Protein Science (1999), *8*:2711–2719. Cambridge University Press. Printed in the USA.

Helix-bundle membrane protein fold templates

JAMES U. BOWIE

Department of Chemistry and Biochemistry and DOE Laboratory of Structural Biology and Molecular Medicine, UCLA, 405 Hilgard Avenue, Los Angeles, California 90095-1570

(RECEIVED June 29, 1999; ACCEPTED August 31, 1999)

Abstract

In the fold recognition approach to structure prediction, a sequence is tested for compatibility with an already known fold. For membrane proteins, however, few folds have been determined experimentally. Here the feasibility of computing the vast majority of likely membrane protein folds is tested. The results indicate that conformation space can be effectively sampled for small numbers of helices. The vast majority of potential monomeric membrane protein structures can be represented by about 30-folds for three helices, but increases exponentially to about 1,500,000 folds for seven helices. The generated folds could serve as templates for fold recognition or as starting points for conformational searches that are well distributed throughout conformation space.

Keywords: fold recognition; helix packing; protein folds; transmembrane helices

With many genome sequencing projects completed or nearing completion, attention is focusing on learning the structures of the protein products (Terwilliger et al., 1998; Montelione & Anderson, 1999). Although high-throughput structure determination will greatly expand our database of known structures, not all protein structures can be determined at atomic resolution. Thus, for most proteins, it will only be possible to visualize their structures using some form of structure prediction. To this end, considerable effort is currently directed toward fold recognition methods that test whether a sequence adopts an already known fold (Bowie et al., 1991; Jones et al., 1992; Marchler-Bauer & Bryant, 1997; Koehl & Levitt, 1999). The fold recognition paradigm greatly simplifies the protein folding problem by limiting the conformational search to regions near the known structures. To the extent that there are a limited number of folds used by nature, a large structure library provided by the high-throughput structure determination projects, combined with fold recognition methods, could yield a practical solution to the protein folding problem for soluble proteins. These efforts, however, will completely miss the roughly 20% of proteins that reside in the membrane (Boyd et al., 1998).

In contrast to soluble proteins, we know only a handful of membrane protein structures and the pace of new structure determination is much slower. Thus, fold recognition methods, as currently formulated, will not be particularly useful for membrane proteins in the near future. But what if it was possible to precalculate the vast majority of membrane protein folds? Then it would not be necessary to wait for experimentally derived structures to apply the fold recognition model. Instead, theoretical folds could provide the structural templates. Here, I test this possibility and find that most of the likely membrane protein folds for up to seven helices can be generated by computer.

Reprint requests to: James U. Bowie, Department of Chemistry and Biochemistry and DOE Laboratory of Structural Biology and Molecular Medicine, UCLA, 405 Hilgard Avenue, Los Angeles, California 90095-1570; e-mail: bowie@mbi.ucla.edu.

Results and discussion

Overview

The vast majority of helix bundle membrane protein folds that could exist were obtained by the following procedure. First, libraries of possible helix-packing arrangements, consistent with geometric constraints on transmembrane helix packings, were created randomly. The number of arrangements needed to obtain 80% saturation of the conformation space was then determined for three to seven helices, where 80% saturation means that there is an 80% chance that any additional helix-packing arrangement generated would already be present in the library. Next, noncompact helix-packing arrangements were eliminated to obtain 80% saturated libraries of compact structures. Duplicate structures were then removed from the compact libraries by clustering the conformations. Finally, the compact helix arrangements in the 80% saturated libraries were converted to folds by adding helix connections. The number of final folds is the number of distinct compact helix-packing arrangements multiplied by the number of ways the helices can be connected. These steps are described in detail below.

Helix-packing arrangements with only geometric constraints

Transmembrane helix-packing arrangements, in which the helices were represented by 30 Å line segments corresponding to the helix

axes, were created randomly. The parameters that define the geometry of these helix packings are shown in Figure 1. If all these parameters were free to take on any value, the number of possible conformations would rapidly explode as the number of helices increase. However, based on observations of known membrane protein structures, many of these parameters could be restricted as follows. (1) For transmembrane helices in known structures, the angle of each transmembrane helix with respect to the bilayer normal τ was found to be less than 40° (Bowie, 1997). Consequently, the accessible conformation space could be reduced by restricting τ to this range. (2) The distribution of interaxial angles between helices (Ω) in known membrane proteins is relatively narrow (Bowie, 1997). A significant reduction in conformational space could therefore be achieved by selecting Ω angles from the observed distribution. (3) Because transmembrane helix assemblies reside in a bilayer, the helices could be restricted to the bilayer plane. (4) To be part of a single structural unit, each helix must contact at least one other helix. (5) Steric overlaps cannot exist. Two parameters could not be restricted in the generation of structures: the point of closest approach of helices, C, and the rotation angle of the point of helix–helix contact, θ.

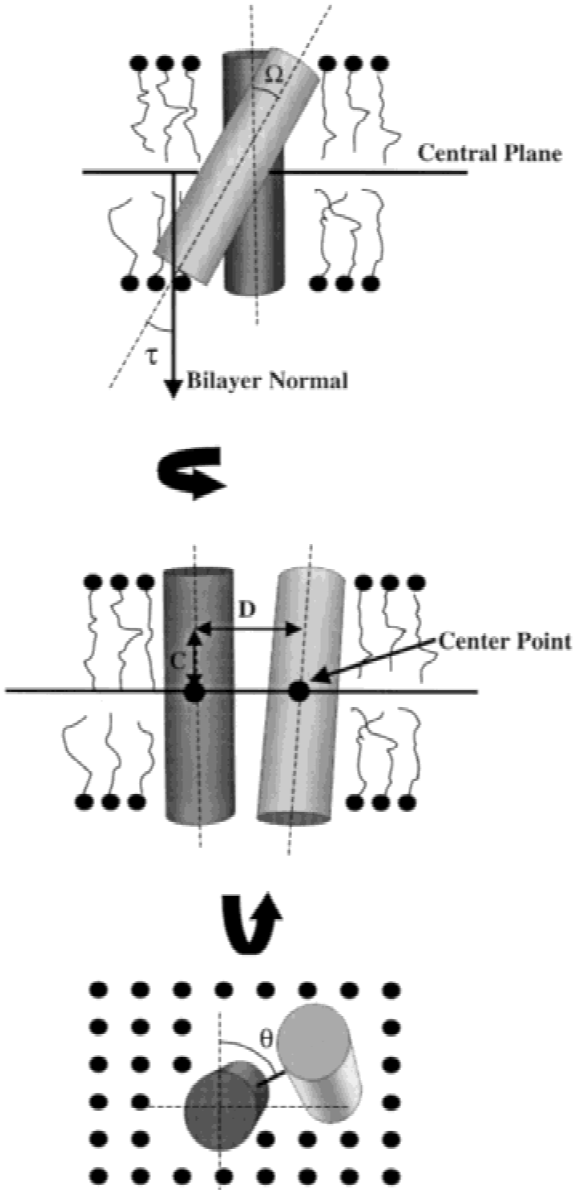

Fig. 1. Geometric parameters defining transmembrane helix assemblies. The central plane represents the center of the bilayer. The Ω angle is the angle between helix axes as defined by Chothia et al. (1981). The angle τ is the angle between the helix axis and the bilayer normal. The parameter *D* is the distance of closest approach of the helix axes (Chothia et al., 1981). The parameter *C* is distance of the point of closest approach of the helix axes to the central plane. The angle θ is the rotation of the point of closest approach about the helix axis.

A set of three helix arrangements, consistent with the geometric restrictions described above, is shown in Figure 2. Some of the structures are relatively compact, while others are quite loosely organized, emphasizing the fact that no structure quality criteria were applied. Thus, the structures generated are possible arrangements, but not necessarily viable natural structures. Addition of a compactness constraint will be described below.

Conformation space coverage

The structure generation protocol will span a certain range of conformation space. How big is the space? This question can be addressed by determining how many structures need to be generated before essentially no new structures can be found, i.e., the space becomes saturated. If a large number of conformations are reasonably probable (conformation space is large), many conformations will be required to reach a point where no new structures can be found. If conformation space is relatively small, few conformations will be needed to saturate the space. The *percent saturation* of a particular library of conformations was defined as the probability that any new structure generated will be found in the library. To determine the percent saturation of a library of generated conformations, an additional 100 conformations were generated and the fraction of the 100 additional structures that could be found in the library was determined.

To measure percent saturation, it was necessary to define criteria for deciding whether two structures were the same. I chose to use the definition of Sander and Schneider, who described proteins as structurally homologous when the root-mean-square deviation (RMSD) of the Cα coordinates is 2.5 Å or less (Sander & Schneider, 1991). Thus, a 2.5 Å RMSD is close enough to imply an evolutionary relationship. The similarity of helix-packing arrangements was defined in a comparable fashion by $RMSD^{Hel}$—the RMSD of the two endpoints and the center point of each helix axis. The Sander and Schneider criterion for structural homology could not be applied directly to the computer-generated helix-packing arrangements, however, because $RMSD^{Hel}$ and the RMSD of a full atom model are not directly comparable. $RMSD^{Hel}$ values are inflated because they are weighted toward the end points, where the deviations are the greatest, and no pair rejection criteria were applied. To compare RMSD values for a full atom model with $RMSD^{Hel}$, the helix axes in the simplified representations were replaced with ideal, polyalanine helices. For 100 structure comparisons with $RMSD^{Hel}$ in the range of 3.9 to 4.1 Å, the average RMSD of the full atom models was 2.6 Å on an average of 84% of the atoms. Thus, helix-packing arrangements with $RMSD^{Hel}$ of 4.0 Å or less were deemed to be similar structures. Examples of three five-helix-bundle arrangements that are within 3.8 Å $RMSD^{Hel}$ are shown in Figure 3.

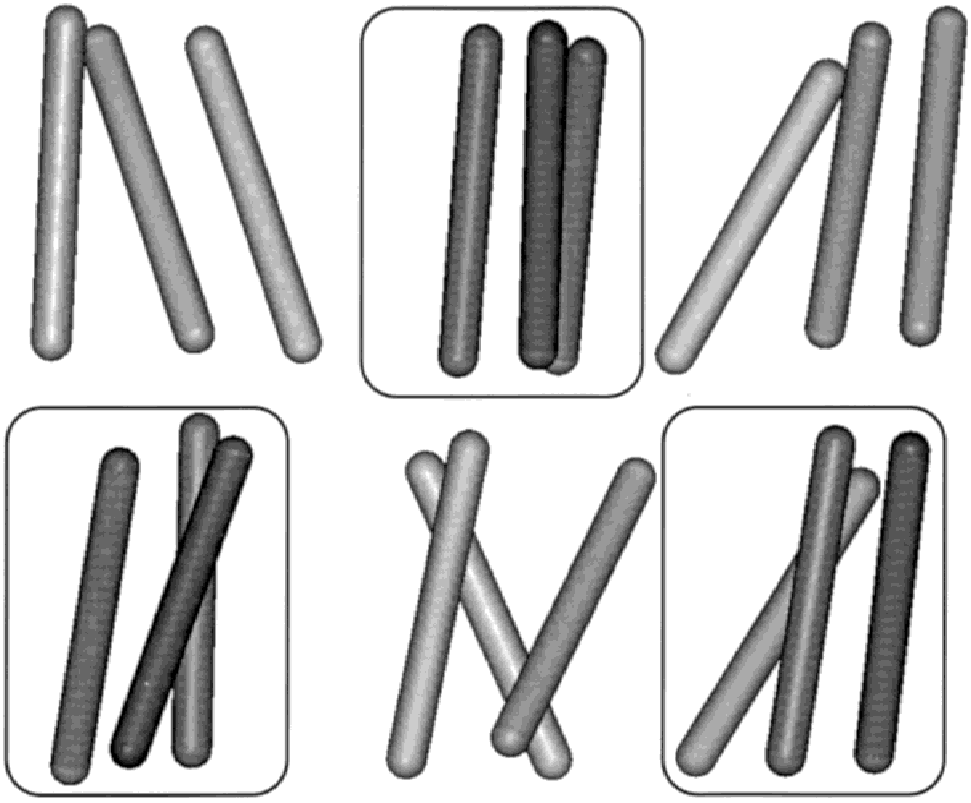

Fig. 2. A collection of three-helix assemblies. The 10 structures needed for 80% saturation of the conformation space when only geometric constraints were applied were clustered into six distinct conformations (see Methods). One representative of each cluster is shown. The boxed structures pass the compactness criteria used in this work. Membrane protein folds can be created from these helix-packing arrangements by connecting the helices in all possible ways.

Figure 4 shows the percent saturation as a function of library size for four helix bundle structures, using only geometric constraints. The curve rises rapidly to about 80% saturation after only 250 structures. After 80% saturation, the curve levels off and few new structures are obtained with increasing library size. Naturally, the size of the space increases as the number of helices increases. Figure 5 shows the number of structures required to achieve 80% saturation as a function of the number of helices. The number needed increases a little less than 10-fold per additional helix, reaching about 150,000 structures for seven helices. Additional reductions in the number of likely helix-packing arrangements will be discussed below.

Correspondence to helix packings in real membrane protein

The results so far indicate that if helix packings are generated using a set of criteria that reflect the geometric constraints on known membrane protein structures, the conformation space is not overwhelmingly large. But does this actually correspond to the conformation space spanned by real membrane proteins? To address this question, I looked for known membrane protein helix-packing arrangements in the computer-generated fold libraries.

Subsets of packed helices were extracted from three known membrane protein structures: bacteriorhodopsin (2BRD), photosynthetic reaction center (1PRC), and the cytochrome bc1 complex (1BCC) (Deisenhofer et al., 1995; Grigorieff et al., 1996; Pebay-Peyroula et al., 1997; Zhang et al., 1998). The only criterion used to extract helix-packing arrangements was that each helix had to contact at least one other helix, but they did not have to be connected or even from the same subunit. All possible sets of three, four, five, six, and seven contacting helix arrangements were extracted from the structures. Many of the helix assemblies would not be independently stable, but they represent the range of transmembrane helix-packing geometries encountered in nature.

For a particular number of helices, two different library sizes were tested: a 1× and a 10× library. The 1× library corresponded to the number of structures needed to obtain 80% saturation for a given number of helices. The 10× library was 10 times the size of the 1× library and should essentially saturate the space spanned by the computer algorithm. If real helix-packing arrangements are drawn from the same pool as the computer-generated arrangements, roughly 80% of the helix-packing arrangements found in known membrane protein structures are expected in the 1× library, and almost all of them should be in the 10× library.

The number of the helix-packing arrangements, extracted from known membrane protein structures, that are found in the computer-generated structure libraries is shown in Table 1. Of the 87 helix-packing arrangements extracted for three to seven helices, 64 (74%) were found in the 1× libraries and 80 (92%) were found in the 10× libraries. Relatively close structures did exist in the computer-generated libraries for many of the missed helix-packing arrangements. For example, for the seven helix arrangement extracted

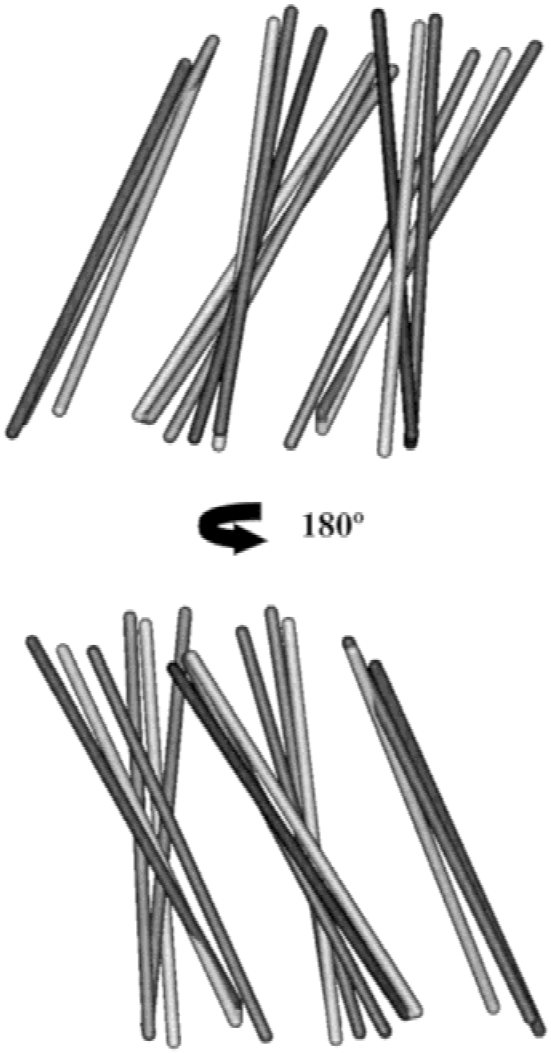

Fig. 3. A collection of five-helix bundles that are considered similar. The structures shown are within 3.8 Å $RMSD^{Hel}$ of each other.

from 1PRC, no structure with an $RMSD^{Hel}$ of less than 4.0 Å was found in the 1× database, but a structure within 4.25 Å $RMSD^{Hel}$ was present. This structure is shown in Figure 6 and appears to be a relatively similar packing arrangement. Thus, the vast majority of transmembrane helix-packing arrangements can indeed be found in the artificially generated structures.

Addition of a compactness constraint

Although the helix-packing arrangements generated at random appear to correspond to the range of helix packings in actual membrane proteins, not all of them would form stable structures independently. In the randomly generated structures shown in Figure 2 and the seven helix arrangement extracted from 1PRC shown in Figure 6, some of the helices splay out from the main body of the structure, resulting in very little contact area. It seems reasonable to expect that most independently folded membrane proteins will tend to be more compact, and these structures should be eliminated from the library. I therefore generated new conformation libraries in which helix-packing arrangements were eliminated if the exposed surface area of any helix was outside the range found for transmembrane helices in known membrane protein structures (see Methods). Three helix-packing arrangements that are judged to be reasonably compact by this definition are circled in Figure 2.

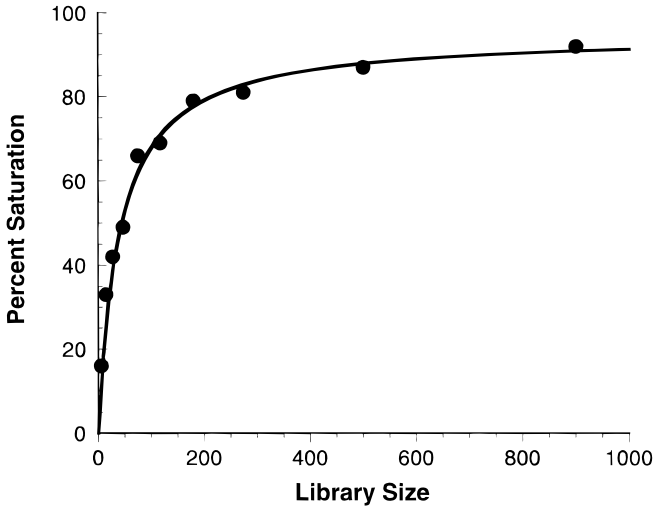

Fig. 4. Percent saturation as a function of library size for four-helix assemblies. For each size of structure library generated by the computer algorithm, 100 additional structures were generated and the plot shows the percent of the additional structures that were present in the library.

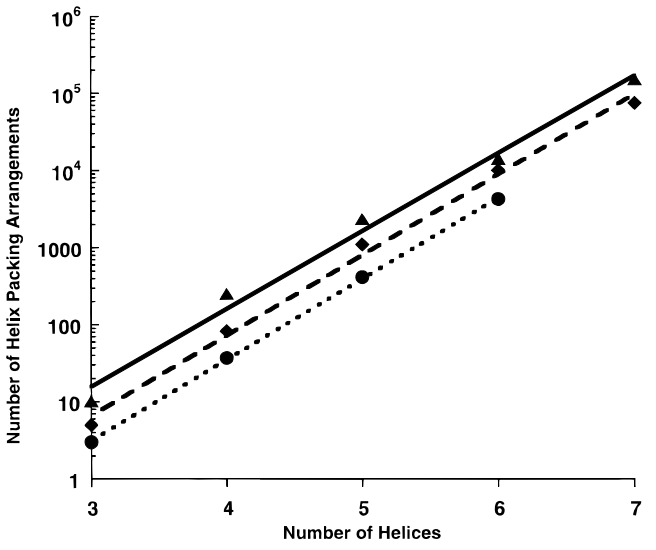

Fig. 5. Number of structures as a function of the number of helices. The number of structures or folds needed to obtain 80% saturation after the application of different constraints is plotted as a function of the number of helices. Triangles: the number of structures for 80% saturation using geometric constraints only. Diamonds: the number of structures for 80% saturation using both geometric constraints and compactness criteria. Circles: the number of different structure clusters obtained after clustering the structures needed for 80% saturation using both geometric constraints and compactness criteria (the number of structures clustered is shown by the diamonds).

Table 1. *Identification of known membrane protein helix packing arrangements in the computer-generated fold libraries*[a]

Number of helices	Library	Number of library structures	Known structures found in library				
			2BRD	1PRC	1BCC	Total	
3	1×	10	13/13	8/15	4/10	25/38	(66%)
	10×	100	13/13	12/15	9/10	34/38	(89%)
4	1×	250	10/10	6/9	5/6	21/25	(84%)
	10×	2,500	10/10	9/9	5/6	24/25	(96%)
5	1×	2,330	6/6	3/5	2/3	11/14	(79%)
	10×	23,300	6/6	5/5	3/3	14/14	(100%)
6	1×	13,900	3/3	2/4	0/1	5/8	(63%)
	10×	139,000	3/3	3/4	1/1	7/8	(88%)
7	1×	150,000	1/1	0/1	—	1/2	(50%)
	10×	1,500,000	1/1	1/1	—	2/2	(100%)

[a]Fold libraries were generated as described in Methods, except the Ω angle distribution used did not include protein from which the packing arrangements were derived. A 1× library corresponds to the number of computer-generated structures needed to achieve 80% saturation. The 10× library contains 10 times the structures used in the 1× library. Helix-packing arrangements were extracted from three known structures, with PDB accession codes 2BRD, 1PRC, and 1BCC. The Ω angle probability distribution that was used to select Ω angles during structure generation was derived from the known membrane protein structures (Bowie, 1997). To eliminate bias, however, the Ω angle distribution used was purged of data derived from the structure being tested. For example, when looking for helix-packing arrangements in 1PRC, the database was generated using an Ω angle distribution derived from all proteins except 1PRC. As a result, the structure of cytochrome *c* oxidase could not be used, because there would be insufficient data for the Ω angle distribution if this structure was eliminated. The numbers below each accession code indicate the number of the known helix-packing arrangements that were found in the library and the number of arrangements tested. For example, there were 15 three-helix arrangements extracted from the 1PRC structure and 8 out of the 15 were found in the 1× library. The last column of the table gives the total number of helix-packing arrangements extracted from the known structures and the number that were found in each of the libraries of computer-generated structures.

The number of compact conformations needed to achieve 80% saturation of the conformation space, as a function of the number of helices, is displayed in Figure 5. The conformation space was reduced about one-half by including the compactness criteria. For seven helices, the number of compact conformations needed for 80% saturation is 75,000 compared to 150,000 without including compactness. Do the compact helix-packing arrangements correspond to real membrane protein conformations? The only membrane protein of known structure that is known to be stable as a monomer is bacteriorhodopsin (Brouillette et al., 1989). I therefore looked for the bacteriorhodopsin fold in a 1× library of compact structures (75,000 structures). Indeed, the bacteriorhodopsin fold occurred 389 times in the 1× compact library. The closest structure had an $RMSD^{Hel}$ of 2.37 Å and is shown in Figure 7.

Clustering of conformations

Not all the helix-packing arrangements in the libraries are distinct. Some structures are more probable than others and will be found with a higher frequency in the fold libraries. Indeed, the fact that some helix-packing arrangements are improbable is precisely why conformation space is limited. For example, structures that contain all $+20°$ helix-packing angles will occur with higher frequency than those with all $-20°$ packing angles, because the positive packing angles are much more likely (Bowie, 1997). Thus, by the time a library is 80% saturated, it can contain many duplicate packing arrangements. To determine the number of unique helical arrangements in the libraries, the conformations were clustered into similar groups.

The results of clustering the 1× libraries of compact conformations are shown in Figure 5. In general, the number of distinct structures in these libraries is about two-thirds the total number of conformations. For example, the compact 1× library for three helices contained five structures and could be represented by only three structures after clustering. Although it was computationally prohibitive to cluster the 75,000 conformations needed for seven helices without a specialized algorithm, linear extrapolation of the data from smaller numbers of helices indicates that the number of different seven helix conformations is about 50,000.

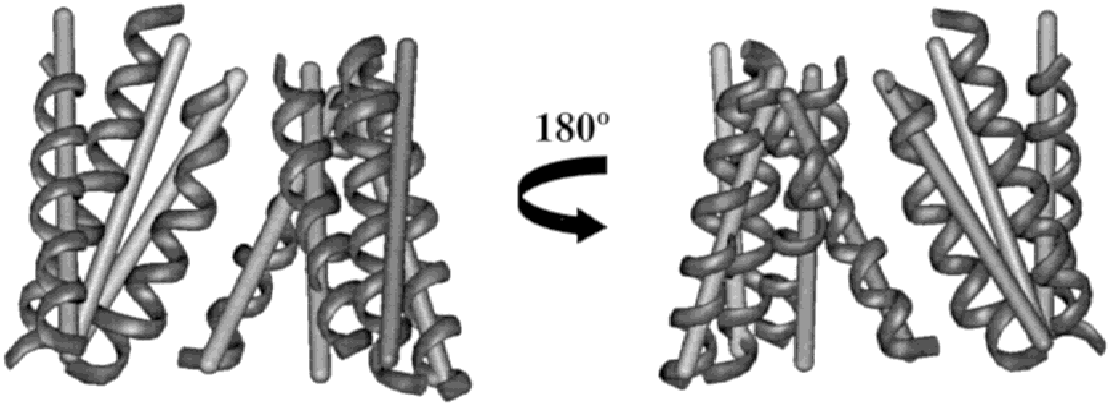

Fig. 6. Comparison of a computer-derived helix-packing arrangement with a seven-helix-packing extracted from 1PRC. The closest match to the seven helix arrangement extracted from the 1PRC structure that was found in the 1× library. Although there is a reasonable correspondence of the helices, the fold shown would not be considered a similar structure by the criterion used, i.e., the $RMSD^{Hel}$ of the 1PRC structure and the computer-generated structure is 4.25 Å. The computer-generated helix-packing arrangement is shown by the rods. The 1PRC structure is shown by the ribbons. The helix-packing arrangement obtained from 1PRC is derived from multiple subunits: four from the L subunit, two from the M subunit, and one from the H subunit.

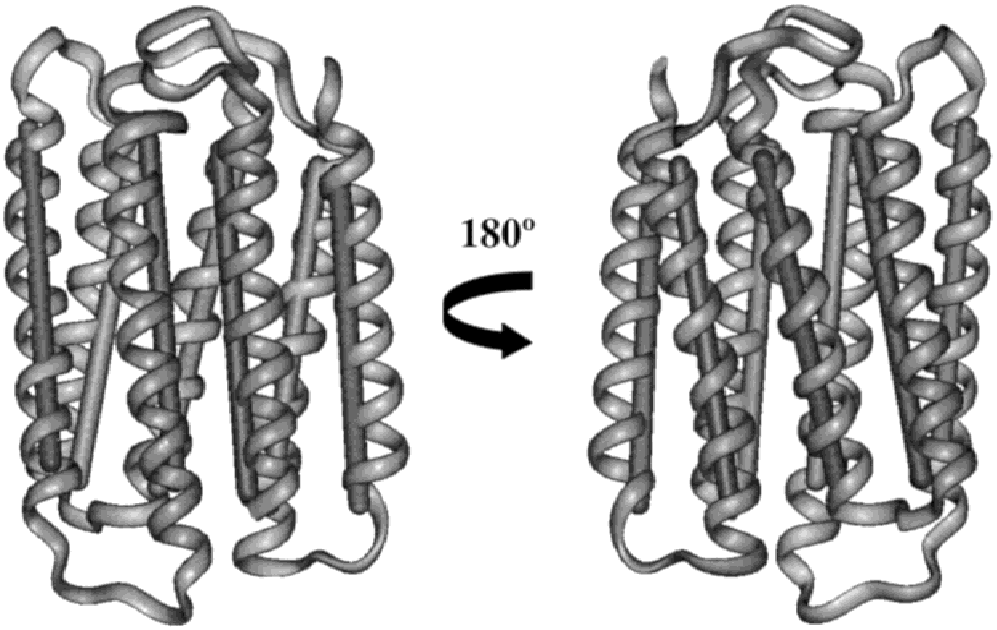

Fig. 7. Comparison of a computer-generated, compact seven-helix arrangement with 2BRD. The closest match to the seven helices of bacteriorhopsin found in the 1× compact database. The $RMSD^{Hel}$ of the structures is 2.37 Å. The computer-generated helix-packing arrangement is shown by the rods and the bacteriorhodopsin structure is shown by the ribbon.

Helix connections

The results presented so far indicate that the vast majority of helix-packing arrangements can be represented by only 3 structures for three helices and 50,000 structures for seven helices. One necessary step in converting these helix-packing arrangements into folds, however, is to specify helix directions and connectivities. Making the reasonable assumption that no connections will pass through the membrane, there are $2 \times N!$ helix connections that are theoretically possible for each helix-packing arrangement, where N is the number of transmembrane helices. For three or four helices, the number of possibilities is still quite modest, but for seven helices there are 10,080 possibilities. If all these threadings were equally probable for every structure, the number of possible structures that need to be considered would rapidly become impractical.

Significant limitations exist on observed helix connectivities in membrane proteins, however, so that not all connections are equally probable. In particular, 97% of helices in membrane proteins are in contact with a neighboring helix in the sequence (Bowie, 1997). This finding greatly reduces the number of likely connections. Moreover, the distances between connected helix end points in known membrane protein structures spans a limited range (see Methods). I therefore applied the following three constraints to decide on acceptable interhelix connections: (1) connected helices must be in contact; (2) the distance between end points must fall in the range observed in known membrane protein structures; and (3) there can be no cross-over connections. Application of these constraints to the randomly generated, compact seven helix arrangements reduces the number of likely connections from 10,080 possible to only 30 on average. This indicates that most seven helix bundle folds can be described by $30 \times 50{,}000 = 1{,}500{,}00$ folds. The number of folds as a function of the number of helices is shown in Figure 8. Naturally, the number of distinct folds decreases dramatically as the number of helices decreases. For three helices, most membrane protein folds can be represented by only 30 structures.

Toward practical fold recognition

I have described a method for exploring the conformation space that is likely to be accessible to membrane proteins and for mea-

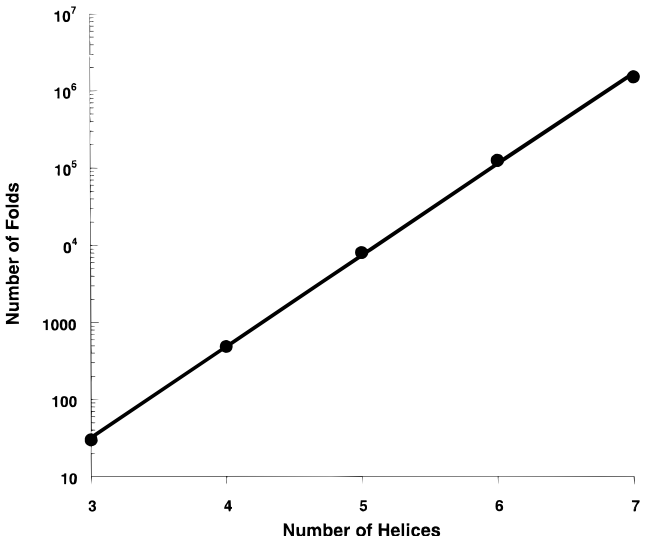

Fig. 8. The number of folds as a function of helix number. The number of folds is obtained by multiplying the number of compact helix arrangements needed for 80% saturation after clustering (circles in Fig. 4), by the average number of ways to connect the helices. The number of seven helix conformations is obtained from the extrapolated value of the curve shown in Fig. 4 (circles).

suring how well conformation space has been sampled. The number of folds needed to saturate the space will naturally depend on the criteria used to judge similarity. Here, I used criteria such that structures judged to be similar would be close enough to be considered easy targets for fold-recognition of soluble proteins, i.e., close enough to imply sequence similarity. By these criteria, it is possible to effectively sample the conformation space accessible to monomeric helix-bundle membrane proteins for up to seven helices. Thus, it seems reasonable to consider a form of fold recognition for smaller membrane proteins in which these computer-generated folding arrangements serve as starting templates for the evaluation of sequence–structure compatibility. Depending on the convergence range of the method used, a finer or coarser sampling of conformation space may be employed.

Because the computer-generated folds are not defined by an atomic model, any practical prediction algorithm will need to deal with additional complexity. In particular, structures based on the folding patterns will have to be constructed that are sufficiently detailed to permit threading of a sequence and scoring by some energy function. Moreover, local conformational searches around the starting template would likely be necessary to optimally orient the sequence in the fold template. Local conformational searching has already been incorporated into some soluble protein fold recognition methods (Godzik et al., 1992).

The level of detail required and the form of the energy function will determine the number of helices that can be handled. Although there has been surprising success in predicting the structure of some helix-bundle membrane proteins using full atom models, these prediction efforts have focused on simple systems with high symmetry constraints or other experimental information (Adams et al., 1995, 1996; Pappu et al., 1999). To deal with more complex systems, it would be advantageous to develop less detailed energy functions for initial model evaluation. Hierarchical methods that can eliminate folds at a very crude level prior to moving to atomic detail will benefit most from these fold libraries and allow the treatment of larger numbers of helices. For example, preliminary orientations of the helices in the structure could possibly be defined using hydrophobicity and sequence conservation (Rees et al., 1989; Cramer et al., 1992; Taylor et al., 1994). Some possible folds could be eliminated by restrictions on loop lengths. Next, residue based energy functions that are often used in fold recognition methods could then be applied (Bowie & Eisenberg, 1993; Fischer et al., 1996; Smith et al., 1997). Similar approaches that start with a large set of possible structures and then cull out the subset of structures compatible with a given sequence have been described for soluble proteins, albeit without the dramatic conformational restrictions that apply to membrane proteins (Cohen et al., 1980; Sternberg et al., 1982; Cohen & Kuntz, 1987; Hinds & Levitt, 1992).

Because the number of conformations increases exponentially with the number of helices (Fig. 8), it is difficult to contemplate handling larger collections of helices without additional constraints. Although the vast majority of membrane proteins contain fewer than seven helices, many important proteins contain larger numbers (Arkin et al., 1997; Wallin & von Heijne, 1998). In such cases, it may be possible to apply experimental constraints to reduce the number of possible folds. For example, knowledge of side chains that bind a ligand or a chromophore could be used to select a subset of reasonable possibilities. Moreover, symmetry constraints can significantly reduce the number of possible conformations and would enable the treatment of larger helical assemblies. Information of this kind could readily be incorporated during fold generation or when selecting possible folds.

The fold recognition paradigm proposed here taps the primary advantage in the structure prediction of membrane proteins compared to soluble proteins, i.e., the reduced conformational possibilities. Sampling the conformation space accessible to soluble proteins at a similar level of detail would be completely intractable. It is expected that the number of soluble protein folds used by nature is only a tiny fraction of the total number possible, however (Chothia, 1992; Zhang & DeLisi, 1998; Govindarajan et al., 1999). Similarly, the number of membrane protein folds that are theoretically possible is unlikely to reflect the number that actually exist in nature. The numbers reported here for seven helix-bundle folds alone greatly exceeds current estimates of the number of extant soluble protein folds (Chothia, 1992; Zhang & DeLisi, 1998; Govindarajan et al., 1999). Nevertheless, the number of theoretically possible folds is sufficiently limited that it may not be necessary to wait for a large library of experimentally determined structures to consider the development of fold recognition methods.

Methods

Structure generation

The parameters describing helix-packing geometry are shown in Figure 1. All helices were represented by a point defining the helix center and a unit vector defining the direction of the helix axis. The helix lengths were set at 30 Å, corresponding to the approximate length of helix needed to span a typical bilayer (Engelman et al., 1986). The central plane of an imaginary bilayer was placed on the *xy* plane so that the membrane normal was along the *z*-axis. The center of the first helix was then set at the origin and rotated in the *yz* plane by a random angle τ between 0 and 40°. Subsequent helices were added with reference to a randomly chosen prior helix. The chosen prior helix was temporarily oriented on the *z*-axis and translated along the z-axis a random distance *C* between -15 and $+15$ Å. The center of the new helix was then placed along the *x*-axis at a distance *D* selected from a Gaussian distribution with a mean of 9.6 Å and a standard deviation of 1.9 Å (Bowie, 1997). The distribution was truncated at 2 standard deviations from the mean. The new helix was then rotated by an helix-packing angle Ω, selected from the distribution of Ω angles seen in known membrane protein structures (Bowie, 1997). The new helix was then rotated by an arbitrary angle θ about the *z*-axis (the axis of the prior helix). The coordinates of the new two helix assembly were then transformed back to the original position of the prior helix. In this new position, the center of the prior helix is back on the central plane, but the center of the new helix is not. The center point of the new helix was then slid back to the central plane along the new helix axis, preserving the packing orientation. Given the finite helix lengths, however, the two helices may no longer be in contact. If, after center adjustment, the distance of closest approach of the new helix was greater than 13.4 Å (2 standard deviations from the mean), the new helix was rejected and the process was repeated. The acceptability of the new helix placement was further evaluated by determining the τ angle and testing for steric conflicts with other helices. If the τ angle was greater than 40°, the new helix was rejected and the process repeated. If the distance of closest approach of the new helix and any of the other helices was less than 6 Å, it was deemed a steric conflict, and the new helix was rejected. In this manner membrane–protein-like helix-packing

arrangements consisting of an arbitrary number of helices could be rapidly generated. A program to generate helix arrangements is available upon request.

Structure comparison

Because helix order is not specified, it was necessary to order the helices appropriately for superposition and RMSD calculation. This was done by first selecting the pair of helices in the first conformation whose centers were farthest apart: the maximum pair. Center-to-center distances of all pairs of helices in the second conformation were then determined, and if they were within 5 Å of the maximum pair from the first conformation, it was considered a possible match. The second conformation was then structurally aligned with the first conformation by superimposing the selected helix pair with the maximum pair in both orientations. The helices that most closely matched in the two structures after superposition were treated as equivalent helices, and the superposition repeated on the full set of equivalent helices. The process was repeated on all helix pairs in the second conformation that could potentially align with the maximum pair of the first conformation, and the lowest RMSD of all the combinations was determined.

Helix-packing arrangements in known membrane protein structures

The transmembrane helix structures were converted into a simplified representation such as the one used for the fold generation algorithm. Each helix was represented by a 30 Å long segment describing the helix axis with a point on the central plane and a unit vector describing the helix direction. The central plane was defined by the coordinates of the center of mass of all the helix atoms and the membrane normal. For 2brd and 1bcc, the membrane normal was defined as the axis of rotation that optimally superimposes the asymmetric units of the oligomers. For 1prc, the membrane normal was defined as the average axis direction of the transmembrane helices.

All helix-packing arrangements were selected from a single asymmetric unit of the oligomeric structures 1bcc and 2brd to avoid double counting. Only helix-packing arrangements were chosen for which each helix contacted at least one other helix. Contact was defined by the method of Chothia (Chothia et al., 1981), with the added criterion that the distance of closest approach was less than 13.4 Å. The latter criterion eliminates errors in the placement of the central plane that could result in truncation of helix segments before the true point of closest approach.

Clustering

A simple clustering algorithm was used that did not employ an exhaustive comparison. A structure was selected at random to represent the first cluster. All other structures were compared to this structure as described above and were added to the cluster if deemed similar. Next the coordinates of the clustered conformations were averaged and structures similar to the averaged structure were added to the cluster. The process was repeated on a new random structure until all structures had been assigned to a cluster. Many clusters are represented by only one conformation.

Compactness

The compactness of a conformation was assessed by determining the fractional area buried of helix, represented by a cylinder. The cylinders were 30 Å long, and the axes corresponded to the helix axes. The fractional area buried was determined by finding what fraction of a set of points, distributed on the cylinder surface, were buried. The points were placed on circles, perpendicular to the cylinder axis, every δ degrees. The starting angle for the placement of points was displaced by 0° or $\delta/2$°. The circles were placed every 2 Å so that each cylinder was sampled by 16 circles of points. Given the circle spacing and the need to use a value of δ evenly divisible into 360°, only certain values of theta and cylinder radii could be used to achieve evenly spaced points. I chose cylinder radii of 6.65 and 13.24 Å and δ angles of 20 and 10°, respectively. With these parameters, all sample points were evenly placed 2.31 Å apart. A sample point was considered buried if it was within another cylinder. A structure generated by the computer algorithm was deemed to be compact if the fractional area buried values for each helix was within the ranges seen in known membrane protein structures. When known membrane protein structures were converted to the helix axis representation as described above, the values of fractional area buried were all in the range of 0.17 to 0.86 with the small cylinder radius. For the large radius, all the helices in known structures were found to have a fractional area buried greater than 0.42.

Helix connections

All possible helix connections that did not pass through the bilayer were considered. Connections were rejected if the connected helices were not in contact, a crossover connection resulted or if the length of the connection was outside the normal range seen in membrane proteins structures. Two connections were considered crossovers if the distance of closest approach of two lines passing through the helix endpoints was within a segment connecting the two endpoints. The range of connection distances in known membrane proteins was determined from the simplified representations of their transmembrane helix domains described above. For these simplified representations, 38 of the 39 connection distances were between 7 and 22 Å. There was one outlier at 37.6 Å. Thus, a connection was rejected if the connection distance was outside the range of 7 to 22 Å.

Acknowledgments

I would like to thank Frank Pettit and Tau Mu Yi for helpful discussions and Chris Thanos for comments on the manuscript. This work was supported by a grant from the DOE, a Pew scholar award, and an NSF National Young Investigator Award to J.U.B.

References

Adams P, Arkin I, Engelman D, Brünger A. 1995. Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nat Struct Biol 2*:154–162.

Adams P, Engelman D, Brünger A. 1996. Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. *Proteins 26*:257–261.

Arkin I, Brünger A, Engelman D. 1997. Are there dominant membrane protein families with a given number of helices? *Proteins 28*:465–466.

Bowie J. 1997. Helix packing in membrane proteins. *J Mol Biol 272*:780–789.

Bowie J, Eisenberg D. 1993. Inverted structure prediction. *Curr Opin Struct Biol 3*:437–444.

Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Boyd D, Schierle C, Beckwith J. 1998. How many membrane proteins are there? *Protein Sci 7*:201–205.

Brouillette C, McMichens R, Stern L, Khorana H. 1989. Structure and thermal

stability of monomeric bacteriorhodopsin in mixed phospholipid/detergent micelles. *Proteins Struct Funct Genet 5*:38.
Chothia C. 1992. One thousand protein families for the molecular biologist. *Nature 357*:543–544.
Chothia C, Levitt M, Richardson D. 1981. Helix to helix packing in proteins. *J Mol Biol 145*:215–250.
Cohen F, Kuntz I. 1987. Prediction of the three-dimensional structure of human growth hormone. *Proteins 2*:162–166.
Cohen F, Sternberg M, Taylor W. 1980. Analysis and prediction of protein β-sheet structures by a combinatorial approach. *Nature 285*:378–382.
Cramer W, Engelman D, von Heinje G, Rees D. 1992. Forces involved in the assembly and stabilization of membrane proteins. *FASEB J 6*:3397.
Deisenhofer J, Epp O, Sinning I, Michel H. 1995. Crystallographic refinement at 2.3 A resolution and refined model of the photosynthetic reaction centre from *Rhodopseudomonas viridis*. *J Mol Biol 246*:429–457.
Engelman D, Steitz T, Goldman A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem 15*:321.
Fischer D, Rice D, Bowie JU, Eisenberg D. 1996. Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J 10*:126–136.
Godzik A, Kolinksi A, Skolnick J. 1992. Topology fingerprint approach to the inverse folding problem. *J Mol Biol 227*:227–238.
Govindarajan S, Recabarren R, Goldstein R. 1999. Estimating the total number of protein folds. *Proteins 35*:408–414.
Grigorieff N, Ceska T, Downing K, Baldwin J, Henderson R. 1996. Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J Mol Biol 259*:393–421.
Hinds D, Levitt M. 1992. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci USA 89*:2536–2540.
Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.
Koehl P, Levitt M. 1999. A brighter future of protein structure prediction. *Nat Struct Biol 6*:108–111.
Marchler-Bauer A, Bryant S. 1997. A measure of success in fold recognition. *Trends Biosci 22*:236–240.
Montelione G, Anderson S. 1999. Structural genomics: Keystone for a human proteome project. *Nat Struct Biol 6*:11–12.
Pappu R, Marshall G, Ponder J. 1999. A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat Struct Biol 6*:50–55.
Pebay-Peyroula E, Rummel G, Rosenbusch J, Landau E. 1997. X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science 277*:1676–1681.
Rees D, DeAntonio L, Eisenberg D. 1989. Hydrophobic organization of membrane proteins. *Science 245*:510.
Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct Funct Genet 9*:56–68.
Smith T, Lo Conte L, Bienkowska J, Gaitatzes C, Rogers RJ, Lathrop R. 1997. Current limitations to protein threading approaches. *J Comput Biol 4*:217–225.
Sternberg M, Cohen F, Taylor W. 1982. A combinatorial approach to the prediction of the tertiary fold of globular proteins. *Biochem Soc Trans 10*:299–301.
Taylor W, Jones D, Green N. 1994. A method for α-helical integral membrane protein fold prediction. *Proteins Struct Funct Genet 18*:281–294.
Terwilliger T, Waldo G, Peat T, Newman J, Chu K, Berendzen J. 1998. Class-directed structure determination: Foundation for a protein structure initiative. *Protein Sci 7*:1851–1856.
Wallin E, von Heijne G. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci 7*:1029–1038.
Zhang C, DeLisi C. 1998. Estimating the number of protein folds. *J Mol Biol 284*:1301–1305.
Zhang Z, Huang L, Shulmeister V, Chi Y, Kim K, Hung L, Crofts A, Berry E, Kim S. 1998. Electron transfer by domain movement in cytochrome bc1. *Nature 392*:S677–S684.