doi:10.1016/S0022-2836(02)00590-9 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 321, 363–378

JMB

A Novel Scoring Function for Predicting the Conformations of Tightly Packed Pairs of Transmembrane α-Helices

Sarel J. Fleishman and Nir Ben-Tal*

Department of Biochemistry George S. Wise Faculty of Life Sciences, Tel-Aviv University 69978 Ramat-Aviv, Israel

*Corresponding author

Pairs of helices in transmembrane (TM) proteins are often tightly packed. We present a scoring function and a computational methodology for predicting the tertiary fold of a pair of α-helices such that its chances of being tightly packed are maximized. Since the number of TM protein structures solved to date is small, it seems unlikely that a reliable scoring function derived statistically from the known set of TM protein structures will be available in the near future. We therefore constructed a scoring function based on the qualitative insights gained in the past two decades from the solved structures of TM and soluble proteins. In brief, we reward the formation of contacts between small amino acid residues such as Gly, Cys, and Ser, that are known to promote dimerization of helices, and penalize the burial of large amino acid residues such as Arg and Trp. As a case study, we show that our method predicts the native structure of the TM homodimer glycophorin A (GpA) to be, in essence, at the global score optimum. In addition, by correlating our results with empirical point mutations on this homodimer, we demonstrate that our method can be a helpful adjunct to mutation analysis. We present a data set of canonical α-helices from the solved structures of TM proteins and provide a set of programs for analyzing it (http://ashtoret.tau.ac.il/~sarel). From this data set we derived 11 helix pairs, and conducted searches around their native states as a further test of our method. Approximately 73% of our predictions showed a reasonable fit (RMS deviation <2 Å) with the native structures compared to the success rate of 8% expected by chance. The search method we employ is less effective for helix pairs that are connected *via* short loops (<20 amino acid residues), indicating that short loops may play an important role in determining the conformation of α-helices in TM proteins.

© 2002 Elsevier Science Ltd. All rights reserved

Keywords: empirical energy function; ridges into grooves; transmembrane helices database; tight packing; structure prediction

Introduction

Transmembrane (TM) proteins are crucial mediators of cell-to-cell signaling and of transport processes. This makes them attractive targets for drug discovery as well as for improving our understanding of cellular processes. Despite their importance, however, only about a dozen distinct folds of TM proteins have been solved to date by such high-resolution methods as crystallography and NMR. Attempts to determine the structure of this class of proteins by these methods are hampered seriously by technical problems related to their purification and crystallization. It would therefore be advantageous if these technical difficulties could be bypassed, and the structure of these proteins inferred by computational means.

Because the number of TM proteins whose structures have been solved at high resolution is small, an energy-like contact potential cannot be constructed by straightforward statistical means. Our approach has been to construct a quasi-energy scoring function based on qualitative analyses of TM protein structures carried out over the past decade. We hope that this work may be used also

Abbreviations used: GpA, glycophorin A; RTK, receptor tyrosine kinase; TM, transmembrane.

E-mail address of the corresponding author: bental@ashtoret.tau.ac.il; http://ashtoret.tau.ac.il

0022-2836/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved

as an evaluation of the current level of understanding of the factors driving helix association in TM proteins.[1]

Structure prediction in soluble proteins by computational methods is considered extremely difficult, largely because of the variety of possible folds, which implies a vast number of degrees of freedom. In contrast, TM proteins may be grouped into two classes, the α-helix bundle and the β-barrel. This considerably reduces the number of degrees of freedom that determine the structures of these proteins. Here, we concern ourselves only with the α-helix bundle class, which is the only one known to inhabit the plasma membrane.

According to the widely accepted two-stage model,[2] the first step in TM protein folding is the insertion into the membrane of the TM domains as α-helices. Only in the second stage do these helices associate to form helix bundles. (For recent reviews of this and other thermodynamic models of membrane protein folding, see Popot & Engelman[3] and White & Wimley.[4])

One of the implications of the two-stage model is that, overall, the stability of individual TM domains is independent of that of other domains. Hence, prediction of TM protein structure may begin with prediction of TM helix locations on amino acid sequences. The past few years have seen much progress in computational methods devised for this purpose.[5] Algorithms for determining the topology of these segments in the membrane, i.e. for establishing whether the N terminus is inside or outside the cell, have been successful.[6] There is room for improvement in the understanding of this stage of protein folding, but essentially it has been well explored. Here, we reduce the problem of TM protein structure prediction to the problem of predicting the correct packing of rigid α-helices. Deviations from ideal α-helicity, such as kinking and uncoiling, are indeed encountered in TM proteins, and are known to have functional importance.[7] However, since there are no known methods for predicting these phenomena from sequences, we do not address them here.

Some early attempts were made to predict helix orientations in relation to each other by using the hydrophobic moment concept.[8,9] However, in view of the hydrophobic nature of the membrane, the hydrophobic driving force is probably less important in this medium than in soluble proteins, and the hydrophobic moment has proved to be of limited use in TM structure prediction.[10,11] The main driving force for the folding process is thus considered to be the efficient packing of helices.[12]

Attempts have been made to predict the structure of specific TM proteins.[13–20] For high-resolution structure prediction of pairs of TM α-helices Adams *et al*.[18] developed a method based on molecular dynamics, utilizing data derived from mutational analyses. Briggs *et al*.[19] extended this method by using phylogenetic data instead of mutational analyses. Pappu *et al*.[20] showed that the computational load associated with searches in conformational space, using models in atomic detail, may be reduced considerably by the use of a potential-smoothing technique. They demonstrated the competence of the approach by successfully retrieving the structure of glycophorin A (GpA). Based on their experience, we explore the possibility of reducing the computational burden further by using low resolution from the outset. This allows us to carry out an exhaustive search of conformational space, and it enables us to systematically test the method on many examples.

The number of solved TM protein structures is relatively small, and the factors driving helix association in the membrane are still poorly understood.[1] Nevertheless, several studies have offered substantial qualitative insight into TM helix–helix dimerization. It was shown that TM helices are at least as tightly packed as helices of soluble proteins, and that small residues (Ala and Gly) and small hydroxyl-containing residues (Ser and Thr) are often buried deeply in TM proteins.[12,21] An important role was ascribed to Gly in mediating helix–helix contacts in TM proteins.[22,23] In an attempt to overcome some of the limitations that are inherent in the analysis of residue propensities in TM proteins because of the small number of solved TM protein structures, Senes *et al*.[24] carried out a statistically more extensive study on the sequences of TM proteins. Their results reinforce the conclusions of the qualitative studies, and suggest that TM helix interactions are often mediated by β-branched amino acid residues (Ile and Val) and, to a lesser extent, by the γ-branched amino acid residue Leu.

Lemmon & Engelman[25] offer an explanation for these dimerization-related phenomena in terms of the so-called lipophobic effect. They argue that the presence of small residues on the face of the helix leads to the formation of cavities, should the helix interact with the "cylindrical" lipid chains. Cavity formation is considered costly in terms of energy. On the other hand, these cavities may be eliminated by another helix with an accommodating pattern of large and small residues. The β-branched amino acids are thought to be preferable for dimerization because their rotamers are constrained within the context of an α-helix.[26] This reduces the entropy loss that usually accompanies the association of protein parts.

Recently, Senes *et al*.[27] showed that hydrogen bonding, with C^{α} acting as hydrogen donor, may be an important factor in driving helix–helix association in TM proteins. Their analysis provides a thermodynamic justification for the important role of amino acid residues with small side-chains (Ala, Gly, Ser and Thr) in mediating helix–helix dimerization; their small volume makes the backbone atoms more accessible. Recent work has shown that hydrogen bonds between polar side-chains, e.g. Asn–Asn, play a significant role in stabilizing helix association in model TM

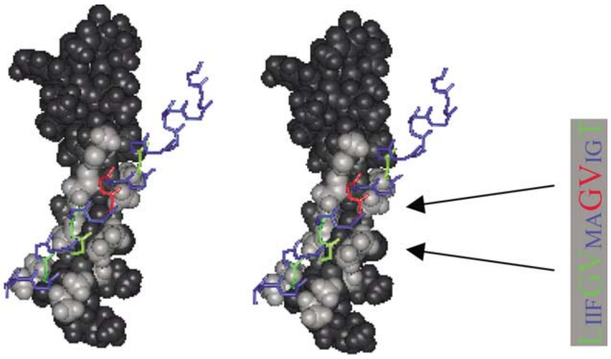


Figure 1. Stereo view of the TM segment of the human homodimer GpA in its native state (PDB code: 1afo). Only the first model of the collection of NMR structures is displayed. One monomer is presented in all-atom CPK rendering and the other shows only the backbone atoms. The light-colored residues on the CPK rendering show the pattern of two ridges on the face of a monomer. The ridge on the right-hand side is formed by the side-chains of Ile76, Val80 and Val84. Gly79 and Gly83 on the other monomer pack against this ridge, allowing for close interactions among backbone atoms. The ridge on the left-hand side is not continuous due to the presence of Gly79 and Gly83. The monomer represented by backbone atoms is colored differentially according to the burial scores of the residues; blue signifies residues that are not buried in the other monomer, green indicates intermediately buried residues, and red indicates residues that are well buried. The burial score of an amino acid residue is computed according to the distance and the angular orientation of the C^α and the axis of the other helix (see Methods). The two helices are packed against each other at a crossing angle of about $-40°$. On the right is a portion of the amino acid sequence of GpA's TM domain in one-letter code, colored according to the burial scores B. Large letters signify residues that mediate contact. Note that the middle part of the sequence contains the GxxxG motif.[24]

proteins.[28–30] Hydrogen bonds have also been found among side-chains in solved TM protein structures.[31]

In a preliminary study, we examined whether the burial of amino acid residues that are frequently observed at helix–helix interfaces may by itself provide a criterion for determining the native-state packing of two helices. We found that this criterion leads, in almost all cases, to the formation of helix dimers with their principal axes essentially parallel with each other, so that the crossing angle between the two helices is negligible. This contrasts with the findings in many solved TM structures, which show a preference for different crossing angles,[32] and predominantly for a crossing angle of approximately $+20°$.

We therefore employed the "ridges-into-grooves" structural motif described by Chothia *et al.*,[33] which is typical for tightly packed helix pairs, e.g. GpA (Figure 1). Chothia *et al.* argue that α-helices are not smooth cylinders, as their side-chains form protrusions on the face of the helix. Residues that are separated by one, three or four residues on the sequence may form continuous ridges on the face of the helix. These ridges are separated by grooves. In order to maximize the hydrophobic surface area making contact between these side-chains, and to minimize cavities, a ridge on one helix may be inserted into a groove on the face of another. For example, a conformation in which a ridge formed by residues separated by four amino acid residues in the sequence of one helix is associated with a ridge separated by three amino acid residues in the sequence of another helix is called 4–3 class packing. Analysis of this model had helped explain why certain crossing angles predominate in the packing of helices.[33]

Recently, Bowie[34] and Walther *et al.*[35] showed that in the case of globular proteins, much of the preference for the packing angles predicted by the ridges-into-grooves model is actually a result of statistical bias. Their results demonstrate that when the packing-angle propensities are normalized, the preference for these packing angles is not as pronounced as expected from the ridges-into-grooves model alone.[33] Nevertheless, the case of TM helices is different, since the area mediating contact between the helices in TM proteins is usually larger than that in globular proteins.[34] Therefore, steric packing is likely to play a more important role in determining the structure of these proteins.[34] We stress that in tightly packed helices (defined here as helices in which the distance between the principal axes is less than 9 Å), steric packing is likely to play an important role. This is because the very short distance between the principal axes of the helices essentially compels the side-chains in the contact region on one helix to be accommodated by the grooves of the other helix.

Our method in constructing the scoring function for discriminating conformations that would allow tight packing from those that would not was to formulate the qualitative insights pertaining to solved TM protein structures, as presented above, in a quantitative manner. We then tested this formulation against a selection of helix pairs from the solved structures of membrane proteins.

Here, we present our scoring function for contact between TM helices. As a case study, we examine the TM homodimer GpA, and discuss at length our computational results on this protein in the light of empirical mutation analyses[36] and its structure determination.[26,37] We present our results of searches for optimal structures of 11 TM helix pairs derived from TM proteins of known structure.

The Proposed Model

Our aim in this work was to construct a scoring function to distinguish conformations that allow tight packing of a pair of helices from conformations that impede such packing. The helices are reduced to their C^α trace. Our function attaches a score based on the amino acid composition of the

Table 1. The maximum score that can be contributed to the total conformational score by a pair of contacting residues

	Gly	β-Branched (I,T,V)	Small (A,C,S)	Constrained (L,N,P)	Others
Gly	−1	−1	−1/2	−1/4	0
β-Branched (I,T,V)	−1	0	−1/2	0	0
Small (A,C,S)	−1/2	−1/2	−1/2	0	0
Constrained (L,N,P)	−1/4	0	0	0	0
Others	0	0	0	0	0

Residues are grouped according to their steric characteristics, and reported in the one-letter code. I, Ile; T, Thr; V, Val; A, Ala; C, Cys; S, Ser; N, Asn; L, Leu; P, Pro. The category Others includes all other amino acids, which contribute zero to the total score. The values reflect the structural analyses described in Introduction.

helices and the space coordinates of their C^{α} atoms to each conformation of two helices. The function is defined such that its minima are associated with tightly packed conformations. For tightly packed pairs of helices, one of these minima should be the native state.

Our approach in computing the score of a given conformation of helices is to maximize the number of contacts between residues that promote helix interactions and to penalize the burial of large amino acid residues. We score any conformation of a given pair of helices as the sum of two terms: a negative term contributing to the score for contact between pairs of residues known to promote close-packing among helices, and a positive term penalizing the burial of large residues in the interface. The optimal score is thus expected to be a global minimum:

$$\text{Score} = \sum (B^i + B^j) M^{(i,j)} + 10 \sum B^l \ : \ (i,j) \in P, l \in L \tag{1}$$

where P is the set of all pairs of residues forming contact in a given conformation, and B^i and B^j are approximations of the burial of the two residues i and j forming that contact between two different helices, as described in Methods. Values for B range from 0, signifying no burial, to 1, signifying complete burial (Figure 1). L is the set of all amino acid residues l with large side-chains (Arg, His, Lys, Met, Phe, Trp and Tyr) that appear on either helix, and are well buried in the interface between the helices ($B^l \geq 0.9$). We penalize the burial of large residues only if they are buried to a large extent in the interface; in other cases, large residues may often assume accommodating conformers, and not form steric hindrances.

$M^{(i,j)}$ is the maximal score contributed by each pair of amino acid residues when mediating contact between a pair of helices (Table 1). To determine these contributions we considered four classes of amino acid residues: Gly; the small residues (Ala, Cys and Ser); the β-branched residues (Ile, Thr and Val); and residues with constrained side-chains (the γ-branched residues Asn and Leu, and Pro). In the absence of a direct statistical method to calculate the relative contribution of each pair of residues to the formation of contacts among helices, we used only four values (0, $-\frac{1}{4}$, $-\frac{1}{2}$ and -1) to reflect the relative contribution of each pair to dimer formation. These values are a crude approximation of the qualitative data available in the literature and presented in Introduction. Thus, since Gly–Gly and Gly–Val contacts have been shown to be favorable for promoting helix contact formation,[23,24,26,38,39] their respective classes contribute substantially to the overall score.

It was recently suggested that the C^{α}–H···O hydrogen bond[27] is a driving force for TM helix contact formation. By promoting contacts between Gly and small residues such as Ala and Cys, as well as contacts between Gly and the hydroxyl-containing amino acid residues Ser and Thr, the scoring function favors contacts between the C^{α} atom of Gly to either the backbone or side-chain hydrogen-bond acceptors. Interhelical hydrogen bonds among polar side-chains were recently shown to strongly promote association of model helices in the membrane.[1,28–30] Though our scoring function is concerned mainly with tight packing of helix pairs, some of the reported hydrogen-bond interactions are included implicitly, e.g. Ser–Ser, and Ser–Thr. Contacts among residues that do not belong to any of the above mentioned classes make no contribution to the overall score.

Results

Glycophorin A

The TM protein on which we have focused most attention as a representative of tightly packed TM proteins is the human GpA.[40] GpA is a monotopic sialoglycoprotein, which is abundant as a homodimer in erythrocyte membranes (Figure 1). In the past decade, the relationship of its amino acid composition to its dimerization characteristics has been scrutinized by a combination of mutational[36,38,41] and computational analyses.[42,43] Recently, its structure was solved both in micelles[26] and in membrane bilayers.[37] The structure conforms to many of the conclusions derived by the mutational analyses. The essential elements of the dimerization of GpA are interactions between Gly and Val residues. Apart from that, the two helices form the ridges-into-grooves class 4–4 packing motif.[33] Because the movement of the two helices comprising the homodimer is not constrained by

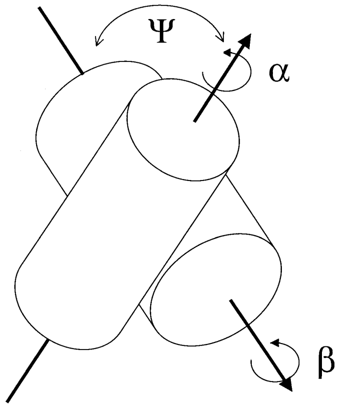


Figure 2. Six degrees of freedom are associated with each helix pair: Three rotational degrees of freedom are indicated in the Figure (α and β represent rotations of the helices around their principal axes; Ψ represents the pair's crossing angle). α and β were set arbitrarily to zero in the native-state conformation. Three translational degrees of freedom set the geometric center of one helix with respect to the other, corresponding to the inter-helical distance (y), the height of one helix with respect to the other (z), and a sliding movement of one helix across the face of the other (x). In all our analyses, y was restricted to the value in the native-state conformation of the helix pair. The large arrows mark the principal axes of the helices. For homodimers such as GpA (Figure 1), we may assume that the structure is symmetrical and therefore force $\alpha = \beta$, and $z = 0$.

an interconnecting loop, the helices may be considered free to sample any configuration. This protein is therefore a particularly suitable example on which to test our scoring function.

The native-state conformation of GpA is situated at a global score optimum

The fact that GpA is a homodimer guarantees that its helices will form a nearly symmetrical tertiary structure.[44] By enforcing symmetry, we substantially diminish the number of degrees of freedom examined in our search method to three: the crossing angle (Ψ); one rotational degree of freedom around the axes of the helices (α); and one translation (x) (Figure 2). This allows us to use an extensive search range and obtain a fine resolution.

Using InsightII/Biopolymer (Accelrys, San Diego), we constructed an approximation of the C^{α} trace of GpA as a homodimer composed of two ideal α-helices. We explored the scoring function for this structure using a high-resolution "score surface" (in analogy with the commonly used term potential surface) in a cross-section, such that x is set at its value in the native state (Figure 3). The local minimum shown in Figure 3 at $\alpha = 0°$,

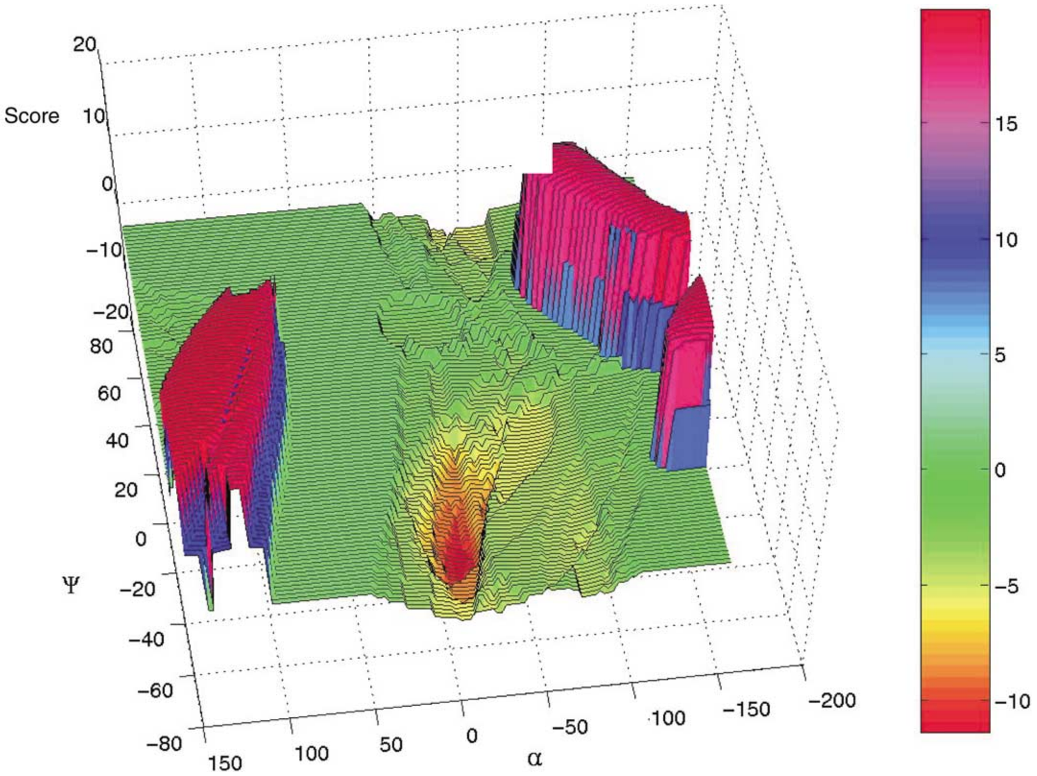


Figure 3. "Score surface" for the homodimeric TM protein GpA around its native state. The structure used for generating this surface is based on two ideal α-helices. The surface was generated by fixing x at its value in the native-state conformation (3.88 Å). While enforcing symmetry ($\alpha = \beta$; see Figure 2), the crossing angle Ψ and the rotation around each monomer's principal axes were modulated. The ranges and step sizes used are: α in the range of −180° to 150° with 2.5° step size; and Ψ in the range of −60° to 80° with 1.5° step size. Note that the native state[26] ($\alpha = 0°$, $\Psi = 40°$) is situated in a score well. The score peaks are associated with the burial of large amino acid residues in the interface. The landscape, as expected, is discontinuous. It is noteworthy, however, that the discontinuity is not very great, and that using a rather coarse step size of 10° for α and 5° for Ψ should suffice to capture its major features.

$\Psi = 40°$ is also the global minimum. Its RMS deviation from the native structure of GpA is 1.41 Å, indicating that ideal α-helices may be used in case the secondary structure is not known with certainty.

It is notable that the region around the minimum in Figure 3 seems to be distinct and large. The scoring function is expected to be discontinuous, but the score surface demonstrates that it is not extremely so, and that searching with a rather coarse resolution (step sizes of 10° for α and 5° for Ψ) would probably not miss the major features of the score surface. We also conducted a search based on the GpA protein database (PDB) structure (1afo)[26] without enforcing symmetry. This resulted in a near-symmetrical structure with an RMS deviation of less than 0.9 Å from the native-state structure (Table 2).

We examined the effect of modulating y, representing the distance between the helices' axes of symmetry, on the optimal structure and its score (Table 3). In general, increasing y results in a less favorable score. These results can be grouped into three categories on the basis of similarity between the optimal conformations: those obtained for interhelical distances between 6 and 7.5 Å; those between 7.5 and 8 Å; and those between 8 and 9 Å. This indicates that in the cases where the interhelical distance is not known with certainty, configuration space can be searched at two or three distinct interhelical distances, e.g. below 7.5 Å and above 8 Å.

Computational results correlate with empirical mutation analysis

We proceeded to determine whether our method could distinguish mutations that hinder dimer formation from those that do not. For this purpose, we analyzed all 106 non-redundant non-polar point mutations carried out by Lemmon *et al.*[36] On the assumption that the mutant GpA monomers form an ideal α-helical secondary structure, we built a C^α trace model for these structures using InsightII/Biopolymer (Accelrys, San Diego). By treating only symmetric conformations, we reduced the number of degrees of freedom to three, as described above. This decrease in the number of degrees of freedom allowed us to carry out a fine-grained search of the structures across much of the conformation space. The search ranges and step sizes are specified in Figure 4.

Figure 4 shows a comparison of our results with the mutation analysis conducted by Lemmon *et al.*[36] We define as a disruption to dimer formation any change in the score of the optimal structure or any deviation of its configuration relative to the optimal structure obtained for the wild-type sequence. Lemmon *et al.*[36] classify their empirical results according to four categories based on the ability of the point mutants to form dimers as well as the wild-type, in significant quantity, in detectable quantity, and no dimer formation. For the purposes of this comparison, we group the classes defined by Lemmon *et al.*[36] as same as wild type and in significant quantity (categories 1 and 2, respectively, in Figure 4), and compare them to our dimer formation class. The other two classes defined by Lemmon *et al.*[36] are compared to our dimer disruption category. It should be noted that our treatment does not allow us to make the distinction made by empirical mutation analyses with regard to the extent of dimer formation.

Our results show a positive correlation with those of Lemmon *et al.*[36] ($r^2 = 0.201$). Significantly, the characteristic mutation Gly83Ala, which abolishes dimer formation *in vitro*,[36] is also disruptive according to our analysis†.

A database of helix pairs

To examine whether our method could distinguish the native-state conformation of pairs of helices from near-native conformations, we analyzed 11 helix pairs chosen from various TM proteins according to automatic procedures as elaborated in Methods (Table 2). The helix pairs were used as they appear in the PDB, i.e. with their deviations from α-helix ideality maintained. We used a five-dimensional lattice to map the conformation space around the native state of each helix pair chosen (Figure 2). The search ranges and step sizes are specified in Table 2.

Our use of a lattice places considerable limitations on the conformation space examined and hence on the range of expected RMS deviation values. We therefore compared the RMS deviation values we obtained for the set of 11 helix pairs to a set of randomly generated structures (Figure 5). We constructed the random set of structures by generating 2000 conformations of the helix pair 1,7 of bacteriorhodopsin (PDB code: 1c3w) throughout the range defined in Table 2 of the five-dimensional lattice with uniform probability. We then calculated the RMS deviation of each of these structures from the helix pair's native-state conformation.

The results presented in Figure 5 indicate that our method yields optimal structures that are close to the native state (<2 Å RMS deviation) in 73% of the cases, as opposed to 8% expected by chance. In some cases (marked with an asterisk) the optimal results are at the end of the search range, and may therefore be underestimates of the real RMS deviation. We did not conduct searches across a larger part of the conformation space, because we treat a helix pair independently of the contacts it forms with other helices. In reality, TM helices in polytopic proteins often form contacts with more than one helix.[21] Such contacts constrain the helix pair from exploring conformations that are far from its native state.

† For supplementary material, see: http://ashtoret.tau.ac.il/~sarel

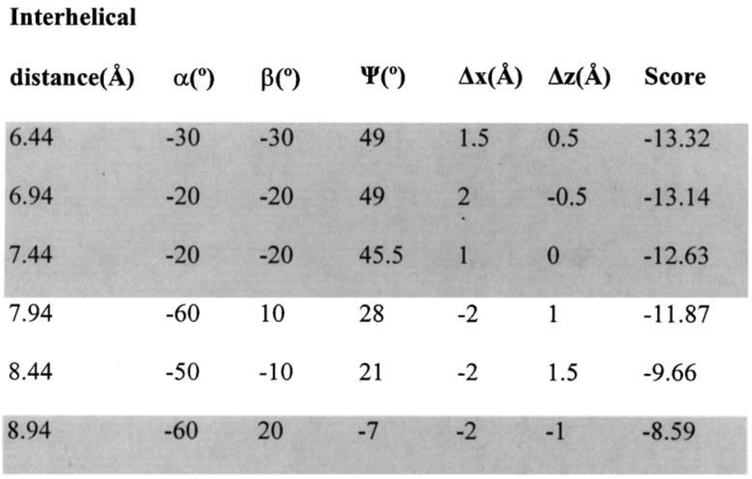
Table 2. Results of a search around the native state conformation of 11 helix pairs sorted according to the RMS deviation of the optimal structure from the native-state structure

											Number of interhelical contacts		
PDB code	Helices	RMS deviation (Å)	Native state crossing angle (deg.)	Interhelical distance (Å)	Δx (Å)	Δz (Å)	α (deg.)	β (deg.)	Ψ (deg.)	Optimal score	True positives	False negatives	False positives
1afo	1, 2	0.89	−40	7.4	−0.50	0.00	−20	−20	−45.5	−12.63	8	0	0
1fx8	2, 11	0.90	−26	7.2	0.00	−1.50	20	0	−31.5	−8.66	8	0	0
1eul	31, 36	1.65	−46	6.4	−1.00	−1.00	30	−40	−38.5	−10.06	6	3	0
1eul	5, 12	1.65	24	8.2	−1.50	−2.00	20	−30	14	−3.62	4	0	0
1occ[a]	108, 131	1.70	24	7.1	−3.00	0.00	40	−20	31.5	−6.12	6	0	3
1occ[a]	32, 54	1.80	14	8.5	1.00	−1.50	20	−60	10.5	−3.97	4	3	1
1c3w[a]	1, 7	1.88	−6	8.9	3.00	−1.50	30	60	21	−4.12	4	2	3
1bl8[a]	10, 12	1.93	13	8.6	−3.00	2.50	−10	60	10.5	−4.46	4	2	5
1qla[a]	1, 8	2.27	23	8.7	−1.50	2.00	0	−60	0	−2.56	2	2	2
1occ[a]	45, 47	2.88	7	7.6	2.50	−2.50	60	60	17.5	−3.89	2	4	2
1fx8[a]	9, 15	5.22	−41	6.5	0.50	−1.00	60	−20	56	−9.51	2	2	5

For each pair of helices, the interhelical distance was maintained at the value in the native-state conformation and the five other degrees of freedom were modulated. Δx and Δz describe the change in the x and z values in the optimal conformation relative to the native-state conformation. x, z, α, β, and Ψ are the degrees of freedom defined for the search in Figure 2. x and z were modulated between −3 and 3 Å with a step size of 0.5 Å; α, β were modulated between −60 and 60° with a step size of 10°; and Ψ was modulated between −77 and 77° with a step size of 3.5°. It should be noted that the crossing angles (Ψ) of the optimal structures are near their native-state values almost throughout the dataset. The native and predicted structures were also compared visually to determine the number of predicted true positive, false positive and false negative residue contacts. Figure 5 shows a distribution of the RMS deviation values reported here.

[a] Structures whose optimal results are at the ends of the search range. The RMS deviations reported for these structures should be regarded as underestimates.

Table 3. Search results for the structure of glycophorin A (GpA) at different interhelical distances (*y*)

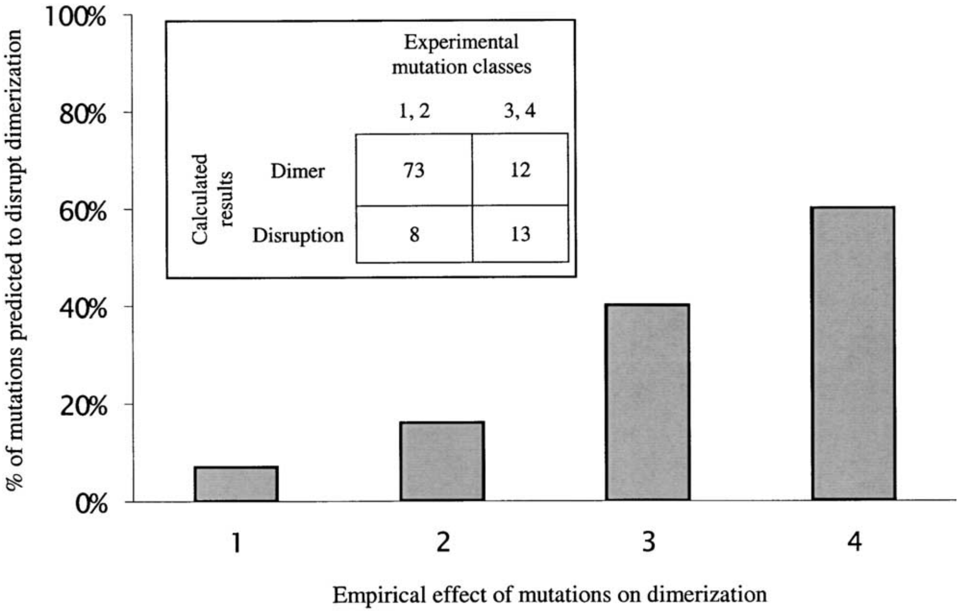


Interhelical distance(Å)	α(°)	β(°)	Ψ(°)	Δx(Å)	Δz(Å)	Score
6.44	-30	-30	49	1.5	0.5	-13.32
6.94	-20	-20	49	2	-0.5	-13.14
7.44	-20	-20	45.5	1	0	-12.63
7.94	-60	10	28	-2	1	-11.87
8.44	-50	-10	21	-2	1.5	-9.66
8.94	-60	20	-7	-2	-1	-8.59

The parameters, their search ranges, and step sizes are defined as in Table 1. The scoring function simply increases with the interhelical distance. Note that the structures obtained throughout the range of 6.44–7.44 Å are essentially similar; as are the structures between 7.94 Å and 8.44 Å.

It is noteworthy that the success of our method is not restricted to a particular packing class. Most helix pairs examined in Table 2 are packed according to the 4–3 class packing,[33] which is more prevalent in TM proteins.[32] Nevertheless, our method shows considerable success with these as well as with pairs packed in the 4–4 packing class.[33]

We also tested a subset of helix pairs whose interhelical distance is beyond the 9 Å range,



	1, 2	3, 4
Dimer	73	12
Disruption	8	13

Figure 4. Comparison of computational and empirical results for 106 non-polar point mutations of GpA. The mutations are classified according to four groups, in keeping with the results reported by Lemmon *et al.*[36] Mutations in group 1 dimerized as well as the wild-type; those in group 2 dimerized significantly; group 3 mutants dimerized in detectable amounts; and group 4 mutations showed no detectable dimerization. Inset: a Table showing the correlation between the experimental results reported by Lemmon *et al.*[36] (horizontal) and our computational results (vertical). The correlation coefficient r^2 for these data is 0.201. Significantly, our computational results found only a small percentage (<10%) of mutations in classes 1 and 2 to be disruptive, whereas a large percentage (>50%) of the mutations in classes 3 and 4 were predicted to be disruptive. The search ranges and step sizes used with each mutant were Ψ in the range of −75 to 75° with a step size of 3°. The rotation around the principal axes (α = β) was carried out throughout the range 0 to 360° with a step size of 5°. *x* was searched in the range of −15 to 15 Å with a step size of 0.5 Å.

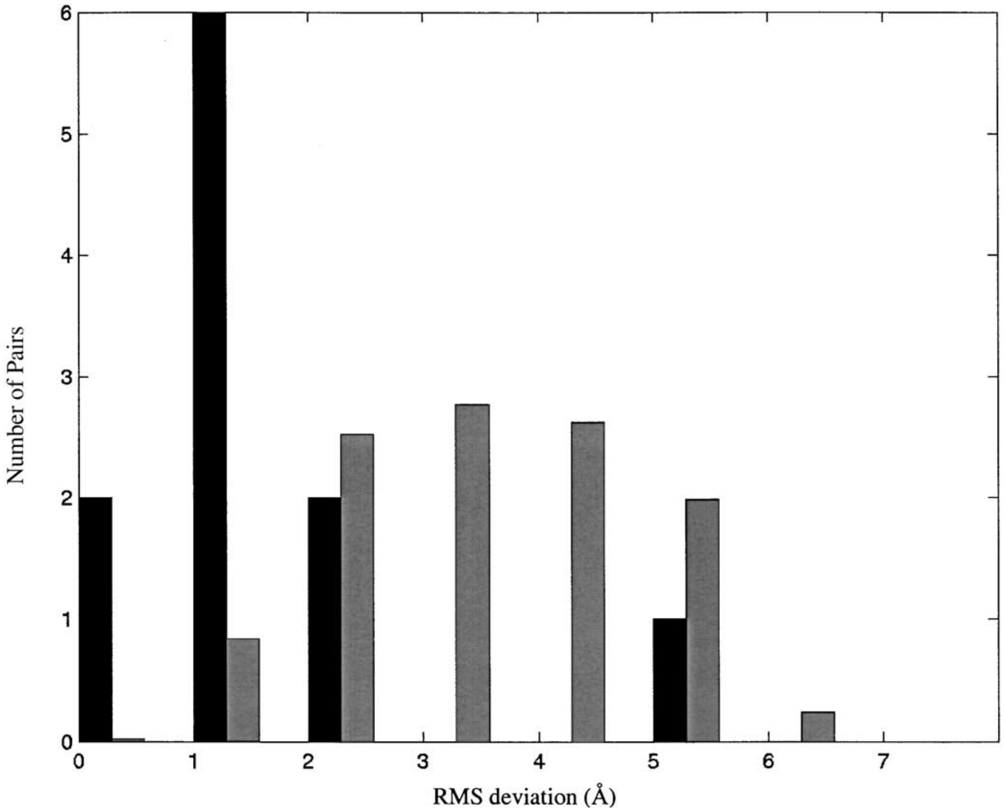


Figure 5. Distribution of RMS deviation values for the search results in Table 4 compared to a random set of structures. Dark bars indicate the distribution of RMS deviation values of a selection of 11 helix pairs to their native-state conformations. Light bars indicate the distribution of RMS deviation values expected by chance (see the text). The expected values were normalized according to the number of helix pairs. Note that 73% of the optimal structures obtained are within 2 Å of the native-state conformation in contrast to the expected value of approximately 8%.

but our results in those cases were much poorer (data not shown). We conclude that the interactions captured by this method are related more closely to those of tightly packed helix pairs.

Discussion

This work had two related goals. The first was to demonstrate the value of a simple rule; small residues go inside, for structure prediction in membrane proteins. To this effect, we used the rule in a simple though exhaustive search method and tested it in 11 carefully chosen TM helix pairs found in the dataset of 11 membrane proteins of known structure (see Methods). As discussed below, the results demonstrated the predictive power of this simple rule. However, they also showed that certain problems in the current implementation of the methodology need to be resolved for it to be potent in structure prediction in polytopic TM proteins. The second goal was to demonstrate the predictive power of the current methodology for tightly packed TM proteins such as GpA.

Ideally, a method for predicting the packing of TM helices would be based on calculating the free energy change occurring upon helix association (ΔG_{ass}). A step in this direction was recently taken by MacKenzie *et al.*[45] who used the data of Lemmon *et al.*[36] concerning the effect of point mutations on the dimerization of GpA to construct an energy-like function for predicting the effects of mutations on dimerization of GpA.

The approach we use here to derive the score function differs from that used by MacKenzie *et al.*[45] in some important respects. It is much humbler, in that it is aimed at discriminating only those conformations of helix pairs in TM proteins that are tightly packed from those that are not. At the same time, it is more ambitious, in that it is derived from general structural considerations, and should therefore be applicable, in principle, to all tightly packed TM helix pairs. The scoring function of Table 1, which constitutes the basis of our method, is a rudimentary construction reflecting contemporary knowledge of tightly packed TM helices; essentially no attempt was made to fit the values to improve the predictions. As anticipated, our results (Figure 4) do not correlate with the data of Lemmon *et al.*[36] as well as do the results of MacKenzie *et al.*[45] ($r^2 = 0.201$ and $r^2 = 0.760$, respectively). We were encouraged to note, however, that the optimal structure of GpA obtained in our calculations was within less than 1 Å RMS deviation from the native structure (Table 2).

To demonstrate the general applicability of our approach, i.e. the predictive power of the simple rule small residues go inside, we compiled a dataset of 11 helix pairs from TM proteins whose 3D structure is known (Table 2). Since we do not consider the current methodology suitable for structure prediction in polytopic TM proteins, we used the known 3D structure of each helix rather than using the corresponding canonical α-helix. The RMS deviations between the optimal and native structures (Table 2) thus reflect purely the quality of the score function, rather than a mixture of the score function with deviations from ideal α-helicity. Our calculations produced conformations that agreed with the known structures significantly better than expected by chance alone (Figure 5). However, a detailed analysis of the success rate is inherently complicated. It is possible that interactions that are not taken into account in our method, such as interactions with other helices, determine the stability of a particular helix pair within the context of the intact protein. In any event, these carefully chosen helix pairs represent only 28% of tightly-packed (distance of 9 Å or less between the principal axes of the helices) helix pairs in the TM proteins in our database, and approximately 10% of all helix pairs forming significant contact in the membrane (see Methods). This illustrates the restricted range of helix pairs our method can currently tackle.

Further complications in analyzing the success rate in Figure 5 arise from the search method we used. Our goal here was to demonstrate the potential use of the scoring function in TM protein structure predictions. Therefore, we did not treat α-helices that deviate markedly from ideality, e.g. pronounced kinks. These deviations may eliminate certain conformations and allow others. Our method is further limited by the fact that we do not model the interconnecting loops. Short loops place a considerable constraint on the conformation space that a pair of helices may explore. This restriction is therefore an important consideration in structure prediction. In the data set presented in Table 2 we included only helix pairs that are at least 20 residues apart from each other (thus excluding approximately 40% of relevant helix pairs from our analysis). Because of the constraints imposed by short loops on the conformations available for a helix pair to explore, the excluded pairs are capable, at least in principle, of exploring a rather restricted range in conformation space. The success rate indeed dropped significantly when helix pairs connected by shorter loops were analyzed also (data not shown). Relaxing these two limitations by using a more sophisticated search methodology may make it possible to study many more helix pairs.

To avoid introducing steric constraints into our calculations, we maintained the distance (y) between the principal axes of the helices at its value in the native state. Obviously, when *de novo* prediction of tertiary structure is attempted, y would also need to be modulated. Our results for the modulation of y (Table 3) indicate, however, that when *de novo* prediction is attempted, it may be possible to set y at two or three different values and obtain different optimal structures. In any case, by using a more detailed model, in which each residue is represented as two or three interaction sites, these limitations may be eliminated altogether.

The overall picture emerging from studies of TM helix dimers is that the specific factors driving contact formation among helices are qualitatively different in various ranges of interhelical distance. Eilers *et al.*[46] recently showed that the distribution of pairwise contacts between helices separated by large distances is different from that of tightly packed pairs of helices. As an example, interactions among aromatic residues, which are known to stabilize contact between some helix pairs,[47] are unlikely to occur in the dimerization region of helix pairs whose axes are separated by short distances, whereas backbone–backbone interactions are not possible when the interhelical distance is large. Other types of contacts that stabilize and specify TM protein structure may emerge in the future.[1] Thus, the scoring function presented in Table 1 may be considered as a basis for improvement as more knowledge about these factors accumulates.

Overall, despite the simplicity of our approach, the results demonstrate that it captures the salient features driving association between tightly packed TM helices. Significantly, the approach we employed uses a lower resolution than that of other methods for TM structure prediction.[18–20] However, in GpA, which is so far the only case in which direct comparison between the different approaches is possible, our results are comparable with those obtained by other methods. In our opinion this shows, above all, that the efforts to develop predictive tools for the tertiary structure of proteins can harness the knowledge derived from structural analyses of proteins in a straightforward manner. Furthermore, the lower computational burden associated with such low-resolution computations allows us to treat asymmetric helix pairs.

In conclusion, while our approach appears to capture at least certain aspects of tight packing between TM helices, many changes need to be introduced for our method to be robust in TM protein structure prediction. Nevertheless, the results indicate that the method can be used for structure predictions of TM dimers that resemble GpA, where the limitations described above are of secondary importance. One important class of proteins for which this is the case is the receptor tyrosine kinases (RTK).[48] It is well known that a critical step in the activation of these receptors is dimerization, and recent evidence has indicated that at least in some cases, e.g. ErbB2 (HER2), this dimerization is mediated by a specific interface on its TM domain.[49] Significantly, the

Sternberg–Gullick motif,[50] which is believed to promote dimerization of TM domains in RTKs, is similar in its general features to the GxxxG motif driving the dimerization of GpA.[24] Recently, Mendrola *et al.*[51] showed *in vivo* that the TM domains of ErbB receptors dimerize in cell membranes, and that the Sternberg–Gullick motifs are mediators of this dimerization. Our initial calculations on the TM domains of ErbB homo- and heterodimers match these observations (unpublished results).

Methods

The search method

Conformation space

We used the coordinates of the C^α traces of the individual helix pairs that were selected on the basis of the criteria specified below. Different conformations were examined by the scoring function defined in Equation (1). As with any two-rigid-body system, any configuration of two helices is defined completely in terms of three translational and three rotational degrees of freedom (Figure 2).

The computational load of the score calculations is relatively low. We therefore used an exhaustive search method rather than other search heuristics. The scoring function is not suitable for modulation of the interhelical distance (y); if the helices were brought closer together it would simply increase the score of a favorable conformation, regardless of steric clashes that would probably form in reality. We therefore searched conformation space, while maintaining the interhelical distance at the value given by the native-state packing of the helices. This reduces the number of degrees of freedom from six to five. We therefore mapped configuration space onto a five-dimensional lattice, such that each coordinate defines a unique conformation of a helix pair.

Comparison of minima with the native state

We estimated the dissimilarity between the score minima obtained and the conformation of the native state by calculating the RMS deviation between the C^α trace of the predicted conformation and that of the native-state conformation using InsightII (Accerlrys, San Diego). We compared by visual inspection the interhelical contacts formed in the native state conformation with those formed in our predicted structure, and classified the latter as true positive, false negative, or false positive contacts (Table 2). Comparison with the RMS deviation measured between the native state and the predicted structures reveals a good correlation between the two criteria. It also shows that a cut-off RMS deviation value of 2 Å constitutes a reasonable threshold, below which the predicted structures fit well with those of the native state.

Implementation

The search for the lowest score was implemented in MatLab (MathWorks, Natick MA), using completely vectorized code to improve performance, and run on parallel Origin 2000 *SGI* processors. The main computational load is the determination of the score of a conformation. This averaged approximately 14 ms per conformation on each of the Origin 2000 processors.

Quantifying the burial of each amino acid residue

The score function defined in equation (1) is based on quantification of the burial of amino acid residues that mediate contact between the helices. In measuring the extent of burial B^i of amino acid residue i we consider two criteria. The first is the distance between the residue and the principal axis of the other helix; the smaller the distance, the more deeply buried the residue. The second criterion is the orientation of the amino acid with respect to the principal axis of the other helix. The more the amino acid residue is directed towards the other helix, the better its burial (Figure 1).

Formally, we consider two parameters: the distance D^i between amino acid residue i and the axis of the other helix, and the angular orientation A^i of amino acid residue i with respect to the axis of the other helix. We define the burial of an amino acid residue as the intersection of these two criteria:

$$B^i = S(D^i)S(A^i) \qquad (2)$$

where $S(D^i)$ and $S(A^i)$ are transformations of the distance and angular criteria as defined below.

The effect of increasing the distance or the angular orientation of an amino acid residue on its burial is quantified as a sigmoidal transformation. Clearly, the burial of a residue at close contact and the correct orientation are not much altered by small changes, as is the burial of an amino acid residue that is poorly buried. However, at a certain cut-off distance and orientation, the extent of a residue's burial changes rapidly.

We therefore use a sigmoidal relation of the form:

$$S(D^i) = \frac{1}{\left(\frac{D^i}{t}\right)^p + 1} \qquad (3)$$

for the distance, and a similar expression for $S(A^i)$.

The above sigmoidal function produces values ranging from 0, signifying no burial, to 1, signifying complete burial. It approaches unity for small values of D^i and zero for large D^i. Note that $S(D^i = t)$ is 0.5, and that p controls the smoothness of the sigmoid, i.e. for large p the function approaches the form of a step function, in which the step occurs at $D^i = t$. Thus, t and p control the position of the threshold, where the function assumes half-value, and the contour of the function, respectively.

We used an ideal model of a helix pair, whose axes are separated by approximately 7.5 Å, to examine different parameter combinations. Thus, we found the parameter values $t = 60°$ and $p = 4$ to be suitable for transformation of the angle A^i. For transformation of the distance, we first subtract 4.3 Å from the value of D^i calculated for the distance between the amino acid residue and the axis of the other helix. This value approximates the smallest distance possible between an amino acid residue and another helix (the radius of an α-helix to its C^α atoms is 2.3 Å plus 2 Å for two exclusion radii), and produces a value of 1 for $S(D^i)$ if the amino acid residue is as close as possible to the axis of the other helix. The parameter values chosen for the transformation of the distance are $t = 2.5$ Å and $p = 6$. Thus the two transformations for amino acid i are:

$$S(D^i) = \frac{1}{\left(\frac{D^i - 4.3}{2.5}\right)^6 + 1} \tag{4}$$

$$S(A^i) = \frac{1}{\left(\frac{A^i}{60}\right)^4 + 1} \tag{5}$$

where A^i and D^i are given in units of degrees and Å, respectively.

It should be noted that B^i (equation (2)) is sensitive to the choice of t and p values in these S relations. Thus, changes in these parameters may lead to substantial differences in the burial function, and hence in the scoring function defined in equation (1).

Measuring the distance and angular orientation of each residue with respect to the helix opposing it

To allow for some deviations from α-helical ideality, we employed a method presented by Chothia *et al.*[33] for defining a local helical axis rather than the global one. Local axes coincide with the actual curvature of the helical axis. Due to local deformations, the curvature may differ in places from that of the helix's principal axis. The local helical axis v^i of residue i is defined as the cross-product of the vectors Q^i and Q^{i+1}:

$$v^i = Q^i \times Q^{i+1} \tag{6}$$

where:

$$Q^i = C^i + C^{i+2} - 2C^{i+1} \tag{7}$$

and C^i is the position vector of the C^α of residue i. At the helix terminus, the local axes are defined as extensions of the last local axis calculated according to equation (6).

For each residue i we determine the space coordinates of a point p^i nearest to it on the helical axis, according to the method of Walther *et al.*[52] by calculating the geometric center of four consecutive C^α coordinates around i (C^{i-1} to C^{i+3}). Points on the local axis in both termini of the helix are calculated by extending a vector of length 1.5 Å, corresponding to the average helical rise, in the direction of the local axis calculated for those termini, v, defined in equation (6).

The distance D^i between residue i and the axis of the opposing helix is defined in our method as the distance between i and the nearest point p^j on the opposing helix's axis. The angular orientation of i (A^i) with respect to the other helix is then measured as the residue's orientation with respect to p^j. For a residue i, let us formally define a set P of all the points on the axis of the helix opposing residue i as defined above. The distance D^i between residue i and the axis of the opposing helix is defined as the distance between this residue and a point $p^j \in P$:

$$D^i = |C^i - p^j| \tag{8}$$

where p^j, a point on the helix axis, is defined as:

$$p^j = \min_{p \in P}(|C^i - p^j|) \tag{9}$$

The angular orientation of residue i with respect to the axis of the other helix is then defined as the angle formed between two vectors: $p^i - C^i$, a vector in the direction assumed in space by residue i, and $p^j - p^i$, the vector connecting residue i to the axis of the other helix.

Finding the dimerizing residues

An important implication of the ridges-into-grooves structural motif is that contact between tightly packed helices is mediated by amino acid residues that are relatively close to each other on the sequence,[33] i.e. the residues forming contact are all contained in a stretch of not longer than ten residues. This is corroborated, at least partially, by the results of the experiments reported by Mingarro *et al.*[41] which showed that an insertion mutation incorporating four Ala residues in the middle of the dimerization motif of the human GpA does not extend the length of its dimerization motif. We used this implied condition as a criterion for deciding which residues actually form close contact. It is interesting to note that without this criterion, the optimal structures obtained by the method consist of helix pairs with their principal axes parallel with each other (results not shown). This is in accordance with the argument made by Chothia *et al.*[33] that the assumption that helices are smooth cylinders leads to parallel orientations of helix pairs as the most favorable conformation.

To find such a stretch of buried amino acid residues, we examine windows of ten residues on the sequence of each helix for all relatively buried amino acid residues ($B^i \geq 0.2$). Of these windows, we pick the one in which the total burial of its residues is maximal. Formally, let us define W as the set of all contiguous ten residue stretches on each of the helices. We first look for $w' \in W$ such that:

$$w' = \max_{w \in W}\left(\sum_{i \in w\ :\ B^i \geq 0.2} B^i\right) \tag{10}$$

Then, the residues that form the contact between the helices w_{con} are the residues within w' whose burial score B^i indicates that they are well buried ($B^i \geq 0.2$):

$$w_{\text{con}} = \{i \in w'\ :\ B^i \geq 0.2\} \tag{11}$$

We thus obtain two such sets of buried residues, one for each helix in the pair.

Finding the pairs of residues that mediate contact

We were interested in finding a set of pairs P of residues, one from each of the w_{con} terms defined above, that form contact between the two helices. Two residues (i, j), such that i is located on one helix and j on the other, are said to form contact if both are buried (i.e. they are both members of the sets w_{con} defined above), and the distance between i and j is not greater than 5.5 Å. This cut-off should be regarded as rather low, since a choice of larger values, e.g. 6 Å, leads at times to structures conforming to class 3–3 helix packing, which were not identified using the assay described by Chothia *et al.*[33] Almost all contacts between residues in this 3–3 class conformation were relatively long-range (over 5.5 Å).

Construction of a database of helix pairs from solved TM protein structures

To test the validity and the applicability of our scoring function, we set up a dataset of helix pairs from the solved structures of TM proteins. The dataset was constructed using tailor-made programs written in MatLab (MathWorks, Natick MA) and Perl, which are available from our website†, and can easily be modified to suit

† http://www.ashtoret.tau.ac.il/usarel

Table 4. Proteins used in this work and their PDB identifiers

Protein name	PDB identifier
Bacteriorhodopsin	1c3w
Calcium ATPase	1eul
Cytochrome *c* oxidase	1occ
Fumarate reductase	1qla
Glycerol facilitator	1fx8
Glycophorin A	1afo
Light-harvesting complex II	1lgh
Mechanosensitive channel	1msl
Photosynthetic reaction center	1prc
Potassium channel	1bl8
Rhodopsin	1f88

other analytical needs besides those described here. We applied strict criteria for the inclusion of pairs of helices in our dataset. Briefly, we constructed an initial data set of 39 non-redundant helix pairs, whose interhelical distance is within the range of 6–9 Å. A further restriction on this data set is that pairs of helices are excluded if they are tilted against each other. This restriction is imposed because the effective contact area made by tilted helix pairs is usually rather small.[33] Of these 39 pairs, ten were eliminated because one or both of the helices did not conform to strict α-helicity; a further 14 pairs were eliminated because the pair constituted sequence neighbors separated by fewer than 20 amino acid residues; and four more pairs were eliminated after visual inspection because the interface actually formed by the helices was judged to be small, or produced by kinking and coiling of the helices. We obtained a total of 11 helix pairs, which are presented in Table 2.

Data

We obtained 11 structures of TM proteins from the Protein Data Bank (PDB‡) (Table 4). The helical parts in each protein were determined automatically according to the data supplied in the PDB, where available, and taken from the literature in the case of 1afo,[26] for which the data were not included in the PDB. All other parts of the proteins were ignored.

Elimination of α-helices that are far from canonical

With the object of excluding from our analysis any helices that deviate significantly from ideal α-helicity, we determined the characteristic helical rise and radius according to the structure of bovine cytochrome *c* oxidase (PDB code 1occ). The average helix radius is 2.52 Å (σ = 0.14 Å) and the average helical rise is 1.56 Å (σ = 0.09 Å). These values are comparable to those obtained by Walther *et al.*[52] For each helix, we also calculated the global geometric center *G* and a vector in the direction of its principal axis **u**.

The search method for optimal conformations is sensitive to deviations from ideal α-helicity. We therefore eliminated substantial deviations from α-helicity by using a 99% confidence limit around the helical rise and radius. Only helices whose rise and radius fell within both limits were allowed into the subsequent analysis.

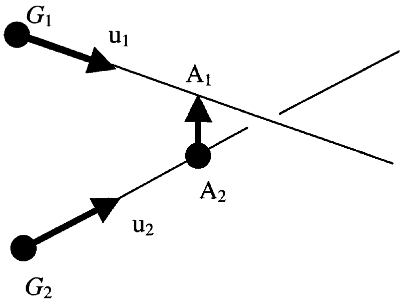


Figure 6. A representation of the method used to find the points of closest approach on two helical axes. G_1 and G_2 are the two helical geometric centers; $\mathbf{u}_1$ and $\mathbf{u}_2$ are two unit vectors pointing in the direction of the helical axes; A_1 and A_2 are the two points of closest approach that we seek. The scheme is adapted from Sunday†.

In addition, we wanted to guarantee that the helical rise and radius were maintained throughout each helix. We therefore selected only those helices in which the standard deviations of the rise and radius did not exceed twice the value of the standard deviation derived above for that parameter.

Selection of pairs of helices making contact

Many of the PDB structures we analyzed contained oligomers of the same subunit. To avoid redundancy, we identified all duplicate helix pairs according to their sequences, and eliminated them. In this way, we obtained, for each protein structure, a non-redundant set of helix pairs that are close to ideal α-helicity. For each possible pair of helices in these sets, we calculated the points of closest approach and the distance between their axes according to the method of Sunday† (Figure 6).

We treat the axes of the helices as infinite lines in 3D space. We know the coordinates of a point on each of the axes (the helix's geometric center *G*) and the direction of the helix axis (**u**). Let us mark the two geometric centers as G_1 and G_2, and unit vectors in the direction of their respective helical axes as $\mathbf{u}_1$ and $\mathbf{u}_2$. Let us also mark A_1 and A_2, the respective points of closest approach, which we seek. Then:

$$A_1 = G_1 + s\mathbf{u_1} \quad \text{and} \quad A_2 = G_2 + t\mathbf{u_2} \tag{12}$$

where *s* and *t* are scalars. By definition, the line connecting A_1 and A_2 is uniquely orthogonal to the two axes, i.e.:

$$\mathbf{u_1}w_c = 0 \quad \text{and} \quad \mathbf{u_2}w_c = 0 \tag{13}$$

where $w_c = A_1 - A_2$. Another way of formulating w_c is:

$$w_c = w_0 + s\mathbf{u_1} - t\mathbf{u_2} \tag{14}$$

where $w_0 = G_1 - G_2$. Substituting equation (14) into the two simultaneous equations defined in equation (13) we obtain:

$$\begin{aligned} (\mathbf{u_1u_1})s - (\mathbf{u_1u_2})t &= -\mathbf{u_1}w_0, \quad \text{and} \\ (\mathbf{u_2u_1})s - (\mathbf{u_2u_2})t &= -\mathbf{u_2}w_0 \end{aligned} \tag{15}$$

‡ http://www.rcsb.org

† http://geometryalgorithms.com

For compactness, let us mark $a = \mathbf{u_1 u_1}$, $b = \mathbf{u_1 u_2}$, $c = \mathbf{u_2 u_2}$, $d = \mathbf{u_1} w_0$ and $e = \mathbf{u_2} w_0$. We can solve equation (15) for s and t:

$$s = \frac{be - cd}{ac - b^2} \text{ and } t = \frac{ae - bd}{ac - b^2} \tag{16}$$

By substituting s and t obtained from equation (16) into equation (12), we finally arrive at the points of closest approach on both helices. In cases where the denominator $ac - b^2$ is zero, the two axes are parallel and the distance between them is simply the distance between a point on one axis and the other axis.

This method thus allows us to limit our dataset to those helices whose distance does not exceed 9 Å. Apart from divulging the distance between the principal axes of the helices, this method allows us to determine whether the points of closest approach (A_1 and A_2) fall inside or outside the span of the helices. We regard pairs whose points of closest approach fall outside the helix span as tilted against each other, forming little if any contact. These pairs are therefore automatically eliminated from the list.

Contact-forming helices of TM proteins are often sequence neighbors.[32] In our initial set of 39 non-redundant helix pairs with interhelical distance in the range of 6–9 Å, we found 15 (38%) that are separated by fewer than 20 amino acid residues. In a preliminary study, we found that our method works considerably better for pairs of helices that are separated by 20 or more residues on the sequence. This is because short loops do not allow the helix pair to explore conformation space freely.[52] We therefore removed from the subsequent analysis all pairs of helices that are connected *via* such short loops.

The structures of all helix pairs were then inspected visually to eliminate helices that were kinked, coiled, or tilted against each other. Helices that exhibited considerable deviations from ideal α-helicity at their ends were split manually or shrunk to produce α-helices closer to the ideal.

To recapitulate: we automatically compiled a non-redundant data set comprised of pairs of helices forming close contact (6–9 Å) that do not deviate considerably from α-helicity, are not tilted against each other, and are separated by loops of 20 or more amino acid residues. We then manually pruned those pairs whose helices were tilted against each other. We also eliminated the ends of helices that deviated from α-helicity.

Calculation of average helix parameters

We computed the average helix rise and radius for each helix by an extension of the method described above for determining the space coordinates of points on the helical axis. For each helix, the average helical rise was computed by taking the average of the distances between subsequent points on the helical axis. The average helical radius was computed by taking the mean of the distances between the space coordinates C^i of residue i and the point on the helical axis p^i associated with it. In helices that contained 20 or more amino acid residues, we disregarded the three terminal residues at both ends, where deviations from ideal α-helicity often occur.

Acknowledgments

We thank Mark A. Lemmon for providing us with a complete list of the dimerization results of GpA point mutants. We acknowledge helpful discussions with Isaiah T. Arkin, Dalit Bechor-Shental and Doron Chema. We thank Shirley Smith for editing the manuscript. This work was supported by a Research Career Development Award from the Israel Cancer Research Fund. We acknowledge the Bioinformatics Unit and the Computation Center at Tel Aviv University for providing us with infrastructure.

References

1. Bowie, J. U. (2000). Understanding membrane protein structure by design. *Nature Struct. Biol.* **7**, 91–94.
2. Popot, J. L. & Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, **29**, 4031–4037.
3. Popot, J. L. & Engelman, D. M. (2000). Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.* **69**, 881–922.
4. White, S. H. & Wimley, W. C. (1999). Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 319–365.
5. von Heijne, G. (1996). Principles of membrane protein assembly and structure. *Prog. Biophys. Mol. Biol.* **66**, 113–139.
6. Tusnady, G. E. & Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **283**, 489–506.
7. Ubarretxena-Belandia, I. & Engelman, D. M. (2001). Helical membrane proteins: diversity of functions in the context of simple architecture. *Curr. Opin. Struct. Biol.* **11**, 370–376.
8. Rees, D. C., DeAntonio, L. & Eisenberg, D. (1989). Hydrophobic organization of membrane proteins. *Science*, **245**, 510–513.
9. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125–142.
10. Pilpel, Y., Ben-Tal, N. & Lancet, D. (1999). kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J. Mol. Biol.* **294**, 921–935.
11. Stevens, T. J. & Arkin, I. T. (1999). Are membrane proteins inside-out proteins? *Proteins: Struct. Funct. Genet.* **36**, 135–143.
12. Eilers, M., Shekar, S. C., Shieh, T., Smith, S. O. & Fleming, P. J. (2000). Internal packing of helical membrane proteins. *Proc. Natl Acad. Sci. USA*, **97**, 5796–5801.
13. Hirokawa, T., Uechi, J., Sasamoto, H., Suwa, M. & Mitaku, S. (2000). A triangle lattice model that predicts transmembrane helix configuration using a polar jigsaw puzzle. *Protein Eng.* **13**, 771–778.
14. Zhdanov, V. P. & Kasemo, B. (2001). Folding of bundles of alpha-helices in solution, membranes, and adsorbed overlayers. *Proteins: Struct. Funct. Genet.* **42**, 481–494.
15. Taylor, W. R., Jones, D. T. & Green, N. M. (1994). A method for alpha-helical integral membrane protein fold prediction. *Proteins: Struct. Funct. Genet.* **18**, 281–294.

16. Tuffery, P. & Lavery, R. (1993). Packing and recognition of protein structural elements: a new approach applied to the 4-helix bundle of myohemerythrin. *Proteins: Struct. Funct. Genet.* **15**, 413–425.
17. Baldwin, J. M., Schertler, G. F. & 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.
18. Adams, P. D., Arkin, I. T., Engelman, D. M. & Brunger, A. T. (1995). Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nature Struct. Biol.* **2**, 154–162.
19. Briggs, J. A., Torres, J. & Arkin, I. T. (2001). A new method to model membrane protein structure based on silent amino acid substitutions. *Proteins: Struct. Funct. Genet.* **44**, 370–375.
20. Pappu, R. V., Marshall, G. R. & Ponder, J. W. (1999). A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nature Struct. Biol.* **6**, 50–55.
21. Adamian, L. & Liang, J. (2001). Helix–helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.* **311**, 891–907.
22. Arkin, I. T. & Brunger, A. T. (1998). Statistical analysis of predicted transmembrane alpha-helices. *Biochim. Biophys. Acta*, **1429**, 113–128.
23. Javadpour, M. M., Eilers, M., Groesbeek, M. & Smith, S. O. (1999). Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophys. J.* **77**, 1609–1618.
24. Senes, A., Gerstein, M. & Engelman, D. M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.* **296**, 921–936.
25. Lemmon, M. A. & Engelman, D. M. (1994). Specificity and promiscuity in membrane helix interactions. *Quart. Rev. Biophys.* **27**, 157–218.
26. MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science*, **276**, 131–133.
27. Senes, A., Ubarretxena-Belandia, I. & Engelman, D. M. (2001). The Calpha–H···O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc. Natl Acad. Sci. USA*, **98**, 9056–9061.
28. Choma, C., Gratkowski, H., Lear, J. D. & DeGrado, W. F. (2000). Asparagine-mediated self-association of a model transmembrane helix. *Nature Struct. Biol.* **7**, 161–166.
29. Zhou, F. X., Cocco, M. J., Russ, W. P., Brunger, A. T. & Engelman, D. M. (2000). Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nature Struct. Biol.* **7**, 154–160.
30. Dawson, J. P., Weinger, J. S. & Engelman, D. M. (2002). Motifs of serine and threonine can drive association of transmembrane helices. *J. Mol. Biol.* **316**, 799–805.
31. Adamian, L. & Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins: Struct. Funct. Genet.* **47**, 209–218.
32. Bowie, J. U. (1997). Helix packing in membrane proteins. *J. Mol. Biol.* **272**, 780–789.
33. Chothia, C., Levitt, M. & Richardson, D. (1981). Helix to helix packing in proteins. *J. Mol. Biol.* **145**, 215–250.
34. Bowie, J. U. (1997). Helix packing angle preferences. *Nature Struct. Biol.* **4**, 915–917.
35. Walther, D., Springer, C. & Cohen, F. E. (1998). Helix–helix packing angle preferences for finite helix axes. *Proteins: Struct. Funct. Genet.* **33**, 457–459.
36. Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J. & Engelman, D. M. (1992). Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry*, **31**, 12719–12725.
37. Smith, S. O., Song, D., Shekar, S., Groesbeek, M., Ziliox, M. & Aimoto, S. (2001). Structure of the transmembrane dimer interface of glycophorin A in membrane bilayers. *Biochemistry*, **40**, 6553–6558.
38. Lemmon, M. A., Flanagan, J. M., Hunt, J. F., Adair, B. D., Bormann, B. J., Dempsey, C. E. & Engelman, D. M. (1992). Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. *J. Biol. Chem.* **267**, 7683–7689.
39. Russ, W. P. & Engelman, D. M. (2000). The GxxxG motif: a framework for transmembrane helix–helix association. *J. Mol. Biol.* **296**, 911–919.
40. Furthmayr, H., Galardy, R. E., Tomita, M. & Marchesi, V. T. (1978). The intramembranous segment of human erythrocyte glycophorin A. *Arch. Biochem. Biophys.* **185**, 21–29.
41. Mingarro, I., Elofsson, A. & von Heijne, G. (1997). Helix–helix packing in a membrane-like environment. *J. Mol. Biol.* **272**, 633–641.
42. Treutlein, H. R., Lemmon, M. A., Engelman, D. M. & Brunger, A. T. (1992). The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. *Biochemistry*, **31**, 12726–12732.
43. Adams, P. D., Engelman, D. M. & Brunger, A. T. (1996). Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. *Proteins: Struct. Funct. Genet.* **26**, 257–261.
44. Branden, C. & Tooze, J. (1999). *Introduction to Protein Structure*, 2nd edit., Garland Publishing Inc, New York.
45. MacKenzie, K. R. & Engelman, D. M. (1998). Structure-based prediction of the stability of transmembrane helix–helix interactions: the sequence dependence of glycophorin A dimerization. *Proc. Natl Acad. Sci. USA*, **95**, 3583–3590.
46. Eilers, M., Patel, A. B., Liu, W. & Smith, S. O. (2002). Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys. J.* **82**, 2720–2736.
47. Doyle, D. A., Morais Cabral, J., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L. *et al.* (1998). The structure of the potassium channel: molecular basis of K + conduction and selectivity. *Science*, **280**, 69–77.
48. Schlessinger, J. (2000). Cell signaling by receptor tyrosine kinases. *Cell*, **103**, 211–225.
49. Burke, C. L. & Stern, D. F. (1998). Activation of Neu (ErbB-2) mediated by disulfide bond-induced dimerization reveals a receptor tyrosine kinase dimer interface. *Mol. Cell. Biol.* **18**, 5371–5379.
50. Sternberg, M. J. & Gullick, W. J. (1990). A sequence motif in the transmembrane region of growth factor receptors with tyrosine kinase activity mediates dimerization. *Protein Eng.* **3**, 245–248.

51. Mendrola, J. M., Berger, M. B., King, M. C. & Lemmon, M. A. (2002). The single transmembrane domains of ErbB receptors self-associate in cell membranes. *J. Biol. Chem.* **277**, 4704–4712.

52. Walther, D., Eisenhaber, F. & Argos, P. (1996). Principles of helix–helix packing in proteins: the helical lattice superposition model. *J. Mol. Biol.* **255**, 536–553.

Edited by G. von Heijne

(Received 19 February 2002; received in revised form 8 June 2002; accepted 11 June 2002)