at http://dx.doi.org/10.1016/j.str.2015.01.009.

## AUTHOR CONTRIBUTIONS

S.Q.Z., D.W.K., I.S., and W.F.D. designed the research; S.Q.Z., D.W.K., M.M., C.A.S., and I.S. performed the research; S.Q.Z., M.M., I.S., and W.F.D. analyzed the data; all authors wrote the paper.

## ACKNOWLEDGMENTS

We thank B.T. Hannigan and G. Gonzales for technical help. M.M. was supported by NIH T32 GM008284. This work was supported by NIH grant GM54616.

Received: July 31, 2014
Revised: December 17, 2014
Accepted: January 6, 2015
Published: February 19, 2015

## REFERENCES

Acharya, A., Rishi, V., and Vinson, C. (2006). Stability of 100 homo and heterotypic coiled-coil a-a' pairs for ten amino acids (A, L, I, V, N, K, S, T, E, and R). Biochemistry *45*, 11324–11332.

Adamian, L., and Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. Proteins *47*, 209–218.

Baker, E., and Hubbard, R. (1984). Hydrogen bonding in globular proteins. Progr. Biophys. Mol. Biol. *44*, 97–179.

Bansal, M., Kumar, S., and Velavan, R. (2000). HELANAL: a program to characterize helix geometry in proteins. J. Biomol. Struct. Dyn. *17*, 811–819.

Barth, P., Wallner, B., and Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. Proc. Natl. Acad. Sci. USA *106*, 1409–1414.

Benjamini, A., and Smit, B. (2012). Robust driving forces for transmembrane helix packing. Biophys. J. *103*, 1227–1235.

Berger, B.W., Kulp, D.W., Span, L.M., DeGrado, J.L., Billings, P.C., Senes, A., Bennett, J.S., and DeGrado, W.F. (2010). Consensus motif for integrin transmembrane helix association. Proc. Natl. Acad. Sci. USA *107*, 703–708.

Borders, C.L., Jr., Broadwater, J.A., Bekeny, P.A., Salmon, J.E., Lee, A.S., Eldridge, A.M., and Pett, V.B. (1994). A structural role for arginine in proteins: multiple hydrogen bonds to backbone carbonyl oxygens. Protein Sci. *3*, 541–548.

Bowie, J.U. (2005). Solving the membrane protein folding problem. Nature *438*, 581–589.

Bowie, J.U. (2011). Membrane protein folding: how important are hydrogen bonds? Curr. Opin. Struct. Biol. *21*, 42–49.

Braun, P., Goldberg, E., Negron, C., von Jan, M., Xu, F., Nanda, V., Koder, R.L., and Noy, D. (2011). Design principles for chlorophyll-binding sites in helical proteins. Proteins *79*, 463–476.

Chothia, C., Levitt, M., and Richardson, D. (1977). Structure of proteins: packing of alpha-helices and pleated sheets. Proc. Natl. Acad. Sci. USA *74*, 4130–4134.

Crick, F.H.C. (1953a). The Fourier transform of a coiled-coil. Acta Crystallogr. *6*, 685–689.

Crick, F.H.C. (1953b). The packing of alpha-helices: simple coiled-coils. Acta Crystallogr. *6*, 689–697.

Cross, T.A., Murray, D.T., and Watts, A. (2013). Helical membrane protein conformations and their environment. Eur. Biophys. J. *42*, 731–755.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., and Michel, H. (1984). X-ray structure analysis of a membrane protein complex. Electron density map at 3 A resolution and a model of the chromophores of the photosynthetic reaction center from *Rhodopseudomonas viridis*. J. Mol. Biol. *180*, 385–398.

Eilers, M., Patel, A.B., Liu, W., and Smith, S.O. (2002). Comparison of helix interactions in membrane and soluble alpha-bundle proteins. Biophys. J. *82*, 2720–2736.

Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. *179*, 125–142.

Gernert, K.M., Surles, M.C., Labean, T.H., Richardson, J.S., and Richardson, D.C. (1995). The Alacoil: a very tight, antiparallel coiled-coil of helices. Protein Sci. *4*, 2252–2260.

Ghirlanda, G. (2009). Design of membrane proteins: toward functional systems. Curr. Opin. Chem. Biol. *13*, 643–651.

Gimpelev, M., Forrest, L.R., Murray, D., and Honig, B. (2004). Helical packing patterns in membrane and soluble proteins. Biophys. J. *87*, 4075–4086.

Gratkowski, H., Dai, Q.H., Wand, A.J., DeGrado, W.F., and Lear, J.D. (2002). Cooperativity and specificity of association of a designed transmembrane peptide. Biophys. J. *83*, 1613–1619.

Greene, L.H., Lewis, T.E., Addou, S., Cuff, A.L., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., et al. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. *35*, D291–D297.

Grigoryan, G., and DeGrado, W.F. (2011). Probing designability via a generalized model of helical bundle geometry. J. Mol. Biol. *405*, 1079–1100.

Han, Q., Aligo, J., Manna, D., Belton, K., Chintapalli, S.V., Hong, Y., Patterson, R.L., van Rossum, D.B., and Konan, K.V. (2011). Conserved GXXXG- and S/T-like motifs in the transmembrane domains of NS4B protein are required for hepatitis C virus replication. J. Virol. *85*, 6464–6479.

Hedin, L.E., Illergard, K., and Elofsson, A. (2011). An introduction to membrane proteins. J. Proteome Res. *10*, 3324–3331.

Herrmann, J.R., Panitz, J.C., Unterreitmeier, S., Fuchs, A., Frishman, D., and Langosch, D. (2009). Complex patterns of histidine, hydroxylated amino acids and the GxxxG motif mediate high-affinity transmembrane domain interactions. J. Mol. Biol. *385*, 912–923.

Javadpour, M.M., Eilers, M., Groesbeek, M., and Smith, S.O. (1999). Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. Biophys. J. *77*, 1609–1618.

Joh, N.H., Oberai, A., Yang, D., Whitelegge, J.P., and Bowie, J.U. (2009). Similar energetic contributions of packing in the core of membrane and water-soluble proteins. J. Am. Chem. Soc. *131*, 10846–10847.

Joo, H., Chavan, A.G., Phan, J., Day, R., and Tsai, J. (2012). An amino acid packing code for alpha-helical structure and protein design. J. Mol. Biol. *419*, 234–254.

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

Kulp, D.W., Subramaniam, S., Donald, J.E., Hannigan, B.T., Mueller, B.K., Grigoryan, G., and Senes, A. (2012). Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). J. Comput. Chem. *33*, 1645–1661.

Langosch, D., and Arkin, I.T. (2009). Interaction and conformational dynamics of membrane-spanning protein helices. Protein Sci. *18*, 1343–1358.

Langosch, D., Herrmann, J.R., Unterreitmeier, S., and Fuchs, A. (2010). Helix-helix interaction patterns in membrane proteins. In Structural Bioinformatics of Membrane Proteins, D. Frishman, ed. (Springer), pp. 165–186.

Lawrie, C.M., Sulistijo, E.S., and MacKenzie, K.R. (2010). Intermonomer hydrogen bonds enhance GxxxG-driven dimerization of the BNIP3 transmembrane domain: roles for sequence context in helix-helix association in membranes. J. Mol. Biol. *396*, 924–936.

Lemmon, M.A., Flanagan, J.M., Hunt, J.F., Adair, B.D., Bormann, B.J., Dempsey, C.E., and Engelman, D.M. (1992). Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. J. Biol. Chem. *267*, 7683–7689.

Lemmon, M.A., Treutlein, H.R., Adams, P.D., Brunger, A.T., and Engelman, D.M. (1994). A dimerization motif for transmembrane alpha-helices. Nat. Struct. Biol. *1*, 157–163.

Liang, J. (2002). Experimental and computational studies of determinants of membrane-protein folding. Curr. Opin. Chem. Biol. *6*, 878–884.

Liu, Y., Engelman, D.M., and Gerstein, M. (2002). Genomic analysis of membrane protein families: abundance and conserved motifs. Genome Biol. *3*, research0054.

Lo, A., Cheng, C.W., Chiu, Y.Y., Sung, T.Y., and Hsu, W.L. (2011). TMPad: an integrated structural database for helix-packing folds in transmembrane proteins. Nucleic Acids Res. *39*, D347–D355.

Lomize, M.A., Pogozheva, I.D., Joo, H., Mosberg, H.I., and Lomize, A.L. (2012). OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. *40*, D370–D376.

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

Martin, J., Letellier, G., Marin, A., Taly, J.F., de Brevern, A.G., and Gibrat, J.F. (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct. Biol. *5*, 17.

Mason, J.M., and Arndt, K.M. (2004). Coiled coil domains: stability, specificity, and biological implications. Chembiochem *5*, 170–176.

McClain, D.L., Binfet, J.P., and Oakley, M.G. (2001). Evaluation of the energetic contribution of interhelical coulombic interactions for coiled coil helix orientation specificity. J. Mol. Biol. *313*, 371–383.

McClain, D.L., Gurnon, D.G., and Oakley, M.G. (2002). Importance of potential interhelical salt-bridges involving interior residues for coiled-coil stability and quaternary structure. J. Mol. Biol. *324*, 257–270.

McDonald, I.K., and Thornton, J.M. (1994). Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. *238*, 777–793.

Mueller, B.K., Subramaniam, S., and Senes, A. (2014). A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Calpha-H hydrogen bonds. Proc. Natl. Acad. Sci. USA *111*, E888–E895.

Ng, D.P., Poulsen, B.E., and Deber, C.M. (2012). Membrane protein misassembly in disease. Biochim. Biophys. Acta *1818*, 1115–1122.

Nugent, T., and Jones, D.T. (2012). Membrane protein structural bioinformatics. J. Struct. Biol. *179*, 327–337.

Oakley, M.G., and Kim, P.S. (1998). A buried polar interaction can direct the relative orientation of helices in a coiled coil. Biochemistry *37*, 12603–12610.

Oberai, A., Joh, N.H., Pettit, F.K., and Bowie, J.U. (2009). Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. Proc. Natl. Acad. Sci. USA *106*, 17747–17750.

Perez-Aguilar, J.M., and Saven, J.G. (2012). Computational design of membrane proteins. Structure *20*, 5–14.

Russ, W.P., and Engelman, D.M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. J. Mol. Biol. *296*, 911–919.

Sal-Man, N., Gerber, D., Bloch, I., and Shai, Y. (2007). Specificity in transmembrane helix-helix interactions mediated by aromatic residues. J. Biol. Chem. *282*, 19753–19761.

Samish, I., Macdermaid, C.M., Perez-Aguilar, J.M., and Saven, J.G. (2011). Theoretical and computational protein design. Annu. Rev. Phys. Chem. *62*, 129–149.

Senes, A., Gerstein, M., and Engelman, D.M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J. Mol. Biol. *296*, 921–936.

Senes, A., Ubarretxena-Belandia, I., and Engelman, D.M. (2001). The Calpha -H…O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. Proc. Natl. Acad. Sci. USA *98*, 9056–9061.

Senes, A., Engel, D.E., and DeGrado, W.F. (2004). Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. Curr. Opin. Struct. Biol. *14*, 465–479.

Shai, Y. (2001). Molecular recognition within the membrane milieu: implications for the structure and function of membrane proteins. J. Membr. Biol. *182*, 91–104.

Sodek, J., Hodges, R.S., Smillie, L.B., and Jurasek, L. (1972). Amino-acid sequence of rabbit skeletal tropomyosin and its coiled-coil structure. Proc. Natl. Acad. Sci. USA *69*, 3800–3804.

Talbot, J.A., and Hodges, R.S. (1982). Tropomyosin: a model protein for studying coiled-coil and a-helical stabilization. Acc. Chem. Res. *15*, 224–230.

Tatko, C.D., Nanda, V., Lear, J.D., and Degrado, W.F. (2006). Polar networks control oligomeric assembly in membranes. J. Am. Chem. Soc. *128*, 4170–4171.

Tusnady, G.E., Dosztanyi, Z., and Simon, I. (2005). PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res. *33*, D275–D278.

Unterreitmeier, S., Fuchs, A., Schaffler, T., Heym, R.G., Frishman, D., and Langosch, D. (2007). Phenylalanine promotes interaction of transmembrane domains via GxxxG motifs. J. Mol. Biol. *374*, 705–718.

Varriale, S., Merlino, A., Coscia, M.R., Mazzarella, L., and Oreste, U. (2010). An evolutionary conserved motif is responsible for immunoglobulin heavy chain packing in the B cell membrane. Mol. Phylogenet. Evol. *57*, 1238–1244.

Wallin, E., and von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci. *7*, 1029–1038.

Walters, R.F., and DeGrado, W.F. (2006). Helix-packing motifs in membrane proteins. Proc. Natl. Acad. Sci. USA *103*, 13658–13663.

Wang, G., and Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589–1591.

Wei, P., Liu, X., Hu, M.-H., Zuo, L.-M., Kai, M., Wang, R., and Luo, S.-Z. (2011). The dimerization interface of the glycoprotein Ibβ transmembrane domain corresponds to polar residues within a leucine zipper motif. Protein Sci. *20*, 1814–1823.

White, S.H. (2009). Biophysical dissection of membrane proteins. Nature *459*, 344–346.

Woolfson, D.N. (2005). The design of coiled-coil structures and assemblies. Adv. Protein Chem. *70*, 79–112.

Zhang, J., and Grigoryan, G. (2013). Mining tertiary structural motifs for assessment of designability. Methods Enzymol. *523*, 21–40.

Zhang, Y., Kulp, D.W., Lear, J.D., and DeGrado, W.F. (2009). Experimental and computational evaluation of forces directing the association of transmembrane helices. J. Am. Chem. Soc. *131*, 11341–11343.

Zhou, F.X., Merianos, H.J., Brunger, A.T., and Engelman, D.M. (2001). Polar residues drive association of polyleucine transmembrane helices. Proc. Natl. Acad. Sci. USA *98*, 2250–2255.