Structural Determinants of Transmembrane Helical Proteins

Susan E. Harrington[1] and Nir Ben-Tal[1,*]
[1]Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
*Correspondence: NirB@tauex.tau.ac.il

DOI 10.1016/j.str.2009.06.009

SUMMARY

We identify a structural feature of transmembrane helical proteins that restricts their conformational space and suggests a new way of understanding the construction and stability of their native states. We show that five kinds of well-known specific favorable interhelical interactions (hydrogen bonds, aromatic interactions, salt bridges, and two interactions from packing motifs) precisely determine the packing of the transmembrane helices in 15 diverse proteins. To show this, we iteratively reassemble the helix bundle of each protein using only these interactions, generic interaction geometries, and individual helix backbone conformations. On average, the representative set of rebuilt structures best satisfying the constraints imposed by the five types of interhelical interactions has an average Cα root-mean-square deviation from the native of 1.03 Å. Implications for protein folding, structure and motion prediction, modeling, and design are discussed.

INTRODUCTION

Transmembrane (TM) helical proteins play critical and diverse roles in the lives of cells. Structure and motion prediction techniques for TM helical proteins are important because of the dynamic nature of these proteins and the difficulty of experimentally solving their structures. There is as yet no sure general method for finding all of the relevant low-energy states of TM helical proteins. Because global minimization is notoriously intractable in general, this would remain true even if we had a perfect potential energy function to minimize (Acton, 1990). To limit the conformational space that needs to be searched, diverse approaches have been developed. These include using various kinds of experimental data like those obtained from electron paramagnetic resonance and cryo-electron microscopy (Liu et al., 2001; Perozo et al., 2002; Sale et al., 2004; Baldwin et al., 1997; Beuming and Weinstein, 2005; Fleishman et al., 2006; Kovacs et al., 2007; Alber et al., 2007), modifying knowledge-based sampling techniques used successfully for soluble proteins (Barth et al., 2007), and characterizing the important conformational restrictions imposed by simple packing conditions (Bowie, 1997; DeGrado et al., 2003; Walters and DeGrado, 2006). In this paper, we describe a consistent structural feature of TM helical proteins that not only greatly limits their conformational space, but also reveals new ways of understanding their folding and stability. Before we describe this feature, we need to give the motivation for our approach.

Central aspects of the folding of TM helical proteins are thought to be well understood (Bowie, 2005; Fleishman and Ben-Tal, 2002; Popot and Engelman, 1990; von Heijne, 1996). The well-accepted two-stage model proposes modular folding where TM helices form first and then associate to form helix bundles (Popot and Engelman, 1990).

The physical forces driving the next stage of folding, helix-helix association, have also been investigated, and some specific interhelical interactions have been shown to play critical roles. Among the critical favorable interhelical interactions that have been discovered are closely packed small residues (Bowie, 2005; Fleishman and Ben-Tal, 2002; Lemmon and Engelman, 1994; Russ and Engelman, 2000; Senes et al., 2000; Schneider and Engelman, 2004; Finger et al., 2006), hydrogen bonds (Bowie, 2005; Fleishman and Ben-Tal, 2002; Zhou et al., 2000; Gratkowski et al., 2001; Dawson et al., 2003), salt bridges (Honig and Hubbell, 1984), aromatic interactions (Dougherty, 1996; Sal-Man et al., 2007; Johnson et al., 2007), and closely packed valines, isoleucines, and leucines, especially when in contact with other valines, isoleucines, and leucines (Fleishman and Ben-Tal, 2002; Lemmon and Engelman, 1994; Senes et al., 2000; Gurezka et al., 1999). These interactions have been experimentally found to drive helix-helix association, sometimes in certain motif contexts (like the famous GxxxG), and additional experiments and theoretical analysis have further supported their particular importance to the stability of TM helical proteins (Bowie, 2005; Fleishman and Ben-Tal, 2002; Lemmon and Engelman, 1994; Lemmon and Engelman, 1994; Russ and Engelman, 2000; Senes et al., 2000; Schneider and Engelman, 2004; Finger et al., 2006; Zhou et al., 2000; Gratkowski et al., 2001; Dawson et al., 2003; Honig and Hubbell, 1984; Dougherty, 1996; Sal-Man et al., 2007; Johnson et al., 2007; Gurezka et al., 1999).

By inspecting solved structures, we found that these five kinds of particularly favorable specific interhelical interactions (hydrogen bonds, salt bridges, aromatic interactions, small residue close packing, and Ile/Val/Leu close packing in Ile/Val/Leu patches) seemed to be consistently distributed so as to nearly fix the helix bundle backbones. We also saw that these five types of interactions appeared to be distributed so that the helix bundles could be iteratively assembled using only these interactions; i.e., by taking the individual helix backbone conformations to start with and by successively putting together helices and later subbundles (helix pairs, triples, etc.) using only these interactions to build the full bundle. Additionally, it seemed that this iterative assembly could be done in a sequence order preserving fashion: i.e., only sequence-adjacent helices or subbundles needed to be assembled together during the assembly process.

This iterative assembly and its order relates to another aspect of TM helical protein folding. Recently, much attention has been focused on the translocon and the mechanism by which it determines which proteins are inserted into the membrane (Bowie, 2005; White and von Heijne, 2008). The specifics of this are being studied, but the translocon can influence the contact order of TM helices as it inserts them into the membrane in sequence order (Sadlish et al., 2005). It is natural to speculate about how active a role it might play in the folding of TM helical proteins and how this affects their final conformations.

These different ideas about folding and stability come together in our study of the five kinds of specific interhelical interactions. Each such energetically favorable interaction can be described in terms of donor/acceptor pairs, where each donor and acceptor must meet a fixed set of geometric conditions. Thus these sets of interhelical interactions can be considered and studied as sets of constraints.

Our hypothesis is that the five types of interactions are distributed so as to highly constrain the backbone in native structures, and we have developed a computational method to test the extent to which they do. At the same time, we test the idea that the helix bundle can be assembled iteratively in a sequence order preserving fashion using only these interactions. This type of iterative assembly process can be seen as consistent with the two-stage model and as a generalized version of helix-helix association (i.e., subbundle-subbundle association) that is driven by the same kinds of interactions. The sequence order preserving property of the assembly is why it can also be considered as consistent with translocon-aided folding.

A priori there is no reason to expect that a native structure must necessarily have even one interhelical interaction of one of the five types. But for our test set we show that the sets of these types of interhelical interactions are in fact highly constraining and that if we significantly perturb the helix positions in the native structures we necessarily break some interactions in the set. Thus one can say a complete set of determining constraints/interactions of the five types has evolved to nearly fix each native state backbone conformation. We call sets of the five types of interhelical interactions “determining sets” when they nearly fix geometrically the packing of the helix backbones as described above. (Our terminology is based on the geometric meaning of “determine”: to specify position, to fix.) Interpretations of the determining sets of interactions can differ; the geometric fact that they exist for every protein in our test set is the chief result of this paper.

The Computational Approach

A detailed description of the computational methods, including the algorithms that were developed, is provided in Experimental Procedures and Supplemental Data (available online), but the general approach is briefly introduced here.

We consider five kinds of specific interhelical interactions, all of which lie within or very close to the inferred hydrocarbon region (see Figure 1 for examples). Three are polar: hydrogen bonds, salt bridges, and some aromatic interactions. Two are packing interactions: Gly, Ala, or Ser close knob-in-hole packing and Ile/Val/Leu close knob-in-hole packing in Ile/Val/Leu contact patches.

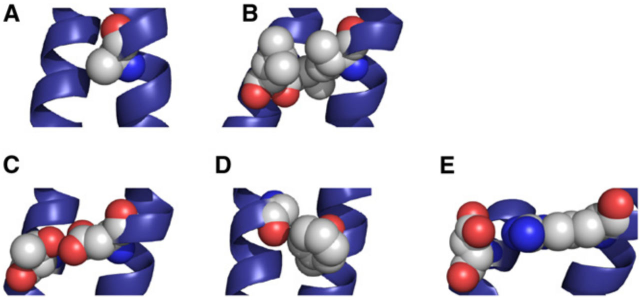

Figure 1. The Five Types of Interactions
Examples from native structures. The helices are shown as cartoons and the atoms of key residues in the interactions are shown as spheres. Carbon atoms are gray, oxygen atoms are red, and nitrogen atoms blue.
(A) Small residue packing, close knob-in-hole. The knob atoms are shown as spheres.
(B) I/V/L packing, close knob-in-hole. The knob is on the right and its atoms are shown as spheres. The side chains displayed on the left are from the surrounding hole residues of I/V/L type.
(C) Hydrogen bond.
(D) Aromatic interaction.
(E) Salt bridge.
This and all subsequent figures were produced using PyMOL (Delano, 2002).

For each interaction type, we have defined a fixed interaction geometry. These interaction geometries are sets of geometric conditions that must be satisfied if the donor and acceptor are interacting (e.g., the usual conditions for a hydrogen bond; for the two packing interactions, the knob is considered the donor and the hole the acceptor). They are not derived from the solved structures and do not vary from protein to protein. For any fixed donor, there is an associated region derived from the geometric conditions where any potential acceptor participating in that given type of interaction must lie. Likewise, for any acceptor, there is an associated region where any potential donor must lie. These interaction regions are used during reassembly.

We selected a diverse set of 15 proteins with known high-resolution structures. They were chosen for their diversity and high resolution, usually less than 3 Å, so that the interactions could be fairly unambiguously read from the structures. The largest number of helices in a protein in our test set is 40 and the smallest is two; the median is seven helices.

From each solved structure, we take the backbone conformations of the individual helices in the hydrocarbon region of the membrane, the set of the five kinds of interhelical interactions (not the geometry of those particular contacts), and the side chain conformations of those residues with a side chain atom explicitly in an interhelical interaction of one of the five types. Those native side chain conformations are fixed and rigid. The side chains are not taken for residues in the two packing interactions. All other side chains have a fixed, rigid, reduced representation based on residue type that is intended to give the obstruction created by a side chain of that residue type irrespective of rotameric state (Figure 2); they are not derived from the native structure. Loops, water, and ligands are not used.

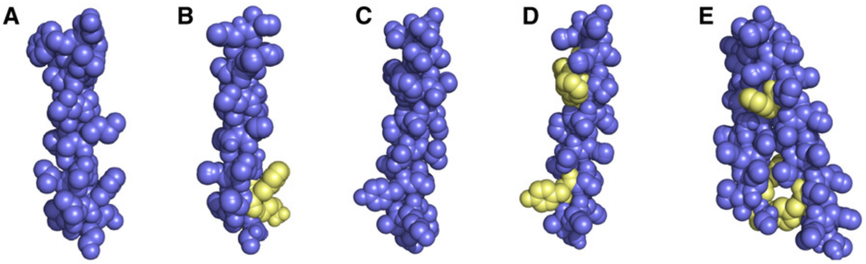

Figure 2. Reduced Representatives of Side Chains

(A) The first helix of bacteriorhodopsin (bR; PDB id 1C3W) with native side chains. All atoms are shown as spheres.

(B) The first helix of bR with the side chain representatives used during reassembly of the helix pair made up of the first two helices. Residues with a reduced side chain representative are shown in blue. The two residues whose native side chains were used are shown in yellow; each acted as a donor in one of the five types of interhelical interactions.

(C) The second helix of bR with native side chains.

(D) The second helix of bR used during reassembly of the first helix pair with reduced side chains shown in blue and the three native side chains used shown in yellow.

(E) Reassembled pair of the first two helices with the only native side chains used shown in yellow.

Sketch of Overall Assembly and the Basic Rigid Motion Assembly of Two Pieces

We use rigid motions to iteratively reassemble the helix bundle backbone of each protein using only its set of the five types of interhelical interactions, predefined interaction geometries, and individual helix backbones. Beginning with N rigid separate pieces, initially the individual helices, we fit two together and so obtain a new set of N-1 rigid pieces. After repeating this N-1 times, we will have one piece at the end, the assembled structure.

The assembly of two rigid pieces using the interactions of the five types between them is based on the following geometry. Any rigid body's position in space can be specified by the positions of any three noncollinear points on the body. If the exact positions of those three points are unknown, but we do know that they must each lie within three given regions in space, then we can obtain an initial ensemble of positions of the body by placing grids on those three regions and systematically selecting these points to give the positions of the three points on the body. These three positions then fix the position of the body itself. If there are additional restrictions on the positions in space of any other points on the body, then we can check the initial ensemble of positions of the body and remove any positions from the ensemble that do not meet those restrictions. By choosing sufficiently fine grids, one can find to any desired accuracy how the specified regions constrain the position of the body (see S1 and Figure S1 in the Supplemental Data for details).

This approach can be adapted to build the combined piece ensemble of two pieces constrained by the set of the five types of interactions between them. We use three of these interactions and their associated interaction regions to position one piece relative to the other and build the initial combined piece ensemble. If there are more than three of the interactions of these types (as is usually the case) or additional geometric conditions (as is the case, e.g., for a hydrogen bond), the initial conformations built using the three interactions are then checked and discarded if they do not have the additional interactions or meet the additional geometric conditions.

Thus three interactions between the two pieces are necessary to build the combined piece ensemble, and it turns out that four interactions of the types we consider will usually constrain the conformations quite well.

Decomposition of the Set of the Five Types of Interactions and Order of Assembly

Which pieces should be joined together and in what order? Our answer is intended to give a well-constructed final structure and a plausible translocon-guided folding pathway, but the answer is far from unique.

For each protein, we decompose the set of the five kinds of interhelical interactions found in the solved structure using the algorithm given in Experimental Procedures. That is, we decide the pieces we will put together and the order in which we will do this. The algorithm is designed to attempt to iteratively reassemble the entire structure by putting together sequence adjacent pieces starting from the beginning of the sequence, where the initial pieces are the individual helices. There must be enough interactions at each step between the two pieces to build the combined piece ensemble. For most pieces, we require four interactions; if one of the pieces is a half-helix, we require three.

For example, here is our iterative assembly for the voltage sensor, 1ORS. At each step, we put two rigid pieces together using the interhelical interactions between them to produce a new fixed piece (actually an ensemble as described before), as shown in Figure 3. We first assemble the first two helices, then add 3-a to the single piece (1 2). At this point, there are insufficiently many interactions to add 3-b to the first piece ((1 2) 3-a), so we next assemble 3-b and 4, and finally put ((1 2) 3-a) and (3-b 4) together to obtain the full structure (((1 2) 3-a) (3-b 4)).

Selection of Substructures to Propagate and Measure of Structural Variability

Once the assembly order is decided, we can iteratively build the structure. When two pieces are put together, the result is an ensemble of conformations that satisfy well the geometric conditions of the interhelical interactions of the five types. It is not possible to use the whole ensemble during the next round of building because we would in some cases ultimately produce millions of structures for the full bundle. Instead, we took a small random subset of the conformations that met a scoring cutoff to carry to the next step, constraining the sampling to preserve diversity. (The scoring depends only on overlaps and the geometric conditions imposed by the interactions. It does not approximate energy.)

We report the maximum Cα root-mean-square deviation (rmsd) between the full structure ensemble members with

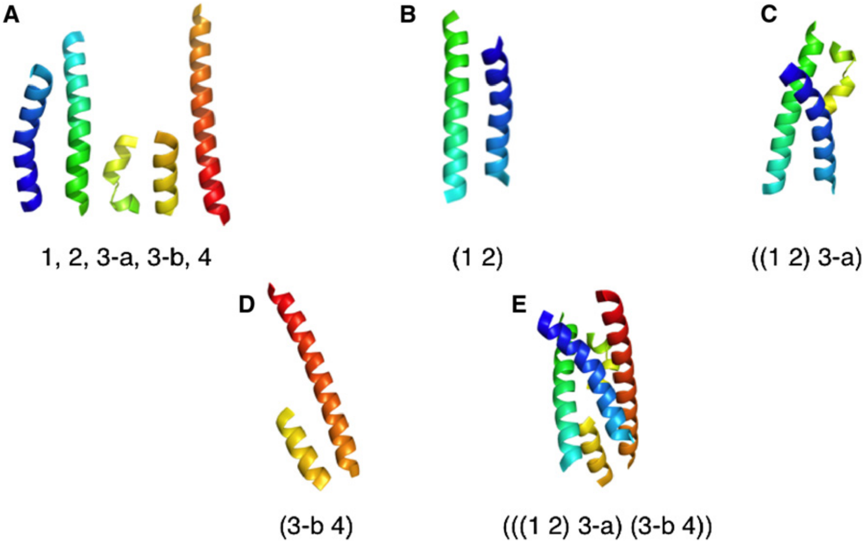

Figure 3. Assembly Order for Voltage Sensor 1ORS

An example of our iterative sequence order respecting assembly. We begin with the individual helices 1, 2, 3-a, 3-b, 4, and assemble iteratively in the order shown.

(A) The helices to be assembled in sequence order.

(B) The first two helices assembled, (1 2).

(C) The third (half) helix is assembled with the first two ((1 2) 3-a).

(D) The last two helices assembled together (3-b 4).

(E) The piece made up of the first three helices and the piece made up of the last two are assembled together to build the full structure (((1 2) 3-a) (3-b 4)).

good scores, and the maximum and minimum Cα rmsds of structures in that ensemble to the native.

RESULTS

The Interhelical Interactions of the Five Types Determine the Packing of Helices in Native Structures

Statistics on the ensembles of reassembled structures constructed for each of the 15 proteins in our data set are given in Table 1. On average for our test set, the ensemble average Cα rmsd from the native is 1.03 Å, with average best and worst Cα rmsd from the native 0.93 Å and 1.17 Å. The average maximum interstructure Cα rmsd for ensemble members is 0.68 Å. Thus the sets of the interhelical interactions of the five types are indeed highly constraining for the proteins in our test set.

For the proteins studied, the average number of interactions between pieces used in reassembly was 6.94. This high number of interactions explains why the substructures are highly constrained on their own and why the iterative approach to rebuilding is successful. The average fraction of helix residues used in the determining set of interactions was 20.9%, and the average fraction of helix residues with interacting side chains taken from the native was 10.6%.

The interhelical interactions of the five types are displayed in three native structures in the upper panels of Figure 4. The lower panels of Figure 4 each show the worst member (in terms of rmsd) of the final ensemble of structures best satisfying the constraints imposed by these types of interhelical interactions aligned with the native. Even for the worst structures, the differences are hardly visible.

The types of interactions used are necessary: if we delete one type from the list, there are many helix bundle backbones of solved structures that would not be determined by the remaining types of interactions. See Experimental Procedures for examples.

Iterative Sequence Order Respecting Assembly

If the last helix is considered adjacent to the first, all of the proteins in our test set could be rebuilt in an entirely sequence order respecting fashion with the exception of aquaporin. In aquaporin, the half-helices interrupt the sequence-order assembly. If we consider helices separated by half-helices to be adjacent, then aquaporin can also be said to be built in a sequence order respecting fashion. Note that this implies that at every point during the reassembly of the proteins in our test set (with the exception of aquaporin) at least one pair of the sequence adjacent pieces had at least four (three if one of the pieces was a half-helix) interactions of the five types between them; otherwise reassembly would have terminated. See S2 in Supplemental Data for assembly order data and S6 for the interactions themselves. The sequence-order style pairing beginning at the start of the sequence is why the assembly can be interpreted as done in a “translocon-aided” style (Sadlish et al., 2005).

Accumulation of Error and Modularity of Folding

For the proteins studied, the accumulation of error is slow during the iterative assembly. This can be seen from the average Cα rmsds of the rebuilt ensembles from the native (best: 0.93 Å, worst: 1.17 Å, and average: 1.03 Å; Table 1). There are two reasons for this. First, the interactions of the five types highly constrain the substructures (in conjunction with the implicit overlap constraints) so that they are almost always very close to the native and not too diverse. This indicates the underlying modularity of the constraints imposed by the determining sets of these types of interactions. But even on the rare occasions when a substructure’s conformations are more variable, subsequent interface interactions will often not be compatible with the conformations less similar to the native. Thus those dissimilar to the native cannot propagate.

The pentamer 2OAR (TuMscL, large-conductance mechanosensitive channel) provides a good example of this phenomenon. The monomer pair of helices is the worst constrained (by its set of interhelical interactions) substructure of all the pieces used for our test set (Figure 5). For helix pairs, the conformations meeting the scoring cutoff usually have a Cα rmsd from the native under 1 Å. But we have one monomer helix pair conformation for 2OAR with a very good score that

Table 1. Ensembles Constructed from the Interhelical Interactions of the Five Types

		Final Ensemble Cα Rmsd from Native (Å)			Max. Interstructure	Interface Interactions				% of Helix Nat. sc.
Protein	Type	Min	Max	Avg	Rmsd (Å)	Min	Max	Avg	Helices	Used
1ORS	Voltage sensor	1.03	1.29	1.17	0.54	5	9	6.25	5	11%
1AFO	Glycophorin A	0.42	0.64	0.51	0.95	8	8	8.00	2	0%
3B9W	Rh protein, poss. ammonia channel	0.94	0.98	0.96	0.46	4	14	7.20	11	12%
2BS2	Fumarate reductase	0.90	1.30	1.01	1.04	4	11	6.75	5	13%
2OAR	Mechanosensitive channel, TuMscL	1.27	1.35	1.30	0.51	4	7	5.00	10	6%
2Z73	Rhodopsin, GPCR	1.09	1.33	1.21	0.70	4	10	6.33	7	12%
2RH1	β2-Adrenergic receptor, GPCR	0.99	1.14	1.05	0.74	4	8	6.00	7	11%
1C3W	Bacteriorhodopsin	1.09	1.19	1.14	0.33	4	10	6.17	7	12%
2QTS	Acid-sensing ion channel	1.15	1.59	1.28	0.83	5	8	5.40	6	14%
2H88	Succinate oxidoreductase	0.92	1.11	1.01	0.42	5	8	6.60	6	18%
2UUH	Leukotriene LTC4 synthase	0.53	0.76	0.63	0.36	7	13	7.93	4	9%
1BL8	Potassium channel	1.05	1.57	1.35	1.1	4	8	5.45	12	7%
2BL2	Rotor of V-type ATPase	0.71	1.01	0.83	0.87	8	17	13.10	40	14%
1OKC	Mitochondrial ADP/ATP carrier	0.81	0.92	0.88	0.39	5	10	6.20	6	12%
2B6O	Aquaporin	1.19	1.34	1.23	0.32	3	11	6.00	8	8%
Average		0.93	1.17	1.03	0.68	5.08	10.08	6.94	9.38	10.60%

The first columns give the minimum (min), maximum (max), and average (avg) Cα rmsd from the native for ensemble members. The next gives the largest Cα rmsd between two ensemble members. The interface interactions columns list the min., max., and avg. number of interactions of the five types between two pieces used during assembly. The helices column lists the number of membrane helices in the structure. The final column lists the fraction of helix residues whose native side chains (nat. sc.) were used in the assembly. GPCR, G-protein-coupled receptor.

has a Cα rmsd of 1.91 Å from the native. It is shown in Figure 5 with a conformation with a comparable score that has a Cα rmsd of 0.43 Å from the native. Both were among the conformations selected to propagate, but the 1.91 Å structure was incompatible with the subsequent inter-monomer interactions of the five types, which force a collision between the monomers with this conformation.

Symmetric oligomers can also create interesting exceptions. When dimer substructures are built independently and identical ones are snapped together, the error can add until it is caught at

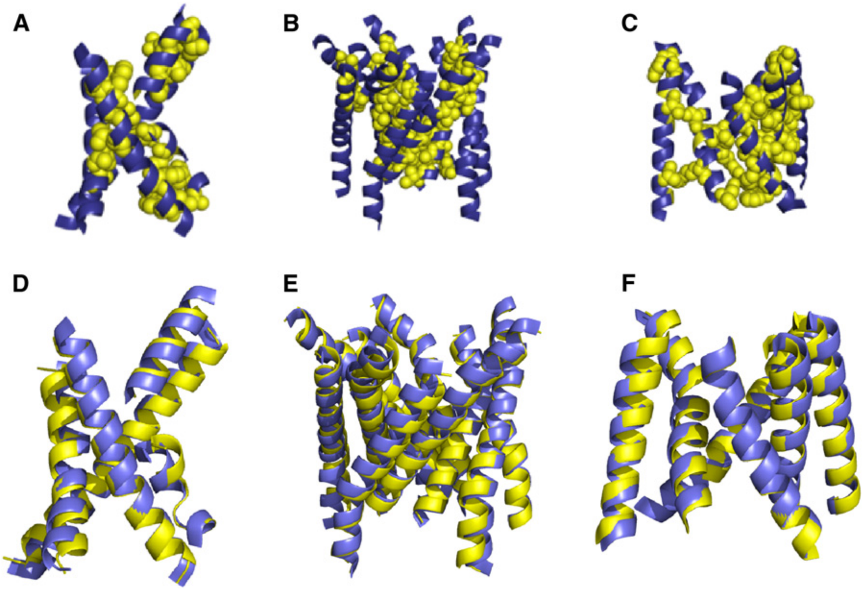

Figure 4. Interhelical Interactions of the Five Types Displayed in Native Structures and Comparison of the Reassembled Structures to Natives

In the upper panels, the interacting side chains are colored yellow and their atoms shown as spheres. From left to right, the proteins are (A) the voltage sensor (PDB id 1ORS), (B) TuMscL (PDB id 2OAR), and (C) succinate oxidoreductase (PDB id 2H88). The lower panels show reassembled structures with the worst Cα rmsds to the native in the final ensemble of those conformations best satisfying the constraints imposed by the interhelical interactions of the five types. These structures (in yellow) are aligned with the native backbones (in blue). The Cα rmsds to the native structures of the reassembled structures shown are 1.3 Å for (D), 1.4 Å for (E), and 1.1 Å for (F), left to right. The differences between the backbones of the worst of the reassembled structures and the native are small and would be barely visible in any figure of this type.

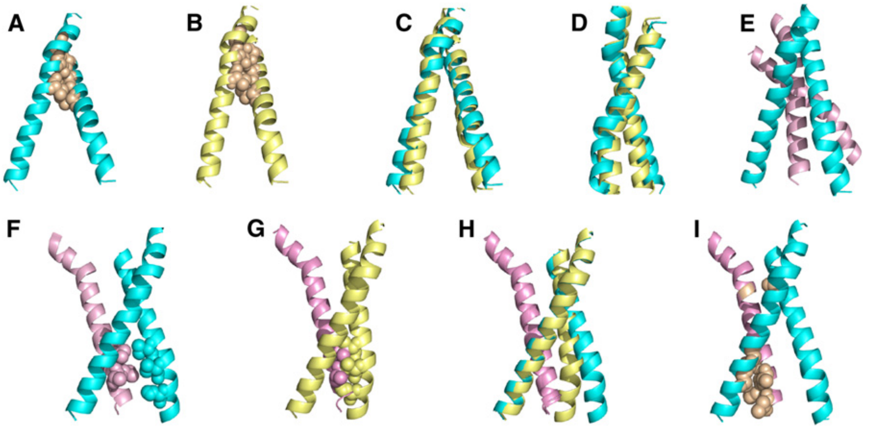

Figure 5. Modularity, Flexibility, and Specificity in TuMscL

In the conformations of the monomer of TuMscL (mechanosensitive channel of large conductance), we see unusual variability under the constraints of the interhelical interactions of the five types derived from the native conformation (PDB iD 2OAR). But the subsequent intermonomer interactions are not compatible with all of the monomer conformations.

(A) One monomer conformation shown in blue. This conformation has a $C\alpha$ rmsd of 0.4 Å from the native. The atoms of the residues in the interhelical interactions of the five types are shown as wheat colored spheres; these interactions were used to build the conformation. This will be called the ''good'' conformation.

(B) Another conformation shown in yellow of the monomer. It has a $C\alpha$ rmsd of 1.9 Å from the native, and was built with the same interhelical interactions as the conformation shown in (A) in blue. Note the narrower region between the helices in this conformation; we shall see that this prevents this monomer conformation from being compatible with the inter-monomer interactions. As before, the atoms of the residues in the interhelical interactions of the five types are shown as wheat colored spheres.

(C) The two conformations aligned.

(D) A side view of the aligned conformations.

(E) The reassembled dimer built with the good 0.4 Å $C\alpha$ rmsd monomer conformation with one monomer shown in blue and the adjacent one in pink (the entire pentamer has $C\alpha$ rmsd 1.3 Å from the native). Note that there is room for the first helix of the adjacent pink monomer between the helices of the blue monomer.

(F) The good monomer in blue shown with the first helix from the adjacent monomer in pink. This is a substructure of the good dimer shown in (E). The atoms shown as spheres are there to compare with those in the conformation shown in the next panel. In this conformation, they do not collide; in the next panel, they do.

(G) The 1.9 Å monomer conformation in yellow with the first helix of the adjacent monomer shown in pink. The helix pair made up of the first helix from each monomer (the two leftmost helices) has the same conformation as in the preceding picture. Note the collision at the bottom of the structure.

(H) The conformations shown in the preceding two panels aligned. The conformation of the leftmost pair of helices is the same for both structures.

(I) The good conformation from panel (F) shown with the intermonomer interactions of the five types colored in wheat. The yellow 1.9 Å monomer conformation is not compatible with this set of interactions because there is not enough room between the monomer helices for the helix from the adjacent monomer.

the last step (Figure 6). It is still possible to build low rmsd structures without imposing symmetry, but far superior ones can be built if the symmetry is used from the start to construct the dimers, as we did for 1BL8 (KcsA, potassium channel) and 2BL2 (rotor of V-type ATPase). For the unsymmetrized 1BL8 and 2BL2, we obtained ensembles with average $C\alpha$ rmsds from the native of 2.17 Å and 1.67 Å; the symmetrized ensembles for 1BL8 and 2BL2 have average $C\alpha$ rmsds from the native of 1.35 Å and 0.83 Å (Table 1). We did not symmetrize 2OAR (TuMscL, large-conductance mechanosensitive channel) because we obtained good structures without doing so (average rmsd of 1.30 Å; Table 1).

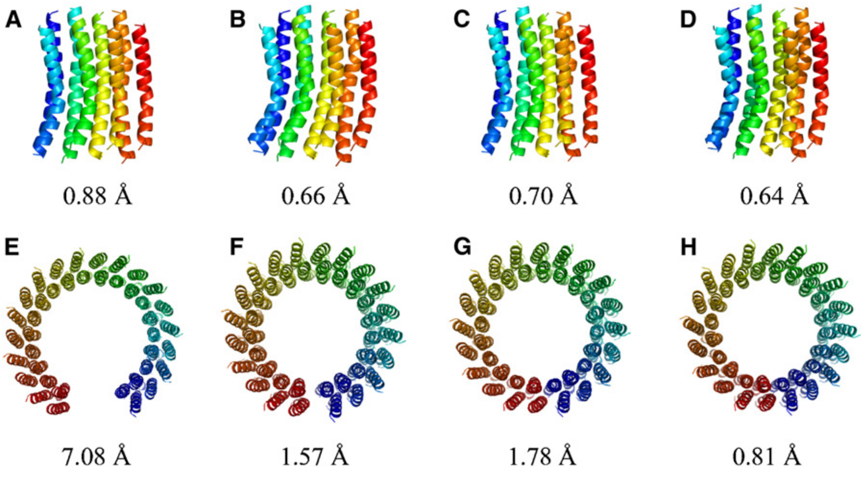

Figure 6. Unusual Accumulation of Error in a Symmetric Decamer

(A–D) The upper panels show four conformations of dimer substructures of the decamer rotor of V-type ATPase (2BL2) built from the interhelical interactions of the five types; their $C\alpha$ rmsds from the native are shown below each conformation. The monomers have four helices, and the dimers are colored according to sequence position with the colors of the spectrum. The rightmost dimer (in D) was built to be compatible with the 10-fold rotational symmetry; the others were not.

(E–H) The lower panels show the full decamers built using the dimers directly above them in the upper panels. Thus, (E) was built from (A), (F) was built from (B), etc. They are shown with their $C\alpha$ rmsds from the native, and are colored according to sequence with the colors of the spectrum. The dimers were all built to satisfy the interhelical interactions of the five types, but note how the structural errors can add until they are caught at the last step where the decamers fail to close properly and so fail to satisfy the interhelical interactions of the five types at this closure (see the bottom part of the structures). When the dimers are symmetrized from the start as in the rightmost structure, this cannot happen.

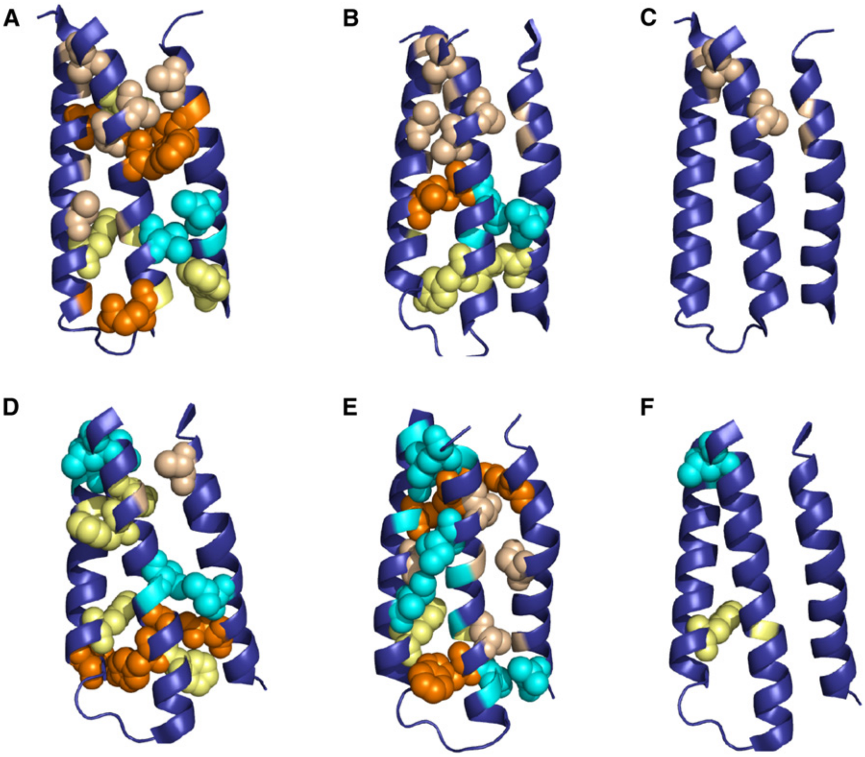

Figure 7. Comparison of Homologs: The Same Fold Can Be Stabilized Using Different Determining Sets of Interactions

The five types of interhelical interactions are displayed as follows. (Only four types occur in these structures: there are no salt bridges.) Residues in a hydrogen bond are colored orange; if a residue's side chain (rather than a backbone atom) is in a hydrogen bond, then its atoms are shown as orange spheres. Residues in aromatic interactions are colored yellow; if their side chains participate in the interaction, then the side chain atoms are shown as spheres. The knob atoms of Gly/Ala/Ser small close knob-in-hole packing are shown as wheat colored spheres. Unless they appear in a different interaction, at least one of the corresponding hole residues is colored in wheat. The knob atoms of Ile/Val/Leu close knob-in-hole packing in I/V/L patches are shown as bright blue (cyan) spheres. The corresponding hole residues are shown as spheres if they are I/V/L in close contact with the knob residue; otherwise if the hole residue has restricted side chain conformations it is also shown in bright blue. The three upper panels show sensory rhodopsins II.

(A) Sensory rhodopsin II, Anabaena (PDB id 1XIO).

(B) Sensory rhodopsin II, N. pharaonis (PDB id 1H68).

(C) The interactions common or closely substituted for the two structures shown in (A) and (B). The lower panels show bacteriorhodopsin and halorhodopsin.

(D) Bacteriorhopsin (PDB id 1C3W). (E) Halorhodopsin (PDB id 1E12).

(F) The interactions common or closely substituted for the two structures shown in (D) and (E).

Statistics on the Interhelical Interactions of the Five Types

The proportions of the kinds of the interactions vary greatly, but there is almost always a mix of polar and packing interactions. Near water channels, for example, the two packing interactions dominate because the polar side chains will tend to interact with the water instead of forming interhelical interactions, and this is reflected in the low percentage of polar interactions in proteins with such channels. Statistics on the proportions of the different types of interactions are shown in S3 in the Supplemental Data.

Homologs with Different Determining Sets of Interactions

Some of the strongest evidence for the structural significance of the determining sets of the five types of interactions comes from homologous proteins. Although many of the most conserved residues are often in the types of interactions we consider, not all such interacting residues are conserved. Determining sets of these interactions are very vulnerable to substitutions, and just one substitution can destroy a determining set, so what happens in homologs?

In the solved structures of homologs, we find distinct determining sets with interactions different in both type and position (Figure 7). Thus we often see compensating substitutions: when one interhelical interaction of one of the five types is lost, another is gained. For bacteriorhodopsin (bR), part of the determining set of interactions is very well conserved, but the interactions among helices 1, 2, and 3 are not at all. Comparing bR (1C3W) to the structurally similar halorhodopsin (1E12), we find only 17% and 11% of their sets for the first three helices in common or closely substituted, respectively. Similarly, comparing sensory rhodopsins II, 1XIO, and 1H68, we find different determining sets stabilizing similar backbone conformations for the first three helices, with 15% and 18% of their sets of interactions common or closely substituted, respectively (Figure 7; also see S4 in the Supplemental Data).

Structures without Determining Sets of Interhelical Interactions of the Five Types

Some solved structures do not have determining sets of these types of interactions, e.g., 1NKZ, a light harvesting complex (Figure 8). In most such proteins, cofactors overwhelm the helix-helix packing as they do in 1NKZ. It also seems likely that loops or other domains outside of the hydrophobic region of the membrane could in some cases impose powerful external constraints that would remove the need for determining sets of interactions in the transmembrane region. However, this does not appear to be common in solved structures at this time: even when there are external domains or loop structures, the interhelical interactions of the five types seem to remain quite constraining. E.g., we have the potassium channel 1BL8 in our test

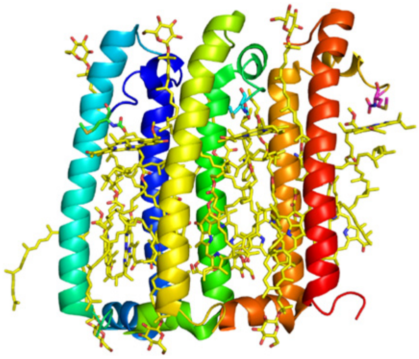

Figure 8. Structure Without a Determining Set of Interhelical Interactions of the Five Types

A light-harvesting complex (PDB id 1NKZ). Note the prosthetic groups and the loose packing of the TM helices.

set, and it is clear that the selectivity filter loop structure constrains the helices. We use only interhelical interactions when we rebuild, so we could not use any information about the selectivity filter. We did manage to successfully rebuild the structure, but found it necessary to impose symmetry, and the resulting ensemble is the most variable in the test set and has the worst average rmsd from the native (1.35 Å; Table 1). This can be attributed in part to the omission of the selectivity filter loop structure.

Retracted structures like 1S7B and 2F2M for EmrE have many structural anomalies, including an absence of determining sets of the five types of interactions (Chang et al., 2006). (See S7 and Figure S2 in the Supplemental Data for an analysis of the retracted structure 2F2M and its insufficient set of interhelical interactions.) This is also true for structures like that of MscS (2OAU) whose biological relevance is debated (Vasquez et al., 2008; Anishkin et al., 2008). And this absence cannot be fixed by changing side chain conformations: there are no possible determining sets of these types of interactions for the backbones in these structures.

Although there are many ambiguous cases at lower resolutions, there are very few clear-cut exceptions known to us. By inspection, we have informally found determining sets of the five types of interactions in most solved structures.

DISCUSSION

The determining sets of the five types of interactions provide a possible partial explanation for the way in which a sequence can specify a stable low-energy structure. To put this another way, amino acid sequences are naturally selected that allow for determining sets in part because the determining sets can act to help create low-energy minima. To see this, imagine a low-energy conformation with a determining set of interactions. First, the abundance of these particularly favorable interactions would tend to act to make the structure a low-energy one. Second, when the backbone positions of the helices are significantly perturbed, some of the determining set of the interactions will be broken. (This is equivalent to what we have shown before: namely, because the set of the interactions of these types highly constrains the positions of the backbones of the helices, significantly perturbing these positions must necessarily break some of the interactions.) At the very least and for very few perturbations, some side chains must be flipped and so rotameric barriers crossed. If we assume the interaction energies are strong enough, it will be difficult to compensate for the lost interactions of the five types given their geometric and partner specificity and the rarity of possible participants. Thus the energies of the perturbed structure would tend to be higher. In contrast, one could not usually say the same of a ‘‘determining set of VDW interactions’’ because of the density and promiscuity of VDW interactions.

The question is what a strong enough interaction energy would be. If the energies were too strong, the backbone would be unable to move, but we know it usually must for the protein to function. If the energies were too weak, then breaking the interactions would not be significant, and they could be easily compensated for by generic VDW interactions not on our list of interactions. The crucial feature is that the interhelical interactions of the five types be very likely superior to alternatives. Of course, the determining sets of the interactions must also act cooperatively: hydropathy, packing, VDW, steric restrictions, loops, ligands, etc. can all also be expected to play important energetic roles.

There is debate about the energies associated with interhelical hydrogen bonds in membrane proteins, which appear to be weaker than was once thought (Joh et al., 2008). The energies of hydrogen bonds can be expected to depend on the length and flexibility of the participating side chains because of the related entropic cost of bond formation. All of the hydrogen bonds found to be especially weak in Joh et al. (2008) are between long side chains, whereas most of our interhelical bonds are between shorter ones, with backbone oxygens the most common acceptors. The side chains in our polar interhelical interactions are also usually well supported by VDW contacts, so that the polar interaction itself is additional to those. If the polar atom interactions are very likely to be better than any possible VDW contacts made by the same atoms or nearby atoms on the side chain, then the overall interaction will likely be better than nonpolar alternatives. (The ubiquity of interhelical polar interactions in high-resolution structures itself suggests this because small perturbations of the side chain would remove the bonds and similar nonpolar substitutions could replace them.) But not all interhelical hydrogen bonds are structurally and energetically significant, and we would omit the insignificant ones from our set of interactions if we knew them. This would likely make little difference to our reassembly because there are usually more than enough interactions.

A much stronger interpretation of the determining sets of interactions is that it is energetically necessary to use at least three such interactions to put two substructures together. That is to say these sets of interactions of the five types drive assembly and nothing else can. Although we are unaware of counterexamples to this, it is a very strong assertion. If true, it would extend the applicability of our approach to abnormal function and mutants. The weaker assumption we have made is that the determining sets of interactions evolved as a useful structural feature, not an absolutely necessary one.

The determining sets of the five types of interactions can do more than act cooperatively to help create energy minima.

Motions in most directions would break both many of the interactions in the determining set and the native-like packing. However, if it is possible to reasonably deform the interactions in certain directions and arrive at a new determining set and conformation (and minimum), then it might be possible to guide motion. This kind of deformation could amount to a low-energy path in conformational space between the two states (Curran and Engelman, 2003).

Our iterative assembly algorithm indicates the important influence that we believe translocon-aided folding can have on final structures. It seems likely to us that a crucial part of the translocon's function is to control contact order of the helices by inserting them sequentially into the membrane, and that this assembly line greatly simplifies the problem nature faces when 'designing' proteins and determining sets of interactions. We ascribe particular structural importance to the interhelical interactions of the five types, and the control of contact order means it is much less likely that these sets of interactions could get scrambled. Membrane proteins often do not fold properly without the translocon, and the potential for scrambling the determining set without sequential insertion gives a good reason for this. If nature wants the first two helices to form a stable pair, it does not have to defend against unwanted competing hydrogen bonds, e.g., between the first helix and the fifth one because the fifth helix is simply not yet in the membrane by the time the first two helices are already there. Without the translocon and sequential insertion, this would not be the case.

The iterative assembly (consistent with translocon-aided folding) and the determining sets of interactions can also be described as a geometric recipe for creating folding funnels. The interactions of the five types are supposed to be able to individually and locally outcompete generic contacts and so can successively funnel and eventually collectively trap the native backbone. That this could be done in a controlled iterative way aided by the translocon makes the process much simpler. Although we do not show this in this paper, the native backbone conformations (and subconformations) usually have many interactions of the five types that can stabilize them in addition to the ones that appear in the solved structures because side chains that participate in these interactions can adopt different conformations and form different interactions. These additional interactions could further aid the folding/funneling process.

An immediate application of the determining sets of interactions is as a test for models. If a model is supposed to be highly accurate (even a backbone model), one can explicitly check for the existence of a native-like determining set of interactions. If the model were lower resolution, one could still check how close one is to having a determining set. Additionally, one can check homologous sequences threaded through the model. All of this requires care, but if the model is supported by a native-like determining set of interactions of the five types, and if diverse homologous sequences also have diverse determining sets supporting the same backbone, we would argue that is strong supporting evidence for the model. The absence of a native-like determining set would be weaker evidence against the model; e.g., as we can see in the erroneous structure for EmrE (see S7 and Figure S2 in Supplemental Data).

The determining sets also suggest an approach to the design of TM helical proteins. For any desired backbone conformation, one would first have to select a sequence order respecting assembly of the helices. Then for each interface dictated by that assembly, one would design a native-like determining set of interhelical interactions of the five types. One could check how well this proposed determining set constrained the backbone using the same algorithm used for the proteins in our test set. In most cases, there would be many sequences that would create possible native-like determining sets of these interactions. The existence of the determining sets of interactions of these types in solved structures does not prove the sufficiency of such a native-like determining set for stability, and this would be an interesting question to investigate. It is clear that the backbone conformation would first need to be compatible with some other features essential to low-energy structures like simple native-like packing conditions, plausible positions in the membrane of TM helices, and surface area restrictions.

The determining sets of the five types of interactions should shed new light on function and motion. The question is how these interaction constraints are tuned to the environment. Various stimuli will lead to a change from one determining set/conformational state to another, and it is not hard to imagine how this could happen. For example, let us say that some hydrogen bonds play a critical structural role in the determining interactions of state S1. Suppose some of the residues in these bonds then break and bind a ligand. Then state S1 no longer has a determining set constraining it and so it can now easily move into state S2 that has a different determining set of interactions that does not need the residues involved with the ligand. Similarly, control by pH could amount to, e.g., having a protonated Asp that can now participate as a donor in our structural interactions, which changes the possible conformational states. Of course, the determining set of interactions analysis does not address delicate energetics, and additional experimental data would be needed to create a convincing picture.

A critical aspect of the structural feature of determining sets of the five types of interactions is its ability to limit conformational space. Conformations that have such a determining set along with native-like packing form a tiny subspace of conformational space. This fact depends upon the kinds of interactions used, as does the energetic significance of the determining set. Crucially, the interactions are particularly energetically favorable, geometrically and partner specific, and the possible participants are quite rare. One could define many hypothetical determining interactions based on different kinds of (e.g., nonspecific) interactions that could be used to rebuild the structures, but the existence of many of these networks would neither much limit conformational space nor have much energetic meaning. Although a more rigorous demonstration of this limiting of conformational space is beyond the scope of this paper, it is easy to see that one can remove the determining sets either by slightly perturbing a small number of side chains or by making a small number of substitutions to break critical interactions. (Also see Supplemental Data for an example of a retracted structure without a determining set.) The limiting of conformational space, the specificity of the determining sets, and the iterative sequence preserving method of assembly create the possibility of enumerating all possible conformations with such determining sets and native-like packing and are the basis for applications in structure and motion prediction that we are developing.

It should be noted that the definitions of the interactions of the five types could be refined in various ways, e.g., by adding conditions to better ensure their likely energetic favorability. It might also be desirable to add some other types of specific favorable interactions to the five types as long as they possess the crucial characteristics listed above. The five types we have considered in this paper are well-known and probably among the most common with the crucial characteristics, but they are not the only possibilities. It would be possible to enlarge the set of interactions considered and make similar energetic and conformational arguments.

Overall, we believe that determining sets of the five types of interactions are a very consistent and comprehensible structural feature of TM helical proteins and that proper understanding of them will greatly simplify structure and motion prediction, as well as the design of helical membrane proteins. Several issues need to be addressed to make use of the determining sets of interactions in structure and motion prediction, e.g., the prediction of the (sometimes irregular) helix conformations themselves. We will present our algorithmic ideas to resolve these issues in future papers. The role of determining sets of interactions should be highly testable using both design and mutational approaches.

EXPERIMENTAL PROCEDURES

Selection of Helices

The helix assignments were made by examining the solved structures, and only helices predominantly in the inferred hydrocarbon region were selected. This includes half-helices. The backbone conformations of these helices are taken from the solved structures and used during reassembly. The helices used are given in S5 in Supplemental Data.

Selection of the Five Kinds of Interactions

All of the interaction types selected have been shown to be energetically significant in dimerization and other studies (Bowie, 2005; Fleishman and Ben-Tal, 2002; Lemmon and Engelman, 1994; Russ and Engelman, 2000; Senes et al., 2000; Schneider and Engelman, 2004; Finger et al., 2006; Zhou et al., 2000; Gratkowski et al., 2001; Dawson et al., 2003; Honig and Hubbell, 1984; Dougherty, 1996; Sal-Man et al., 2007; Johnson et al., 2007; Gurezka et al., 1999), but there are gray areas in terms of possible definitions. For example, usually Ser knob-in-hole packing is unnecessary, but on occasion it is useful, and we also had to decide which aromatic interactions to include.

The selected types were chosen because they seemed to be necessary to rebuild some structures. Examples of structures whose determining sets are dependent on a type of interaction are 2B6O (aquaporin) for small packing, 1AFO (glycophorin) for I/V/L packing, 2QTS (an acid-sensing ion channel) for salt bridges, 1P49 (estrone sulfatase) for aromatic interactions, and 1ORS (voltage sensor) for hydrogen bonds. Likewise, it was necessary and seemed energetically reasonable to take some interactions out of their motif contexts. Structural motifs consist of multiple interactions according to our definition.

Side Chain Representatives and Bumps

The side chain conformations are taken from the solved structures for donor/acceptor pairs in hydrogen bonds, salt bridges, and aromatic interactions. When the acceptor is a backbone oxygen, its side chain is not included unless used in another interaction.

We use a reduced representative of the other side chains, which we call "bumps." They are designed to give the obstacle created by that residue type's side chain irrespective of rotamer state; one usually can do better than Cβ. In helix interiors, some residues have restricted side chain conformations, e.g., Val.

It is convenient for bumps to be written in terms of the usual atoms of the side chains so they can be put into Protein Data Bank (PDB) format. We took an average conformation for the rotamers, and then discarded any atom in the averaged conformation that was not within a cutoff distance to some side chain atom, not necessarily the same one, for every rotamer. The resulting reduced side chain with reduced VDW radii is used as our side chain representative. The coordinates are given in S9 in the Supplemental Data.

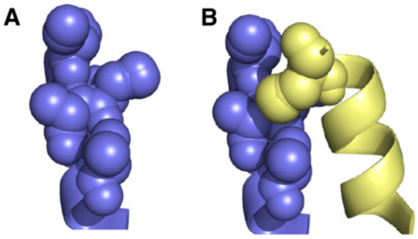

Figure 9. Knob-in-Hole

(A) The residues j, j+3, j+4 and j+7 on an alpha helix with reduced representations of their side chains with their atoms shown as spheres. The numbering starts from the top in this picture; the space surrounded by these residues is called a hole.

(B) An example of interhelical knob-in-hole packing.

Geometric Definition of the Five Types of Interactions

For each type of interaction, there are associated geometric conditions, most of which involve distances between atoms. We need to define the geometry of interactions both to assemble structures and to find the interactions in solved structures. Once the standard geometry has been defined, we associate a quadratic penalty function with it. We use the penalty functions and associated error terms to find the interactions in solved structures. By definition, interhelical interactions of the five types in solved structures all have penalties less than the associated error term. We chose small error terms to be consistent with the errors expected in structures with the resolutions of the structures found in our test set.

Please see S8 in the Supplemental Data for details on the distance cutoffs for the definitions sketched below. See Figure 1 for examples.

Hydrogen Bonds

The definition is adapted from McDonald and Thornton, as is the hydrogen addition (McDonald and Thornton, 1994). Water-bridged bonds count only if the water does not distort the usual geometry. Cα donation is allowed for Gly. All protonation states are considered. Backbone oxygens can act as acceptors.

Salt Bridges

The distance cutoff between the presumed positively and negatively charged atoms is 4 Å, with a small error tolerance added. To avoid protonation ambiguities, salt bridges were omitted from the sets of interactions when they were unnecessary constraints.

Aromatic Interactions

The outside of the aromatic ring is (partially) positively charged and the ring atoms are treated as weak donors, where acceptor atoms must lie near (under about 3.7 Å) the donor ring atom close to the plane of the ring (Dougherty, 1996; Nanda and Schmiedekamp, 2008). Their acceptors are usual hydrogen bond acceptors that either have short side chains, are backbone oxygens, or are possibly charged. Overwhelmingly, the acceptors are backbone oxygens. The center of the ring can act as an acceptor, but does not for this test set.

The Two Packing Interactions

In order to define the two packing motif interactions, we first define a knob-in-hole geometry (Figure 9). The hole is defined by a set of helix residues i, i+3, i+4, i+7 and atoms in those residues O (in i), C (in i+3), Cα (in i+4), Cβ (Cα for Gly) (in i+7), respectively. The hole is divided into three regions: lower (closest to i, denoted "l"), middle (closest to i+3 and i+4, denoted "m") or upper (closest to i+7, denoted "u"). The knob atom must lie within a set of distance cutoffs to the surrounding hole atoms, with the maximum distance to the closest hole atom usually around 5.5 Å. Beta-branched residues can lie slightly farther out. The knob atom position is taken from the reduced side chain representative. It is the most distal symmetrically placed atom in the fixed side chain representative.

Gly|Ala|Ser Knob-in-Hole

Here we take Cα for Gly and Cβ for Ser and Ala as the knob atom.

Many of these small close knob-in-hole interactions in our test set do occur in (smallres)xxx(smallres) motifs, but on occasion it was necessary to use those outside of this kind of motif. Additional conditions could be imposed, but we have chosen this definition for the sake of simplicity.

Ile|Val|Leu Patch Interaction

These residues, especially Val and Ile, are sterically restricted by the helix backbone, and so tight interhelical contacts between these residues are entropically favored (Walters and DeGrado, 2006; Popot and Engelman, 1990; Lemmon and Engelman, 1994).

If Ile, Val, or Leu acts as a knob to a hole, where it makes a substantial contact with at least one of the surrounding hole residues that is an Ile, Val or Leu, and if at least one of the other surrounding side chains has a restricted conformation, i.e., G, A, I, V or L, then the contact is an Ile/Val/Leu patch interaction. Here we take Cβ for Val and Ile, and Cγ for Leu as the knob atom.

Determining Set Decomposition Details

We begin with the helices in sequence order H1, …, Hn, and score the set of the five types of interactions between H1 and H2 and between H2 and H3. We need to decide which, if either, of these two pairs we will assemble into a piece. Our assembly pair score is based on the number (more being better) of the interface interactions of the five types, length of the connecting loop, and position in the sequence. We reward short connecting loops and earlier position in the sequence; short interacting residues and backbone oxygen acceptors are also preferred. If there are at least four interactions for at least one of the pairs, we can build with the better scoring pair of substructures, and revise the list. If our initial list is H1, H2, H3, H4, H5, H6 and we assemble H2 and H3 (the better pair), our new list will be: H1, (H2 H3), H4, H5, H6. If we cannot build with either interface interactions, we advance two steps down the list to the next unchecked triple; here, H3, H4, H5.

We repeat this process with the first three elements of the list whenever two pieces have been assembled (i.e., the list has been updated). For the updated list given above, H1, (H2 H3), H4, H5, H6, we then score the interactions between H1 and (H2 H3) and between (H2 H3) and H4. When we score the interactions between H1 and (H2 H3), e.g., we consider all of the interactions between H1 and H2 and H1 and H3, because (H2 H3) is now a single piece. In general, when we score the interactions between two preassembled pieces, we consider the interactions between the helices in the first piece with the helices in the second piece. Ultimately, there should be only two pieces left to put together, and we just need to check that there are sufficiently many interactions between them to assemble. If we are at the list's end and there is still more than one piece, we repeat the above procedure, but we can now try to assemble using second neighbors. For our test set, this procedure always completely assembled the protein. See Figure S10 in Supplemental Data for more details on the scoring used above.

Three-Dimensional Assembly Details

Once the interactions between two pieces are known from the solved structure, we can reassemble them into a combined piece ensemble. (For additional background on the underlying geometry, see Figure S1 in the Supplemental Data.) We select three of the interhelical interactions of the five types: this is done automatically based on their dispersion across the interface and type. Shorter side chains and backbone oxygens are preferred as donors and/or acceptors, and the greater the area of the triangle formed by the three donors and/or acceptors the better their score.

For each interaction, there is a region (based on the interaction type) associated with either the donor or acceptor where we place a grid. One can always assume it is the acceptor, as we will in this discussion. (For the two packing interactions, the hole is considered the acceptor.) The grid points in this interaction region give the possible positions of the donor (which is on the other piece). All of the interaction geometries have a simple donor-acceptor distance cutoff in addition to other requirements. The interaction region associated with the acceptor can be simply generated using this donor-acceptor cutoff. The grid points in this interaction region, which give the possible donor positions, can then be checked for any additional geometric criteria that can be applied to just this donor position. In this way, we obtain a set of grid points (possible donor positions) of the interaction region associated to the acceptor.

Cycling through these grid points gives an ensemble of structures for the combined piece. This depends on the fact that the positions of any three noncollinear points on a body determine the position of the body. We cycle through the possible grid points in order as follows. For the first interaction, the donor atom is placed on a grid point; for the second interaction, the dimension of its associated grid drops by one because of the distance constraint imposed by the rigid bodies. (The description of this distance depends on the way the donors and acceptors are distributed between the bodies. For example, if the first two donors lie on one body, then the distance between these donors must match the distance between the two grid points associated with the acceptors on the other body.) This is done by keeping the first two coordinates of the grid and selecting the third (by solving the associated quadratic) so that the distance matches correctly. Likewise, the dimension of the associated third grid drops again once the first two grid points are selected (for now two distances need to be matched, and two simultaneous quadratics solved). In this way, we find grid points that are consistent with the distance constraints imposed by the bodies. By running through these grid points, we obtain our ensemble. If there are additional interactions of the five types or additional geometric conditions (as is the case, for example, for a hydrogen bond), the conformations are checked for those and discarded if they don't have the additional interactions or meet the additional geometric conditions.

The geometric conditions and the overlaps have associated quadratic penalty functions, so screening the conformations for the geometric conditions amounts to screening according to the assigned penalty scores. The overlap penalties are based on the sums of the VDW radii, and side chain/bump overlaps count less than those of backbone atoms. (The coefficients of quadratic terms for side chain/bump overlaps are half of those for the others for each side chain/bump involved.)

The two important parameters for assembly are grid size and overlap cutoff for the resulting structures. There were usually between 400 and 1500 grid points in a region, depending on how the parameter was set.

If a run failed to produce a good solution for the set of interactions or produced too few, the job was rerun with finer grids and a higher overlap cutoff. If the job still failed, a new set of three interactions was chosen, and the job repeated. This is sometimes necessary because the input structures can be distorted enough so that not all the native interactions are possible: the best you can have is a slightly distorted version of them. This can be scored, but distortions can change the interactions that can be used for assembly. The interactions that are used for construction must exist with their usual geometries. On occasion it was useful to change the three used for construction.

If a run had produced more than 10,000 structures, it was stopped and rerun with a slightly coarser grid.

Traditional clustering was not used on the subensembles because the resulting clusters tended to consist of one native-like cluster (with most of the structures) and other relatively poorly scoring, non-native-like atypical boundary structures. Consequently, only the native-like cluster structure would be compatible with the determining set of interactions in subsequent steps.

Instead, for each ensemble produced at each step, we select good conformations according to a scoring cutoff from our ensemble of good structures and measure the largest Cα rmsd between them. The cutoff is based on the best score plus an error term. If fewer than five structures met this cutoff, then the five structures with the lowest scores were chosen to be the members of the ensemble of good substructures. The scoring depends only on overlaps and the geometric conditions imposed by the five kinds of interactions. It does not approximate energy. Three conformations are then randomly selected to propagate such that the largest rmsd among them is at least two thirds of the largest rmsd among the structures in the full ensemble. This is intended to ensure some structural diversity.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, two figures, and four tables and can be found with this article online at http://www.cell.com/structure/supplemental/S0969-2126(09)00251-2.

ACKNOWLEDGMENTS

This work was supported by grants 222/04 and 611/07 from the Israel Science Foundation.

Received: November 6, 2008
Revised: June 15, 2009
Accepted: June 16, 2009
Published: August 11, 2009

REFERENCES

Acton, F.S. (1990). Numerical methods that work. In Minimum Methods (Washington, DC: The Mathematical Association of America), pp. 448–476.

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprato, A., Karni-Schmidt, O., Williams, R., Chait, B.T., et al. (2007). Determining the architectures of macromolecular assemblies. Nature *450*, 683–694.

Anishkin, A., Akitake, B., and Sukharev, S. (2008). Characterization of the resting MscS: modeling and analysis of the closed bacterial mechanosensitive channel of small conductance. Biophys. J. *94*, 1252–1266.

Baldwin, J.M., Schertler, G.F., and Unger, V.M. (1997). An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. J. Mol. Biol. *272*, 144–164.

Barth, P., Schonbrun, J., and Baker, D. (2007). Toward high-resolution prediction and design of transmembrane helical protein structures. Proc. Natl. Acad. Sci. USA *104*, 15682–15687.

Beuming, T., and Weinstein, H. (2005). Modeling membrane proteins based on low resolution electron microscopy: a template for the TM domains of the oxalate transporter OxlT. Protein Eng. Des. Sel. *18*, 2613–2627.

Bowie, J.U. (1997). Helix packing in membrane proteins. J. Mol. Biol. *272*, 780–789.

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

Chang, G., Roth, C.B., Reyes, C.L., Pornillos, O., Chen, Y.J., and Chen, A.P. (2006). Retraction. Science *314*, 1875.

Curran, A.R., and Engelman, D.M. (2003). Sequence motifs, polar interactions and conformational changes in helical membrane proteins. Curr. Opin. Struct. Biol. *13*, 412–417.

Dawson, J.P., Melnyk, R.A., Deber, C.M., and Engelman, D.M. (2003). Sequence context strongly modulates association of polar residues in transmembrane helices. J. Mol. Biol. *331*, 255–262.

DeGrado, W., Gratkowski, H., and Lear, J.D. (2003). How do helix–helix interactions help determine the folds of membrane proteins? Perspectives from the study of homo-oligomeric helical bundles. Protein Sci. *12*, 647–665.

Delano, W. (2002). Pymol. www.pymol.org.

Dougherty, D. (1996). Cation-pi interactions in chemistry and biology: a new view of Benzene, Phe, Tyr, and Trp. Science *271*, 163–168.

Finger, C., Volkmer, T., Prodohl, A., Otzen, D.E., Engelman, D.M., and Schneider, D. (2006). The stability of transmembrane helix interactions measured in a biological membrane. J. Mol. Biol. *358*, 1221–1228.

Fleishman, S.J., and Ben-Tal, N. (2002). A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. J. Mol. Biol. *321*, 363–378.

Fleishman, S.J., Unger, V., and Ben-Tal, N. (2006). Transmembrane protein structures without X-rays. Trends Biochem. Sci. *31*, 106–113.

Gratkowski, H., Lear, J.D., and DeGrado, W.F. (2001). Polar side chains drive the association of model transmembrane peptides. Proc. Natl. Acad. Sci. USA *98*, 880–885.

Gurezka, R., Laage, R., Brosig, B., and Langosch, D.A. (1999). Heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments. J. Biol. Chem. *274*, 9265–9270.

Honig, B.H., and Hubbell, W.L. (1984). Stability of ‘‘salt bridges’’ in membrane proteins. Proc. Natl. Acad. Sci. USA *81*, 5412–5416.

Joh, N.H., Min, A., Faham, S., Whitelegge, J.P., Yang, D., Woods, V.L., and Bowie, J.U. (2008). Modest stabilization by most hydrogen-bonded side chain interactions in membrane proteins. Nature *453*, 1266–1270.

Johnson, R.M., Hecht, K., and Deber, C.M. (2007). Aromatic and cation-pi interactions enhance helix-helix association in a membrane environment. Biochemistry *46*, 9208–9214.

Kovacs, J.A., Yeager, M., and Abagyan, R. (2007). Computational prediction of atomic structures of helical membrane proteins aided by EM maps. Biophys. J. *93*, 1950–1959.

Lemmon, M.A., and Engelman, D.M. (1994). Specificity and promiscuity in membrane protein interactions. Q. Rev. Biophys. *27*, 157–218.

Liu, Y.S., Sompornpisut, P., and Perozo, E. (2001). Structure of the KcsA channel intracellular gate in the open state. Nat. Struct. Biol. *8*, 883–887.

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

Nanda, V., and Schmiedekamp, A. (2008). Are aromatic carbon donor hydrogen bonds linear in proteins? Proteins *70*, 489–497.

Perozo, E., Cortes, D.M., Sompornpisut, P., Kloda, A., and Martinac, B. (2002). Open channel structure of MscL and the gating mechanism of mechanosensitive channels. Nature *418*, 942–948.

Popot, J.L., and Engelman, D.M. (1990). Membrane protein folding and oligomerization: the two-stage model. Biochemistry *29*, 4031–4037.

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

Sadlish, H., Pitonzo, D., Johnson, A.E., and Skach, W.R. (2005). Sequential triage of transmembrane segments by Sec61alpha during biogenesis of a native multispanning membrane protein. Nat. Struct. Mol. Biol. *12*, 870–878.

Sale, K., Faulon, J.L., Gray, G.A., Schoeniger, J.S., and Young, M.M. (2004). Optimal bundling of transmembrane helices using sparse distance constraints. Protein Sci. *13*, 2613–2627.

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

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.

Schneider, D., and Engelman, D.M. (2004). Motifs of two small residues can assist but are not sufficient to mediate transmembrane helix interactions. J. Mol. Biol. *343*, 799–804.

Vasquez, V., Sotomayor, M., Cortes, D.M., Roux, B., Schulten, K., and Perozo, E. (2008). Three dimensional architecture of membrane-embedded MscS in the closed conformation. J. Mol. Biol. *378*, 55–70.

von Heijne, G. (1996). Principles of membrane protein assembly and structure. Prog. Biophys. Mol. Biol. *66*, 113–139.

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

White, S.H., and von Heijne, G. (2008). How translocons select transmembrane helices. Annu Rev Biophys *37*, 23–42.

Zhou, F.X., Cocco, M.J., Russ, W.P., Brunger, A.T., and Engelman, D.M. (2000). Interhelical hydrogen bonding drives strong interactions in membrane proteins. Nat. Struct. Biol. *7*, 154–160.