Method To Assess Packing Quality of Transmembrane α-Helices in Proteins. 1. Parametrization Using Structural Data

Anton O. Chugunov,*[,†,‡] Valery N. Novoseletsky,[†,§] Dmitry E. Nolde,[†] Alexander S. Arseniev,[†] and Roman G. Efremov[†]

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, ul. Miklukho-Maklaya 16/10, GSP Moscow 117997, Russia, Department of Bioengineering, Biological Faculty, M.V. Lomonosov Moscow State University, Vorobiovy gory, 119899 Moscow, Russia, and Moscow Institute of Physics and Technology (State University), Institutskii per. 9, Dolgoprudny, Moscow Region 141700, Russia

Received November 17, 2006

Integral membrane proteins (MPs) are pharmaceutical targets of exceptional importance. Modern methods of three-dimensional protein structure determination often fail to supply the fast growing field of structure-based drug design with the requested MPs' structures. That is why computational modeling techniques gain a special importance for these objects. Among the principal difficulties limiting application of these methods is the low quality of the MPs' models built in silico. In this series of two papers we present a computational approach to the assessment of the packing "quality" of transmembrane (TM) α-helical domains in proteins. The method is based on the concept of protein environment classes, whereby each amino acid residue is described in terms of its environment polarity and accessibility to the membrane. In the first paper we analyze a nonredundant set of 26 TM α-helical domains and compute the residues' propensities to five predefined classes of membrane-protein environments. Here we evaluate the proposed approach only by various test sets, cross-validation protocols and ability of the method to delimit the crystal structure of visual rhodopsin, and a number of its erroneous theoretical models. More advanced validation of the method is given in the second article of this series. We assume that the developed "membrane score" method will be helpful in optimizing computer models of TM domains of MPs, especially G-protein coupled receptors.

1. INTRODUCTION

Integral membrane proteins (MPs) are objects of exceptional biological importance known to mediate transmembrane (TM) signal transduction, light absorption, generation of TM potential, etc. G-protein coupled receptors (GPCR), the most broad and important class of MPs, for instance, are targets for more than 50% of all currently marketed drugs.[1] Functioning of MPs depends primarily on the TM domain, which often binds a ligand and accommodates conformational reorganization initiating intracellular response. Information on the structure and functioning of TM domains is highly required for pharmaceutical applications, such as structure-based drug design (SBDD). Modern experimental techniques of three-dimensional (3D) structure determination, such as X-ray crystallography and NMR spectroscopy, on the other hand, often fail to solve the problem due to technical difficulties related to protein purification and crystallization.[2] Only a few tens of MPs' structures have been determined to this day, making up <1% of the total number of known structures in the Protein Data Bank (PDB),[3] although the genomes sequenced so far encode at least 15−30% of MPs.[4]

Molecular modeling, however, could go a long way in predicting the 3D structure of membrane-spanning proteins, complementarily to experimental methods. Since most TM domains of MPs are formed of α-helices (they and β-barrels are the only two folds discovered in MPs to date),[5] and therefore, given that exactly this class is pharmacologically important, we will further focus on α-helical TM proteins. The modeling includes the following stages: (i) TM segments are identified in the protein's amino acid sequence; (ii) the optimal mutual arrangement of the helices is predicted; (iii) specific structural/functional features, e.g., kinks and other deviations from the ideal α-helicity, are delineated.[6] In this protocol, stage (ii) poses the greatest challenge—even in the case of the simplest helix−helix systems the computational procedure is not straightforward and requires exhaustive sampling of the peptides' conformational space in a heterogeneous membranelike environment.[7,8] Obviously, for multihelix complexes the task becomes much more complicated due to a higher number of degrees of freedom. Modeling approaches used to build such models can be divided into two groups.

One group includes homology-based techniques, the most reliable way to construct a model for subsequent applications in SBDD.[9] In this case a TM helix bundle is modeled using atomic coordinates of a suitable template with known 3D structure. The other group of methods, in contrast, does not directly employ the structural data obtained for homologous proteins. Some of these techniques are based on simulations

* Corresponding author phone/fax: +7(495)3362000; e-mail: volster@nmr.ru.
† Russian Academy of Sciences.
‡ M.V. Lomonosov Moscow State University.
§ Moscow Institute of Physics and Technology (State University).

10.1021/ci600516x CCC: $37.00
Published on Web 03/20/2007

with empirical force fields, such as Monte Carlo (MC) and molecular dynamics (MD) methods.[10,11] Calculations are carried out using all-atom, coarse grained, or lattice protein models in explicit and/or implicit membranes. Recent adaptation of the ab initio folding program ROSETTA[12] to TM α-helical proteins by inclusion of a statistical membrane-mimic energy term has demonstrated relative progress of ab initio methods for MPs.[13] However, apart from small globular proteins,[14] this low-resolution approach is still unable to handle such big proteins as GPCRs or bacteriorhodopsin.[13] The TASSER (Threading/ASSEmbly Refinement) approach, bridging the gap between traditional homology modeling and ab initio techniques, has recently been applied to structure prediction of all 907 probable human GPCRs.[15] Using three templates with a homology level less than 30%, it was able to reconstruct the structure of rhodopsin with C_α-rmsd of 2.1 Å. Another high-throughput GPCR modeling approach was used to predict the structures of 277 human nonolfactory receptors, starting from either rhodopsin crystal structure or several homology-built models of various GPCRs, previously reported by the authors.[16] In this method, the backbone of a model was taken from the most homologous template, and the side chains conformations were transferred from a position-dependent rotamer library, derived from the same set of structures.[16] The simulations of TM domains are often supplemented by mutagenesis and other experimental data serving as constraints in modeling protocols. Sometimes, the body of various biochemical and biophysical data might be large enough to predict the structure of a particular MP from the derived set of spatial constraints only.[17] Furthermore, in many techniques of this group helix packing is optimized using a number of criteria reflecting the general principles derived from the analysis of the available spatial structures of TM domains of MPs (see recent reviews by, e.g., Bowie[18] and White and von Heijne[19]).

Advantages and limitations of both groups of methods are discussed in detail elsewhere.[20] However, once a model is proposed, a common problem arises, namely that of assessing its quality for subsequent refinement and application to biomedical tasks. This problem, albeit of critical importance, is yet to be solved. 3D models of GPCRs, for example, are routinely built with recourse to fully automated WEB services (e.g., SWISS-MODEL).[21] Although some of the constructed models have been successfully employed in drug design,[20] the application of homology-built models is still very limited, principally because the models generated directly on the rhodopsin structural template, the only available high-resolution structure from the GPCR family, are uncertain. Thus, apart from the extramembrane regions (modeling of which still poses a great challenge), even prediction of the structure of TM domains is far from being straightforward, since mutual positions of α-helices may vary from receptor to receptor. Therefore, additional information is required to refine in silico models of TM parts of MPs.

One of such criteria is based on the analysis of hydrophobic moments[22] of individual segments. MPs reside in a highly nonpolar environment, and their lateral surface residues are, on average, more hydrophobic than those buried inside the bundle,[23] whereas globular proteins usually form a hydrophobic core exposing polar residues into water.[24] It was, therefore, suggested that TM helices in the models of MPs should be arranged in such a way that their hydrophobic moments point to the lipid environment.[23] However, applicability of this criterion is limited, since hydrophobic moments of many MPs do not conform to this rule.[25,26] This can seemingly happen when hydrophobicity is more or less uniformly distributed between the buried and exposed sides of a helix. For instance, nonhydrophobic residues can often be found in membrane-exposed positions.[27] These preferences, quantitatively estimated via the frequency of residue occurrence in mono- and polytopic α-helical MPs, may also be helpful for adjustment of helices' orientations in the models.[28]

Another criterion for assembling of TM segments in MPs is based on the analysis of orientations of variability moments of the helices. The most variable residues in a family of homologous proteins tend to be exposed to the membrane.[23,29−32] This may mean that the interactions between conserved buried residues are especially important for the maintenance of the spatial structure, whereas the contacts of the exposed residues with the membrane are less specific. Variability moments of TM α-helices are, therefore, most likely to point out of the lipid-exposed helix side. This finding may be used to assemble simple TM proteins (e.g., dimers)[33] or to optimize low-resolution MP structures obtained, e.g., by electron cryomicroscopy. In the latter case neither the exact positions of side chains and kinks nor even the orientation of the helices can be determined from the electron density data.[34,35]

Furthermore, folding of MPs may also be driven by favorable van der Waals[26] and/or highly specific polar (e.g., H-bonds)[36] interactions. As a result, the optimal packing density in MPs may be significantly higher as compared to that in globular proteins.[37] This is especially true for interhelical interfaces, often formed by small residues,[38] like the famous GXXXG motif (X denotes any residue).[39−42] This principle can be applied to the prediction of arrangement of α-helices in TM dimers,[43,44] oligomers,[34] and α-helical bundles.[45] The aforementioned methods of construction of spatial models of TM domains and some additional criteria are often used in a concerted manner to minimize modeling errors and uncertainties. Thus, the hybrid method PREDICT,[46] designed for GPCR modeling, utilizes a battery of molecular modeling tools, such as hydrophobic moments, MC and MD simulations in explicit full-atom lipid bilayers.

For a long time, the problem of assessment of the packing quality has been attracting serious attention in the case of globular proteins. Several efficient methods have been developed to solve it. The most successful of these are based on microscopic parameters characterizing the environment of a particular residue in the 3D structure. These techniques[47,48] express a residue's environment in terms of its polarity and accessibility to the surrounding and describe its "compatibility" with a given environment (in other words, similarity of the residue's microenvironment with those in the training set). These approaches (especially the algorithm of 3D-1D profiles)[47] have been shown to perform well in searches of structural inconsistencies of protein structures and in discrimination between correct protein models (or their selected parts) and the misfolded ones. At the same time, since these methods were refined for globular proteins, their applicability to TM parts of MPs is questionable, and the approach requires parametrization on a suitable set of spatial structures of membrane-spanning domains.

Here we present a series of two articles dedicated to development and validation of a novel method of assessing the packing quality of TM α-helical proteins. In the first paper, we describe an analysis of a nonredundant set of 26 high-resolution X-ray structures of MPs that allowed delineation of the residues' propensities to five predefined classes of membrane-protein environment as well as optimization of parameters for these classes. The developed "membrane score" method circumscribes the residues' compatibilities with their microenvironments in the properly folded conformations. As a result, strong correlation between the TM sequence length and the corresponding score was obtained for structures from the training and test sets. Additionally, the "membrane score" method successfully delimits well-packed (e.g., X-ray) rhodopsin structure and a set of its 13 erroneous theoretical models, yielding a good correlation between the root-mean-square deviation (rmsd) from the "native" structure and the score impairment. (Hereinafter the term "correct" is used to refer to the X-ray structures; it should be borne in mind, however, that such models do not necessarily reside under the native conditions—in the biological membranes). The more versatile validation of the method, including the testing of its ability to distinguish between the "native" structures and the "decoy" conformations as well as to detect alignment errors, is given in the second article of this series. The developed "membrane score" method is believed to be useful in optimizing computer models of TM domains of MPs, especially those of GPCRs.

2. RESULTS AND DISCUSSION

2.1. Characteristics of Residues in TM Helical Bundles. There are almost 200 spatial structures of MPs solved in experiments to this day. However, this list is highly redundant, and after low-resolution and NMR-derived structures, β-barrel proteins, and closely homologous proteins were excluded, the training set of 21 proteins was established (see Methods). A separate set was composed of photosynthetic proteins, since their TM domains abound in light-absorbing pigments that are thought to substitute interhelical protein−protein interactions by protein-pigment ones. Another 5 structures with a considerable degree of homology to MPs from the training set were combined within the test set (see Methods).

Development of a scoring function for TM α-helical domains requires accurate delineation of the "optimal" boundaries of membrane-spanning segments in each protein structure. This was done using a simple computational procedure of scanning a given MP model along the membrane normal (*Z*-axis) with an implicit hydrophobic slab of varying thickness (see Methods). Two parameters were adjusted during this search: *Z*-coordinate and the thickness of the slab. As a result, the best arrangement of the protein (corresponding to the minimal energy of protein interaction with the slab, E_{solv}, Figure 1) was obtained, and the boundaries of TM segments were determined. For all proteins in the databases of MPs used in this study it was possible to find a prominent minimum of the function $E_{solv}(Z)$, which corresponded well to the approximate location of the TM helix bundle described in the literature. Variation of the slab thickness (the resulting average value was 27 ± 4 Å) allowed further optimization of the boundaries of TM helices (Figure 1, Table S1 in the Supporting Information). (One should bear in mind, however, that the boundaries of TM domain are quite vague rather than sharply defined because of the dynamic nature of water−lipid interface.) Based on this data, the final database (training set) composed of 175 TM α-helices and 4067 residues was constructed. As it was reasonable to expect, the same procedure, if applied to globular proteins, yields significantly different solvation energy profiles with an energy maximum at the membrane region (data not shown), thus confirming their incompatibility with the membrane environment.

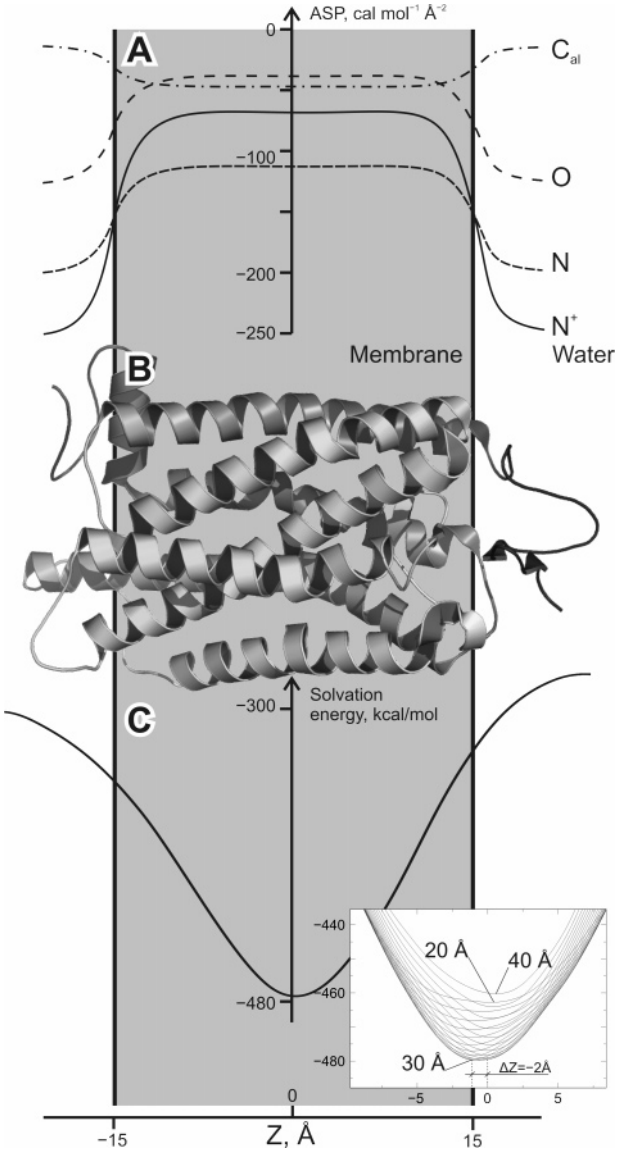

Figure 1. The scheme illustrating optimal positioning of a protein with respect to the membrane (by the example of rhodopsin). A. Implicit membrane model presented as a hydrophobic slab (colored in gray) described by the atomic solvation parameters (ASP). Only four types of ASPs are shown for clarity (full description of the model is given elsewhere).[49,50] B. The structure of bovine visual rhodopsin (PDB code 1U19) aligned along the membrane normal and corresponding to the minimal energy of protein-membrane interaction (E_{solv}). C. Dependence of E_{solv} on the shift of the rhodopsin model from its optimal position (B) in the slab along the axis *Z*. *Inset:* The minima of E_{solv} obtained for various values of the membrane thickness. $Z = 0$ indicates the optimal position of the center of mass of the protein model.

Obviously, the approach cannot be used to study the atomic-scale details of protein-membrane interactions due to the implicit nature of the slab. Also, rigidity of the protein structure and a number of other factors were not taken into account. On the other hand, the approach allows efficient determination of the boundaries of TM helices, and, hence, the main objective of this stage of the work was realized. Furthermore, the results obtained agree well with the data reported in the literature, though different approaches and parameters were used to find optimal protein orientations with respect to the membrane. Among these databases are OPM ("Orientation of Proteins in Membranes")[51] and PDB_TM,[52] along with the TMDET[53] and IMPALA[54] methods.

To understand how appropriate the resulting data set of residues in TM helical domains was, a number of residues' parameters were calculated and compared with those reported in the literature. Among these are the distributions of different types of residues (Figure S1) along the membrane normal and residues' preferences to either buried or exposed positions (Figure S2) as well as the profiles of mean hydrophobicity along the membrane normal obtained for lipid-exposed and buried residues (Figure S3). As seen in Figure S1, the probabilities to find different types of residues with a particular coordinate Z agree well with the data reported previously. Such distributions can be unimodal—with a maximum near the membrane center (nonpolar nonaromatic residues, Ser and Thr) or bimodal—with maxima in the interfacial region (charged, polar or aromatic residues, Met and Pro). Some of the distributions are nonsymmetric: e.g., the cytoplasmic side is rich in Arg, Asn, Asp, Glu, Lys, and His, while Trp more frequently occurs at the extracellular side. Also, the inner membrane interfaces are enriched with positively charged residues—this coincides with the well-known "positive-inside" rule.[30,55] Bimodality is much less pronounced for the distribution of Phe residues than for its more polar aromatic analogues—Trp, Tyr, and His. Similar observations were made by Ulmschneider et al., who derived implicit membrane-mimic potentials from such distributions.[56] These and other[30] findings correlate well with the data on the "translocation code" defining the correct folding of MPs and extracellular secretion of proteins.[19,57,58]

The calculated residues' preferences for buried positions versus the membrane-accessible ones are in a slightly worse agreement with the data reported previously.[59] For example, Trp, Phe, Arg, and Ile residues, like in other studies, prefer exposed positions, whereas residues Gly, Ser, Thr, Pro, Ala, and Cys are mostly found in the buried ones (Figure S2). On the other hand, our results do not confirm a conclusion made by Eyre et al. that histidines strongly prefer buried positions, while the opposite is true for glutamines. These discrepancies may arise from the differences in the data sets as well as in the methods used to characterize the residues' environments.

Finally, TM segments selected for the data sets were characterized using the well-known residues' hydrophobicity scales by Eisenberg[23] and by Wimley and White.[60] In both cases the results were very similar (data not presented). The average hydrophobicity profiles calculated along the membrane normal demonstrated that the "central" parts of TM regions (that span nonpolar acyl chains of lipids) are considerably more hydrophobic than the peripheral ones. This coincides with the main principle of popular algorithms used to predict TM helices based on the amino acid sequence. Furthermore, lipid-exposed residues are significantly more hydrophobic than the buried ones (Figure S3), which corroborates with the observations reported earlier.[23,30,56]

To summarize, statistical analysis of the residues' environments and their locations in membranes reveals that the results obtained for both training and test sets of TM helical structures of MPs are consistent with the data reported by other researchers. The data sets were further employed to develop a scoring function destined to quantitative estimation of the quality of helix packing.

2.2. Packing Quality of Helices in TM Domains of Proteins. Resulting distributions of residues in the training data set over the values of their polarity properties (F_p^1 and F_{np}^1, see Methods) are shown in Figure 2. The parameters F_p^1 and F_{np}^1 represent the fractions of the side-chain area forming polar and nonpolar contacts with other TM helices in MPs with known structure. In other words, coordinates of each point on a related diagram characterize both the polarity of the protein and/or membrane environment of a given residue in the data base. Proximity to the coordinate's origin indicates the membrane exposure, whereas location near the line defined by the equation $F_p^1 = 1 - F_{np}^1$ means that the residue is deeply buried in the protein interior. As seen in Figure 2, distribution may be drastically different for different types of residues, like, e.g., for isoleucines and arginines. Arginines are concentrated near the membrane boundaries, while almost all isoleucines are found in the hydrophobic membrane core. Moreover, isoleucines mainly occur in nonpolar surroundings (see definitions of environment classes in Figure 3), while arginines prefer more polar positions. Arginine residues situated in the central membrane regions are usually buried in order to avoid unfavorable contacts with hydrophobic acyl chains of lipids. Finally, interfacial isoleucine residues have a higher affinity to polar environments than the "central" ones.

Such differences in preferential location of residues in TM helices make it possible characterization of residues' environments in a quantitative manner and definition of environment classes. To identify these, a pictorial scheme was proposed (Figure 3). Class 1 corresponds to membrane-exposed residues, classes 2 and 3 include partially buried residues, while classes 4 and 5 consist of completely buried ones. Furthermore, classes 2 and 4 are populated with residues found in nonpolar environments, and classes 3 and 5 are composed of residues mainly forming interhelical polar contacts (Figure 3). As described in "Methods", boundaries of the classes are defined by three numerical parameters, namely a specifying the borderline value between the partially and completely buried classes, b, the borderline value between the partially buried and exposed classes, and the angle α delimiting the polar and nonpolar classes. The "exposed" class is not further subdivided. The final values of these parameters optimized for the set of 26 MPs (united "training" and "test" set) are $a = 0.70$, $b = 0.05$, and $\mathrm{tg}\alpha = 2$, resulting in a total database score of 475.11. Corresponding individual scores for all residue types in particular environment classes are given in Table 1. It is worth noting that the score values presented in the table were derived only from analysis of residues' distributions between classes in the set of experimental structures. No "weighting factors" were used,

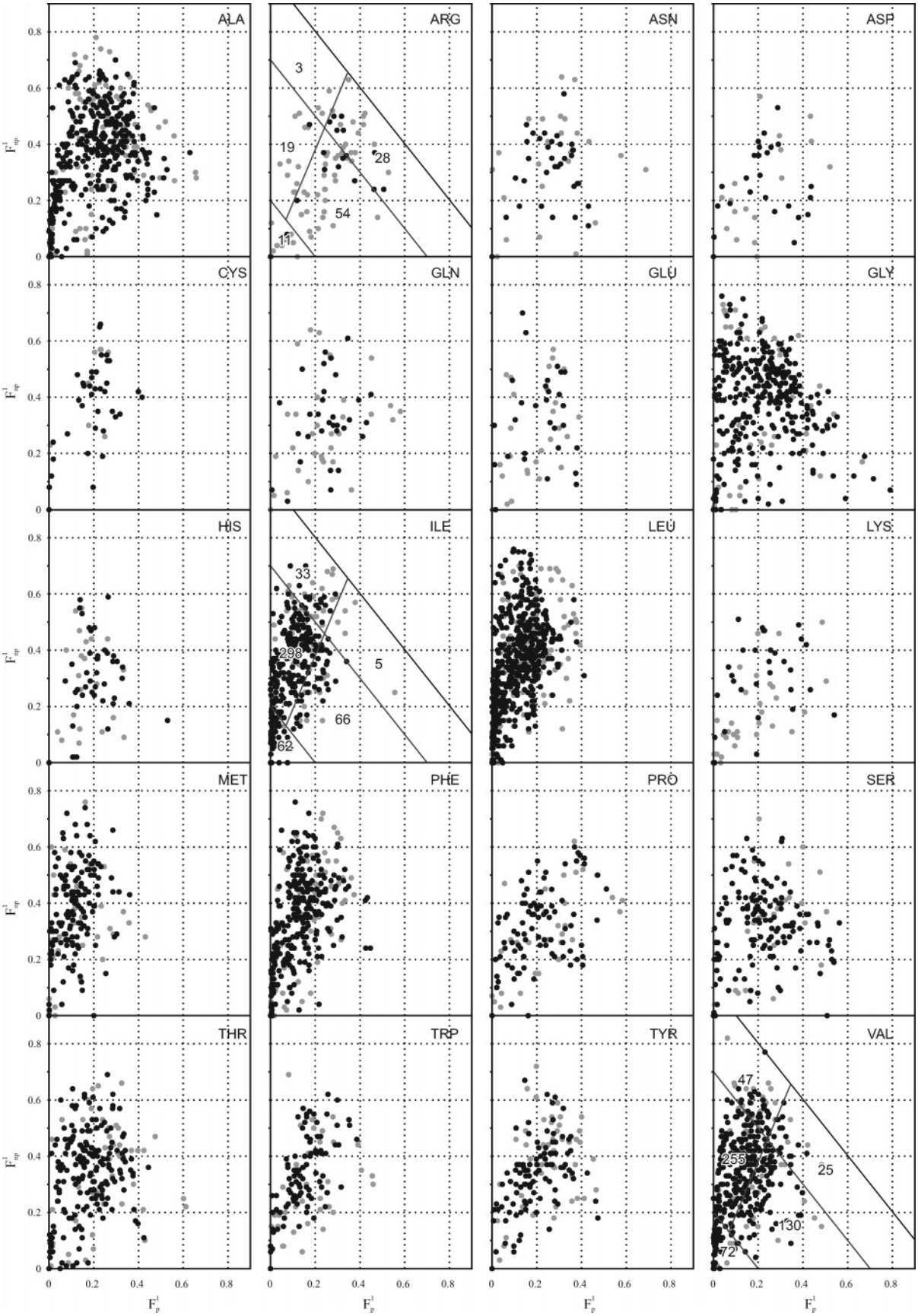

Figure 2. Polarity properties of residues in TM α-helical domains of proteins from the training set. Distributions of residues over the values of F_p^1 and F_{np}^1 − fractions of the surface area covered by polar and nonpolar atoms of neighboring helices, respectively. Each data point corresponds to a single residue of a given type in the database. Gray and black dots denote residues on the interface ($|Z| \geq 15$ Å) and in the membrane core ($|Z| < 15$ Å), respectively. For three residue types (namely − arginine, isoleucine, and valine) the environment classes scheme is given (it is introduced in Figure 4). Values at every environment class correspond to residue count in these classes.

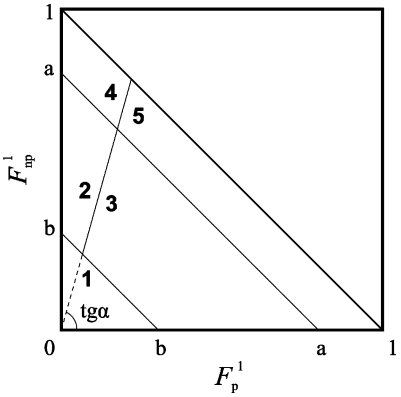

Figure 3. Definition of environment classes for residues in TM helix bundles. Depending on the values of F_p^1 and F_{np}^1, each residue is attributed to one of five environment classes. Class 1 corresponds to residues preferentially exposed to the membrane; classes 2, 3 and 4, 5 are composed of residues partly and completely buried inside the bundle, respectively. Classes 2 and 4 denote nonpolar, while classes 3 and 5 − polar environment. Other details are the same as in the legend to Figure 2.

Table 1. Residues Preferences for Environment Classes[a]

	classes[b]				
residues	1	2	3	4	5
ALA	0.09	−0.33	−0.10	0.29	0.64
ARG	−0.07	−0.96	0.54	−1.20	0.82
ASN	0.21	−1.05	0.56	−1.13	0.60
ASP	−0.04	−1.12	0.86	−1.67	−0.09
CYS	−0.42	−0.23	0.25	0.71	−0.41
GLN	−0.22	−1.02	0.66	−0.18	0.38
GLU	−0.56	−0.77	0.57	−0.34	0.63
GLY	−0.16	−0.32	−0.15	0.36	0.79
HIS	−0.72	−0.55	0.90	−0.97	−1.87
ILE	0.26	0.40	−0.65	−0.19	−2.30
LEU	0.10	0.34	−0.53	0.01	−0.97
LYS	0.48	−0.78	0.64	−2.02	−0.15
MET	−0.21	0.44	−0.74	0.19	−1.27
PHE	−0.20	0.25	−0.28	0.26	−0.64
PRO	−0.36	−0.38	0.50	−0.87	0.35
SER	−0.25	−0.51	0.54	−0.37	0.23
THR	−0.19	−0.19	0.22	−0.11	0.30
TRP	−0.31	0.11	0.10	−0.11	−0.32
TYR	−0.58	−0.40	0.44	−0.21	0.46
VAL	0.28	0.11	−0.10	0.03	−0.82

[a] The values are given for the "final" training data set. [b] The scores were obtained for the optimal borders between the classes $a = 0.70$, $b = 0.05$, and $\mathrm{tg}\alpha = 2$.

and the only prerequisite was the choice of the "basic" parameters (F_p^1 and F_{np}^1) and the classes scheme.

Positive values indicate that residue falls into a favorable class, which is typical for it in the spatial structures of the training set and vice versa (see Methods). Isoleucines, for instance, tend to reside in environments of the membrane-exposed and both buried nonpolar classes; they are rarely found in either of the polar classes. Conversely, arginines prefer to be in environments corresponding to the polar classes (Figure 2). One can quantitatively verify this observation: isoleucine falls to nonpolar classes (2 or 4) in (298 + 33) = 331 cases of 464 (71%) and to polar ones—in (66 + 5) = 71 cases of 464 (15%) (where 464 is the total number of isoleucines in the database). For arginines these numbers are (19 + 3)/115 = 19% and (54 + 28)/115 = 71%, respectively. These data permit clear and vivid delimitation of these residues by quantitative preferences to different environment classes, thus confirming the well-known regularities following from the physicochemical nature of amino acid residues and the membrane. Furthermore, the resulting 2D distributions obtained for all residue types were compared one with another using the Kolmogorov-Smirnov (KS) statistics.[61] To check, whether two 2D distributions differ or not, the significance level (*probe*) for the KS test was calculated. The lower the *probe* value is, the greater the difference is between the two distributions (identical distributions yield a value of 1.0). For example, the calculated *probe* values for pairs of residues isoleucine−arginine and isoleucine−leucine are 2.97×10^{-14} and 5.00×10^{-2}, respectively. This is considered as an evidence that the distributions for both pairs of residues are different, although the environmental characteristics of isoleucine and leucine are much more similar than those for arginines (this is quite expectable!).

If residues with poor scores form a compact spatial cluster in MP model, this may point out a problematic region that requires special attention while optimizing the model. However, similar distribution of residues with low scores may indicate some functional or oligomerization sites of the protein, as well. For instance, in the crystallographic structure of rhodopsin the retinal-binding pocket is lined with "poorly packed" residues (including Lys296 that covalently binds chromophore), according to the membrane score criterion. The possibility of distinguishing between these two cases is yet to be established.

The proposed method of the "membrane score" is conceptually similar to the 3D environment profiles technique,[47] although there are some important differences. First, unlike the 3D profiles algorithm developed for globular proteins, the "membrane score" is optimized for TM α-helical proteins. Second, the method only takes into account one secondary structure type (α-helix), since random coil conformation is highly unfavorable in a membrane milieu. We do not consider here *β*-sheet TM domains, since it would require a special study and is, therefore, beyond the scope of the present work. Finally, definitions of the environment classes are drastically different from those proposed for water-soluble proteins by Bowie et al.[47]

The capacity of the residue scores to assess TM helix packing quality as well as to differentiate between correct and misfolded 3D models of rhodopsin is described in the next section. A number of additional tests to evaluate practical applicability of the method are given in the second article of this series.

2.3. Validation of the Method. In this paper we present only a part of the validation tests for the "membrane score" method. These are the approbation on different data sets to confirm self-consistency and stability of the method and application to a set of 13 theoretical models of visual rhodopsin (most of them contain modeling errors) that were collected from literature or specially built for this purpose using conventional modeling tools. The latter test is designed to check the method's ability to detect the "nativelike" structures.

2.3.1. Validation Using Different Data Sets. Assuming that proteins from the training and test sets (see Methods) are organized in a similar manner with respect to the residues' environment classes, it would be reasonable to expect that proteins of similar sequence length get close scores. The

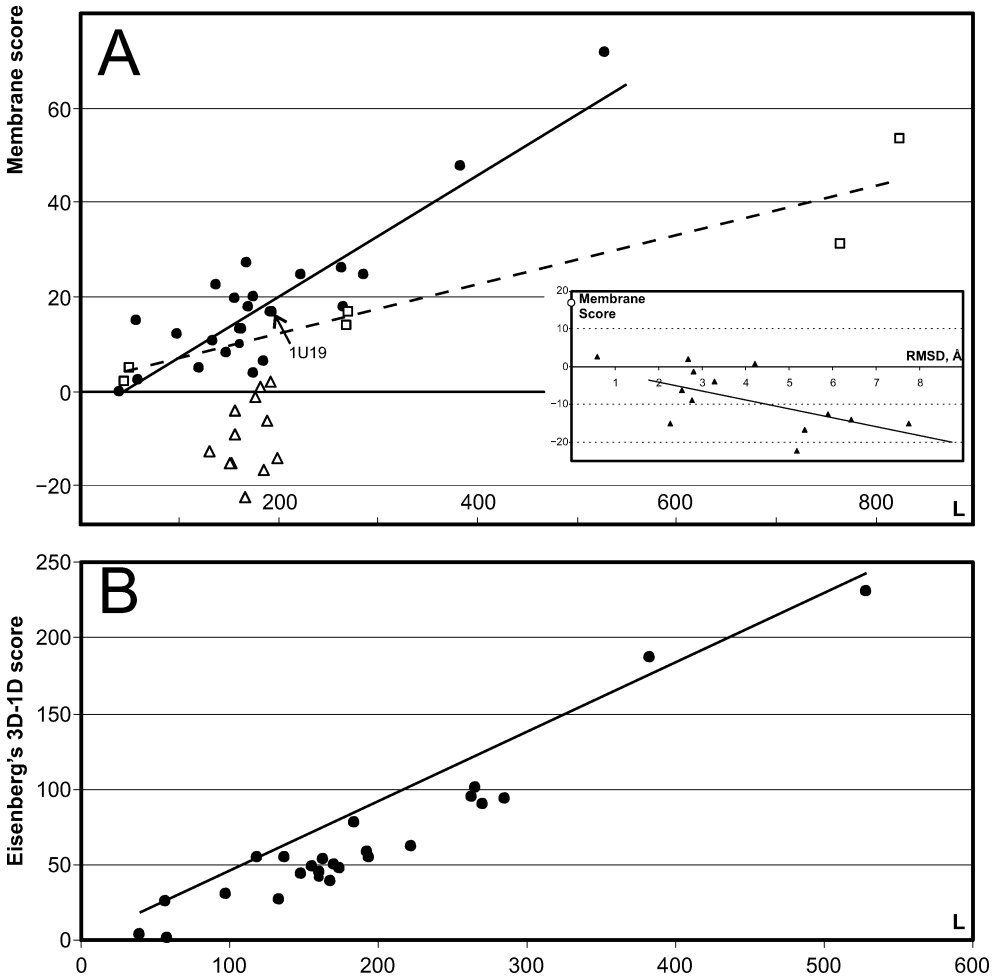

Figure 4. Membrane scores (S^{mem}) and 3D-1D profile scores (S^{3D-1D})[47] for membrane proteins, as a function of sequence length. A. S^{mem} values vs TM domain length for the training set (black circles), photosynthetic proteins (white squares), and computer models of rhodopsin (white triangles). The solid and dashed lines represent the least-squares fits obtained for the training set and photosynthetic proteins, respectively. *Inset:* Dependence of the values S^{mem} on the root-mean-square deviation from the crystal structure (empty circle) for theoretical models of rhodopsin. The regression trend-line is given. B. The values S^{3D-1D} for the final set of membrane proteins. The solid line shows the "ideal" score value for a protein with a given length (L):[47] $S^{3D-1D}_{ideal} = \exp(-0.83 + 1.008 \times \ln(L))$.

training set proteins reveal a linear dependency of the membrane score (S^{mem}) on TM domain sequence length (L), described by the equation $S^{mem} = 0.13 \times L - 5.21$ with the regression quality parameter $r^2 = 0.81$. Test set structures fit well the regression line for the training set, providing evidence that proteins from both sets are organized in a similar way and that the training set is sufficiently representative (data not shown). If one (e.g., rhodopsin) or several structures were excluded from the training set (sets S2 and S3, respectively), the results did not change significantly, thus confirming the stability and self-consistency of the method. Furthermore, the calculated parameters of the class borderline values for the corresponding training sets did not change (data not shown). The aforementioned results demonstrate high robustness of the method and allow combination of the data sets into the "final" data base for subsequent application. In the latter case the scores and border parameters listed in Table 1 were employed.

The corresponding plot $S^{mem}(L)$ for the "final" data set is shown in Figure 4a. The least-squares regression yields a linear function described by the equation $S^{mem}(L) = 0.13 \times L - 5.54$ ($r^2 = 0.76$). Analogous plots for a set of photosynthetic proteins (PhPs) and theoretical models of rhodopsin are also given in the same figure. It is seen that the least-squares lines obtained for the "final" data set and PhPs are different. For the latter ones, the membrane score values are systematically lower than those for the "final" set of MPs. This can be illustrated by the proteins with the largest TM domains, photosystems I and II (PDB codes 1JB0 and 2AXT, respectively). Such a discrepancy is not caused by the absence of these proteins in the training set: even if they are taken into account by parametrization, they still fit to a separate line ($S^{mem}(L) = 0.038 \times L + 3.010$, $r^2 = 0.96$), lying below the regression line obtained for the training set (data not shown). One possible explanation of this fact is a high abundance of such TM domains in photosynthetic pigments. This considerably impedes interhelical interactions, substituting them by protein-pigment contacts.

Figure 4b shows how different the packing quality results for MPs are if assessed using the Eisenberg's 3D-1D method and the scoring function adapted for TM helical bundles (scores are derived for the "final" data set). The values of the 3D-1D score (S_{3D-1D}) deviate considerably from the corresponding optimal values envisaged for a protein of a given length (S_{corr}).[62] Thus, the mean value and standard deviation for the ratio S_{3D-1D}/S_{corr} are 0.68 and 0.23, respectively. At the same time, a spatial model is considered as misfolded if this ratio is less than 0.45.[62] Furthermore,

Table 2. Structural Data Sets Used in This Work

no.	PDB ID	resolution, Å	name and organism	no. of helices	no. of residues
			Training Set		
1	1C3W	1.55	Bacteriorhodopsin (BR)/*Halobacterium salinarium*	7	160
2	1E12	1.80	Halorhodopsin (HR)/*Halobacterium salinarium*	7	170
3	1H68	2.1	sensory rhodopsin II (SRII)/*Natronobacterium pharaonis*	7	163
4	1U19	2.2	visual rhodopsin/*Bos Taurus*	7	194
5	1K4C	2.0	KcsA H+ gated potassium channel/*Streptomyces lividans*	2	58
6	1MSL	3.5	MscL mechanosensitive channel/*Mycobacterium tuberculosis*	2	40
7	1KPL	3.0	H+/Cl- exchange transporter/*Salmonella typhimurium*	28	529
8	1FX8	2.2	GlpF glycerol facilitator channel/*E. coli*	8	174
9	1XQE	1.8	AmtB ammonia channel/*E. coli*	11	263
10	1IWG	3.5	AcrB bacterial multidrug efflux transporter/*E. coli*	12	266
11	1PW4	3.3	GlpT glycerol-3-phosphate transporter/*E. coli*	12	270
12	1L7V	3.2	BtuCD vitamin B12 transporter/*E. coli*	9	184
13	1SU4	2.6	calcium ATPase/rabbit sarcoplasmic reticulum	10	192
14	2BL2	2.1	rotor of V-type Na+-ATPase/*Enterococcus hirae*	4	119
15	1YCE	2.4	rotor of F-type Na+-ATPase/*Ilyobacter tartaricus*	2	57
16	1QLA	2.2	fumarate reductase complex/*Wolinella succinogenes*	5	137
17	1NEK	2.6	succinate dehydrogenase (complex II)/*E. coli*	6	168
18	1EHK	2.8	cytochrome *c* oxidase, ba3/*T. thermophilus*	15	383
19	1EZV	2.4	cytochrome bc1 complexes/*S. cerevisiae*	12	310
20	1KQF	2.3	formate dehydrogenase/*E. coli*	4	97
21	1Q16	1.6	nitrate reductase A/*E. coli*	5	133
			total:	175	4067
			Test Set		
22	1XIO	2.0	sensory rhodopsin II (SR)/*Anabaena Nostoc*	7	161
23	1J4N	2.2	AQP1 aquaporin water channel/bovine red blood cell	8	148
24	1RC2	2.5	AQPZ aquaporin water channel/*E. coli*	8	156
25	1BCC	3.16	cytochrome bc1 complex/chicken heart mitochondria	12	285
26	1OKC	2.2	mitochondrial ADP/ATP carrier/bovine heart mitochondria	6	174
			total:	42	924
			Photosynthetic Proteins		
27	1NKZ	2.0	light-harvesting complex/*Rhodopseudomonas acidophila*	2 (6)	44
28	1LGH	2.4	light-harvesting complex/*Rhodospirillum molischianum*	2 (4)	50
29	1JB0	2.5	photosystem I/*Thermosynechococcus elongates*	30	766
30	1PRC	2.3	photosynthetic reaction center/*Rhodopseudomonas virdis*	11	271
31	1OGV	2.35	photosynthetic reaction center/*Rhodobacter sphaeroides*	11	268
32	2AXT	3.0	photosystem II/*Thermosynechococcus elongates*	33	827
			total:	89	2226

the distributions $S_{corr}(L)$ and $S_{3D-1D}(L)$ are different at $p = 3 \times 10^{-8}$ significance level (as calculated by *t*-test). Therefore, the structural and hydrophobic organization is different in globular and membrane proteins. Moreover, the standard 3D-1D profiles technique is inappropriate for TM α-helix bundles. This conclusion is illustrated below by the example of visual rhodopsin and several of its computer models. As discussed above, the main reason for this is the drastically different environmental characteristics of residues in MPs and globular proteins.

2.3.2. Validation Using Incorrect Computer Models of TM Domains. One of the most important requirements to any method destined to assess the protein packing quality is its ability to distinguish between well-folded ("correct") and misfolded 3D models. For instance, as applied to the present work, the method should attribute the highest values of S^{mem} to crystallographic structures. On the contrary, the misfolded structures, like computer models, containing certain errors and therefore inconsistent with the main principles of protein organization, should be unambiguously eliminated by such an algorithm due to their poor score. Furthermore, if the correct structure is unknown, such techniques could be used to select the "best" models from a list of alternatives built, e.g., via homology modeling. Both tasks are very important for the structure-based drug design applications. Therefore, the "membrane score" approach presented in this work should be rigorously tested.

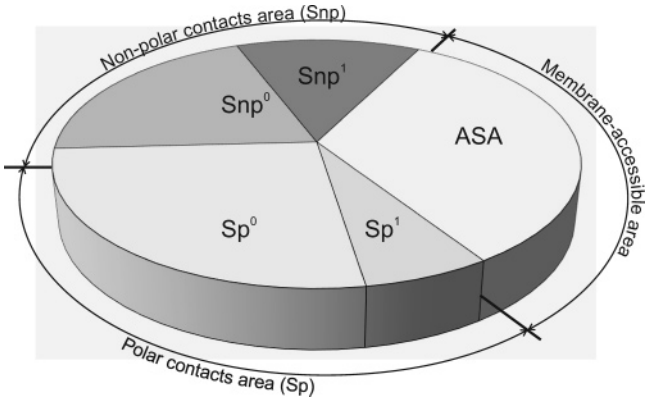

Figure 5. The scheme of residues' surface area division. The total residue surface area is subdivided into the following parts: ASA is the surface area accessible to membrane; S_p and S_{np} are the areas covered by polar and nonpolar protein atoms, respectively. The two latter are further subdivided into S_p^0 and S_{np}^0 − the contact areas with residues of the same TM segment, and S_p^1 and S_{np}^1 − the contact areas with other helices.

To evaluate the method's ability to distinguish a deliberately "correct" structure among misfolded ones, we chose bovine visual rhodopsin and a set of its 13 computer models, mostly with modeling errors (see Methods). Rhodopsin was

Table 3. Computer Models of Bovine Rhodopsin Used in This Work

ID	source (ref)[a]	description/probable modeling errors	rmsd,[b] Å	S^{mem}
1U19/1F88	[63]/[68]	reference structure	0.00	16.81
OPSD_BOVIN_br	www.gpcr.org[69]	template: bacteriorhodopsin (bRh). Incorrect amino acid alignment in 1 region.	5.17	−22.24
OPSD_BOVIN_br-modeler	homology model, built using MODELLER[70]	template: bRh[71] (PDB ID: 1C3W). Handmade alignment.	5.9	−12.56
OPSD_BOVIN_br1-modeler	homology model, built using MODELLER[70]	template: bRh (PDB ID: 1C3W). Alignment is taken from DBAli.[72]	7.75	−15.18
opsd_bovin_herzyk	[73]	template: 2D electron density maps of frog rhodopsin[74]	2.80	−1.17
OPSD_BOVIN_baldwin	www.gpcr.org[58]	template: C_α template for GPCR modeling[75]	2.76	−8.95
OPSD_BOVIN_rh	www.gpcr.org[58]	template: bovine rhodopsin. Certain alignment errors in 4, TM5 regions.	3.64	−15.11
opsd_bovin_shieh	[76]	templates: C_α template for GPCR modeling[75] and 2D electron density maps of frog rhodopsin[74]	4.22	0.90
rhod_human_donelly	*Human* rhodopsin model[77]	template: 2D electron density maps of frog rhodopsin[74]	5.35	−16.82
opsd_bovin_lin	[78]	templates: C_α template for GPCR modeling[75] and 2D electron density maps of frog rhodopsin[74]	6.42	−14.05
Pdb1boj	[79]	theoretical model; Plenty of distance restraints are taken into account. Photoactivated state.	2.67	1.98
Pdb1bok	[79]	as above, "dark-adapted" state	2.54	−6.18
Pdb1ov0	[80]	theoretical model of meta-2 state of rhodopsin	3.28	−3.90
opsd_bovin_TASSER	built from TASSER model of human rhodopsin[15]	see text	0.58	2.61

[a] Few models describe entire structures (sometimes with bound retinal); others correspond only to TM domain. [b] Root-mean-square deviations from the crystal structure are calculated for C_α atoms considering a "consensus" TM domain of 156 residues.

chosen, as it is the only GPCR-member with the experimentally determined 3D structure,[63] and many modeling attempts have been made before the crystal structure became available.[73,75−80] It is impossible to collect from public domain a substantial number of models of any other membrane protein. Application of the standard 3D-1D method to the set of rhodopsin models failed to delineate at a statistically significant level the crystal structure among the models built in silico. The "correct" structure was either not the top-scoring one or the gap between this structure and the models was insignificant and irregular (data not shown). Attempts were also made to "adopt" the 3D-1D profiles method to TM protein domains. For instance, the environment classes for membrane-exposed residues were changed to take into account their hydrophobic surrounding. In addition, only buried residues were considered in the calculations of 3D-1D scores. The improvements of the results were only minor (data not shown), and, regardless of the modifications made, it was still impossible to efficiently recognize the X-ray structure among the erroneous models. This clearly shows that further parametrization of the residues' propensities to be in TM protein domains is required.

Indeed, the "membrane score" method performs well and permits statistically significant delineation of the crystallographic model from all the computer models under consideration (Figure 4a). It can be seen that the majority of the models receive negative membrane scores, and these are mainly "automatically" built models or models based on incorrect sequence alignment. Interestingly, the best-scoring models from this set revealing positive values of S_{ij}^{mem} are known to be carefully optimized by their authors. These models demonstrate post factum the lowest RMSDs from the crystal structure (see Methods): there is a prominent anticorrelation (with a coefficient −0.6) between the membrane score and rmsd value (Figure 4a, inset).

2.4. Limitations of the Approach. It is worth noting that photosynthetic proteins, whose TM domains are enriched with various pigments and cofactors, conceivably follow the packing rules distinct from those observed for most other MPs. The results presented here demonstrate that such proteins score systematically lower than the models of similar length from the training and test sets. This is most probably caused by protein−pigment interactions, in many respects substituting interhelical protein−protein contacts.

Furthermore, application of the "membrane score" method to assess the packing quality of NMR- or cryo-EM derived MPs yielded seriously decreased scores, as compared to X-ray structures of the same length (data not shown). This may suggest that "resolution" of the former two methods is not high enough to place the residues' side chains correctly. Consequently, the "membrane score" approach is not applicable to comparative studies of structures determined by different methods and may not be able to compare several NMR or EM-structures.

CONCLUSIONS

We presented a novel method (the "membrane score") of assessing the packing quality of TM α-helical domains in integral MPs. It allows efficient differentiation between the crystal structure of a protein and its incorrect models. It was shown that rmsd between a given model and the X-ray structure anticorrelates with the membrane score, a property that may be useful in optimization of computer models built for MPs.

Prominent correlation between the length of a protein sequence and the corresponding score values obtained for both training and test sets of X-ray structures of MPs allows quantitative estimation of the packing quality for a particular protein model. This will hopefully be used in development of an automated criterion for delineation of "correct" MPs' structures. Although solution of this problem lies beyond the scope of the present work, some preliminary criteria could already be formulated. Thus, properly folded ("correct") structures of MPs should at least have a positive score, and

this value should be roughly proportional to the sequence length, according to the law defined for the training and test sets. However, the exact threshold value discriminating between "correct" and erroneous models is yet to be defined.

At the moment, we still cannot offer a universal strategy that incorporates the membrane scoring methodology in MPs models' optimization pipeline. However, some interesting results were obtained in this work that raise hopes that these strategies will emerge and will be successfully employed in production of realistic models of membrane proteins—in the first hand, GPCR ones.

In the second article of this series we present a more versatile evaluation of practical applicability of the "membrane score" method. It engages ensembles of "decoy" conformations representing either alternative TM α-helices packing or homology models with errors in alignment and illustrates how our method delimits decoy and "native" conformations. Also, some issues related to refinement of theoretical models are addressed.

3. METHODS

3.1. Selection of MPs for the Data Sets. The data sets of MPs used in this study were constructed based on the list of 191 structures published on the Web site "Membrane proteins of known 3D structure" (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html). All nonhelical proteins, low-resolution structures (with resolution > 3.5 Å), NMR-derived models, and homologous proteins were omitted. In each family of closely related proteins only one member solved to the highest resolution was selected for the final set. This procedure yielded 21 X-ray structures of nonhomologous α-helical TM proteins—the training set. The test set comprised 5 proteins, related to some structures from the training set. Photosynthetic proteins comprised a separate group, because their TM domains actively interact with photosynthetic pigments and, therefore, interhelical interactions are replaced by protein-pigment contacts. The "final" set (S1) combines the training and test sets. Test sets 2 (S2) and 3 (S3) were obtained by taking out the rhodopsin structure (entry 1U19 in the Protein Data Bank)[64] and the models 1EHK, 1XQE, 1NEK, and 1KQF from the training set, respectively. Some characteristics of these sets are given in Table 2.

3.2. Spatial Arrangement of Proteins in Membranes. Atomic coordinates of all the selected structures of MPs (monomers only) were taken from the Protein Data Bank.[64] To make the models suitable for further work, their orientation with respect to the membrane should be delineated. This is necessary for determination of the boundaries of the TM helices and membrane exposure of individual residues. The following automated computational procedure was proposed. (i) Each model was arranged in such a way that the principal moment of inertia of its TM domain was aligned along the *Z*-axis corresponding to the membrane normal. Directionality of the axis was chosen as follows: smaller values of *Z* corresponded to the cytoplasmic side of the plasma membrane or the matrix side of inner mitochondrial membrane. (ii) The model was then "scanned" along the *Z*-axis with a "hydrophobic slab" mimicking lipid bilayer, and the energy of "protein-membrane" interaction (E_{solv}) was calculated at each step. The slab was represented by two parallel planes XY separated with a distance *D* (the "membrane thickness") and described by an effective potential based on the atomic solvation parameters (ASP) formalism. E_{solv} was calculated according to the formula

$$E_{solv} = \sum_{i=1}^{N} \Delta\sigma_i \mathrm{ASA}_i \qquad (1)$$

where ASA_i and $\Delta\sigma_i$ are accessible surface area and ASP of atom *i*, respectively, and *N* is the number of atoms. Depending on the *Z*-coordinate of atom *i*, the values of ASPs were taken to approximate cyclohexane and water inside ($|Z| < D/2$) and outside the slab, respectively.[65] Other details of the implicit membrane model were described earlier.[50] (iii) The optimal position of a protein with respect to the slab was defined as corresponding to the minimum of E_{solv}, and the *Z*-coordinate of the center of the slab was assigned a value $Z = 0$ Å. Atomic coordinates were then recalculated in this coordinate system.

In order to determine the optimal "membrane" thickness for each protein, the values of *D* were varied in the range $20 \div 40$ Å with 2 Å increment. A prominent minimum of E_{solv} was found in each case, indicating the most favorable membrane/protein position and optimal membrane thickness (see an example in Figure 1). The optimal parameters of the slab, along with the identified boundaries of TM regions for proteins in the data sets, are given in the Supporting Information (Table S1).

3.3. Environmental Characteristics of Residues in TM Helices. Only residues in the delineated TM regions were further used to assess their environmental characteristics in different data sets. To prevent boundary effects, the calculations were done taking into account the atoms located within 5 Å along the *Z*-axis, outside the terminal residues of a given TM segment. In total, the training set comprised 175 TM α-helices containing 4067 residues (see Table 2). Residues in α-helical conformation were identified using the DSSP software.[66] It is worth noting that no residues in regions of local destabilization of α-helices have been omitted from the training set. Only lengthy regions of a disordered protein (such as a half-membrane spanning part in a glycerol channel structure) were excluded from the database.

The environment of each residue in the 3D model was characterized by two parameters—the fractions of its side chain surface area that are covered by polar (nitrogen or oxygen) and nonpolar atoms of neighboring helices in the TM domain, F_p^1 and F_{np}^1, respectively

$$F_p^{\,1} = \frac{S_p - S_p^{\,0}}{S^0}$$

$$F_{np}^{\,1} = \frac{S_{np} - S_{np}^{\,0}}{S^0} \qquad (2)$$

where S_p and S_{np} are polar and nonpolar side-chain areas of the residue, respectively (Figure 5). S_p^0 and S_{np}^0 are the corresponding values calculated for the isolated TM α-helix containing the residue under consideration. S^0 is the solvent-accessible area of the side chain in a Gly-X-Gly tripeptide.

The values of S^0 were taken from ref 67. As seen in Figure 5, the values of polar/nonpolar areas unambiguously define the lipid exposure of a given residue.

3.4. Calculation of the "Membrane Score". Depending on its values of F_p^1 and F_{np}^1, each residue in the training data set was assigned to one of five environment classes. The latter were defined by three parameters—the values *a*, *b*, and tgα (Figure 2). The compatibility of each of the 20 amino acids with each of five environment classes was expressed in terms of the membrane score values calculated thus

$$S_{ij}^{\text{mem}} = \ln(P_{ij}/P_j) \tag{3}$$

Here P_{ij} is the probability of finding residue of type *i* in environment *j*. P_j is the probability of finding any residue in environment *j*. The borderline values for the environment classes (Figure 2) were adjusted iteratively, the matrix S_{ij}^{mem} being recalculated each time, so as to maximize the total membrane score for the whole training set

$$S_{\text{total}}^{\text{mem}} = \sum_{ij} N_{ij} S_{ij}^{\text{mem}} \tag{4}$$

where N_{ij} is the number of residues of type *i* in environment *j*. If there were no residues *i* in environment *j*, N_{ij} was set to one in (3) to avoid critical points of the logarithmic function. In this case the term $N_{ij}S_{ij}^{\text{mem}}$ was set to zero and, therefore, was not accounted for in the sum (4). During the optimization procedure, the parameters *a*, *b*, and tgα were varied in the ranges 0.10−0.95, 0.05−*a*, and 0.0−6.0, respectively. The optimal borderline values for the classes and the corresponding scores S_{ij}^{mem} are given in Table 1.

3.5. Computer Models of TM Domain of Visual Rhodopsin. The data set of erroneous and misfolded 3D structures of bovine visual rhodopsin contained 13 models. Most of them have been predicted before the appearance of the crystal structure.[63] Some of these models are available in the public domain, whereas some others were built in this work using popular homology modeling tools. Characteristics of the models are given in Table 3. We will place a little emphasis only on the last model in the table. It is built with the MODELLER software[70] by homology with another model of *human* visual rhodopsin, available at the Web page, dedicated to the modeling of all human GPCRs with TASSER.[15] It has been shown that TASSER is able to reconstruct bovine rhodopsin structure to as low as 2.1 Å C_α-rmsd from the crystal structure, starting from three low-homology templates (<30% identity).[15] However, during the modeling of all human GPCRs the condition of using only low-homology templates was not implied, so TASSER used coordinates of C_α atoms from bovine rhodopsin crystal structure for this model. Resulting RMSDs are ∼0.6 Å over C_α atoms and ∼3.6 Å over all heavy atoms.

ACKNOWLEDGMENT

This work was supported by the Russian Foundation for Basic Research (grants 05-04-49283-a, 06-04-49194-a), by the Programme RAS MCB, and by the Russian Federation Federal Agency for Science and Innovations (The State contract 02.467.11.3003 of 20.04.2005, grant SS-4728.2006.4 "The leading scientific schools"). We thank Dr. Pavel E. Volynsky for his assistance with the implicit membrane model.

Supporting Information Available: Residues' distributions along the membrane normal (Figure S1), their preferences for buried or exposed positions (Figure S2), distribution of average hydrophobicity along the membrane normal (Figure S3), and detailed description of the "training set" of MPs (Table S1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Klabunde, T.; Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem* **2002**, *10*, 928−944.
(2) Torres, J.; Stevens, T. J.; Samso, M. Membrane proteins: the "Wild West'" of structural biology. *Trends Biochem. Sci.* **2003**, *28*, 137−144.
(3) Fleishman, S. J.; Unger, V. M.; Ben-Tal, N. Transmembrane protein structures without X-rays. *Trends Biochem. Sci.* **2006**, *31*, 106−113.
(4) Wallin, E.; von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **1998**, *7*, 1029−1038.
(5) Cowan, S. W.; Rosenbusch, J. P. Folding pattern diversity of integral membrane proteins. *Science* **1994**, *264*, 914−916.
(6) Ubarretxena-Belandia, I.; Engelman, D. M. Helical membrane proteins: diversity of functions in the context of simple architecture. *Curr. Opin. Struct. Biol.* **2001**, *11*, 370−376.
(7) Arkin, I. T. Structural aspects of oligomerization taking place between the transmembrane alpha-helices of bitopic membrane proteins. *Biochim. Biophys. Acta* **2002**, *1565*, 347−363.
(8) Vereshaga, Y. A.; Volynsky, P. E.; Nolde, D. E.; Arseniev, A. S.; Efremov, R. G. Helix interactions in membranes: lessons from unrestrained Monte Carlo simulations. *J. Chem. Theory Comput.* **2005**, *1*, 1252−1264.
(9) Hillisch, A.; Pineda, L. F.; Hilgenfeld, R. Utility of homology models in the drug discovery process. *Drug Discovery Today* **2004**, *9*, 659−669.
(10) Sperotto, M. M.; May, S.; Baumgaertner, A. Modelling of proteins in membranes. *Chem. Phys. Lipids* **2006**, *141*, 2−29.
(11) Efremov, R. G.; Nolde, D. E.; Konshina, A. G.; Syrtcev, N. P.; Arseniev, A. S. Peptides and proteins in membranes: what can we learn via computer simulations? *Curr. Med. Chem.* **2004**, *11*, 2421−2442.
(12) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **2004**, *383*, 66−93.
(13) Yarov-Yarovoy, V.; Schonbrun, J.; Baker, D. Multipass Membrane Protein Structure Prediction Using Rosetta. *Proteins* **2006**, *62*, 1010−1025.
(14) Bradley, P.; Misura, K. M.; Baker, D. Toward high-resolution *de novo* structure prediction for small proteins. *Science* **2005**, *309*, 1868−1871.
(15) Zhang, Y.; Devries, M. E.; Skolnick, J. Structure modeling of all identified G-protein-coupled receptors in the human genome. *PLoS Comput. Biol.* **2006**, *2*, e13.
(16) Bissantz, C.; Logean, A.; Rognan, D. High-throughput modeling of human G-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1162−1176.
(17) Sorgen, P. L.; Hu, Y.; Guan, L.; Kaback, H. R.; Girvin, M. E. An approach to membrane protein structure without crystals. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14037−14040.
(18) Bowie, J. U. Solving the membrane protein folding problem. *Nature* **2005**, *438*, 581−589.
(19) White, S. H.; von Heijne, G. Transmembrane helices before, during, and after insertion. *Curr. Opin. Struct. Biol.* **2005**, *15*, 378−386.
(20) Archer, E.; Maigret, B.; Escrieut, C.; Pradayrol, L.; Fourmy, D. Rhodopsin crystal: new template yielding realistic models of G-protein-coupled receptors? *Trends Pharmacol. Sci.* **2003**, *24*, 36−40.
(21) Kopp, J.; Schwede, T. The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Res.* **2006**, *34*, D315−D318.
(22) Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* **1982**, *299*, 371−374.
(23) Rees, D. C.; DeAntonio, L.; Eisenberg, D. Hydrophobic organization of membrane proteins. *Science* **1989**, *245*, 510−513.

(24) Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **1976**, *105*, 1−12.
(25) Efremov, R. G.; Vergoten, G. Hydrophobic organization of alpha-helix membrane bundle in bacteriorhodopsin. *J. Protein Chem.* **1996**, *15*, 63−76.
(26) Stevens, T. J.; Arkin, I. T. Are membrane proteins "inside-out" proteins? *Proteins* **1999**, *36*, 135−143.
(27) Samatey, F. A.; Xu, C.; Popot, J. L. On the distribution of amino acid residues in transmembrane alpha-helix bundles. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 4577−4581.
(28) 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.
(29) Donnelly, D.; Overington, J. P.; Ruffle, S. V.; Nugent, J. H.; Blundell, T. L. Modeling α-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. *Protein Sci.* **1993**, *2*, 55−70.
(30) Wallin, E.; Tsukihara, T.; Yoshikawa, S.; von Heijne, G.; Elofsson, A. Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria. *Protein Sci.* **1997**, *6*, 808−815.
(31) Stevens, T. J.; Arkin, I. T. Substitution rates in alpha-helical transmembrane proteins. *Protein Sci.* **2001**, *10*, 2507−2517.
(32) 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.
(33) Park, Y.; Helms, V. Assembly of transmembrane helices of simple polytopic membrane proteins from sequence conservation patterns. *Proteins* **2006**, *64*, 895−905.
(34) Fleishman, S. J.; Harrington, S.; Friesner, R. A.; 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.
(35) Beuming, T.; Weinstein, H. Modeling membrane proteins based on low-resolution electron microscopy maps: a template for the TM domains of the oxalate transporter OxlT. *Protein Eng. Des. Sel.* **2005**, *18*, 119−125.
(36) Sal-Man, N.; Gerber, D.; Shai, Y. Hetero-assembly between all-L- and all-D-amino acid transmembrane domains: forces involved and implication for inactivation of membrane proteins. *J. Mol. Biol.* **2004**, *344*, 855−864.
(37) Eilers, M.; Shekar, S. C.; Shieh, T.; Smith, S. O.; Fleming, P. J. Internal packing of helical membrane proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5796−5801.
(38) Eilers, M.; Patel, A. B.; Liu, W.; Smith, S. O. Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys. J.* **2002**, *5*, 2720−2736.
(39) Senes, A.; Engel, D. E.; DeGrado, W. F. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr. Opin. Struct. Biol.* **2004**, *14*, 465−479.
(40) Curran, A. R.; Engelman, D. M. Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr. Opin. Struct. Biol.* **2003**, *13*, 412−417.
(41) Gimpelev, M.; Forrest, L. R.; Murray, D.; Honig, B. Helical packing patterns in membrane and soluble proteins. *Biophys. J.* **2004**, *87*, 4075−4086.
(42) Walters, R. F.; DeGrado, W. F. Helix-packing motifs in membrane proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13658−13663.
(43) Fleishman, S. J.; Ben-Tal, N. A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J. Mol. Biol.* **2002**, *321*, 363−378.
(44) 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.
(45) Liu, W.; Eilers, M.; Patel, A. B.; Smith, S. O. Helix packing moments reveal diversity and conservation in membrane protein structure. *J. Mol. Biol.* **2004**, *337*, 713−729.
(46) Becker, O. M.; Shacham, S.; Marantz, Y.; Noiman, S. Modeling the 3D structure of GPCRs: advances and application to drug discovery. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 353−361.
(47) Bowie, J. U.; Lüthy, R.; Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **1991**, *253*, 164−170.
(48) Delarue, M.; Koehl, P. Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J. Mol. Biol.* **1995**, *249*, 675−690.
(49) Efremov, R. G.; Nolde, D. E.; Volynsky, P. E.; Arseniev, A. S. Modeling of peptides in implicit membrane-mimetic media. *Mol. Simul.* **2000**, *24*, 275−291.
(50) Efremov, R. G.; Volynsky, P. E.; Nolde, D. E.; Arseniev, A. S. Implicit two-phase solvation model as a tool to assess conformation and energetics of proteins in membrane-mimic media. *Theor. Chem. Acc.* **2001**, *106*, 48−54.
(51) Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinformatics* **2006**, *22*, 623−625.
(52) Tusnady, G. E.; Dosztanyi, Z.; Simon, I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.* **2005**, *33*, D275−D278.
(53) Tusnady, G. E.; Dosztanyi, Z.; Simon, I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics* **2005**, *21*, 1276−1277.
(54) Basyn, F.; Spies, B.; Bouffioux, O.; Thomas, A.; Brasseur, R. Insertion of X-ray structures of proteins in membranes. *J. Mol. Graphics Modell.* **2003**, *22*, 11−21.
(55) von Heijne, G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* **1986**, *5*, 3021−3027.
(56) Ulmschneider, M. B.; Sansom, M. S.; Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* **2005**, *59*, 252−265.
(57) Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S. H.; von Heijne, G. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **2005**, *433*, 377−381.
(58) Hessa, T.; White, S. H.; von Heijne, G. Membrane insertion of a potassium-channel voltage sensor. *Science* **2005**, *307*, 1427.
(59) Eyre, T. A.; Partridge, L.; Thornton, J. M. Computational analysis of alpha-helical membrane protein structure: implications for the prediction of 3D structural models. *Protein Eng. Des. Sel.* **2004**, *17*, 613−624.
(60) Wimley, W. C.; White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **1996**, *10*, 842−848.
(61) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. Statistical Description of Data. In *Numerical Recipes in Fortran. The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: New York, 1992; pp 640−644.
(62) Lüthy, R.; Bowie, J. U.; Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **1992**, *356*, 83−85.
(63) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000**, *289*, 739−745.
(64) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(65) Efremov, R. G.; Nolde, D. E.; Vergoten, G.; Arseniev, A. S. Peptides in membranes: assessment of the effects of environment via simulations with implicit solvation model. *Theor. Chem. Acc.* **1999**, *101*, 170−174.
(66) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(67) Eisenberg, D.; Wesson, M.; Yamashita, M. Interpretation of protein folding and binding with atomic solvation parameters. *Chem. Scr.* **1989**, *29A*, 217−221.
(68) Okada, T.; Sugihara, M.; Bondar, A. N.; Elstner, M.; Entel, P.; Buss, V. The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. *J. Mol. Biol.* **2004**, *342*, 571−583.
(69) Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F.; Vriend, G. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **1998**, *26*, 277−281.
(70) Marti-Renom, M. A.; Stuart, A.; Fiser, A.; Sánchez, R.; Melo, F.; Šali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.
(71) Luecke, H.; Schobert, B.; Richter, H. T.; Cartailler, J. P.; Lanyi, J. K. Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.* **1999**, *291*, 899−911.
(72) Marti-Renom, M. A.; Ilyin, V. A.; Šali, A. DBAli: a database of protein structure alignments. *Bioinformatics* **2001**, *17*, 746−747.
(73) Herzyk, P.; Hubbard, R. E. Combined biophysical and biochemical information confirms arrangement of transmembrane helices visible from the three-dimensional map of frog rhodopsin. *J. Mol. Biol.* **1998**, *281*, 741−754.
(74) Schertler, G. F.; Hargrave, P. A. Projection structure of frog rhodopsin in two crystal forms. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 11578−11582.
(75) Baldwin, J. M.; Schertler, G. F.; Unger, V. M. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.* **1997**, *272*, 144−164.

(76) Shieh, T.; Han, M.; Sakmar, T. P.; Smith, S. O. The steric trigger in rhodopsin activation. *J. Mol. Biol.* **1997**, *269*, 373−384.
(77) Donnelly, D.; Findlay, J. B.; Blundell, T. L. The evolution and structure of aminergic G protein-coupled receptors. *Recept. Channels* **1994**, *2*, 61−78.
(78) Lin, S. W.; Imamoto, Y.; Fukada, Y.; Shichida, Y.; Yoshizawa, T.; Mathies, R. A. What makes red visual pigments red? A resonance Raman microprobe study of retinal chromophore structure in iodopsin. *Biochemistry* **1994**, *33*, 2151−2160.
(79) Pogozheva, I. D.; Lomize, A. L.; Mosberg, H. I. The transmembrane 7-alpha-bundle of rhodopsin: distance geometry calculations with hydrogen bonding constraints. *Biophys. J.* **1997**, *72*, 1963-1985.
(80) Nikiforovich, G. V.; Marshall, G. R. Three-dimensional model for meta-II rhodopsin, an activated G-protein-coupled receptor. *Biochemistry* **2003**, *42*, 9110−9120.

CI600516X