WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Predicting transmembrane helix pair configurations with knowledge-based distance-dependent pair potentials

Christina Wendel and Holger Gohlke*

Department of Biological Sciences, Molecular Bioinformatics Group, J. W. Goethe-University, Frankfurt, Germany

ABSTRACT

As a first step toward a novel de novo structure prediction approach for α-helical membrane proteins, we developed coarse-grained knowledge-based potentials to score the mutual configuration of transmembrane (TM) helices. Using a comprehensive database of 71 known membrane protein structures, pairwise potentials depending solely on amino acid types and distances between C_α-atoms were derived. To evaluate the potentials, they were used as an objective function for the rigid docking of 442 TM helix pairs. This is by far the largest test data set reported to date for that purpose. After clustering 500 docking runs for each pair and considering the largest cluster, we found solutions with a root mean squared (RMS) deviation <2 Å for about 30% of all helix pairs. Encouragingly, if only clusters that contain at least 20% of all decoys are considered, a success rate >71% (with a RMS deviation <2 Å) is obtained. The cluster size thus serves as a measure of significance to identify good docking solutions. In a leave-one-protein-family-out cross-validation study, more than 2/3 of the helix pairs were still predicted with an RMS deviation <2.5 Å (if only clusters that contain at least 20% of all decoys are considered). This demonstrates the predictive power of the potentials in general, although it is advisable to further extend the knowledge base to derive more robust potentials in the future. When compared to the scoring function of Fleishman and Ben-Tal, a comparable performance is found by our cross-validated potentials. Finally, well-predicted "anchor helix pairs" can be reliably identified for most of the proteins of the test data set. This is important for an extension of the approach towards TM helix bundles because these anchor pairs will act as "nucleation sites" to which more helices will be added subsequently, which alleviates the sampling problem.

Proteins 2008; 70:984–999.

Key words: knowledge-based potentials; protein structure prediction; membrane proteins; helix docking; coarse-graining.

INTRODUCTION

Accounting for more than 75% of all therapeutic drug targets,[1] the prominent role of α-helical transmembrane (TM) proteins is clearly evident. Estimated 20–30% of the open reading frames in sequenced genomes code for helical membrane proteins,[2] and these proteins are involved in most of the signal transduction and metabolic pathways of a cell. But due to major problems in high-resolution structure determination of TM proteins, they are only poorly represented (≈0.5%) in the Protein Data Bank (PDB).[3] Bypassing difficulties in experimental structure determination, reliable methods to computationally obtain structure models of TM proteins will thus aid in understanding and further investigating this important class of proteins.

Provided that suitable templates exist, homology modelling has proven to produce the most exact structure models of soluble proteins.[4] However, conditions necessary for successful homology modelling may not be met in the case of TM proteins. For example, for modelling G-protein coupled receptors (GPCRs), bovine rhodopsin provides the only template, with a sequence identity often less than 25% to the target sequence.[5] Therefore, it is questionable whether generated GPCR models are of sufficient quality for use in structure-based drug design.[6,7] Despite these limitations, homology modelling is widely applied to predict TM protein structures.[8] This fact reflects the lack of reliable, generally applicable, and efficient de novo prediction methods that do not require experimental constraints.

De novo structure prediction methods of small soluble proteins experienced significant improvements during the last years.[4] While not directly transferable to TM proteins, at least general strategies can be adopted. De novo structure prediction approaches usually consist of a sampling method to explore the conformational space of the considered protein and an energy function to distinguish native-like conformations from decoys. An effective solution to the sampling problem, originating from the large conformational space accessible to proteins, is given by a hierarchical approach: Coarse-

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat.

Grant sponsors: J.W. Goethe-University, Boehringer Ingelheim Pharma.

*Correspondence to: Holger Gohlke, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany.
E-mail: gohlke@bioinformatik.uni-frankfurt.de

Received 10 November 2006; Revised 19 March 2007; Accepted 16 April 2007
Published online 10 September 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21574

grained models are first used to determine a set of good starting structures, which are subsequently refined to obtain detailed structural models.[9] However, energy functions, especially coarse-grained ones, are often too inaccurate to place the native structure within a distinct minimum compared to non-native conformations.[9] Therefore, representative structures out of the largest cluster of generated conformations are often considered in subsequent steps,[10,11] based on the assumption that the native structure occupies a broader energy minimum than decoys.[12] These considerations should be equally applicable to soluble as well as membrane proteins.

An additional framework to structure prediction approaches for TM proteins is given by the two-stage model of membrane protein folding.[13] The two-stage model assumes that, first, single stable helices are formed across the lipid bilayer. The helices then associate to form the functional TM protein structure in the second stage.[14,15] For TM protein structure prediction according to this model, methods that predict the localization of helical regions in the sequence[2,16–18] provide solutions for the first stage of folding. These methods already accomplish success rates of more than 90%.[19,20] Therefore, we focus on the second stage in this study: predicting the association of helices to form the correct three-dimensional (3D) structure.

This task can be regarded as a docking problem from an algorithmic point of view. To accomplish it, a suitable energy function to score mutual helix orientations is needed. In some cases of existing approaches, the inclusion of experimental information is still required to overcome deficiencies or lack of information in the scoring functions. For example, knowledge-based scoring functions (derived from ≈30 3D TM protein structures[21,22] or TM protein sequences[23]) that predict the lipid-facing side of a TM helix[21–23] alone are incapable of predicting how helices associate with each other. But when combined with cryo-EM data to approximately position TM helices in the beginning, reasonable structure models can be generated by predicting lipid-facing sides of TM helices.[24] Similarly, a scoring function combining experimental distance constraints and terms derived by a statistical analysis of 14 TM protein structures was used to predict the structure of the GPCR rhodopsin with a root mean squared (RMS) deviation of 3.2 Å.[25]

For de novo structure prediction excluding any experimental constraints, two types of scoring functions can be distinguished: those based on physical force fields and knowledge-based ones. The methods using physical force-fields mainly optimize van der Waals (vdW) interactions between helices.[11,26,27] Reasonable predictions were achieved in validation studies considering up to five helix pairs (for some cases, tetramers or pentamers were generated applying symmetry operations). However, considering only vdW interactions does not sufficiently take into account the complex interplay of forces occurring between TM helices[28,29] (e.g., also polar interactions often play a critical role[30,31]). Hence, these scoring functions cannot be considered generally applicable.

A knowledge-based approach developed specifically to predict GPCR structures is Predict,[32] which generates structure models using knowledge-based contact potentials derived from soluble proteins. These potentials were augmented amongst others by terms representing the membrane environment. Predict was shown to predict rhodopsin with a RMS deviation of 2.9 Å and has also been validated on three other GPCRs.[32] But assumptions that apply specifically to GPCR structures rule out a general applicability to other TM proteins. In another knowledge-based approach from Dobbs *et al.*,[33] optimized C_β pair-contact potentials were derived from one or two TM protein structures and were subsequently shown to reproduce the structures they were trained from. However, any wider applicability is very unlikely. Fleishman and Ben-Tal[34] developed a scoring function based on the qualitative analysis of known TM protein structures. To the best of our knowledge, they used the largest data set of helix pairs for validation so far, consisting of 11 helix pairs. Seventy three percent of the helix pairs were predicted with a RMS deviation <2 Å in a systematic search with five degrees of freedom.[34]

The possibility to derive a reliable statistical scoring function from TM proteins alone has been regarded very unlikely so far.[33,34] But in a recent adaptation of the Rosetta de novo structure prediction method to TM proteins,[35] statistically derived amino acid pair propensities were used as part of a coarse-grained scoring function and yielded promising results. The corresponding knowledge-base consisted of 28 helical TM protein structures and additional information gained from multiple sequence alignments. Tested on 12 multipass TM proteins the method generated reasonable predictions with RMS deviations <4 Å for large parts of the proteins.[35]

In the present study we set out to derive coarse-grained distance-dependent knowledge-based pair potentials from a database of 71 known helical TM protein structures. This step is the first one in an endeavor to predict helical TM protein structures without any additional experimental information. The knowledge-base from which the potentials were derived is the largest one used so far. The predictive power of the C_α-atoms-only potentials was thoroughly evaluated by rigidly docking TM helices using a large test data set of 442 helix pairs. Cross-validation studies were performed in addition to assess possible training effects due to the still limited number of structures in the knowledge-base. Several results stand out: First, the size of structural clusters could be devised as a measure of significance for successful dockings. Second, in nearly 3/4 of all cases helix pair orientations with a RMSD <2 Å were obtained, if only clusters that contain at least 20% of all generated decoys were considered. Finally, we were able to identify so-

called "anchor helix pairs" for most of the tested proteins. These anchor helix pairs will be used in future extensions of the method as "nucleation sites" for adding other helices to ultimately predict whole helix bundles.

RESULTS AND DISCUSSION

In the following, we will first describe the derivation and properties of the knowledge-based potentials. Then, we will present results obtained by rigid docking of a large test data set of helix pairs. Subsequently, leave-one-protein-family-out cross-validation studies will be described. Finally, for further validation, helix pair predictions obtained with our scoring function will be compared to those obtained with the scoring function from Fleishman and Ben-Tal.[34]

Knowledge-based potentials

The two-stage model assumes for the first stage of TM protein folding that single helices remain stable after insertion into the lipid bilayer.[14] Thus, further folding of helical TM proteins is determined by the association of the single helices. The free energy of helix association can be divided into contributions by lipid–lipid, helix–lipid, and helix–helix interactions.[36,37] Helix–helix interactions account for a large part of the driving force to helix association. Therefore, we focus on a proper representation of helix–helix interactions in this study.

Using knowledge-based potentials to score helix–helix interactions appeared to be particularly promising as they provide an implicit representation of the various interactions involved in TM helix association. That way, the problem of correctly representing and weighting many different contributions as required by physics-based scoring functions is avoided. As contact potentials have been shown to be less successful in distinguishing native structures from non-native decoys,[38–41] we decided to derive distance-dependent pair potentials instead. Facing the trade-off between accuracy and efficiency, we chose to consider only distances between C_α-atoms of pairs of amino acids on interacting TM helix pairs, which yields coarse-grained potentials suitable for a fast sampling of the conformational space.

Parameter settings

Starting from a knowledge base of 71 helical TM protein structures, we derived 210 pair potentials. Since empirical methods can only be validated by their predictive power and ability to reproduce experimental data, we selected parameters for the derivation of the potentials according to the prediction results for a test data set of helix pairs.

Crucial to the predictive power of knowledge-based potentials is the reference state, which represents an a priori distribution of common and redundant pair interactions. Two different reference states from Sippl[42,43] and Gohlke *et al.*[44] were tested. The reference state of Gohlke *et al.* was shown to yield better results in the case of scoring protein-ligand interactions,[44] whereas potentials derived with the reference state of Sippl were successfully applied to predict protein structures and protein stability and to detect errors in protein structures.[42,43,45,46] Therefore, our finding that Sippl's reference state[42,43] leads to potentials with superior predictive power for the association of TM helices is in good agreement with previous experiences.

A scaling factor of $4\pi r_b^2 dr$ (with r_b being the inner radius of the spherical shell and dr the width of the shell) is sometimes used to account for differences in the volumes of the spherical shells when sampling the radial pair distributions.[44,47] In contrast, we obtained better results without scaling the radial pair distributions (data not shown). This can be explained by the fact that the evaluated helix pairs are found in only a small portion of the volume of a spherical shell. Pairs with large distances will therefore be under-represented if the scaling is applied, as the volume of a spherical shell increases with the square of r_b, while the volume that is occupied by the interacting helices differs only marginally. It has already been recognised along these lines that the assumption of an infinitely large system—implied by the scaling factor of $4\pi r_b^2 dr$—might not be appropriate regarding protein systems. Zhou *et al.* used optimized scaling factors roughly proportional to $r_b^{1.6}$ for protein structures instead.[48,49] We have not tested such an adapted scaling factor yet.

Knowledge base

To derive meaningful potentials, the knowledge base must be sufficiently large. This is because the radial pair-distributions derived from a database will only represent the "true" pair-distributions of the considered atoms if the database represents all possible pairwise interactions according to their thermodynamic probability distribution. We therefore need to address the question whether our database fulfils this requirement.

Our database consists of 71 TM protein structures, from which we extracted 969 TM helices to determine pairwise interhelical distances between C_α-atoms. In all but five cases for which on average less than 10 counts per distance bin were sampled at least one of the amino acids involved was Arg, Asn, Asp, Cys, Gln, Glu, His, Lys, or Trp. However, the rare occurrence of these amino acids is also reflected in the test data set. It can thus be assumed that potentials that are based upon insufficient amounts of data will not provide important contributions to the score of a TM helix pair.

We did not apply methods that account for distributions with low occurrence frequencies, as, for example, proposed by Sippl.[42] In those methods statistically insig-

 DOI 10.1002/prot

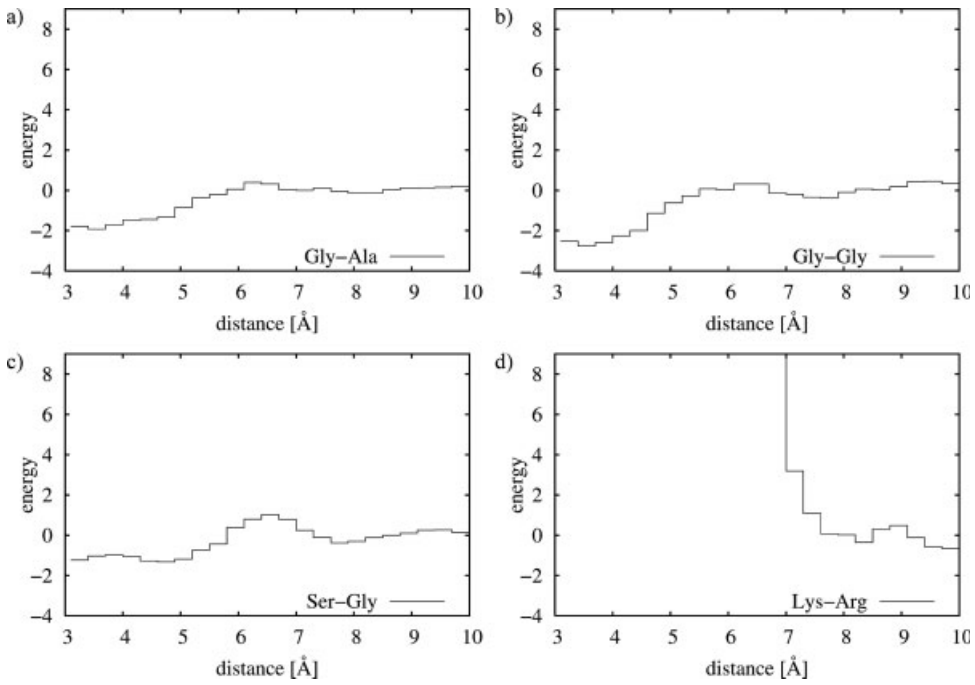

Figure 1

*Potentials of the amino acid pairs Gly-Ala (**a**), Gly-Gly (**b**), Ser-Gly (**c**), and Lys-Arg (**d**). Potentials involving Glycin have a clear minimum at short distances (a–c), whereas for the Lys-Arg potential large distances are found to be more favorable than short distances, as expected for like-charged amino acids. The energy is given in arbitrary units.*

nificant distributions will converge to the reference state. Instead, we intended to fully incorporate the information contained in the few observations extracted from the database rather than to reduce the information content. For all other distributions, on average 764 pair interactions could be sampled from the database. The database of 969 TM helices thus seems to be sufficiently large to apply statistical analysis. Cross-validation studies to test the robustness of the derived potentials reported below will readdress this question.

Derived potentials

The derived potentials represent known interactions between amino acids in TM helices, as demonstrated in four cases in Figure 1. Glycin, which allows for tight packing between two helices,[29] often plays an important role in TM helix association. It is therefore expected that pair potentials involving glycin show favorable interactions at short distances. Indeed, minima are found at distances <5 Å for all these potentials [e.g. Fig. 1(a–c)], with the exception of Gly-Asp, Gly-Glu, and Gly-Lys interactions (data not shown). The minimum of the Gly-Gly potential [Fig. 1(b)], located at a distance between 3.4 and 3.7 Å, is particularly deep. This reflects the fact that pairs of glycin are known to pack tightly against each other, such as in GxxxG motives.[50,51] Thereby, they not only contribute to the association of TM helices by favorable vdW interactions[29,52] but also by C_α—H···O hydrogen bonds,[31] which can be formed when backbone atoms of two helices get into close contact.

An example of unfavorable interactions at short distances is given by the Lys-Arg potential [Fig. 1(d)] as expected for interactions between like-charged amino acids. The minimum is found at the largest distance sampled (between 9.7 and 10 Å), whereas no pair interaction between C_α-atoms of Lys and Arg was observed below 7.5 Å.

A comparison to the kPROT scale,[23] which is used to predict lipid-facing sides of TM helices, demonstrates that general properties of amino acids in TM proteins are reflected in the derived potentials. Amino acids are devided in two groups by kPROT in order to predict lipid exposure. Ala, Arg, Cys, Ile, Leu, Lys, and Val belong to the group of amino acids that are more likely to occur on the lipid exposed surface of TM proteins and are therefore only rarely buried inside the protein.[23] Interactions between those amino acids are represented by 28 possible amino acid pairs. Assuming that these amino acids prefer to face the lipid region, it can be expected that they are only seldomly found in the inter-

Table I
Docking Results of Four Helix Pairs

Helix pair[a]	Protein	Cluster size[b]	Rank[c]	RMSD[d]	Energy of predicted structure	Energy of native structure
1yce (8, 10)	Rotor of F-type Na^+-ATPase	457	1	0.33	−40.11	−38.72
2bbj (3, 7)	CorA Mg^{2+} transporter	181	62	0.49	−8.17	−5.70
2axt (16, 17)	Photosystem II	372	6	0.71	−16.84	−12.22
1occ (45, 50)	Cytochrome c oxidase	268	21	0.13	−8.71	−8.67

[a]Helix pairs are denoted by their PDB code and the indices of the helices. See Methods section for a detailed description.
[b]Corresponds to the largest cluster found for the particular helix pair.
[c]All clusters are sorted according to the lowest-energy structure they comprise.
[d]RMSD of the lowest energy configuration of the largest cluster in Å.

face of interacting helices. Therefore, less favorable potential values can be expected compared to other pairs. Neglecting pairs with very few (<50) observations in the database (Arg-Arg, Cys-Arg, Cys-Cys, Lys-Arg, Lys-Cys, Lys-Lys), the lowest potential value is >-0.4 in 59% of the remaining potentials, while this is the case in only 28% of all the 210 potentials. Furthermore, only one (Val-Val) of the pair potentials involving the aforementioned amino acids belong to those 25% with a minimal potential value <-1.0. Hence, interactions between these amino acids do not obtain very favorable potential values, supporting the findings of kPROT.

Docking helix pairs

In the following sections, we will address the predictive power of the derived pair potentials. For that, we rigidly docked 442 helix pairs derived from a test data set of 58 different TM proteins (see Methods section for how the test data set proteins were chosen) using only the knowledge-based pair potentials as a scoring function to distinguish between near-native geometries and decoys. We note that this validation data set is by far the largest reported to date, allowing us to thoroughly assess the scope and limitations of our approach.

If helix association were only determined by helix–helix (and helix–lipid and lipid–lipid) interaction energies, the native TM protein state should be the conformation of lowest energy.

As the derived potentials represent mainly helix–helix interactions and in addition are only coarse-grained potentials, we do not expect that they represent the actual energy surface well enough to be able to identify a conformation of the native state ensemble as the lowest energy conformation. In fact, when scoring the native configurations of four helix pairs discussed further later, the experimental structure in no case receives a better score than the decoys (Table I). However, as the narrow well of the native state is surrounded by a broad, more shallow minimum in which near-native conformations can be found, it has been hypothesized that native-like structures reside in a broader energy minimum than non-native structures.[12] When carried over to the field of protein structure prediction, this implies that instead of focussing solely on the conformation of lowest energy, the lowest energy structure of the largest cluster of structurally related low energy solutions should be considered the most probable conformation.[12,53] This strategy is also applied in the present study.

Energy landscapes of four helix pairs demonstrate the advantage of considering the largest cluster (Fig. 2). The helix pairs are predicted with a RMSD value <1 Å (Table I). In the case of 1yce (8, 10) [Fig. 2(a)], the largest cluster also comprises the lowest energy configuration (see Methods section for a detailed description of the nomenclature of helix pairs). In the other cases, however, the RMSD values of the overall lowest energy configurations [marked with an arrow in Fig. 2(b–d)] are considerably higher than those of the lowest energy configurations of the largest cluster. The largest clusters are represented by the cumulative occurrences of points below 2 Å RMSD. In contrast, the overall lowest energy configurations are comprised by much smaller clusters or even isolated points [Fig. 2(b–d)], indicating that the corresponding energy minima are very narrow. Thus, as it has already been found in the case of soluble protein structure prediction,[12] native-like structures of TM proteins also reside in broader energy minima compared to non-native states.

Figure 3 depicts predicted helix pair configurations with a RMSD to the native structure between 2 and 5 Å. In the field of small-molecule docking to proteins solutions with a RMSD value below 2 Å are usually considered as good,[54] and solutions between 2 and 3 Å RMSD are still seen as partial success.[55,56] We note, however, that in the case of small molecule docking sometimes 10–15% of randomly generated orientations already yield RMSDs <2 Å.[54] In contrast, it is very unlikely to find such good orientations by chance when docking two helices together, as the search space is rather unlimited in this case. Furthermore, we apply a coarse-grained helix representation (considering only C_α-atoms) for the bene-

DOI 10.1002/prot

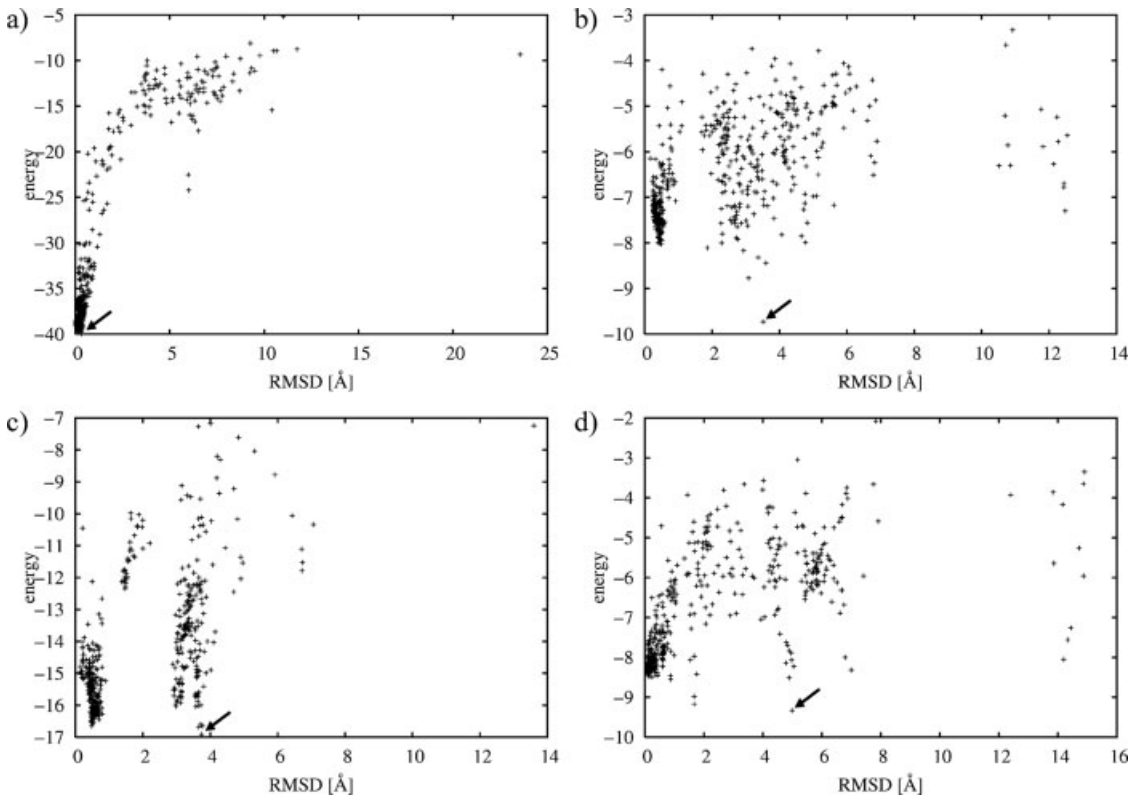

Figure 2

*Energy landscapes of four helix pairs. (**a**) 1yce (8, 10), (**b**) 2bbj (3, 7), (**c**) 2axt (16, 17), and (**d**) 1occ (45, 50). Helix pairs are denoted by their PDB code and the indices of the helices. The energy of each conformation is plotted against the RMSD to the native orientation. To achieve a broader sampling, the docking runs were repeated with only 30,000 energy evaluations and without a local search in this case. The largest clusters found for the helix pairs are depicted by the cumulative occurrences of points below 2 Å RMSD. The overall lowest energy configuration is marked with an arrow.*

fit of sampling efficiency, but at the expense of sampling accuracy. At the current stage, our primary goal thus is to identify configurations good enough to serve as starting structures for subsequent refinement stages. Given that the problem of refining initial models has received increasing attention in recent years and encouraging signals of progress are visible in this area,[4] even solutions with a RMSD of 5 Å [Fig. 3(e)] may serve this purpose in the future.

In Figure 4, the proportion of the 442 helix pairs for which a configuration below a given RMSD was predicted is plotted against the cluster size of the largest cluster. Of all helix pairs 31.8% could be predicted with a RMSD <2 Å. For all RMSD cut-offs considered, the proportion of correctly predicted pairs clearly increases with increasing cluster size. However, the slopes of the curves decrease considerably for cluster sizes >100 decoys. If only those pairs are considered for which the size of the largest cluster was at least 100 (i.e., the cluster comprises at least 20% of all decoys), the success rate of correctly predicted helix pairs with a RMSD <2 Å increases to 71.3%.

The cluster size can thus be used as a measure of significance for the identification of good docking solutions (Table II). This is an important finding in view of an extension of our approach to predict TM helix bundles:

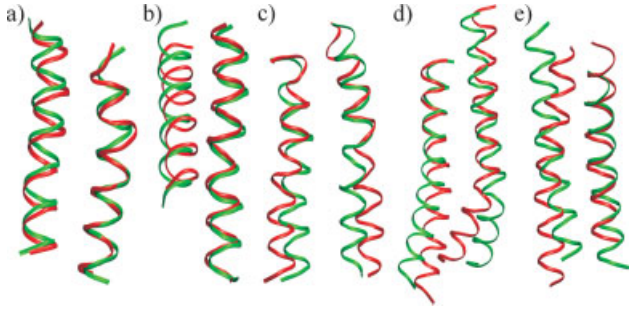

Figure 3

*Five examples of docked helix pairs. Shown in green is the native orientation, in red the predicted orientation. From left to right, the RMSD values increase from 2 to 2.5, 3, 4, and 5 Å. The helix pairs are (**a**) 1xio (1, 2), (**b**) 1c3w (0, 1), (**c**) 1pw4 (6, 9), (**d**) 1u19 (2, 5), and (**e**) 1ogv (4, 6).*

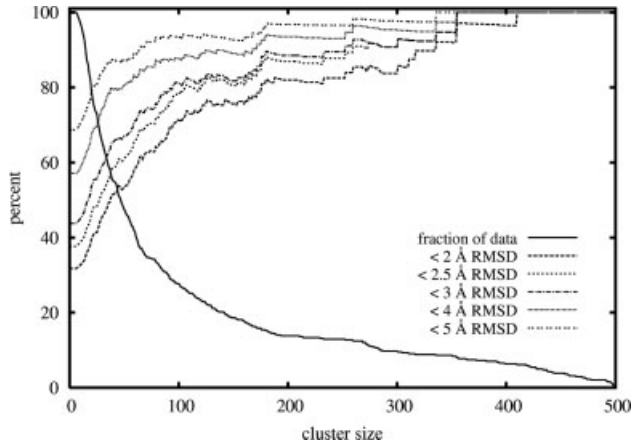

Figure 4

Docking results of 442 TM helix pairs with the original potentials. The percentage of helix pairs for which an orientation below a given RMSD was predicted is plotted versus the size of the largest cluster. In all cases, 500 docking runs were performed and clustered. In addition, the straight line indicates the proportion of all 442 helix pairs that are still considered for a given cluster size cut-off.

if all possible helix pairs of a bundle are docked initially, those for which the largest clusters could be generated are the ones that are most likely docked correctly. These helix pairs can then serve as "anchor pairs" (see later) to which more helices are added in subsequent steps.

In the following, the influence of the interhelical distances and the crossing angles between the helices in the native state as well as the loop lengths between neighboring helices on the success rate of helix pair predictions is investigated.

Interhelical distance

More than 75% of all helix pairs have an interhelical distance <10 Å [Fig. S1(a)]. The interhelical distance is measured as the shortest distance between the two helix axes. Not unexpectedly, smaller interhelical distances correlate with better helix pair predictions, as shown in Figure S1(b). As such, for less than 10% of the pairs with an interhelical distance between 10 and 12 Å in the native state, solutions with RMSD <2 Å were predicted. This is possibly because interactions between such distant helices are not sufficiently captured by the potentials, as we only consider pairs of C_α-atoms with a maximal distance of 10 Å during the derivation. As only very few pairs contribute to the overall score value in the case of an interhelical distance >10 Å, those pairs cannot be expected to be predicted correctly. Neglecting these pairs increases the overall success rate to 38.6% (<2 Å RMSD), and 76.3% are correctly predicted if only clusters with >100 docking solutions are considered. Nevertheless, we kept the pairs in the test data set because we were also interested to see whether it is possible to predict loosely packed helix pairs. For that, we also derived and tested potentials with a maximal distance up to 12 Å to incorporate long-range interactions into the potentials. However, no improvement was observed (data not shown). This can be explained by the fact that considering long distance interactions tampers the potentials, as then pairs that do not contact each other in the interface of two helices but rather lie on the "outer" sides of two helices are also incorporated.

Crossing angle

Considering the crossing angles of all helix pairs, there are more pairs that show a right-handed crossing angle than a left-handed one. Overall there are only few cases with nearly parallel or anti-parallel orientation [Fig. S2(a,b)]. While ≈30% of all pairs can be predicted with RMSD values <2 Å, the success rates for nearly parallel or anti-parallel orientations are significantly higher [Fig. S2(c)]. For pairs with crossing angles between −15° and 0°, the success rate is 75%. However, such helix orientations are only found in ≈4% of all pairs [Fig. S2(b)]. In contrast, success rates below average are obtained for pairs with crossing angles close to the limits of the ranges considered (e.g., for ranges −150° to −135°, 30° to 45° and 135° to 150°). On the one hand, this could be due to insufficient sampling during the docking (only deviations up to ± 45° from parallel or antiparallel orientation are considered, see Methods section). On the other hand, the interface size between helices decreases with increasing deviations from an (anti-)parallel orientation. This leads to fewer interactions that can be evaluated to distinguish good docking solutions from decoys, as also found in the case of helix pairs with large interhelical distances.

Loop length

Short structured loops between helices are expected to influence the mutual orientation.[15] Without incorporating

Table II

Success Rates of Original and Cross Validated Potentials

RMSD (Å)	Original (%)	Cross validated (%)	Δ (%)
All pairs			
<2	31.75	24.04	7.71
<2.5	37.64	30.39	7.25
<3	43.76	39.68	4.08
<4	57.14	55.32	1.82
<5	68.71	68.25	0.46
Cluster size >100			
<2	71.31	55.84	15.47
<2.5	78.69	66.23	12.46
<3	81.15	72.73	8.42
<4	87.71	83.12	4.59
<5	93.44	89.61	3.83

DOI 10.1002/prot

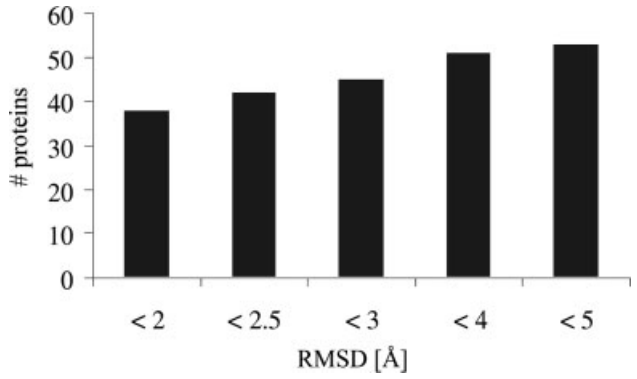

Figure 5

Proteins with more than two TM helices for which at least one anchor helix pair exists. The number of proteins (in total 56) is depicted for which at least one anchor helix pair was predicted with a RMSD value below a given value.

any knowledge about such loops, we predicted the orientation of 142 TM helix pairs with loop lengths <20 amino acids. The corresponding success rates show only a marginal decrease compared to the overall success rates (Table S1). But only 23% of pairs with a loop <10 amino acids can be predicted with a RMSD value <2 Å, resulting in a decrease of almost 9% compared to the overall success rate. Since the success rates for pairs with short loops (<20 amino acids) are also slightly worse than for all pairs, incorporating information about loop lengths in the future and, thereby, restricting the available configurational space is expected to further improve the results.

Anchor pairs

The 442 helix pairs of the test data set are derived from 58 different TM proteins. Since these proteins are very different in size (from 2 to 66 TM helices) and have quite different biological functions, the question arises whether the well predicted helix pairs are sufficiently distributed among all proteins. This question is important with respect to the concept of anchor pairs introduced earlier: as we want to predict TM protein structures by successively adding more helices to a well predicted anchor pair in the future, it is necessary that anchor pairs can be found across many different proteins. Neglecting proteins with only two TM helices, for 75% of the proteins in our test data set at least one helix pair was predicted with a RMSD value <2.5 Å (Fig. 5), indicating that well predicted pairs are sufficiently distributed among the different proteins. Again, such anchor helix pairs can be identified using the cluster size as measure of significance of the docking result (see earlier).

Cross-validation

Since the TM helix pairs of the test data set were also included in the knowledge base from which the potentials were originally derived, we performed a leave-one-protein-family-out cross-validation to assess whether the prediction results arise from training effects. Therefore, new potentials were derived for each protein to be predicted from a database from which proteins of the same family (i.e., with a sequence identity of >30% of any chain) were omitted (see Methods section).

Docking results

Each of the 442 helix pairs was docked again, using the respective cross-validated potentials as a scoring function (Fig. 6, Table II). The largest difference with respect to the results of the original potentials is found for docking solutions with a RMSD <2 Å. While the success rate was 31.8% before, only 24.0% were predicted with <2 Å RMSD with the cross-validated potentials. Considering only pairs with a cluster size of at least 100, this decrease becomes even more obvious: the success rate drops by 15.5% to a value of 55.8%. The correlation between cluster size and success rate also becomes weaker in the range of cluster sizes between 100 and 200. For cluster sizes >200, success rates cannot be judged reliably anymore, because the statistics is based on at most 6.5% of all helix pairs in this case.

When analyzing these results on a case-by-case basis, the deterioration in RMSD values between docking with the original and cross-validated potentials is on average 0.38 Å, whereby for 250 pairs only minimal differences are observed [Fig. S3(a)]. Only for 80 pairs is the deviation ΔRMSD larger than 1 Å. Figure S3(b) confirms the observation (Table II) that especially pairs that were pre-

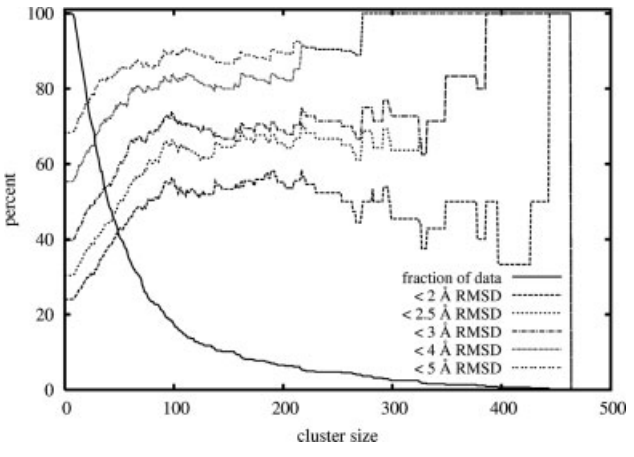

Figure 6

Docking results of 442 TM helix pairs with cross-validated potentials. The percentage of helix pairs for which an orientation below a given RMSD was predicted is plotted versus the size of the largest cluster. In all cases, 500 docking runs were performed and clustered. In addition, the straight line indicates the proportion of all 442 helix pairs that are still considered for a given cluster size cut-off.

dicted with RMSD <2 Å before become worse when the cross-validated potentials are applied. Relatively large positive deviations frequently occur in the region of RMSD <1 Å (>50% of all deviations are >2 Å), whereas significant negative ΔRMSD values, indicating better predictions than before, mostly occur for RMSD values >4 Å. Considering the cluster sizes of the docking solutions, large deviations are found predominantly for cluster sizes <50 obtained with the original potentials [Fig. S3(c)]. As these docking runs cannot be considered converged, however, the deteriorations observed cannot be attributed to the use of the cross-validated potentials. However, large deviations are also found for especially large clusters. As these pairs were in general predicted with very small RMSD values before, the deviations may be attributed to over-training the original potentials.

Overall, the results of the cross-validation suggest that the database from which the potentials are derived is still not large enough to yield sufficiently robust potentials. Thus, knowledge about newly determined TM protein structures should be incorporated in the future. Here, a clear advantage of the knowledge-based approach comes into play as these potentials can be easily rederived. We note, however, that the cross-validation we performed was quite strict. Instead of omitting one helix pair at a time, we omitted all members of a protein family that have chains with a sequence identity >30% to any of the chains of the protein under consideration from the database. This can lead to a loss of up to 14% of all proteins. It is for this reason that we consider the results obtained with the cross-validated potentials still very promising: Even if only C_α-atoms are considered, 2/3 of the helix pairs can be predicted with a RMSD value <2.5 Å (cluster sizes ≥100).

Comparison of potentials

To evaluate the differences between the original and cross-validated potentials, we calculated rank correlation coefficients. Interestingly, they indicated very good correlations between the original and the cross-validated potentials (data not shown). Two explanations may be given for this observation that contrasts with the decreased success rates of the docking results. First, the docking results are probably sensitive even to small changes in the potentials that are not reflected in the correlation coefficients. Second, the correlation coefficients only compare potentials of the same atom types. Therefore, no information is contained in the correlation coefficients about the mutual relationship of the potentials, for example, the relative positions and depths of minima of potentials of different interactions.

Thus, we compared cross-validated potential fields with original ones for a particular helix pair. We chose to use the helix pair 1occ (30, 31) as an example, as it is the one with the largest positive deviation in RMSD value (2.16 Å) for which the deterioration cannot be attributed to insufficient convergence of the docking. For the derivation of the respective cross-validated potentials, 11% of all helices were missing. The helix pair was predicted with a RMSD of 0.45 Å before, whereas a RMSD of 2.62 Å was obtained with the cross-validated potentials. Both docking runs are converged with cluster sizes of 184 and 134, respectively. Hence, the reason for the deviation must be due to differences in the potentials.

The overall scores of the ligand helix in both cases are similar (−9.11 and −8.45, respectively). Considering contributions of single residues, the largest difference is found for Ser156 in the ligand helix. While its contribution is very favorable in the case of the original potentials (with a value of −2.52), a significantly less favorable contribution is found for Ser156 in the case of the cross-validated potentials (−1.25). This is also depicted in Figure 7(a), where differences in the potential grids occur at positions that are occupied by the serine C_α-atom in the docking solution of the original potentials.

Scoring the configuration predicted by the original potentials with the cross-validated potentials gives a score of −6.10. Off the total deviation of 3.01 to the score with the original potentials, the contributions of the two serines Ser149 and Ser156 account for 2.32. The contribution of Ser156 is only −0.77 with the cross-validated potentials, compared to −2.52 before. Breaking the serine's contribution further down finally indicates the pairwise interaction that leads to the different potential values: Ser-Ser accounts for the largest difference. Although this pair potential is very similar overall [Fig. 7(b)], as indicated by the correlation coefficients, for small distances differences between the original and cross-validated potentials are pronounced. As Ser156 closely interacts with Ser101 of the receptor helix at a distance of 4.35 Å in the actual configuration, the difference in potentials at this respective bin obviously has a great influence on the docking result. This demonstrates that even small differences between the potentials may occasionally yield large deviations in the resulting predictions.

Comparison to the scoring function of Fleishman and Ben-Tal

For further evaluation of the predictive power of our potentials, we compared our results with those obtained by the scoring function developed by Fleishman and Ben-Tal.[34] This function was derived by a qualitative analysis of TM protein structures. For the comparison, the same 11 helix pairs as reported in Ref. 34 were used, and the configurational search in the vicinity of the native state of the respective helix pairs was performed analogous to Fleishman and Ben-Tal. For those pairs that belong to one of the 71 proteins used to derive the potentials, we applied the cross-validated potentials for scoring. We note at this point that the results obtained in

 DOI 10.1002/prot

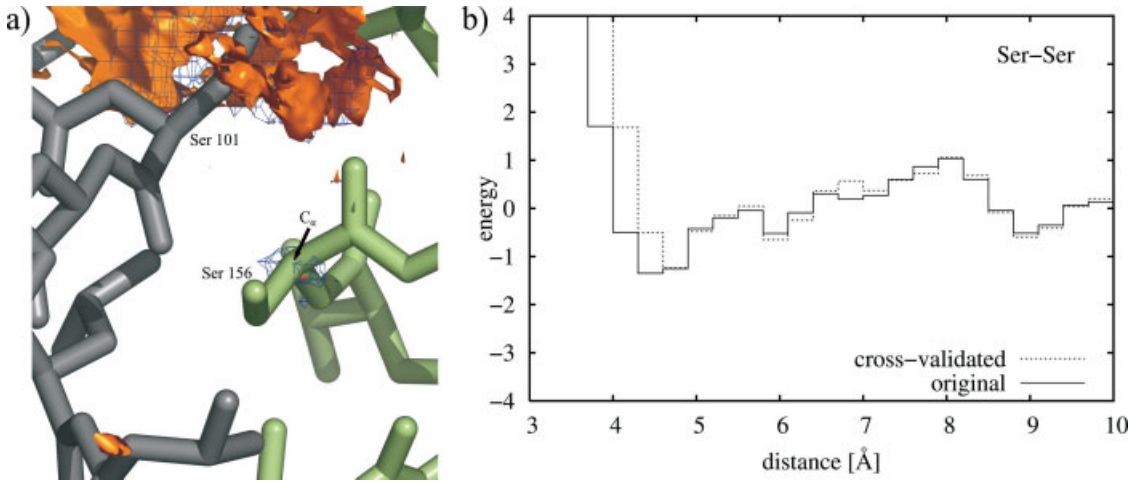

Figure 7

*Serine potentials in the case of helix pair 1occ (30, 31). (**a**) The receptor helix is shown in grey. The ligand helix, docked with the original potentials, is shown in green. The blue isopleths depict the original serine potential, the orange ones the cross-validated serine potential. The contour surfaces are displayed for values ≤ -2.2. The C_α-atom of Ser156 lies in a favorable region of the original potential (blue), but this position is not covered by the cross-validated potential (orange). (**b**) Ser-Ser potential of the original and cross-validated potentials used for docking the helix pair 1occ (30, 31). While the potentials are overall very similar, the greatest deviation is found for small distances, that is, the distance bins 3.7–4.0, 4.0–4.3, and 4.3–4.6 Å. As the distance between Ser101 and Ser156 is only 4.35 Å in the native state, the deviation of the Ser-Ser potentials at small distances has a major impact on the docking result.*

this comparison cannot be compared to the docking results presented earlier. Keeping the interhelical distance fixed at the native value during the configurational search considerably simplifies the search problem.

The RMSD values of the lowest energy configurations found for the 11 helix pairs with both methods are shown in Figure 8. Fleishman and Ben-Tal regarded pairs with RMSD values <2 Å as correctly predicted. While they found eight pairs that fulfil this criterion, only six pairs are found when using the knowledge-based potentials. Encouragingly, in all six cases with a RMSD <2 Å, our predictions deviate less from the native state than the ones of Fleishman and Ben-Tal. The most obvious case is that of 1fx8 (9, 15), for which we find a RMSD of 1.32 Å compared to 5.22 Å. Our worst prediction is 1occ (32, 54) with a RMSD of 3.46 Å. However, as shown in Figure S4, helix 54 is truncated in this test data set. This is due to the noncanonical character of the TM helix. The helix pair in its full length is part of our test data set of 442 helix and was docked with a RMSD value of

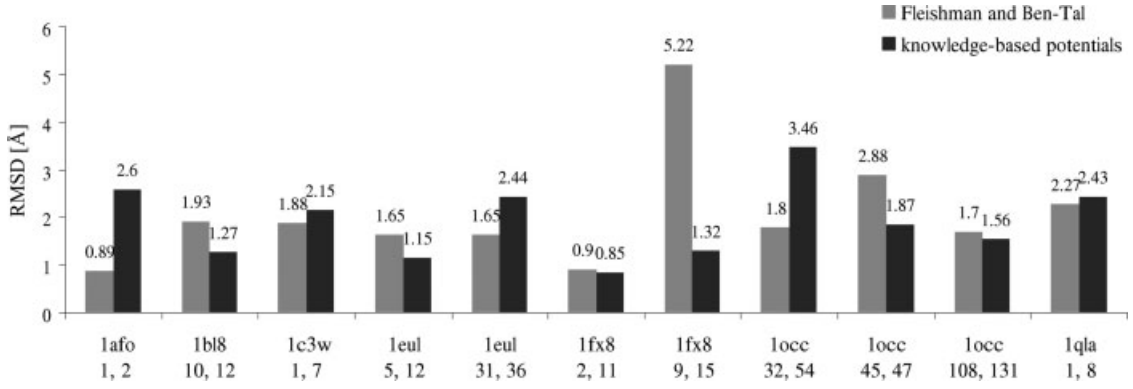

Figure 8

Comparison of the RMSD values for 11 helix pairs, obtained by either scoring with the function of Fleishman and Ben-Tal[34] or the knowledge-based potentials derived in this study. In both cases, a systematic conformational search was applied. The indices of the TM helices correspond to the definition of helical regions in the PDB-file, as reported in Ref. 34.

0.16 Å. As Figure S4 also shows, helix pair 1occ (45, 47) does not lie in the membrane-spanning region of the protein and should therefore be omitted from the test data set. Neglecting this pair, on average, we predicted helix pair orientations with a RMSD of 1.92 Å compared to 1.98 Å of Fleishman and Ben-Tal. Taken together the knowledge-based potentials described in this study yielded comparable results to the scoring function of Fleishman and Ben-Tal on the test data set of 11 helix pairs.

It may be interesting to note in this context that we are able to predict 5 of the 11 helix pairs [1bl8 (10, 12), 1eul (5, 12), 1eul (31, 36), 1fx8 (9, 15), 1occ (108, 131)] with an RMSD <2 Å when docking with cross-validated potentials, allowing six degrees of freedom.

CONCLUDING REMARKS

In this study, distance-dependent knowledge-based pair potentials have been derived from a large database of 71 known TM protein structures. For the first time, mutual TM helix pair configurations were scored solely using such potentials. The helix pair configurations have been generated using a rigid docking approach. Neither information about helix-lipid interactions nor any other constraints to restrict the available configurational space such as, for example, provided by loops connecting the helices were incorporated. The evaluation was performed on the largest test data set reported to date (442 TM helix pairs), comprising also more difficult to predict helix pairs with short loops and interhelical distances up to 12 Å. Notably, only interactions between C_α-atoms of amino acid pairs were considered. These coarse-grained potentials can be evaluated efficiently. This is important in view of the sampling problem in initial stages of TM helix bundle predictions, because the number of likely folds increases exponentially with the number of helices considered.[57]

Encouragingly, in more than 71% of the cases a "good" (using a strict RMSD cut-off of 2.0 Å) helix pair configuration could be identified as the best scored solution of the largest cluster generated, if only clusters were considered that contained at least 20% of the decoys. Here two remarks are in order: First, the cluster size could be devised as a measure of significance for the identification of good docking solutions. This is important when it comes to extending the approach to predict TM helix bundles. Second, the finding that successful dockings usually have a larger number of structural neighbors implies that native-like structures of TM proteins reside in broader energy minima than non-native states.

The robustness of the potentials was tested by a rigorous leave-one-TM-protein-family-out cross-validation, where up to 14% of the proteins of the knowledge base were omitted at once. When considering only clusters that contained at least 20% of the decoys, convincingly, more than 2/3 of the helix pairs were still predicted with a RMSD <2.5 Å. Although this demonstrates the predictive power of the potentials in general, the performance decrease compared to the original potentials advises to rederive the potentials with an increased knowledge base in the future. Finally, when compared to the scoring function of Fleishman and Ben-Tal[34] on a limited test data set of 11 helix pairs, comparable prediction results were obtained with the potentials derived here.

Despite the success in predicting TM helix pair configurations, certain limitations of our approach exist. Foremost, only helix conformations extracted from the experimental TM protein structures were used in the rigid docking, although in blind predictions either ideal helix geometries or predicted conformations of kinked or bended helices will be applied. That way, however, our prediction results reflect purely the quality of the scoring function rather than influences due to nonideal helix geometries. Fortuitous influences on the success rate due to close-packing effects of side-chains can be ruled out because we only consider C_α-atom positions in our coarse-grained approach. Nevertheless, we consider this a major restriction of our method, and work in this area is currently underway.

At first glance, the obtained success rates may come as a surprise given that no interactions between sidechains are considered. Regarding, however, that predominantly small and β-branched amino acids with limited conformational variability reside in TM helix interfaces,[28,58,59] our potentials of C_α-atom interactions may implicitly contain already sufficient knowledge of more detailed sidechain interactions. Along these lines, it is intriguing to note that the scoring function of Fleishman and Ben-Tal, which was tailored to favor the close-packing of small helices and penalize the burial of large residues, already performs very well.[34]

As we only use pair potentials, many body effects are not captured by our approach, that is, the interaction between two C_α-atoms is not modulated by the presence of other amino acids in the neighborhood. This may become particularly pronounced for buried helices within larger TM protein structures. We note, however, that due to the derivation process the statistical potentials implicitly incorporate knowledge about an average molecular environment of a specific residue pair[44] and, thus, are qualitatively different from potentials describing pair interactions in the gas-phase, such as vdW potentials.[11,26]

Finally, our potentials show overall good predictive power although the influence of the membrane is not explicitly taken into account. On the one hand, this may be explained by the suggestion that TM helix association is not driven by hydrophobic interactions.[60] On the other hand, the decrease of the success rate observed for helix pairs with large crossing angles may indicate a missing

 DOI 10.1002/prot

scoring term accounting for the membrane field, which provides a counterbalance to the pair potentials that work particularly well in the case of (anti-)parallel configurations. More work in this area is clearly needed. Furthermore, the knowledge-based potentials may be augmented by terms that take into account the lengths of the loops between two helices, consider explicit packing rules from small sequence motifs,[50,51] or include knowledge about the location of an amino acid in a helix and/or the deepness of the amino acid's burial in the membrane.[35]

The presented potentials are the first step in an endeavor to ultimately predict TM helix bundles. In view of this, it was encouraging to find well-predicted anchor helix pairs for most of the proteins contained in the test set. These anchor pairs will be used as nucleation sites to which more helices will be added subsequently. That way it is our hope to bypass the combinatorial explosion one would face when testing all possible combinations systematically.

METHODS

Distance-dependent pair potentials

Database of membrane proteins

For the derivation of the potentials structures of helical TM proteins were selected from the database of Steven White (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html, version of Februray 2007). Structures with a resolution lower than 4.0 Å and structures with a sequence identity of more than 60% were neglected. To determine the sequence identity we aligned each protein with all of its family members. The classification in protein families was adopted from Steven White's database, thereby combining bacterial rhodopsins and GPCRs; oxygenases and photosynthetic reaction centers; P-type ATPases, V-type ATPases, and F-type ATPases, respectively. This resulted in a database of 71 helical TM structures (PDB codes: 1a91, 1afo, 1bgy, 1c3w, 1e12, 1ehk, 1ezv, 1fft, 1fx8, 1h2s, 1j4n, 1jb0, 1k4c, 1kqf, 1l0v, 1l7v, 1lgh, 1lnq, 1nek, 1nkz, 1occ, 1oed, 1ogv, 1okc, 1orq, 1ots, 1p7b, 1prc, 1p7v, 1pw4, 1q16, 1rc2, 1rhz, 1rwt, 1u19, 1vf5, 1vgo, 1wpg, 1xfh, 1xio, 1xl4, 1xqf, 1yce, 1yew, 1ygm, 1ymg, 1z98, 1zcd, 1zll, 1zoy, 2a65, 2a79, 2ahy, 2axt, 2b2f, 2bbj, 2bl2, 2bs2, 2d57, 2dhh, 2f2b, 2gfp, 2gsm, 2hi7, 2hyd, 2ic8, 2j58, 2j7a, 2nq2, 2oar, 2oau).

All but six of the structures were solved by X-ray crystallography; 1a91, 1afo, 1ygm, and 1zll were determined by NMR spectroscopy, 1oed and 2d57 by cryo-electron microscopy. Initially, secondary structure was assigned by DSSP.[61] After identifying TM regions by visual inspection, the corresponding TM helices were determined using the DSSP assignments. To take into consideration that kinks often occur in TM helices[62,63] but should not break a helix into small parts, all sequences either fully assigned as H (α-helix), H interrupted only by G (3_{10}-helix) residues, or H interrupted by no more than two residues with other assignments than H or G were regarded as one TM helix. Some further adjustments of the TM helices (shortening of helices and consideration of obvious kinks that are sometimes assigned by DSSP to have more than two T residues) were performed manually. In total, this led to 969 TM helices from which the knowledge-based potentials were derived. The dataset is available from the authors upon request.

Derivation of the potentials

Distances between C_α-atoms from 3.1 Å up to 10.0 Å were considered for the derivation of the potentials. As a compromise between a sufficiently high resolution and the number of observed pair interactions available for each potential type, the bin size was set to $dr = 0.3$ Å. Frequencies $N_{ij}(r_b)$ were counted for each amino acid pair ij occurring in a distance range between r_b and $r_b + dr$.

$$N_{ij}(r_b) = \sum_i \sum_j \delta\left(\|\vec{r}_i - \vec{r}_j\|, r_b\right), \tag{1}$$

whereby $\vec{r}_i$ and $\vec{r}_j$ are the Cartesian coordinates of the corresponding C_α-atoms. To generate smoother potential functions, an observed pair interaction is not completely assigned to just one bin, but "smeared" to neighboring bins proportional to the location of the pair interaction distance in the central bin. For this, an isosceles triangle function is implemented as a smoothing function δ, similar to the one used by Gohlke *et al.*[44] for the derivation of the DrugScore potentials. The triangle width was set to 0.7 Å. The normalized radial pair distribution functions $g_{ij}(r_b)$ of the 210 different pairs are then calculated as follows:

$$g_{ij}(r_b) = \frac{N_{ij}(r_b)}{\sum_b N_{ij}(r_b)}. \tag{2}$$

For obtaining netto potentials $\Delta W_{ij}(r_b)$ that represent only the specific interactions we are interested in, a reference state $g(r_b)$ is required that removes the contributions from zero-interaction contacts. We tested both reference states proposed by Gohlke *et al.*[44] and by Sippl.[42,43] As initial tests gave better results with the latter one (data not shown), this reference state was finally used for the derivation of the potentials:

$$g(r_b) = \frac{\sum_i \sum_j N_{ij}(r_b)}{\sum_b \sum_i \sum_j N_{ij}(r_b)}. \tag{3}$$

Considering the reference state finally yields the netto potentials:

$$\Delta W_{ij}(r_b) = -\ln \frac{g_{ij}(r_b)}{g(r_b)}. \tag{4}$$

The potential value is set to 0 in bins for which no contacts were found in the database. Setting those bins to interpolated values of neighboring occupied bins did not change the results when docking helix pairs (data not shown). An additional repulsive term was added to the short-distance region of the potentials to avoid clashes of helices upon docking. For that, a tolerance of 0.3 Å is subtracted from the smallest observed interaction distance of a particular residue pair. Below this distance the respective potential is set to a large positive value (1000 in our case). All potential values are given in arbitrary units in this study.

Assembly of transmembrane helix pairs

Docking of helix pairs

We used the Lamarckian genetic algorithm of Autodock 3.0[64] for the rigid docking of TM helix pairs. Autodock has an efficient grid-based energy evaluation and allows to consider user provided potential grids. This makes this application ideally suited to test the predictive power of the generated potentials on a large test data set.

For a given "receptor" helix (i.e., the first one in the TM protein sequence) of each helix pair, 20 potential grids g^i are precalculated according to:

$$g^i_{\vec{r}} = \sum_j \Delta W_{ij}\left(\|\vec{r} - \vec{r}_j\|\right), \qquad (5)$$

where $\vec{r}$ is the location of a particular grid point. Thus, $g^i_{\vec{r}}$ is the sum over all interactions of receptor helix atoms located at $\vec{r}_j$ with a probe atom of type i placed at the grid point $\vec{r}$. The distance between two grid points is set to 0.3 Å. The energy contribution of an atom located at an arbitrary position is then determined by a trilinear interpolation between the values of the eight nearest grid points. Atoms of the "ligand" helix lying outside of the potential grids are penalized by disfavorable contributions. Thus, a compromise must be found between grids large enough to not restrict the sampling of the ligand helix on the one hand, but small enough to fit into the available computer memory on the other hand. On the basis of the fact that most TM helices adopt a near parallel or antiparallel orientation[31,65,66] we ensured that the grids cover the most populated ranges of interhelical crossing angles ($-45°$ to $45°$, $-135°$ to $-180°$, and $135°$ to $180°$). For that, the receptor helix is placed in the center of the grid, with its helix axis oriented in parallel to the z-axis. Grid borders were initially determined by adding a 10 Å margin in each direction to the largest extension of the helix. If this is still insufficient to cover the desired crossing angle ranges, the grid sizes are adapted accordingly. The grid size is identical in x and y-direction. As only rigid helix conformations are considered, no contributions from internal energies are considered for scoring.

For the docking of each helix pair 500 runs of the genetic algorithm were performed. The population size was set to 100 and the maximal number of energy evaluations to 3×10^5. The 500 resulting helix pair orientations were clustered, with conformations with a mutual C_α-atom RMSD value <1 Å assigned to the same cluster. The conformation with the lowest energy of the largest cluster is regarded as the prediction result, instead of the overall lowest energy conformation. RMSD values given for the largest cluster correspond to this prediction result.

Calculation of helix axes and crossing angles

The helix geometry, especially the helix axes and crossing angles, are used not only to select the test data set of 442 helix pairs for docking, but are also needed for the global search performed to compare our potentials with the scoring function of Fleishman and Ben-Tal[34] (see later).

Considering the helix as a rigid body, the helix axis is given by the main spatial expansion of the atomic coordinates. The helix axis corresponds to the eigenvector with the smallest eigenvalue derived from

$$I = \begin{bmatrix} \sum_i y_i^2 z_i^2 & \sum_i -x_i y_i & \sum_i -x_i z_i \\ \sum_i -x_i y_i & \sum_i x_i^2 z_i^2 & \sum_i -y_i z_i \\ \sum_i -x_i z_i & \sum_i -y_i z_i & \sum_i x_i^2 y_i^2 \end{bmatrix}. \qquad (6)$$

Thereby, the sum over i is the sum over all points with coordinates (x_i, y_i, z_i) of a body. The direction of the helix axis points from the N-terminus to the C-terminus.

The crossing angle is defined as the angle around the line of closest approach between two helices.[67] It ranges from $-180°$ to $180°$, being negative if the forward helix is rotated clockwise around the backwards one.

To calculate the line of closest approach, the helix axes are given as two lines $P(s) = P_0 + s\vec{u}$ and $Q(t) = Q_0 + t\vec{v}$. The vector which connects the two points $P(s_c)$ and $Q(t_c)$ that are closest to each other on these lines is defined as the line of closest approach $\vec{w}_c = P_0 - Q_0 + s_c\vec{u} - t_c\vec{v}$. It is orthogonal to the line direction vectors $\vec{u}$ and $\vec{v}$. Therefore, $\vec{w}_c$ has to fulfil the linear equalities

$$(\vec{u}\cdot\vec{u})s_c - (\vec{u}\cdot\vec{v})t_c = -\vec{u}(P_0 - Q_0) \quad \text{and}$$
$$(\vec{u}\cdot\vec{v})s_c - (\vec{v}\cdot\vec{v})t_c = -\vec{v}(P_0 - Q_0).$$

Solving this system of linear equations gives the line of closest approach, for which a unique solution exists only if the axes are not parallel to each other (if they are, either s_c or t_c must be chosen arbitrarily). The length of the line of closest approach gives the interhelical distance between two helices.

Test data set of 442 TM helix pairs

The test data set of helix pairs for docking was created from the database that was also used for deriving the

 DOI 10.1002/prot

potentials. However, only those helix pairs were included in the test data for which: (1) both helices are at least 10 amino acids long, (2) the interhelical distance is at most 12 Å, (3) the line of closest approach between the helices lies within the inner third of the helices (a violation of this condition could produce artefacts when determining the crossing angle, because helix axes are treated as infinitely long and the crossing angle is defined around the line of closest approach), and (4) the crossing angle is in the range of -45° to 45°, -180° to -135°, or 135° to 180°. These criteria resulted in a test data set of 442 TM helix pairs from 58 proteins. Helix pairs are denoted by their PDB code and the indices of the helices throughout the study (e.g., 1c3w (0, 1)). After identifying helices as described earlier, indices are assigned to all TM helices of each protein in ascending order from N- to C-terminus, starting at 0. Chains are thereby traversed in alphabetical order. The dataset is available from the authors upon request.

Cross validation

To test the robustness of the potentials, a leave-one-protein-family-out cross-validation was performed. Every TM helix pair was docked again with potentials that were derived from a knowledge base in which not only the protein under consideration was omitted but all proteins of its family with a sequence identity >30% of any two chains. This resulted in a much more stringent cross-validation than if only one helix pair was excluded from the knowledge base in each case. All other parameters used to derive the cross-validated potentials were identical to the ones used to derive the original potentials.

Comparison to the scoring function of Fleishman and Ben-Tal[34]

Fleishman and Ben-Tal developed a scoring function based on the qualitative analysis of TM protein structures.[34] To compare the knowledge-based potentials with this scoring function, we used the same test data set of 11 helix pairs and performed a global search as described by these authors.

Test data set of 11 TM helix pairs

The dataset consists of 11 helix pairs, whereby the indices, in contrast to our numbering scheme, correspond to the definition of helical regions in the PDB-file: 1afo (1, 2), 1bl8 (10, 12), 1c3w (1, 7), 1eul (5, 12), 1eul (31, 36), 1fx8 (2, 11), 1fx8 (9, 15), 1occ (32, 54), 1occ (45, 47), 1occ (108, 131), and 1qla (1, 8).

Global search around the native state

Starting from the native orientation of a helix pair, a global search is performed by systematically changing the mutual orientation of the helices in a defined region around the native structure. After each step, an energy evaluation is performed. Finally, the lowest-energy orientation is regarded as the predicted orientation. The global search is performed according to the description of Fleishman and Ben-Tal.[34] Five of the six degrees of freedom are varied; the interhelical distance is kept fixed at its native value. The step sizes and ranges of the crossing angle, the translation in *x*- and *z*-direction, and the rotations around the two helix axes were taken from Table II of the study of Fleishman and Ben-Tal.[34] Differing from their approach we calculated crossing angles not only between -90° and 90°, but between -180° and 180°, thus distinguishing between parallel and antiparallel orientations. Accordingly, we adapted the range of crossing angles to search for antiparallel orientations to -180° to -103° and 103° to 180°.

ACKNOWLEDGMENT

We are grateful to Sarel Fleishman for helpful discussions.

REFERENCES

1. Watts A. Solid-state NMR in drug design and discovery for membrane-embedded targets. Nat Rev Drug Discov 2005;4:555–568.
2. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567–580.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
4. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15:285–289.
5. Forrest LR, Tang CL, Honig B. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J 2006;91:508–517.
6. Hillisch A, Pineda LF, Hilgenfeld R. Utility of homology models in the drug discovery process. Drug Discov Today 2004;9:659–669.
7. Oliveira L, Hulsen T, Hulsik DL, Paiva ACM, Vriend G. Heavier-than-air flying machines are impossible. FEBS Lett 2004;564:269–273.
8. Fanelli F, De Benedetti PG. Computational modeling approaches to structure-function analysis of G protein-coupled receptors. Chem Rev 2005;105:3297–3351.
9. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress in modeling of protein structures and interactions. Science 2005;310:638–642.
10. Ginalski K, Grishin NV, Godzik A, Rychlewski L. Practical lessons from protein structure prediction. Nucleic Acids Res 2005;33:1874–1891.
11. Kim S, Chamberlain AK, Bowie JU. A simple method for modeling transmembrane helix oligomers. J Mol Biol 2003;329:831–840.
12. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
13. Bowie J. Solving the membrane protein folding problem. Nature 2005;438:581–589.
14. Popot JL, Engelman DM. Membrane protein folding and oligomerization: the two-stage model. Biochemistry (Mosc) 1990;29:4031–4037.

15. Popot JL, Engelman DM. Helical membrane protein folding, stability, and evolution. Annu Rev Biochem 2000;69:881–922.
16. Juretic D, Zoranic L, Zucic D. Basic charge clusters and predictions of membrane protein topology. J Chem Inform Comput Sci 2002;42:620–632.
17. Persson B, Argos P. Prediction of membrane protein topology utilizing multiple sequence alignments. J Protein Chem 1997;16:453–457.
18. Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849–850.
19. Chen CP, Kernytsky A, Rost B. Transmembrane helix predictions revisited. Protein Sci 2002;11:2774–2791.
20. Cuthbertson JM, Doyle DA, Sansom MS. Transmembrane helix prediction: a comparative evaluation and analysis. Protein Eng Des Sel 2005;18:295–308.
21. Adamian L, Nanda V, DeGrado WF, Liang J. Empirical lipid propensities of amino acid residues in multispan α helical membrane proteins. Proteins 2005;59:496–509.
22. Beuming T, Weinstein H. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. Bioinformatics 2004;20:1822–1835.
23. Pilpel Y, Ben-Tal N, Lancet D. kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. J Mol Biol 1999;294:921–935.
24. Fleishman SJ, Harrington S, Friesner RA, Honig B, Ben-Tal N. An automatic method for predicting transmembrane protein structures using cryo-EM and evolutionary data. Biophys J 2004;87:3448–3459.
25. Sale K, Faulon JL, Gray GA, Schoeniger JS, Young MM. Optimal bundling of transmembrane helices using sparse distance constraints. Protein Sci 2004;13:2613–2627.
26. Pappu RV, Marshall GR, Ponder JW. A potential smoothing algorithm accurately predicts transmembrane helix packing. Nat Struct Biol 1999;6:50–55.
27. Park Y, Elsner M, Staritzbichler R, Helms V. Novel scoring function for modeling structures of oligomers of transmembrane α-helices. Proteins 2004;57:577–585.
28. Adamian L, Liang J. Helix–helix packing and interfacial pairwise interactions of residues in membrane proteins. J Mol Biol 2001;311:891–907.
29. Eilers M, Patel AB, Liu W, Smith SO. Comparison of helix interactions in membrane and soluble α-bundle proteins. Biophys J 2002;82:2720–2736.
30. Curran AR, Engelman DM. Sequence motifs, polar interactions and conformational changes in helical membrane proteins. Curr Opin Struct Biol 2003;13:412–417.
31. Senes A, Ubarretxena-Belandia I, Engelman DM. The Cα—H···O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. Proc Natl Acad Sci USA 2001;98:9056–9061.
32. Shacham S, Marantz Y, Bar-Haim S, Kalid O, Warshaviak D, Avisar N, Inbal B, Heifetz A, Fichman M, Topf M, Naor Z, Noiman S, Becker OM. PREDICT modeling and in-silico screening for G-protein coupled receptors. Proteins 2004;57:51–86.
33. Dobbs H, Orlandini E, Bonaccini R, Seno F. Optimal potentials for predicting inter-helical packing in transmembrane proteins. Proteins 2002;49:342–349.
34. Fleishman SJ, Ben-Tal N. A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane α-helices. J Mol Biol 2002;321:363–378.
35. Yarov-Yarovoy V, Schonbrun J, Baker D. Multipass membrane protein structure prediction using Rosetta. Proteins 2006;62:1010–1025.
36. Fleming KG, Engelman DM. Specificity in transmembrane helix–helix interactions can define a hierarchy of stability for sequence variants. Proc Natl Acad Sci USA 2001;98:14340–14344.
37. White SH, Wimley WC. Membrane protein folding and stability: physical principles. Annu Rev Biophys Biomol Struct 1999;28:319–365.
38. Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996;258:367–392.
39. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
40. Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101–11108.
41. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins 2000;38:134–148.
42. Sippl MJ. Calculation of conformational ensembles from potentials of mean force—an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
43. Sippl MJ. Boltzmann principle, knowledge-based mean fields and protein-folding—an approach to the computational determination of protein structures. J Comput Aided Mol Des 1993;7:473–501.
44. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 2000;295:337–356.
45. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
46. Wiederstein M, Sippl MJ. Protein sequence randomization: Efficient estimation of protein stability using knowledge-based potentials. J Mol Biol 2005;345:1199–1212.
47. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 1997;266:195–214.
48. Zhou HY, Zhou YQ. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
49. Zhang C, Liu S, Zhu QQ, Zhou YQ. A knowledge-based energy function for protein-ligand, protein–protein, and protein-DNA complexes. J Med Chem 2005;48:2325–2335.
50. Russ WP, Engelman DM. The GxxxG motif: a framework for transmembrane helix–helix association. J Mol Biol 2000;296:911–919.
51. Senes A, Engel DE, DeGrado WF. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. Curr Opin Struct Biol 2004;14:465–479.
52. Schneider D. Rendezvous in a membrane: close packing, hydrogen bonding, and the formation of transmembrane helix oligomers. FEBS Lett 2004;577:5–8.
53. Zhang Y, Skolnick J. SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.
54. Cole JC, Murray CW, Nissink JWM, Taylor RD, Taylor R. Comparing protein-ligand docking programs is difficult. Proteins 2005;60:325–332.
55. Bursulaya BD, Totrov M, Abagyan R, Brooks CL. Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des 2003;17:755–763.
56. Vieth M, Hirst JD, Kolinski A, Brooks CL. Assessing energy functions for flexible docking. J Comput Chem 1998;19:1612–1622.
57. Bowie JU. Helix-bundle membrane protein fold templates. Protein Sci 1999;8:2711–2719.
58. Senes A, Gerstein M, Engelman DM. Statistical analysis of amino acid patterns in transmembrane helices: The GxxxG motif occurs frequently and in association with β-branched residues at neighboring positions. J Mol Biol 2000;296:921–936.
59. Jiang SL, Vakser IA. Side chains in transmembrane helices are shorter at helix–helix interfaces. Proteins 2000;40:429–435.
60. Stevens TJ, Arkin IT. Are membrane proteins "inside-out" proteins? Proteins 1999;36:135–143.
61. Kabsch W, Sander C. Dictionary of protein secondary structure—pattern-recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.

 DOI 10.1002/prot

62. Riek RP, Rigoutsos I, Novotny J, Graham RM. Non-α-helical elements modulate polytopic membrane protein architecture. J Mol Biol 2001;306:349–362.
63. Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU. The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc Natl Acad Sci USA 2004;101:959–963.
64. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 1998;19:1639–1662.
65. Bowie JU. Helix packing in membrane proteins. J Mol Biol 1997;272:780–789.
66. Bywater RP, Thomas D, Vriend G. A sequence and structural study of transmembrane helices. J Comput Aided Mol Des 2001;15:533–552.
67. Chothia C, Levitt M, Richardson D. Helix to helix packing in proteins. J Mol Biol 1981;145:215–250.