Topology of Membrane Proteins

Gábor E. Tusnády and István Simon*

Institute of Enzymology, BRC, Hungarian Academy of Sciences, Budapest, Hungary, and
Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary

Received August 25, 2000

Integral membrane proteins play important roles in living cells. Due to difficulties of experimental techniques, theoretical approaches, i.e., topology prediction methods, are important for structure determination of this class of proteins. Here we show a detailed comparison of transmembrane topology prediction methods. According to this comparison, we conclude that the topology of integral membrane proteins is determined by the maximum divergence of the amino acid composition of sequence segments. These segments are located in different areas of the cell, which can be characterized by different physicochemical properties. The results of these prediction methods compared to the X-ray diffraction data of several transmembrane proteins will also be discussed.

INTRODUCTION

Integral membrane proteins form about 25% of all protein sequences.[1,2] They are of vital importance for living cells, playing a role in communication and in the transport of the cells with the outside world. Due to difficulties of experimental techniques, theoretical approaches, i.e., topology prediction methods, are important for structure determination of this class of proteins. The comparison of various prediction methods may also help to identify the principles governing the structure formation of these proteins.

Early transmembrane structure prediction methods identify the most hydrophobic residue clusters as the approximate location of the transmembrane segments. These approaches were based on the hydrophobicity of amino acids determined by various physicochemical measures.[3−9] It was shown, however, that from the viewpoint of protein structure formation, parameters obtained by statistical analysis of protein sequence databases are more reliable[10−12] than parameters based on hydrophobicity measures. Therefore, transmembrane segment predictions based on statistical parameters are more accurate than those methods based on physicochemical parameters.[12−14]

Predicting only the approximate location of the transmembrane segments in the primary structure is an incomplete topology prediction, as it does not tell anything about the orientation of transmembrane segments. It has been shown that the positively charged residues prefer to appear at the cytoplasmic site of the membrane due to the asymmetric lipid distribution.[15−17] Methods based on this principle result in three types of sequence segments: inside, transmembrane, and outside segments. Statistical analyses of transmembrane protein sequences revealed that the two ends of transmembrane helices also have characteristic amino acid distribution.[12,18] Using these five types of segments (inside, outside, transmembrane helix, inside helix cap, and outside helix cap) in a dynamic programming algorithm resulted in a more accurate topology prediction.[12] Topology prediction based on hidden Markov models[19,20] used five and seven types of sequence segments.

The crudest description of folding both globular and transmembrane proteins is based on the hydrophobicity of the residues. In globular proteins, clusters of hydrophobic residues tend to stay inside the protein and are not exposed to the aqueous medium, while polar and charged residues appear on the surface and make the globular protein soluble in electrolyte solvent. The same applies to transmembrane proteins: hydrophilic or polar and charged residues prefer to be exposed to the aqueous medium, while hydrophobic patches of transmembrane protein residues stay mainly in the nonpolar part of the lipid membrane. In some cases, however, these hydrophobic segments can be found in buried areas, in the hydrophobic core of globular domains connecting transmembrane segments.

Concerning only hydrophobicity, folding follows the principle “similis simili gaudet” as a first approach. The segregation of polar, nonpolar, and positively and negatively charged residues, however, is far from complete in the primary structure. As a consequence, the residues are not necessarily located in their preferred environment. Therefore, the formation of the structure is a result of compromise, and the above-mentioned principle can never be perfectly fulfilled. In transmembrane proteins the hydrophobic environment of the transmembrane segments is the result of the bulk lipid tails. In this case altering the structure of transmembrane segment connecting loops does not change the environment of the transmembrane segments. In globular proteins, however, the environment of any segment of the polypeptide chain is determined by other segments of the polypeptide chains as well as by solvent atoms. Therefore, the structure prediction methods based on local sequence information are more accurate for transmembrane proteins than the more advanced prediction methods for globular proteins.

Comparing the Results of Prediction Methods on a Test Set. Sequences for comparing 11 topology prediction meth-

* To whom correspondence should be addressed: Institute of Enzymology, BRC, Hungarian Academy of Sciences, P.O. Box 7, H-1518 Budapest, Hungary. Phone: (36-1) 466-9276. Fax: (36-1) 466-5465. E-mail: simon@enzim.hu.

10.1021/ci0001280 CCC: $20.00
Published on Web 02/14/2001

Table 1. Prediction Accuracy of Various Transmembrane Topologies and α-Helix Prediction Methods[a]

method	N_{prd}	N_{cor}	Q_P (%)	N_{TM}	Q_{TM} (%)	N_{TT}	Q_{TT} (%)
TMAP	429	398	91	70	68	43	42
PRED-TMR	423	406	94	73	71		
PHDhtm	425	409	94	77	75	66	64
SOSUI	428	411	94	79	77		
DAS	462	412	91	81	79		
SPLIT	430	401	92	83	81	68	66
TMPRED	464	418	92	84	82	63	61
TMHMM	430	420	96	89	86	80	78
MEMSAT	442	419	95	90	87	78	76
TOPPRED	485	429	93	92	89	69	67
HMMTOP	458	432	96	93	90	79	77

[a] References to the methods are given in the text. N_{prd} and N_{cor} are the numbers of predicted and correctly predicted transmembrane helices, respectively; $Q_P = 100[(N_{cor}/N_{obs})(N_{cor}/N_{prd})]^{1/2}$, where $N_{obs} = 442$ is the number of observed transmembrane helices in the test set. N_{TM} (Q_{TM}) is the number (and percent) of proteins for which all transmembrane segments were predicted correctly. N_{TT} (Q_{TT}) is the number (and percent) of proteins for which both the topology and the transmembrane segments were predicted correctly. The 100% corresponds to 103 proteins.

ods have been selected as follows: A similarity filtering was applied on the 158 transmembrane proteins, which were collected in our earlier work.[19] The filtering procedure resulted in a nonredundant data set with 103 proteins and 442 transmembrane α-helices. The Swiss-Prot sequence identifiers together with the established transmembrane topology data of the 103 selected protein can be found at the following URL: http://www.enzim.hu/tusi/tm/testset.html.

The 11 transmembrane α-helix and topology prediction programs use different principles for transmembrane segments and topology prediction. The methods are the following: TOPPRED[15] and PRED-TMR[21] apply hydrophathy analysis. SOSUI,[22] SPLIT,[23] and TMPRED[24] use various propensity scales. in the DAS[14] method transmembrane segments are identified by each other. TMAP[18] is based on statistical data of transmembrane proteins. MEMSAT[12] predicts topology using a dynamic programming algorithm. PHDhtm[25] applies the neural-network algorithm. HMMTOP[19] and TMHMM[20] use various hidden Markov models.

Most of these programs predict the localization of the helical transmembrane segments in the primary structure as well as the topology, i.e., the localization of transmembrane segments connecting loops relative to the membrane (inside or outside). Three of these methods (DAS, SOSUI, PRED-TMR) predict only the positions of the transmembrane α-helices without the topology. TOPPRED, PHDhtm and SPLIT methods apply the “positive-inside” rule, while others use fixed (TMAP, MEMSAT) or variable amino acid compositions (TMHMM, HMMTOP) for topology prediction.

To obtain comparable results, only single sequence information was used for prediction by all methods. In all cases the default parameters (if any) were applied.

The results of the predictions are presented in Table 1. The transmembrane α-helix prediction accuracy of all methods is above 90% (column Q_P (%) in Table 1), measured according to Cserző et al. (1997).[14] Since these accuracies vary within a small range (91−96%), two other measurements of accuracy were also introduced. These are the number (and percent) of proteins for which (i) all the transmembrane segments and (ii) all the transmembrane segments with the topology are correctly predicted. These new measures are noted as N_{TM} (Q_{TM}) and N_{TT} (Q_{TT}) in Table 1, respectively. The prediction methods are sorted according to the value of N_{TM} in Table 1. Q_P values are above 90%, as mentioned above, while Q_{TM} values show a broader spectrum, which ranges from 68% (TMAP) to 90% (HMMTOP).

Interestingly, the first developed TOPPRED algorithm performs better topology prediction than the other methods, which are also based on the positive-inside rule. This may be due to the different algorithms applied in these programs. TOPPRED selects from potential transmembrane segments to maximize the charge difference, while the others use the positive-inside rule to make a decision between the only two possible topologies, N-in and N-out, respectively. Topology predictions taking into account the distribution of all amino acids and not only the distribution of positively charged ones (MEMSAT, HMMTOP, and TMHMM) result in higher prediction accuracy. TMHMM results in the best topology prediction on this test set. It reaches a worse transmembrane α-helix prediction accuracy, however, than HMMTOP using a different hidden Markov model. This may be the consequence of the seven states used in the TMHMM model, which can make the model more sensitive to the topogenic signals, i.e., the variation of the amino acid compositions.

Transmembrane Proteins with Known 3D Structure. Due to the difficulties in crystallizing transmembrane proteins, only a few 3D structures have been determined by X-ray crystallography. Therefore, there is a large gap between the number of solved structures of globular proteins and the number of solved structures of transmembrane proteins. While the first group contains more than 1000 nonhomologous proteins, the latter one consists of only 13 nonhomologous integral membrane proteins. In this study eight of these structures were selected, which are built up by transmembrane α-helices and have a resolution better than 3.5 Å. The selected proteins are the following (entries of which sequence was used are shown in bold): bacteriorhodopsin (PDB code: 1at9, 1ap9, 2brd, 1brr, 1brx, **1c3w**, 1qhj, 1qko, 1qkp), bacterial photosynthetic reaction centers (**2prc**, 1pss, 2rcr), light-harvesting complexes (**1kzu**, 1lgh), cytochrome *c* oxidase (**1occ**, 1qle), cytochrome bc_1 complex (1bcc, **1bgy**, 1qcr), potassium channel (**1bl8**), mechanosensitive channel (**1msl**), and fumarate reductase complex (1fum, **1qla**, 1qlb).[26]

The crystal structures of transmembrane proteins show a common arrangement of the transmembrane α-helices: they cross the membrane nearly perpendicular to the membrane surface. In channel and transporter proteins, the transmembrane α-helices form regular, often symmetric “pipes”. In receptors and proteins dealing with electron transport, the helices are often tilted or form a supercoil structure. These arrangements have been shown to be favored energetically.[27]

Where is the Membrane? The crystal structures of transmembrane proteins do not elucidate the exact placement of the protein in the lipid bilayers. Although some approaches have been developed based on the 3D structure of transmembrane proteins to find the position of the boundaries of the lipids relative to the protein, no general definition exists to determine these boundaries.

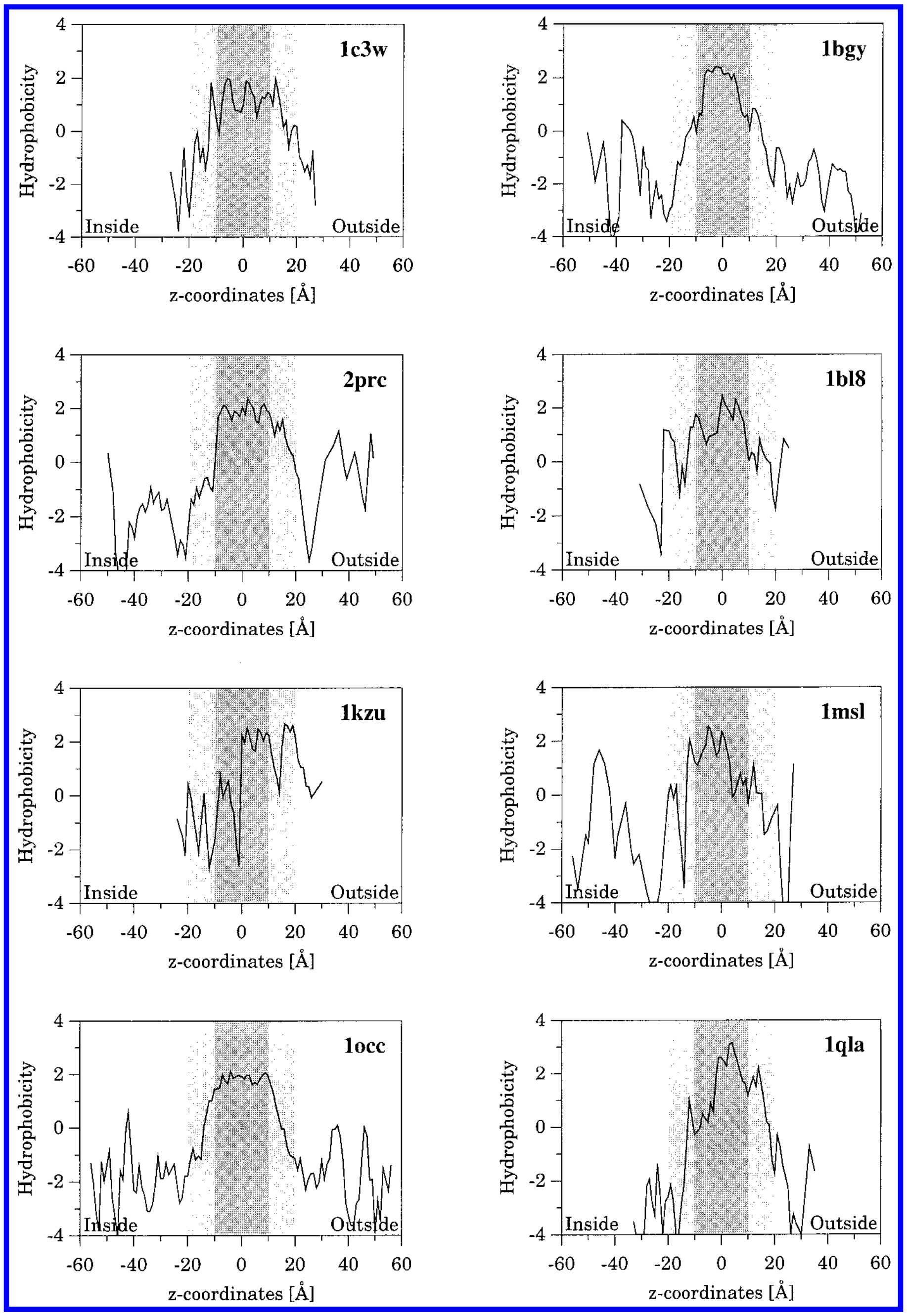

Figure 1. Average hydrophobicity of various transmembrane proteins measured parallel to the average direction of the transmembrane helices in 1 Å slices. Zero is set to the average middle point of the transmembrane helices. The dark gray area marks the region defined by distances of ±10 Å, while the lighter gray area shows the region defined by distances of ±20 Å from the zero point.

Unger and co-workers define the cytoplasmic and extracellular surfaces of the membranes as the hydrophobic−hydrophilic interfaces, determined from the distribution of nitrogen and oxygen atoms in the hydrophilic amino acid side chains.[28]

Baldwin and co-workers have made a detailed sequence analysis on the rhodopsin family of G-protein-coupled receptors. They deduced the inner and outer boundaries of the membrane using the sequence alignment of about 500 sequences by considering the positions of sites where no polar or charged residues occur on the region of the helix surface where the variation within groups occurs.[29]

Wallace and Janes (1999) investigated the distribution and the possible role of tryptophan in the membrane environment. They found that tryptophans tend to be clustered at the interfacial region of the bilayer, and often form two parallel aromatic bands around the girth of the protein.[30]

Wallin and co-workers investigated the structure of cytochrome *c* oxidase[31] by special hydrophobicity measurements. They assigned the *z*-coordinate as the average axis of the transmembrane helices. This direction is perpendicular to the membrane surface. Then they cut the protein in 1 Å slices perpendicular to the *z*-coordinate and calculated the average hydrophobicity of the residues in each slice. They found that these average hydrophobicity values vary little over the innermost lipid region (about 20 Å); then they decrease over the next 10 Å.

In this study we have performed the same calculation for the eight selected transmembrane proteins, as shown in Figure 1. The average hydrophobicity profile of 1bgy is the

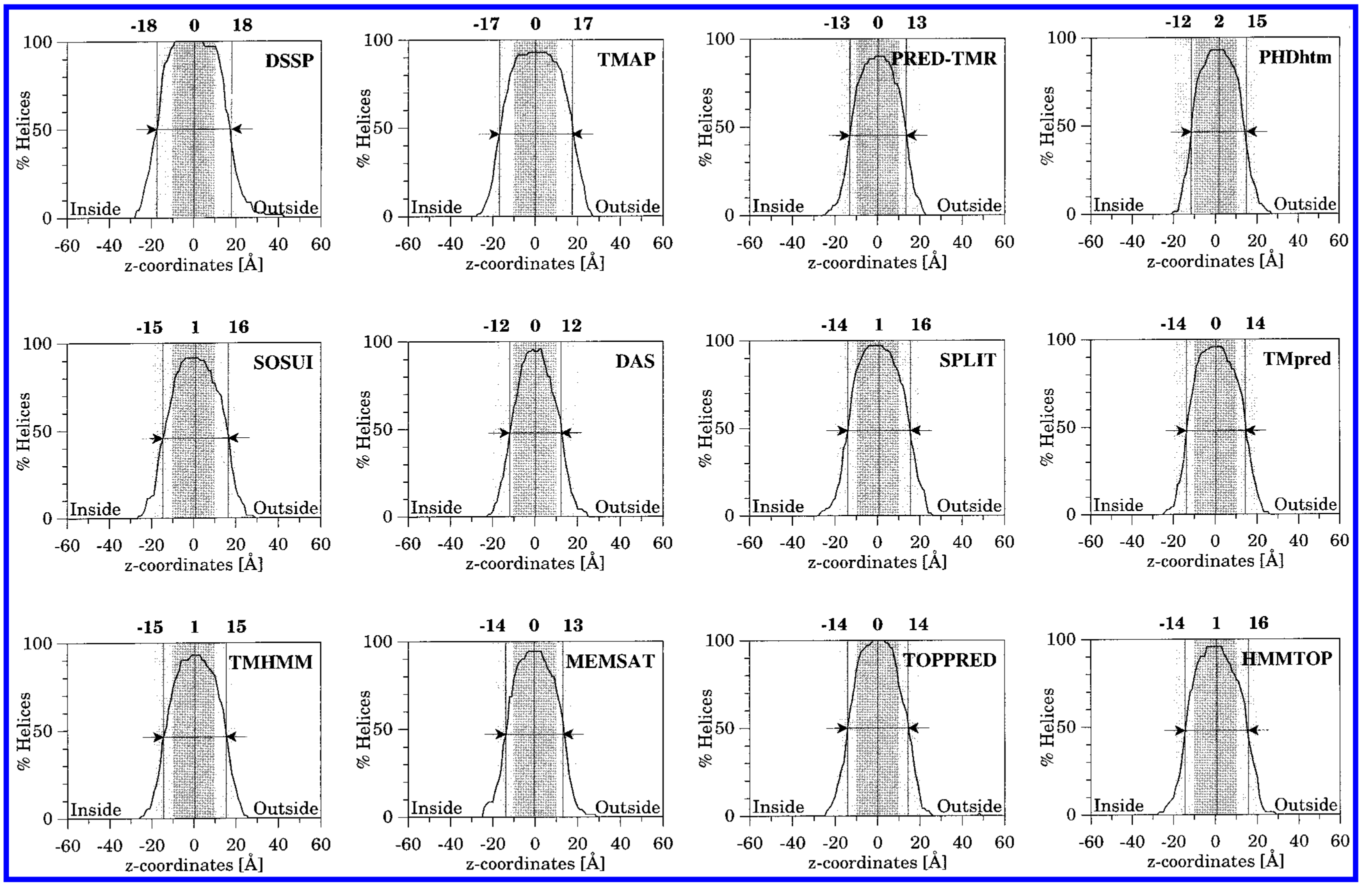

Figure 2. Distribution of transmembrane helices measured parallel to the average direction of the transmembrane helices in 1 Å slices in the eight selected proteins (see the text). Transmembrane helices are defined by the DSSP program and predicted by various transmembrane helix prediction methods. The *z*-coordinates at the value of 50% and at the mean of the curves are shown above the graphs. The gray areas show the same regions as in Figure 1.

same as was shown by Wallin et al. (1997),[31] and we see similar smooth plots with high values in the region of -10 to $+10$ Å in the case of 2prc, 1occ, and 1qla. In some cases, however, we obtained noisy (1c3w), shifted (1msl), or differently shaped (1kzu, 1bl8) plots resulting from the special structure of these proteins. Hence, in these cases the innermost part of the membrane cannot be determined as easily as in the case of cytochrome *c* oxidase.

Comparison of Results of Prediction Methods with X-ray Diffraction Data. The distributions of the transmembrane α-helices perpendicular to the membrane in the eight α-helical transmembrane proteins listed above are shown in Figure 2, panel DSSP. The half-width of the curve is 36 Å, which corresponds to an α-helix formed by 36 residues. The longest helix consists of 43 residues in the cytochrome bc_1 protein (1bgy), while the shortest helix is formed by 17 residues in the photosynthetic reaction center (2prc) determined by the DSSP program.[32] The length distribution indicates that the ends of the helices can most frequently be found in the interfacial area, but for about 35% of the helices, the ends are in the intra- or extracellular aqueous medium.

The average helix length is somewhat longer than is generally used in the various prediction programs. This raises the question of which part of the helices is regarded as "transmembrane" by the prediction methods. Therefore, we predicted the transmembrane segments of the eight selected proteins by the eleven prediction methods used on the test set, and calculated the distribution of these segments similar to the DSSP curve.

The distributions of the predicted transmembrane segments are shown in Figure 2. The mean values are close to zero, indicating that most probably the centers of the helices are predicted by the various prediction methods. The length of the predicted segments varies: DAS predicts the shortest helices (the half-width of the curve is 24 Å), while the TMAP method predicts the longest ones (the half-width of the curve is 34 Å). These values show that the lengths of the predicted helices are comparable to the width of the apolar part of the membrane, rather than the full width of the membrane. There is a small offset to the extracytoplasmic area in the case of the the PHDhtm, SOSUI, SPLIT, TMHMM, and HMMTOP methods, though they are not significant.

CONCLUSION

The comparison of the results of the topology and transmembrane α-helix prediction methods shows that two methods based on different hidden Markov models perform the highest prediction power. Knowing the mathematical details of the hidden Markov model, these results suggest that the topologies of the transmembrane proteins are determined by the maximum divergence of the amino acid composition of sequence segments which are located in different areas of the cell. These parts of the cell can be characterized by different physicochemical properties.

The 3D structure of transmembrane proteins does not disclose the exact location of the protein in the lipid bilayers. The hydrophobicity plots perpendicular to the membrane assign the apolar part of the membrane only, when the protein contains many transmembrane helices, holding enough information for the assignment. Although the studied prediction methods are based on different principles, these methods undoubtedly predict those helical parts as transmembrane segments that occupy the most apolar parts of the membrane.

ACKNOWLEDGMENT

We thank Dr. Monika Fuxreiter for helpful discussion of this paper. This work was supported by Grants T 30566 from the Hungarian Scientific Research Found (OTKA) and AKP 98-13 3,3 from the Hungarian Academy of Sciences. G.E.T. acknowledges the support of the Foundation for Hungarian Higher Education and Research (Zoltán Magyary Fellowship).

REFERENCES AND NOTES

(1) Jones, D. T. Do transmembrane protein superfolds exist? *FEBS Lett.* **1998**, *423*, 281−285.

(2) Wallin, E.; von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organism. *Protein Sci.* **1998**, *7*, 1029−1038.

(3) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105−132.

(4) Eisenberg, D.; Schwartz, E.; Komáromy, M.; Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **1984**, *179*, 125−142.

(5) Engelman, D. M.; Steitz, T. A.; Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Chem.* **1986**, *15*, 321−353.

(6) Cornette, J. L.; Cease, K. B.; Margalit, H.; Spouge, L.; Berzofsky, J. A.; DeLisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **1987**, *195*, 659−685.

(7) Esposti, M. D.; Crimi, M.; Venturoli, G. A critical evaluation of the hydropathy profile of membrane proteins. *Eur. J. Biochem.* **1990**, *190*, 207−219.

(8) Ponnuswamy, P. K.; Gromiha, M. M. Prediction of transmembrane helices from hydrophobic characteristics of protein. *Int. J. Pept. Protein Res.* **1993**, *42*, 326−341.

(9) Gromiha, M. M.; Ponnuswamy, P. K. Prediction of protein secondary structures from their hydrophobic characteristics. *Int. J. Pept. Protein Res.* **1995**, *45*, 225−240.

(10) Tusnády, G. E.; Tusnády, G.; Simon, I. Independence divergence-generated binary trees of amino acids. *Protein Eng.* **1995**, *8*, 417−423.

(11) Tüdős, É.; Cserző, M.; Simon, I. Predicting isomorphic residue replacements for protein design. *Int. J. Pept. Protein Res.* **1990**, *36*, 236−239.

(12) Jones, D. T.; Taylor, W. R.; Thorton, J. M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemisty* **1994**, *33*, 3038−3049. http://insulin.-brunel.ac.uk/ jones/memsat.html.

(13) Cserző, M.; Bernassau, J−M.; Simon, I.; Maigret, B. Unusual alignment strategy for transmembrane proteins. *J. Mol. Biol.* **1994**, *243*, 388−396.

(14) Cserző, M.; Wallin, E.; Simon, I.; von Heijne, G.; Elofsson, A. Prediction of transmembrane α-helices in prokariotic membrane proteins: the dense aligment surface method. *Protein Eng.* **1997**, *10*, 673−676. http://www.sbc.su.se/-miklos/DAS/.

(15) von Heijne, G. Membrane protein structure prediction. *J. Mol. Biol.* **1992**, **225**, 487−494. http://www.sbc.su.se/-erikw/toppred2/.

(16) Sipos, L.; von Heijne, G. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* **1993**, *213*, 1333−1340.

(17) van Klompenburg, W.; Nilsson, I.; von Heijne, G.; de Kruijff, B. Anionic phospholipids are determinants of membrane protein topology. *EMBO J.* **1997**, *16*, 4261−4266.

(18) Milpetz, F.; Argos, P. TMAP: a new email and WWW service for membrane-protein structural predictions. *Trends Biochem. Sci.* **1995**, *20*, 204−205.

(19) Tusnády, G. E.; Simon, I. Principles governing amino acid composition of integral membrane proteins: Application to topology prediction. *J. Mol. Biol.* **1998**, *283*, 489−506. http://www.enzim.hu/hmmtop.

(20) Sonnhammer, E. L. L.; von Heijne, G.; Krogh, A. A hidden Markov model for predicting transemembrane helices in protein sequences. In *Proceedings of Sixth International Conference on Intelligent Systems for Molecular Biology*; AAAI/MIT Press: Menlo Park, CA, 1998; Vol. 6, pp 175−182. http://www.cbs.dtu.dk/services/TMHMM-1.0/.

(21) Pasquier, C.; Promponas, V. J.; Palaios, G. A.; Hamodrakas, J. S.; Hamodrakas, S. J. Predicting transmembrane segment in proteins. *Protein Eng.*, submitted for publication. http://o2.db.uoa.gr/PRED-TMR.

(22) Hirokawa, T.; Boon-Chieng, S.; Mitaku, S. SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **1998**, *14*, 378−379. http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html.

(23) Juretić, D.; Lučin, A. The preference functions method for predicting protein helical turns with membrane propensity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 575−585. http://pref.etfos.hr/split/.

(24) Hofmann, K.; W. Stoffel. TMbase−A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler* **1993**, *347*, 166. http://www.ch.embnet.org/software/TMPRED_form.html.

(25) Rost, B.; Fariselli, P.; Casadio, R. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **1996**, *5*, 1704−1718. http://www.embl-heidelberg.de/predictprotein.

(26) White, S. H.; Wimley, W. C. Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 319−365. http://blanco.biomol.uci.edu/MembraneR._proteins_xtal.html.

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

(28) Kimura, Y.; Vassylyev, D. G.; Miyazawa, A.; Kidera, A.; Matsushima, M.; Mitsuoka, K.; Murata, K.; Hirai, T.; Fujiyoshi, Y. Surface of bacteriorhodopsin revealed by high-resolution electron crystallography. *Nature* **1997**, *389*, 206−211.

(29) 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.

(30) Wallace, B. A.; Janes, R. W. Tryptophans in membrane proteins: X-ray crystallographic analyses. *Adv. Exp. Med. Biol.* **1999**, *467*, 789−799.

(31) 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.

(32) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

CI0001280