Article No. jmbi.1999.3257 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 294, 921–935

JMB

kPROT: A Knowledge-based Scale for the Propensity of Residue Orientation in Transmembrane Segments. Application to Membrane Protein Structure Prediction

Yitzhak Pilpel[1]*, Nir Ben-Tal[2] and Doron Lancet[1]

[1]*Department of Molecular Genetics and the Crown Genome Center, The Weizmann Institute of Science, Rehovot 76100, Israel*

[2]*Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel*

**Corresponding author*

Modeling of integral membrane proteins and the prediction of their functional sites requires the identification of transmembrane (TM) segments and the determination of their angular orientations. Hydrophobicity scales predict accurately the location of TM helices, but are less accurate in computing angular disposition. Estimating lipid-exposure propensities of the residues from statistics of solved membrane protein structures has the disadvantage of relying on relatively few proteins. As an alternative, we propose here a scale of knowledge-based Propensities for Residue Orientation in Transmembrane segments (kPROT), derived from the analysis of more than 5000 non-redundant protein sequences. We assume that residues that tend to be exposed to the membrane are more frequent in TM segments of single-span proteins, while residues that prefer to be buried in the transmembrane bundle interior are present mainly in multi-span TMs. The kPROT value for each residue is thus defined as the logarithm of the ratio of its proportions in single and multiple TM spans. The scale is refined further by defining it for three discrete sections of the TM segment; namely, extracellular, central, and intracellular. The capacity of the kPROT scale to predict angular helical orientation was compared to that of alternative methods in a benchmark test, using a diversity of multi-span α-helical transmembrane proteins with a solved 3D structure. kPROT yielded an average angular error of 41°, significantly lower than that of alternative scales (62°-68°). The new scale thus provides a useful general tool for modeling and prediction of functional residues in membrane proteins. A WWW server (http://bioinfo.weizmann.ac.il/kPROT) is available for automatic helix orientation prediction with kPROT.

Keywords: knowledge-based potential; structure prediction; membrane proteins; helical moments; hydrophobicity scales

Introduction

Structural exploration of integral membrane proteins is difficult, and currently the high-resolution structure of only a few proteins is known (von Heijne, 1996; Preusch *et al.*, 1998). In the absence of experimental structural evidence, modeling the structure of the transmembrane (TM) portion of membrane proteins consists of predicting the location of the TM segments along the amino acid sequence, and establishing their intracellular/extracellular topology. This is currently accomplished with a very high level of accuracy, based on hydrophobicity scales and knowledge-based statistical propensities (Kyte & Doolittle, 1982; Engelman *et al.*, 1986; von Heijne, 1992; Jones *et al.*, 1994a; Persson & Argos, 1994; Rost *et al.*, 1995, 1996; Cserzo *et al.*, 1997; Tusnady & Simon, 1998). Modeling proteins with multiple TM segments requires, in addition, to predict the angular orientation of each TM segment, i.e. to determine which residues are exposed to the lipid phase and which are buried in the interior of the TM bundle.

Hydrophobicity moments of TM helices are currently the main *ab initio* chemically related method for predicting the relative angular orientations of TM segments (Eisenberg *et al.*, 1982, 1984; Rees *et al.*, 1989). In these methods, the angular orien-

Present address: Y. Pilpel, Department of Genetics, Harvard Medical School, 200 Longwood Ave., Boston, MA 02115, USA.

E-mail address of the corresponding author: bnpilpel@membran1.weizmann.ac.il

0022-2836/99/490921–15 $30.00/0

tations are predicted by directing the helical hydrophobic moments to the lipid phase. However, tests on membrane proteins with known 3D structures have shown that hydrophobicity moments are poor indicators of the solvent-exposed face of TM helices (Cronet *et al*., 1993; Stevens & Arkin, 1999). This may be partly because some hydrophobic residues tend to face both the lipid and the protein core.

Methods based on the statistics of known high-resolution structures of integral membrane proteins have been used to derive lipid exposure propensities of the different residues (Cronet *et al*., 1993; Donnelly *et al*., 1993). However, relying on the small data set of known 3D-structures may generate a biased view of the true structure space, and it would be preferable to extract information from a larger data set. The multitude of sequences of integral membrane proteins, e.g. more than 10,000 in the SWISS-PROT database (Bairoch & Boeckmann, 1991), may serve as a source for deriving a transmembrane helix orientation scale. Such a sequence-derived scale may be representative of nearly all membrane proteins and would be potentially endowed with a considerably greater accuracy and statistical significance than that of current structure-based scales.

Samatey *et al*. (1995) have taken such an approach and derived a sequence-based scale in which they utilized the fact that a multitude of transmembrane spans have an α-helical periodicity. They used a power spectrum method to select TM sequences that display an α-helical periodicity and derived a scale of the propensity of the residues to be buried *versus* membrane-exposed in the central portion of the spans. The use of a periodicity-based method required that the analyzed portion of the TM segment will be exposed to a more or less uniform lipid environment. This limited their analysis to the central section of the TM spans, excluding the residues which face the polar lipid headgroups. Taylor *et al*. (1994) have predicted helical orientations using an additional sequence-based scale, originally derived for locating the TM segments along the primary structure (Jones *et al*., 1994a). In their scale, the tendency of the residues to be exposed to the membrane or to be buried in the interior of the protein was evaluated from a set of single-span proteins, as the preference of the residue to be in the middle TM section, compared to its relative abundance in the non-transmembrane protein segments.

We describe here an alternative scale for predicting TM angular orientations that is derived from information on all α-helical transmembrane protein sequences in the SWISS-PROT database. Our scale stems from known differences in the frequencies of amino acid (Jones *et al*., 1994b), including conserved proline residues (von Heijne, 1991), in the membrane-spanning helices of proteins with single and multiple TM segments. The present scale is based on the idea that a higher abundance of a residue in the TM segments of multi-span proteins indicates a tendency to face the protein's interior. In contrast, a higher abundance of a residue in the TM segments of single-span proteins indicates that it has a higher tendency to be exposed to the lipid phase. In the proposed knowledge-based scale for Propensities Residue Orientation in Transmembrane segments (kPROT), the transmembrane helix orientation propensity of each residue is related to the ratio of the two abundances. We show that kPROT, compared to other scales, has a higher capacity to predict TM helix angular orientations.

Results

The kPROT scale

The kPROT value for residue i is defined as:

$$\mathrm{kPROT}^i = \ln\left[\frac{f_{\mathrm{s}}^i}{f_{\mathrm{m}}^i}\right] \qquad (1)$$

where f_{s}^i and f_{m}^i are the proportions of the residue in the total set of TM segments of proteins with single and multiple spans, respectively. A logarithmic relation is used in order to convert frequencies in the database into free-energy-like scores, assuming that the database constitutes a statistical ensemble (Jernigan & Bahar, 1996; Vajda *et al*., 1997; Zhang & Skolnick, 1998).

Table 1 lists the kPROT scale derived from sequences of entire TM segments. In Figure 1(a), we compare between kPROT and the Eisenberg *et al*. (1982) hydrophobicity scale (Eisenberg *et al*., 1982). The kPROT values of the aliphatic residues Val, Leu, Ile, and Ala are positive, implying a higher tendency to face the membrane, while the negatively charged and polar residues Asp, Glu, His, Asn, Gln, Ser, and Thr display a higher inferred preference to face the protein interior. The kPROT values of these residues are generally in agreement with hydrophobicity scales and they presumably reflect the lipophobic effect. Consistently, the average value of hydrophobicity is found to be somewhat higher in the TM segments of the single-span proteins (0.56 kcal/mol in single-span proteins compared to 0.5 kcal/mol in multiple span proteins).

In contrast, the kPROT values of some other residues deviate qualitatively from the propensities derived from hydrophobicity scales. The aromatic residues Phe, Trp and Tyr, as well as Met, Pro and Gly, display a preference to be buried in the protein interior, although usually considered hydrophobic to various degrees. Cys and the two positively charged residues, Arg and Lys, display a high propensity to be exposed to the membrane.

Figure 2(a) depicts the fraction of the membranophilic ($\mathrm{kPROT} > 0$) and membranophobic ($\mathrm{kPROT} < 0$) amino acid residues in the TM segments as a function of the number of TM segments in the integral membrane protein set. It may be seen that the frequency of the membranophilic/membranophobic residues initially changes with

Table 1. The generation of the kPROT scale

A. *Number and fraction of the residues in single and multi-span sets*

Residue		Total TM		Extracellular		Central		Intracellular		Both termini	
		Numb	Frac	Numb	Frac	Numb	Frac	Numb	Frac	Numb	Frac
A	S	4888	0.105	532	0.103	3038	0.12	399	0.077	1850	0.087
	M	45,179	0.103	1517	0.1	26,134	0.109	1403	0.093	19,048	0.095
C	S	1038	0.022	47	0.009	588	0.023	213	0.041	454	0.021
	M	7488	0.017	233	0.015	4895	0.02	338	0.022	2593	0.013
D	S	214	0.005	12	0.002	71	0.003	27	0.005	143	0.007
	M	4793	0.011	135	0.009	2002	0.008	270	0.018	2793	0.014
E	S	220	0.005	20	0.004	73	0.003	23	0.004	147	0.007
	M	4875	0.011	161	0.011	2077	0.009	254	0.017	2798	0.014
F	S	3714	0.08	346	0.067	1907	0.075	436	0.084	1807	0.085
	M	39,170	0.089	1533	0.101	21,021	0.088	1133	0.075	18,149	0.091
G	S	3402	0.073	370	0.072	2101	0.083	228	0.044	1301	0.061
	M	36,316	0.083	1103	0.073	20,892	0.087	928	0.061	15,426	0.077
H	S	289	0.006	48	0.009	86	0.003	71	0.014	203	0.01
	M	3835	0.009	174	0.011	1621	0.007	168	0.011	2214	0.011
I	S	6216	0.133	794	0.154	3455	0.136	569	0.11	2761	0.13
	M	51,703	0.118	1651	0.109	28,959	0.121	1531	0.101	22,744	0.114
K	S	678	0.015	17	0.003	92	0.004	278	0.054	586	0.028
	M	5000	0.011	117	0.008	1517	0.006	423	0.028	3483	0.017
L	S	9456	0.203	965	0.187	5653	0.223	964	0.186	3803	0.179
	M	73,628	0.168	2523	0.167	40,961	0.171	2395	0.158	32,668	0.163
M	S	1164	0.025	141	0.027	591	0.023	132	0.026	577	0.027
	M	14,984	0.034	549	0.036	8211	0.034	535	0.035	6773	0.034
N	S	457	0.01	46	0.009	147	0.006	59	0.011	310	0.015
	M	8464	0.019	268	0.018	4205	0.018	367	0.024	4259	0.021
P	S	901	0.019	144	0.028	405	0.016	60	0.012	496	0.023
	M	14,127	0.032	494	0.033	7393	0.031	374	0.025	6734	0.034
Q	S	369	0.008	33	0.006	131	0.005	62	0.012	238	0.011
	M	5947	0.014	211	0.014	2836	0.012	268	0.018	3111	0.016
R	S	593	0.013	23	0.004	59	0.002	295	0.057	537	0.025
	M	4676	0.011	115	0.008	1292	0.005	559	0.037	3384	0.017
S	S	2307	0.049	255	0.049	1312	0.052	200	0.039	995	0.047
	M	26,929	0.061	835	0.055	15,359	0.064	881	0.058	11,573	0.058
T	S	2299	0.049	254	0.049	1269	0.05	171	0.033	1030	0.048
	M	22,018	0.05	767	0.051	12,372	0.052	769	0.051	9646	0.048
V	S	6128	0.131	749	0.145	3761	0.148	522	0.101	2375	0.112
	M	45,966	0.105	1683	0.111	25,960	0.109	1526	0.101	20,008	0.1
W	S	891	0.019	186	0.036	230	0.009	157	0.03	661	0.031
	M	9426	0.021	423	0.028	4152	0.017	344	0.023	5274	0.026
Y	S	1383	0.03	188	0.036	372	0.015	304	0.059	1011	0.047
	M	14,657	0.033	658	0.043	7088	0.03	684	0.045	7571	0.038
Total	S	46,607		5170		25,341		5170		21,285	
	M	439,181		15,150		238,947		15,150		200,249	

B. *The kPROT scale*

Residue	Total TM	Extracellular	Central	Intracellular	Both termini
A	0.02±0.02	0.03±0.01	0.09±0.01	−0.18±0.00	−0.09±0.01
C	0.27±0.03	−0.53±0.01	0.12±0.03	0.61±0.02	0.50±0.01
D	−0.87±0.08	−1.35±0.08	−1.10±0.04	−1.23±0.05	−0.73±0.07
E	−0.86±0.08	−1.01±0.08	−1.10±0.04	−1.33±0.04	−0.70±0.06
F	−0.11±0.02	−0.41±0.01	−0.16±0.01	0.12±0.01	−0.07±0.01
G	−0.12±0.02	−0.02±0.01	−0.05±0.01	−0.33±0.01	−0.23±0.01
H	−0.34±0.06	−0.21±0.05	−0.69±0.04	0.21±0.04	−0.15±0.05
I	0.12±0.01	0.34±0.01	0.12±0.01	0.09±0.00	0.13±0.01
K	0.25±0.04	−0.85±0.03	−0.56±0.02	0.66±0.04	0.46±0.03
L	0.19±0.01	0.11±0.01	0.26±0.01	0.17±0.00	0.09±0.00
M	−0.31±0.03	−0.28±0.02	−0.39±0.02	−0.32±0.01	−0.22±0.01
N	−0.68±0.05	−0.69±0.04	−1.11±0.03	−0.75±0.03	−0.38±0.04
P	−0.51±0.03	−0.16±0.03	−0.66±0.02	−0.75±0.02	−0.37±0.02
Q	−0.54±0.05	−0.78±0.05	−0.83±0.04	−0.39±0.03	−0.33±0.04
R	0.18±0.04	−0.53±0.03	−0.84±0.02	0.44±0.04	0.40±0.04
S	−0.21±0.02	−0.11±0.01	−0.22±0.02	−0.41±0.01	−0.21±0.01
T	−0.02±0.02	−0.03±0.01	−0.03±0.02	−0.43±0.01	0.00±0.01
V	0.23±0.01	0.27±0.01	0.31±0.01	0.00±0.00	0.11±0.01
W	−0.12±0.03	0.25±0.03	−0.65±0.02	0.29±0.02	0.16±0.02
Y	−0.12±0.03	−0.18±0.02	−0.70±0.02	0.26±0.01	0.23±0.02

A, The number of occurrences (numb), and fraction (frac) of each amino acid in the non-redundant protein set in single (S) and multi (M)-span proteins in the total TM, in the five positions of the extracellular/intracellular termini, in the central portion of the TM, and in the ten grouped positions of both termini. Note, that number of residues in the both termini category is larger than the sum of the extracellular and intracellular categories. This is because the latter are derived from topologically annotated sequences only, whereas the former is computed for all transmembrane proteins in the database. B, The kPROT values derived from A with standard deviation margins. The Total TM column constitute the one-way scale, the columns Central, and Both termini are the two-way scale, columns Extracellular, Central, and Intracellular constitute the three-way scale.

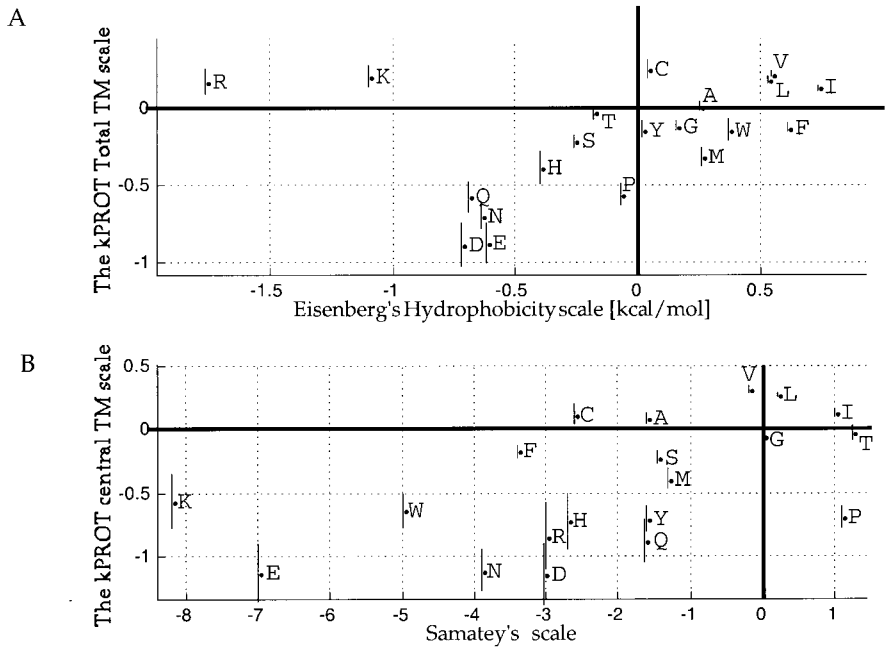

Figure 1. (a) A comparison of the total-TM kPROT scale and Eisenberg's hydrophobicity scale (Eisenberg *et al*., 1982). The error bars depict the statistical errors on the kPROT values. The correlation between the two scales is 0.28, (and 0.8 if Arg and Lys are omitted). The correlation between the total TM-kPROT and other hydrophobicity scales are: kPROT-Engelman *et al*. (1986) 0.37; kPROT-Kyte & Doolittle (1982) 0.55; and kPROT-Taylor *et al*. (1994) 0.55. (b) A comparison between the TM center kPROT scale and the normalized Samatey *et al*. (1995) scale. The correlation between the two scales is 0.57, the average correlation between the Samatey scale and a set of other hydrophobicity scales is 0.52 (Samatey *et al*., 1995).

the number of TM segments. This behavior likely reflects an increase in the fraction of the helix that is buried in the interior of the bundle as the number of spans increases. The observed plateau at a value of three TM segments suggests that helical bundles with more than three spans may be concave, and composed of small internal bundles of typically 3 TM segments, as observed, for example in the structure of bacteriorhodopsin (Henderson *et al*., 1990) and rhodopsin (Scherlter *et al*., 1993).

For each of the 20 residues we also calculated a probability function for the number of appearances (White, 1994) in the TM segments of single and multiple-span proteins. In general, the densities display monotonic decrease with number of appearances for the least represented residues and a more symmetrical form for the more abundant ones (see Figure 2(b) for four representative residues: Gln, Ser, Val and Leu, and the kPROT WWW server† for all 20 residues). Gln and Ser (two residues with negative kPROT values) display a higher probability for a high number of appearances in multi-span proteins than in single spans. On the other hand, Val and Leu (which have a positive kPROT value) display a higher probability to occur multiple times in a TM in single-span proteins than in multi spanners. Taking for each of the 20 residues the ratio of the means of the two densities, in single and multiple spans, results in an alternative scale that is highly similar to the kPROT scale (correlation=0.97, not shown).

The position-dependent kPROT scale

A refinement of kPROT is achieved in the ''three-way'' position-dependent scale in which each residue is assigned a different value depending on its belonging to each of three TM sub-sections: intracellular terminus, extracellular terminus and the remaining center of the TM segment (see Materials and Methods). Figure 3 and Table 1 display the three-way kPROT scale. In the two-way kPROT scale, the TM segment is divided into two sections: the TM center and both intracellular and extracellular termini.

Figure 1(b) displays a comparison between the kPROT of the central sub-section of the TM segment and the scale used by Samatey *et al*., also derived from this TM sub-section (Samatey *et al*., 1995). The two scales display an overall agreement on the orientation propensity of many of the residues. In particular, the tendency of the aromatic residues to face the protein interior, seen in the kPROT scale, is clearly seen in the Samatey scale. The two scales do not agree, however, with respect to the orientation propensities of several residues. Val, Ala and Cys are assigned by kPROT a high tendency to face the lipid and an opposite tendency by the Samatey scale. On the other hand, Thr and Pro appear as having a higher tendency to be buried by kPROT while by the Samatey scale

† http://bioinfo.weizmann.ac.il/kPROT

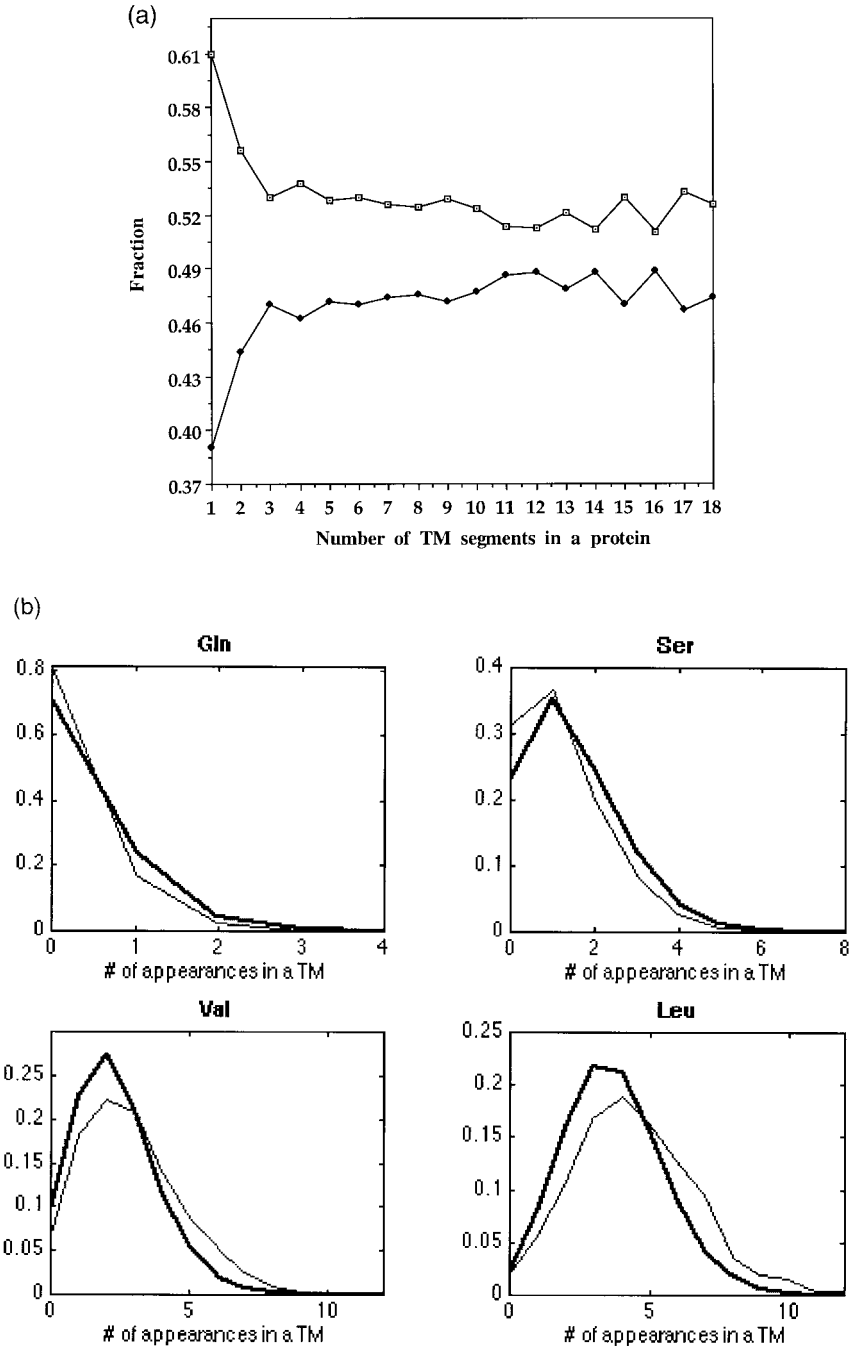

Figure 2. (a) Fraction of all residues with kPROT value larger than zero, i.e. Ala, Cys, Ile, Lys, Leu, Arg, Val (open squares) and all residues with kPROT values smaller than zero, i.e. Asp, Glu, Phe, Gly, His, Met, Asn, Pro, Gln, Set, Thr, Trp, Tyr (filled squares) in the TM segments of all proteins in the non-redundant set, as a function of the number of TM segments in the protein. The very few proteins with more than 18 TM segments are not included. (b) Density functions depicting observed probabilities of number of occurrences in a TM segment of individual residues. The densities were separately calculated for the sets of single (thin line) and multiple (thick line) span proteins. Four representative residues are shown here, the rest are available on the kPROT WWW server.

they appear as having a higher preference to face the lipid.

Regarding the three-way kPROT scale, several residues change their transmembrane helix orientation propensities as a function of their location along the TM segment. In particular, the aromatic residues display an interesting behavior: while all three clearly prefer facing the interior of the protein in the central portion of the TM, they show a higher propensity to face the lipid head-groups at either or both TM termini. Although derived from SWISS-PROT annotation, which itself is partially based on predictions, the kPROT propensities of the three aromatic residues at both TM termini are qualitatively in agreement (data not shown) with propensities estimated from a set of proteins with

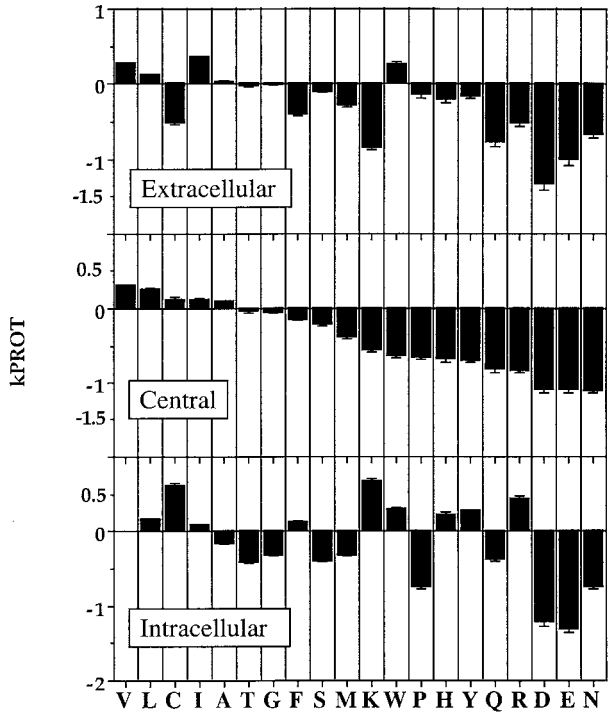

Figure 3. The position-dependent kPROT scales. The transmembrane helix orientation propensity values of each residue in the extracellular terminus of the TM, the central portion of the TM, and the intracellular terminus are shown with their statistical errors.

experimentally determined TM boundaries and topologies (Jones *et al.*, 1994a). The propensity of Trp and Tyr to face the lipid at the TM ends is in agreement with observations from solved membrane protein structures, where they constitute the so-called aromatic belt (von Heijne, 1996; Preusch *et al.*, 1998). The high affinity of Trp to lipid-water interfaces was ascribed to its pi electronic structure and associated quadrupolar moment, which favor residing in the electrostatically complex environment of the interface (Yau *et al.*, 1998).

The positively charged residues Arg, Lys and His display a high preference to face the lipid only when located at the cytoplasmic end of the TM segment. This tendency may be ascribed to electrostatic anchoring of the TM segments onto negatively charged lipid head-groups, which in eukaryotes occurs mainly on the intracellular leaflet (Monne *et al.*, 1998).

The derived kPROT values are insensitive to the somewhat arbitrary definition of the boundaries between TM sections. Considering TM ends of four residues (as previously suggested (Jones *et al.*, 1994a)) or six residues, as an alternative to the five residues definition chosen here, results in almost identical scales (correlations >0.98). In addition, the positive kPROT values observed for Lys and Arg at the intracellular terminus are observed even when extending the TM segments of the multi-span proteins towards the intracellular side by up to three residues beyond the SWISS-PROT annotation of helix ends (not shown).

We compared the position-dependent three-way kPROT scale with four classes of propensity scales (including three hydropathy scales) by drawing a similarity dendrogram (Figure 4). These include: (1) classical hydrophobicity scales (Eisenberg *et al.*, 1982; Kyte & Doolittle, 1982; Engelman *et al.*, 1986); (2) sequence-based scales (Taylor *et al.*, 1994; Samatey *et al.*, 1995); (3) a structure-based scale for facing the interior of water-soluble globular proteins (Miller *et al.*, 1987); (4) a scale based on partitioning between water and water-lipid interfaces (Wimley & White, 1996). It is apparent that while the TM-terminal kPROT scales cluster with the water-membrane interface scale, kPROT for the TM-center is more akin to the other three scale classes. Within the latter branch of the dendrogram, the central kPROT is positioned closer to the other scale based on multi-span membrane protein sequence statistics, suggesting that both capture properties beyond simple hydropathy. This is in contrast to the scale for globular proteins, which is almost indistinguishable from hydrophobicity scales. It may thus be suggested that the position-dependent kPROT, which includes all of the kPROT segmental scales, reflects a balanced combination of physicochemical properties, so as

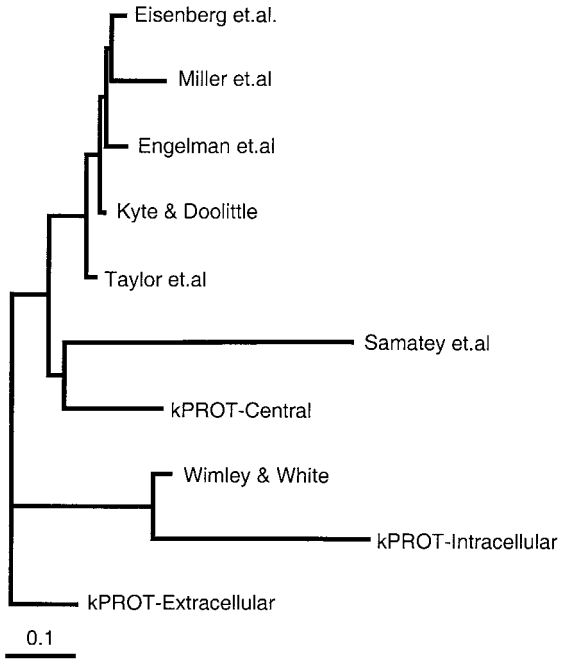

Figure 4. A dendrogram depiction of mutual similarities between kPROT and other propensity scales. The "distance" between pairs of scales was defined as $1 - c$, where c is the correlation coefficient between scales. The data for the dendrogram were generated with the FITCH program, of the PHYLIP package (Kuhner & Felsenstein, 1994), which implements the Fitch-Margolias algorithm for tree construction by a least-squares fit to a distance matrix. The tree was rendered with the TREEVIEW program (Page, 1996).

to render it an optimal tool for TM orientation prediction.

Benchmark testing

We assessed the accuracy of the kPROT and several other scales in predicting the helical angular orientation of the TM segments in seven proteins with experimentally determined structure. Figures 5 and 6 show in detail the moment analyses for three selected membrane protein families. Results for the rest of the benchmark proteins are available at the kPROT WWW server. Figure 5(a) shows the results obtained for bacteriorhodopsin with a set of its homologs displayed as helical moments superimposed on the true structure. Despite a scatter in the position-dependent kPROT moment directions among individual homologs, the average moment shows a good agreement with the experimental membrane-facing vectors (mean error of 25°). An alternative method, in which a family moment is obtained based on a consensus sequence of each helix, may be used (Figure 5(a)).

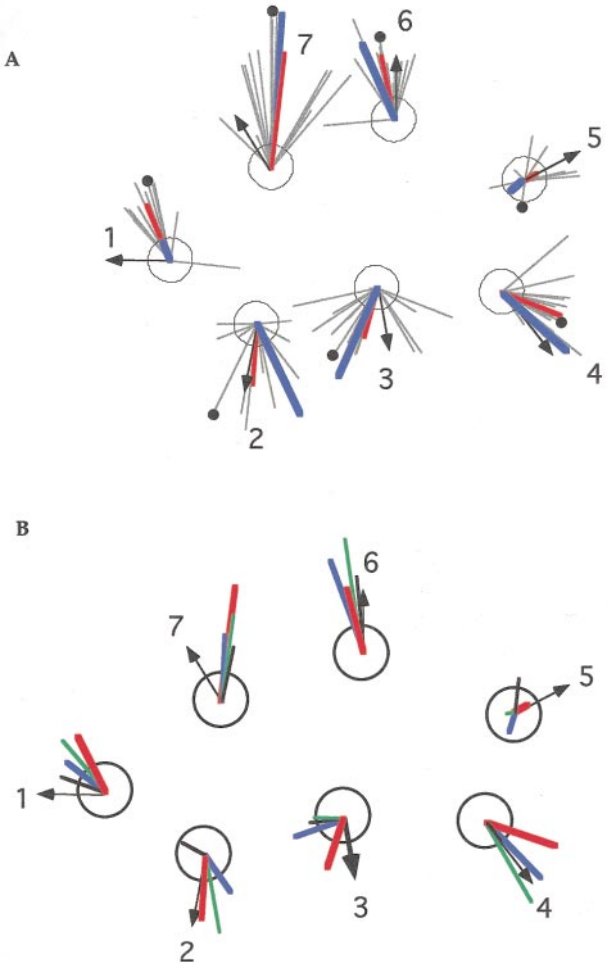

Figure 5. A depiction of the benchmark test for bacteriorhodopsin. The ''correct'' structure-based membrane-facing vectors are indicated in each TM as thin arrows. (a) Helical moments calculated by the three-way kPROT scale for bacteriorhodopsin (SWISS-PROT ID: BACR_HALHA, marked with a filled circle) and 18 other homologs. Moments are shown for each homolog (black thin lines), for the average of all 19 sequences (red thick lines), and for the consensus sequence of the alignment (blue thick lines). (b) Helical moments computed on the average of the 19 homologs by kPROT (red lines) and alternative scales, (Samatey *et al.*, 1995) (black), (Kyte & Doolittle, 1982) (blue), and (Taylor *et al.*, 1994) (green). Such maps are available on the kPROT WWW server for the rest of the benchmark proteins.

Figure 5(b) depicts a comparison of the kPROT scale to other representative published scales. While in all cases the moments generated using the kPROT scale face the membrane, one or two helices are wrongly oriented when using each of the alternative scales. Figure 6(a) and (b), respectively show the kPROT prediction for individual TM segments of two additional proteins, glycophorin homodimer and the mechanosensitive ion channel. The predicted helical moments, generated with the tested scales, are superimposed on a helical wheel. The seven amino acid positions implicated in forming the helix-helix interactions in the glycophorin A dimer, which should ideally be facing away from the predicted helical moment, are highlighted (Figure 6(a)).

Figure 7(a) shows a summary of the prediction accuracy of different propensity scales. The average error angles obtained by kPROT is 41(±16)°. The Samatey scale obtained an average error of 61(±27)°, while the errors obtained with the rest of the hydrophobicity scales are in the range of 65°-68°. Next, we repeated the benchmark test and omitted from the sequence of each TM segment the five residues at each of its termini. We compared the prediction accuracy of these central segments as obtained by two scales derived for the central section of TM segments, namely the TM center kPROT scale and the Samatey scale. While the prediction accuracy obtained by the TM center kPROT was lower than that obtained with the three-way kPROT scale 46(±13)°, the prediction accuracy obtained by the Samatey scale has improved to 56(±29)°.

Figure 7(b) shows the average percentage of TM segments in each protein correctly predicted to face the lipid by each of the alternative scales. It is seen that with the kPROT scale almost all helical moments are correctly predicted to face the membrane, while with other scales more helices are incorrectly predicted to face the protein interior.

The ideal propensity scale should maximize the amplitude of the helical periodicity moments of interfacial helices (Eisenberg *et al.*, 1984; Cornette *et al.*, 1987). This provides another means for comparing the different scales, which may be applied to all membrane protein sequences, rather than only to those with experimentally determined structures. We have calculated the α-helical periodicity value (AP), a measure for the intensity of the helical periodicity relative to all other periodicities (Komiya *et al.*, 1988; Donnelly *et al.*, 1993) according to the different scales, for the TM segments of the multi-span set of topologically annotated pro-

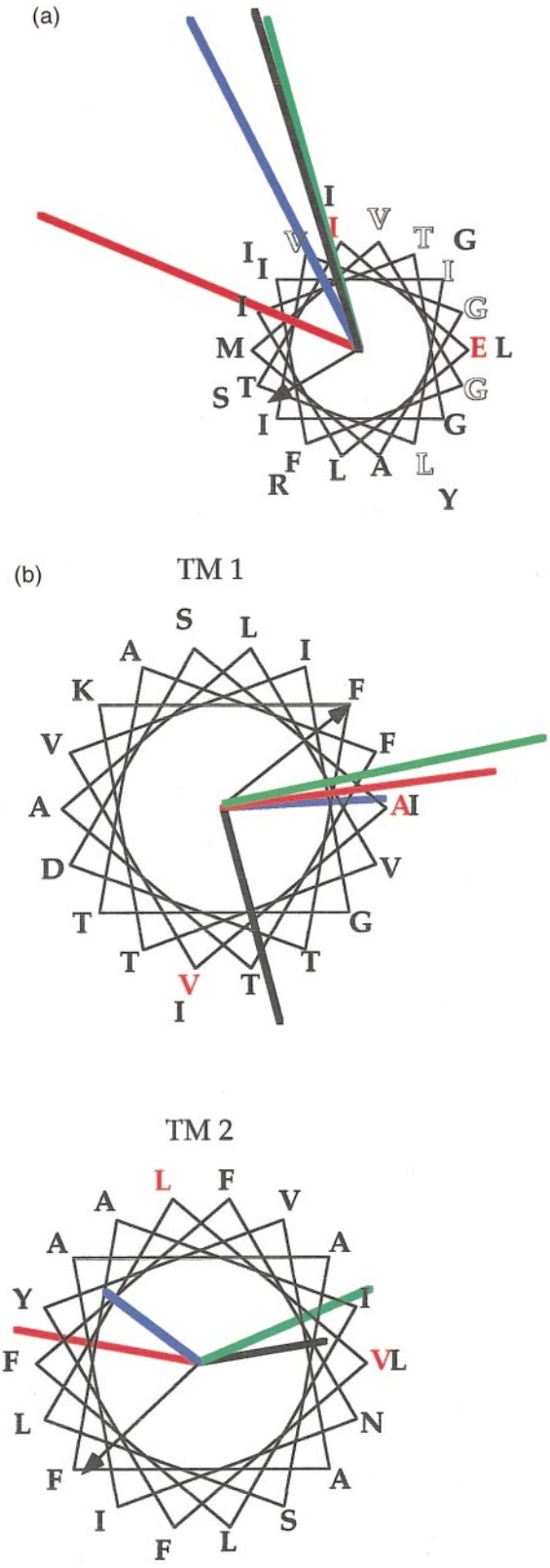

Figure 6. Benchmark test on (a) glycophorin A dimer, and (b) the mechanosensitive ion channel, applied to individual TM segments. The results for these proteins are shown in the output format of the kPROT WWW server. The helical moments of the various scales (colored as in Figure 5(b)) are drawn on a clockwise helical wheel for TM segments with extracellular N terminus and counterclockwise wheel for helices with the opposite topology. Predictions are done for each protein and its set of homologs, and shown for the averaged family moment only, as in Figure 5(b). The sequence shown is that of the first protein in the alignment, and the first two residues are colored red. The correct membrane-facing vectors are superimposed as black arrows as in Figure 5. The seven amino acid residues implicated in forming the helix-helix interactions in the glycophorin A dimer are highlighted (a).

teins in the SWISS-PROT database (Figure 8, black bars). It may be seen that the average AP values calculated using kPROT are larger than those calculated using all the other scales. This indicates that the kPROT scale accounts better for the helical distribution of the residues in TM segments. Because the Samatey scale was derived using only the central portion of TM segments, Figure 8 shows a comparison in which AP values are computed for the central TM region only, using the kPROT scale of the TM center sub-segment (Table 1), and the other propensity scales (Figure 8, gray bars). Under these more restrictive conditions, the AP values for all scales are more similar, with insignificantly higher AP values obtained with kPROT and the Kyte & Doolittle and Samatey scales.

Discussion

Amino acid residue hydrophobicity is well correlated with aqueous exposure in globular proteins (Eisenberg *et al.*, 1982; Honig & Nicholls, 1995), as may be seen from the dendrogram in Figure 4. According to a central paradigm, membrane proteins were considered ''inside-out'' proteins, in that they have a polar core while apolar residues are exposed to the membrane (Rees *et al.*, 1989). This notion has recently been challenged, based on an analysis of several available membrane protein structures (Stevens & Arkin, 1999). The reason for this may be the fact that for integral membrane proteins both the interior and the lipid-exposed regions tend to be hydrophobic. Therefore, other properties of the residues should be searched, that are sensitive to the differences between these two environments.

The chemical environment of the residues in TM segments is highly complex (White & Wimley, 1994). When exposed to the membrane, residues may interact with the different bilayer hydrophobic core components, or with diverse polar lipid headgroups. When facing the protein interior, amino acid side-chains may interact with those on other TM segments, with water molecules, or with functional ligands. Knowledge-based approaches constitute an effective way to capture such intricacies. However, because the number of solved membrane protein structures is rather limited, an alternative strategy that utilizes the vast number

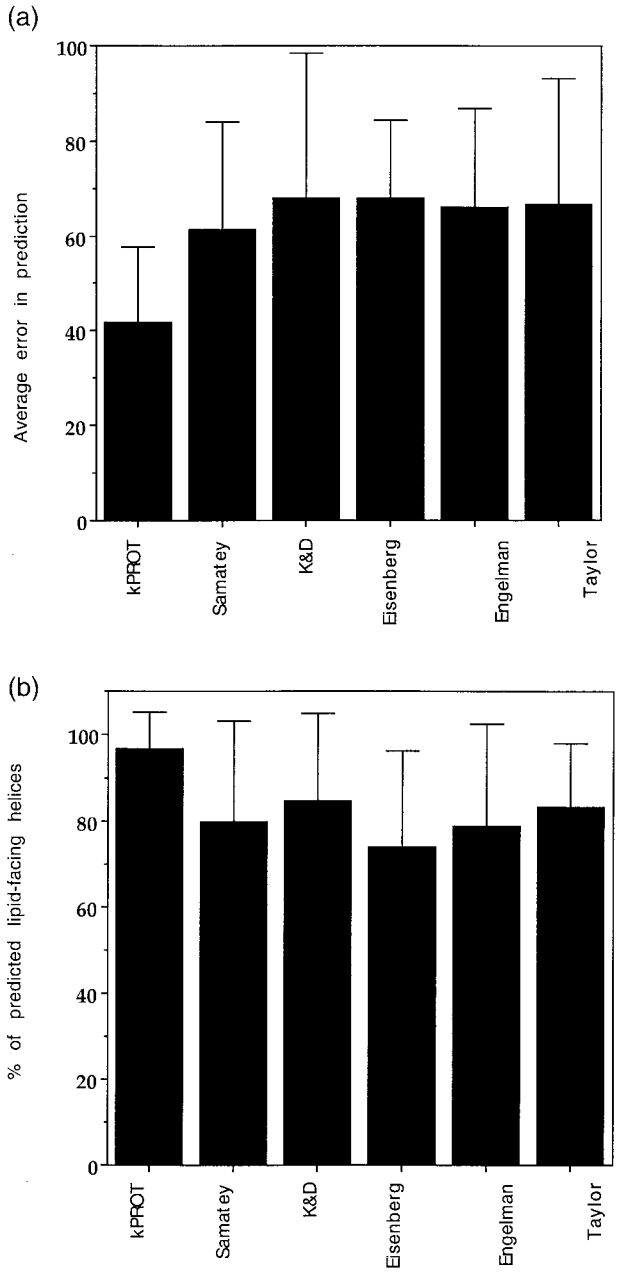

Figure 7. Summary of the benchmark. Helical moments, averaged over the benchmark proteins, using the kPROT scale, and the scales of Samatey *et al.*, Kyte & Doolittle, Taylor *et al.*, and Eisenberg *et al.* (a) The averaged error in predictions is indicated by degrees of deviation between the predicted and membrane facing moments, with error bars for standard deviations. (b) Average percentage of helices in each protein with helical moments correctly predicted to face the membrane, with error bars for standard deviations. A helical moment was considered as ''correctly predicted'' if it was directed towards the membrane-facing wheel section of the angle formed between the center of the helix and the centers of its ''left'' and ''right'' helix neighbors.

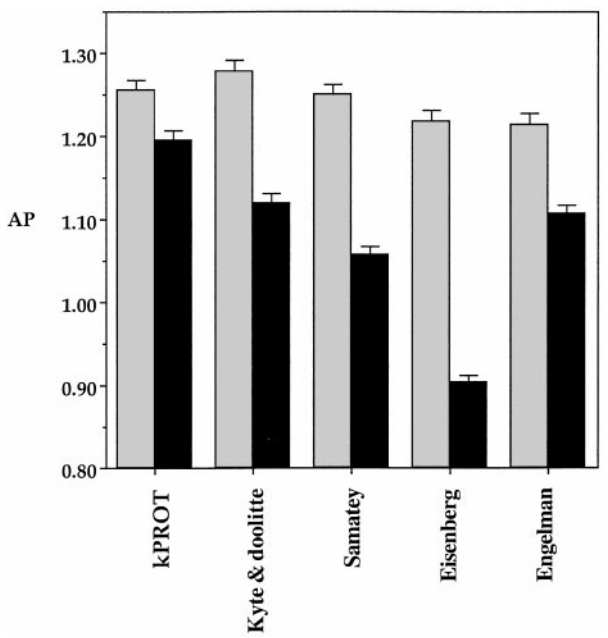

Figure 8. The average α-helical periodicity (AP) index of all 3041 TM segments of the topologically annotated multi-span set are shown for each of the prediction scales, for whole TM segments (black bars) and for TM centers (gray bars). The standard deviations are shown as error bars.

of available protein sequences would be highly beneficial.

The kPROT scale reported here is aimed at providing such an alternative route. It is based on the idea that, since a transmembrane bundle core exists only in multi-span, but not in single-span membrane proteins, frequency ratios may be used to predict lipid exposure, as was reasoned for proline-induced kinks (von Heijne, 1991). When applied to the sequences of all multi-span α-helical membrane proteins with known structure, the kPROT prediction was in good agreement with the experimentally determined helical orientations. A comparative benchmark demonstrated a pronounced advantage of kPROT over existing propensity methods, including both hydropathy-based and spectrum-based scales. The degree of accuracy obtained by kPROT is comparable to that reported by Donnelly *et al.* (1993) in the prediction of helix orientation for bacteriorhodopsin, using helical conservation moments or a substitution matrix for lipid-exposed residues in photosynthetic reaction centers. Two clear limitations of the alternate methods are that they are applicable only when multiple homologous sequences are available, and that judgement must be exerted on where to point the conservation moments. While for bacteriorhodopsins and some seven other TM protein families, a conservation-inside assumption is natural, there are other examples, e.g. the olfactory receptors, in which the bundle interior may be highly variable (Pilpel & Lancet, 1999).

While the kPROT scale does not rely on the availability of multiple homologs, its performance may improve when such protein families are analyzed, as is the case for other predictive schemes

(Jones *et al*., 1994a; Persson & Argos, 1994; Rost *et al*., 1995). Notably, as exemplified in Figure 5 for bacteriorhodopsins, and observed in the other proteins in the benchmark test, the scattering seen in the values of the kPROT moments for aligned homologous sequences is rather small, suggesting correlated multiple sampling. Significantly, the pairwise distances in the BLOSUM matrix used for the alignment is not necessarily correlated with the metric of kPROT.

An application of the kPROT scale for *de novo* prediction may ideally be attempted for proteins whose low-resolution two-dimensional map of the TM bundle is available, cases that are now accumulating rapidly (Heymann *et al*., 1997). One such case is the aquaporin water channel for which a low-resolution structure is available (Cheng *et al*., 1997; Walz *et al*., 1997). We computed a model of helical orientations using the kPROT scale for this protein (unpublished). This model is in agreement with sequence-based predictions of functional residues in the protein (Froger *et al*., 1998; Heymann *et al*., 1998). Thus, by combining kPROT analysis with previously available low-resolution structure and with topological constraints, it is possible to infer an adequate structural model.

Even when low-resolution data are not available, an orientation prediction for each individual helix may still be done. Such predictions lead to the identification of functional sites, usually located within the transmembrane bundle interior. In cases in which even the TM topology is not known, the three-way position-dependent kPROT may not be used, and the two-way scale has to be applied.

Limitations of the kPROT scale approach

One major difficulty of the present method is that the kPROT scale is derived based on all single-span proteins. This includes an unknown proportion of proteins that form homo- or hetero-oligomers in the membrane, e.g. fibroblast growth factor, nerve-growth factor and T-cell receptor families. Residues that tend to appear at the oligomerization interface, and which are thus markers of buried helix faces, could erroneously be seen as lipid-exposed. In the future, it would be important to perform fine tuning for the kPROT scale by including specific information stemming from experimental data on oligomerizing bitopic proteins (MacKenzie *et al*., 1997; MacKenzie & Engelman, 1998). Still, a large fraction of the circumference of dimerizing helices are probably exposed to lipid, supporting the inclusion of such proteins in the set of single-spanners.

The second difficulty is in the application of the kPROT scale to helix orientation prediction of oligomerizing multi-span proteins. In such cases the dichotomy between being buried and lipid-exposed may be obscured, and difficulties could arise in determining the orientation of at least some of the helices. Bacteriorhodopsin is an example of such a protein that appears as a trimer in the crystalline purple membrane. Interestingly, despite this difficulty, the kPROT prediction accuracy for bacteriorhodopsin was not lower for the helices (2, 3 and 4) at the protein-protein interface. This may reflect an equilibrium between the monomeric and the trimeric forms of this protein (Gulik-Krzywicki *et al*., 1987), or the inclusion of small amounts of lipid that was suggested to act as a glue in the formation of the trimer (Essen *et al*., 1998; Sato *et al*., 1999). A general implication of this limitation to the kPROT scale performance is that for oligomerizing proteins the prediction for some helices may be compromised.

An additional difficulty, inherent to the logic of the kPROT approach, would arise if certain residues had a significant tendency to be exposed to the membrane preferentially in multi-spanning proteins. Such residues would appear as having a tendency to be buried. While such events cannot be totally eliminated, we have taken measures to minimize their effect. This is done through the use of a stringent elimination of homologs ($\sim$40 % identity cutoff) in the non-redundant protein set used for deriving the kPROT scale. This way, the influence of conserved exposed residues in protein families would be negated. Examples of potential kPROT biases of this kind might be manifested for the residues Gly and Pro, both appearing as buried in kPROT. Proline is an example of how an amino acid could potentially be misassigned in kPROT: it might be essential for the construction of a multi-span bundle, but potentially excluded from its interior. However, in agreement with the kPROT prediction, proline is actually found to be buried in several helices of solved membrane protein structures, as rationalized by the requirement that the unsatisfied hydrogen-bond at position $i-4$, on the same helix face, should not be membrane-exposed (von Heijne, 1991).

An additional potential problem in kPROT approach is that it relies on sequence-based prediction of TM boundaries by other algorithms. Future improvement of the computational and experimental TM annotation will therefore allow the generation of more accurate kPROT scales.

Comparison with alternative scales

Although developed by a completely different mathematical approach, the power spectrum-based scale (Samatey *et al*., 1995), and the kPROT scale of the central TM portion agree on the orientation propensities of many of the residues. In particular, both scales predict that the aromatic residues should be buried at the central section of the TM segment. Such preferences were explained by the incompatibility of these bulky residues with the aliphatic environment of the membrane. In addition, it was argued that the aromatic residues may be preferred in the interior of the protein, as they can form intra-protein interactions or interact with other aromatic ligand counterparts (Samatey *et al*., 1995).

The major likely source of the discrepancies between the kPROT and the Samatey scale is in the set of proteins used to derive each of the scales. The kPROT set is highly heterogenic, it includes eukaryotic proteins from the various organelles, in addition to prokaryotic and archeal proteins, while the set of proteins used by Samatey *et al*. is composed of eukaryotic plasma membrane proteins only. The number of unique proteins used in kPROT is >100 times larger than the number used by Samatey *et al*. In addition, Samatey *et al*. selected for their analysis only segments that display high lateral asymmetry. We tested whether the discrepancies between the two scales partially arise from these differences. For that we derived a kPROT scale from sequences of eukaryotic plasma membrane proteins only (not shown). The Samatey scale was found to be more similar to the plasma membrane proteins kPROT than to the general kPROT scale (correlation coefficients of 0.7 and 0.57, respectively). Another potential source of difference is the reliance of the kPROT scale on multi-span proteins and on single-spanners. Possible inaccuracies in kPROT due to the use of dimerizing single-spanners, e.g. regarding Gly, as discussed above, may also account for differences between the two scales. On the other hand, a potential source of distortion in the Samatey scale may be errors in determining helix angular orientations, which are not inferred directly from their data.

A crucial factor that allows the proposed kPROT scale to account for the complex membrane environment is the assignment to each residue a propensity value for each of the three positional segments along the helix. Such refinement, in which we use sub-segments as short as five residues, or potentially even shorter, could not be attained in power spectrum-based scales that require a minimal segmental length (more than two helical turns).

Taylor *et al*. (1994) have proposed the use of a scale for the preference of residues to be present in the middle section of single TM segments as compared to the entire sequence of the single-span protein set. This scale was found optimal for predicting the location of TM segments along the primary structure (Jones *et al*., 1994a). Thus, while in the kPROT scale the preference of the residues to be buried is estimated by their enhanced presence in multi-span TM segments, in the Taylor scale such preferences are estimated from the residue composition in the extramembrane loops and in the TM segment termini. Consequently, residues that show enhanced tendency to be buried according to the kPROT and Samatey scales, e.g. the aromatic, have a considerably lower propensity to be buried according to the Taylor scale. Indeed the latter scale is very similar to the classical hydrophobicity scales (Figure 4).

An attempt to use kPROT for generating straightforward hydrophobicity profiles for locating TM segments along several sequences (data not shown) suggested that kPROT was less accurate than classical hydrophobicity scales. Thus, the picture that emerges is that of scale specialization, whereby hydrophobicity scales are better tuned to distinguish between lipid and water exposure propensities, while the kPROT scale is better at differentiating membrane-exposed from protein-buried residues.

A potential physico-chemical correlate for such a specialization of the different scales may be discussed in view of the ''two-stage model'' for the folding of integral membrane proteins (Popot & Engelman, 1990). According to this model, in the first stage of the folding process hydrophobic α-helices are established across the lipid bilayer. In the second stage they interact to form functional transmembrane bundles. While hydropathy scales are obviously related to the first stage, kPROT and the Samatey power spectrum-based scale are likely related to the second stage: negative values, of both scales, may indicate more than a mere tendency to avoid lipid exposure. Such propensities may reflect a specific role in the second stage of the folding process; namely, assembly of the polypeptide in the membrane by directing molecular recognition events between transmembrane elements. Residues such as the aromatic ones may participate in such intra-protein helix association (Samatey *et al*., 1995).

Further refinements for the kPROT approach would be to generate separate scales for membrane proteins from different cellular organelles, different phylogenetic kingdoms or different functional classes. Preliminary versions of such specific scales have been developed and some of these are publicly available on the kPROT WWW server.

Materials and Methods

Derivation of the kPROT scale

Transmembrane segment sequences were extracted from the SWISS-PROT database (release #35) (Bairoch & Boeckmann, 1991) with the Sequence Retrieval System (SRS) (Etzold *et al*., 1996) using the TRANSMEM key word in the feature (FT) annotation filed. For the derivation of the three-way position-dependent kPROT scale we retrieved protein entries for which the TM topology was annotated. Topology was inferred from the annotation of the extra/intracellular segments, flanking TM segments, using the key words EXTRACELLULAR and CYTOPLASMIC in the feature (FT) annotation filed. The SWISS-PROT accession numbers of all the analysed sequences is available at the kPROT WWW server.

For the derivation of the three-way position-dependent kPROT scale, we demarcated each TM segment into three sub-segments; namely, the five amino acid positions at the intracellular terminus, the five positions at the extracellular terminus and the remaining central portion of the TM segment. In the two-way kPROT scale, each residue is assigned with a propensity value in the center of the TM segment and with a value reflecting an averaged propensity to face the membrane at the two TM termini grouped together. The two-way scale did not

Table 2. Protein sequence statistics

Protein set	Single TM		Multiple TM		Single+Multiple	
	All	Topol	All	Topol	All	Topol
Original set	3682	2003	6634 (40,100)	2294 (14,333)	10,316 (43,782)	4297 (16,336)
Non-redundant set	**2164**	1034	**3322 (20,124)**	498 (3041)	**5486** (22,288)	1532 (4075)

Number of proteins (and number of TM segments in multi-span proteins) in the single and multiple TM categories and their sum. In each category the number of sequences in the original ''redundant set'', and in the non-redundant set are shown. The total number of sequences is designated as All and the sequences that are topologically annotated are designated Topol. In bold are the number of sequences (and transmembrane segments) used for the derivation of the one-way, two-way and the central TM portion of the three-way scales, which are not based on topological annotations; underlined are the number of sequences that contributed to derivation of the Extracellular and Intracellular components of the three-way scale.

The representation of the various protein families in the set of analysed proteins may be seen in the kPROT WWW server. The most highly represented annotated families in the set of single-span proteins are: cell adhesion proteins 97; glycosyltransferase 86; serine protease 24; metalloprotease 23; glycosidase 77; oxidoreductase 77 electron transport 76; tyrosine-protein kinase 67; EGF-like domain 38; calcium-binding proteins 38; respiratory chain proteins 37; extracellular matrix 23; MHC class I 23; mono-oxygenase 23; zymogen 22; serine/threonine-protein kinase 19. In the set of multi-spanning proteins, the most highly represented families are: transport proteins 801; G-protein coupled receptors 145; oxidoreductases 110; symporters 82; ionic channels 74; electron transport 72. A large portion of the proteins that served for this analysis are annotated in SWISS-PROT as hypothetical proteins.

require proteins with annotated topology. The number of proteins used in each category is listed in Table 2.

The average length of the TM segments in the SWISS-PROT sets of single and multi-span proteins was found here to be almost identical (25.81 ± 3.3 and 25.82 ± 2.2, respectively, the entire distribution of lengths is available on the kPROT WWW server). These observed lengths are in good agreement with an average of 26.4 observed in a set of 45 TM segments of multi-span proteins with experimentally determined 3D structure (Bowie, 1997).

Yet, many of the SWISS-PROT annotations are based on predictions, either by means of homology to known structures or by *ab initio* methods. This is clearly a source of error, especially regarding exact location of TM ends. The fact that SWISS-PROT annotation is based, as a standard, on a set of several independent, highly accurate (~95% accuracy) TM segment prediction schemes (Apweiler *et al*., 1997), minimizes the possibility of consistent errors in TM annotation.

Despite that, we introduced a change in the TM boundary annotation by adding one extra residue position to each TM segment at its cytoplasmic terminus. This modification was done because it resulted in a kPROT scale with enhanced prediction accuracy. This is mainly attributed to an enhanced tendency of Lys and Arg to face the membrane at the intracellular TM end when using the modified TM end definition. This was the only significant difference from the scale derived with exact annotated TM termini.

To avoid compositional bias due to an unequal representation of protein families in the database, we created a non-redundant set of TM sequences. This was done with the program PURGE of the BLAST package (Neuwald *et al*., 1995) with a similarity threshold of 100, which was found to correspond to ~40% identity in the TM segments.

In order to assess the statistical error of the kPROT values, we start by estimating the standard deviation of the means in the sets of single-span and multi-span proteins: $\sigma(f_s)$ and $\sigma(f_m)$ respectively, from their densities (Figure 2(b) and kPROT WWW server), as $\hat{\sigma}(f_s) = s_s/\sqrt{n_s}$ and $\hat{\sigma}(f_m) = s_m/\sqrt{n_m}$ where s_s and s_m are the observed standard deviations of each distribution, and n_s and n_m are the respective number of protein segments of each set in the sample (listed in Table 2).

For further propagating the errors from the frequencies to the kPROT formula we apply the relation:

$$\sigma(x \pm y) = \sqrt{(\sigma(x))^2 + (\sigma(y))^2} \tag{2a}$$

to the kPROT definition:

$$\text{kPROT} = \log\left[\frac{f_s}{f_m}\right] = \log(f_s) - \log(f_m) \tag{2b}$$

and thus evaluate the standard deviation on kPROT as:

$$\sqrt{(\sigma(\log(f_s))^2 + (\sigma(\log(f_m))^2} \tag{2c}$$

The standard deviation of the logarithm of f_s (and equally for f_m) was approximated by a first-order Taylor series:

$$\sigma(\log(f_s)) = \frac{\hat{\sigma}(f_s)}{f_s} \tag{2d}$$

For each of the benchmark proteins, we created a separate ''jack-knifed'' kPROT scale, in which we omitted from the initial set of multi-span sequences the sequence of the tested protein. The reported benchmark results of each protein were obtained with its respective jack-knifed scale. These modified scales are practically identical (correlation >0.99) with the general scale shown in Table 1B and in Figure 3, since they result from the deletion of only one sequence at a time from a set of 5486 non-homologous proteins.

Benchmark analysis of protein structures

A set of helical multi-spanning proteins with high-resolution determined 3D structures, used for the benchmark, was retrieved from a published list† by Stephen White. The proteins, with their PDB codes, are: bacteriorhodopsin 1BRD (Henderson *et al*., 1990); bacterial light harvest proteins II, 1LGH (Koepke *et al*., 1996) and 1KUZ (Prince *et al*., 1997); mitochondrial cytochrome oxidase, 1OCC (Tsukihara *et al*., 1996); potassium channel, 1BL8 (Doyle *et al*., 1998); mechanosensitive ion channel 1MSL (Chang *et al*., 1998); and glycophorin A dimer

† http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html

1AFO (MacKenzie *et al*., 1997). Two structures, of cytochrome *bc*1, (Iwata *et al*., 1998) and the photosynthetic reaction center (Yeates *et al*., 1987), were not included because of their more complex arrangement of TM helices. These structures are the subject of a separate in-depth kPROT analysis to be performed in our laboratory.

Moment calculations

The helical moment vector (Eisenberg, 1984) **M** was computed for the different propensity scales as the moment length |**M**| and the moment direction Θ relative to the angular direction of the α-carbon atom of the first amino acid residue in the TM segment, as follows:

$$| \mathbf{M} | = \sqrt{\sum_{i=1}^{n} (\mathbf{M}_i \cos \theta_i)^2 + \sum_{i=1}^{n} (\mathbf{M}_i \sin \theta_i)^2} \qquad (3)$$

and:

$$\Theta = \text{arctg}\left[\frac{\sum_{i=1}^{n} \mathbf{M_i} \sin \theta_i}{\sum_{i=1}^{n} \mathbf{M_i} \cos \theta_i}\right] \qquad (4)$$

where $\mathbf{M}_i$ are the propensity values according to a given scale of residue i in the sequence and the summation is done over the n amino acid residues in the TM segment. In using the two-way or three-way kPROT scales, the values **M** were taken from the corresponding columns in Table 1 according to the location of the residue in the TM segment.

We calculated the location of the centers of the TM helices in the two-dimensional (usually the *XY*) plane of the TM bundle (e.g. Figure 5) from the PDB file, as the mean of the *X* and *Y* coordinates of the α-carbon atoms of the TM-constituent amino acid residues.

Determination of angular orientations

For the purpose of benchmarking and for future *de novo* prediction of helical orientations, we define a ''membrane-facing vector''. For proteins with known 3D structure, this was previously defined as the solvent exposure moment (Donnelly *et al*., 1993). In order to have a standard that will apply equally to both benchmarking and *de novo* predictions, we use here an alternative definition whereby the membrane-facing vector of a given TM is the outward-facing bisector of the angle formed between the TM center and those of its two closest neighbors (which coincide in the case of proteins with only two spans). In the benchmark, the error is obtained as the angle between the predicted moment and the membrane-facing vector. In *de novo* orientation prediction (unpublished results), the helical TM segment is rotated so that the calculated moment coincides with the membrane-facing vector.

In cases of proteins containing fully buried TM segments, such as in light harvest protein, and the mitochondrial cytochrome oxidase, only interfacial TM segments were subject to benchmarking.

Availability

We have generated a WWW server (http://bioinfo.weizmann.ac.il/kPROT) that offers automatic prediction of TM helical orientations using the kPROT scale.

Acknowledgments

We thank Ephraim Katchalsky-Katzir, Moises Eisenberg, Daniel Segré, Gustavo Glusman, Shai Rosenwald, Ora Furman-Schueler, Yossef Kliger, Tidhar Zifer, Jean-Luc Popot, Burkhard Rost, Stephen White and Shmuel Pietrokovski for helpful discussions. We thank Jaime Prilusky for help in establishing the WWW server.

Doron Lancet holds the Ralph and Lois Silver Chair in Neurogenomics. Supported by grants to D.L. from the Ministry of Science (National Laboratory for Genome Infrastructure), the National Institutes of Health (DC00305), the Krupp foundation, the German-Israeli Foundation for scientific research and development. Also from by the Weizmann Institute's Crown Human Genome Center, and Glasberg, Levy, Nathan Brunschwig and Levine funds. Y.P was partially funded by the John F. Kennedy Memorial Fund Scholarship; N. B.-T.'s research was supported by grant number 96-228 from the United States-Israel Binational Science Foundation and by fellowships from the Wolfson and Alon Foundations.

References

Apweiler, R., Gateau, A., Contrino, S., Martin, M. J., Junker, V., O'Donovan, C., Lang, F., Mitaritonna, N., Kappus, S. & Bairoch, A. (1997). Protein sequence annotation in the genome era: the annotation concept of SWISS-PROT + TREMBL. *Intelligent Sys. Mol. Biol.* **5**, 33-43.

Bairoch, A. & Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **19**, 2247-2249.

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

Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. (1998). Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. *Science,* **282**, 2220-2226.

Cheng, A., van Hoek, A. N., Yeager, M., Verkman, A. S. & Mitra, A. K. (1997). Three-dimensional organization of a human water channel. *Nature,* **387**, 627-630.

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

Cronet, P., Sander, C. & Vriend, G. (1993). Modeling the transmembrane seven helix bundle. *Protein Eng.* **6**, 59-64.

Cserzo, M., Wallin, E., Simon, I., von Heijne, G. & Elofsson, A. (1997). Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng.* **10**, 673-676.

Donnelly, D., Overington, J. P., Ruffle, S. V., Nugent, J. H. A. & Blundel, T. L. (1993). Modeling α-helical transmembrane domains: the calculation and use of substitution tables for lipid facing residues. *Protein. Sci.* **2**, 55-70.

Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T. & MacKinnon, R. (1998). The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science,* **280**, 69-77.

Eisenberg, D. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl Acad. Sci. USA,* **81**, 140-144.

Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1982). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature,* **299**, 371-374.

Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125-142.

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

Essen, L., Siegert, R., Lehmann, W. D. & Oesterhelt, D. (1998). Lipid patches in membrane protein oligomers: crystal structure of the bacteriorhodopsin-lipid complex. *Proc. Natl Acad. Sci. USA,* **95**, 11673-11678.

Etzold, T., Ulyanov, A. & Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.* **266**, 114-128.

Froger, A., Tallur, B., Thomas, D. & Delamarche, C. (1998). Prediction of functional residues in water channels and related proteins. *Protein Sci.* **7**, 1458-1468.

Gulik-Krzywicki, T., Seigneuret, M. & Rigaud, J. L. (1987). Monomer-oligomer equilibrium of bacteriorhodopsin in reconstituted proteoliposomes. A freeze fracture electron microscope study. *J. Biol. Chem.* **262**, 15580-15588.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E. & Downing, K. H. (1990). Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **213**, 899-929.

Heymann, J. B., Muller, D. J., Mitsuoka, K. & Engel, A. (1997). Electron and atomic force microscopy of membrane proteins. *Curr. Opin. Struct. Biol.* **7**, 543-549.

Heymann, J. B., Agre, P. & Engel, A. (1998). Progress on the structure and function of aquaporin 1. *J. Struct. Biol.* **121**, 191-206.

Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science,* **268**, 1144-1149.

Iwata, S., Lee, J. W., Okada, K., Lee, J. K., Iwata, M., Rasmussen, B., Link, T. A., Ramaswamy, S. & Jap, B. K. (1998). Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science,* **281**, 64-71.

Jernigan, R. L. & Bahar, I. (1996). Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **6**, 195-209.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994a). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry,* **33**, 3038-3049.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994b). A mutation data matrix for transmembrane proteins. *FEBS Letters,* **339**, 269-275.

Koepke, J., Hu, X., Muenke, C., Schulten, K. & Michel, H. (1996). The crystal structure of the light-harvesting complex II (B800-850) from *Rhodospirillum molischianum*. *Structure,* **4**, 581-597.

Komiya, H., Yeates, T. O., Rees, D. C., Allen, J. P. & Feher, G. (1988). Structure of the reaction center from *Rhodobacter sphaeroides* R-26 and 2.4.1: symmetry relations and sequence comparisons between different species. *Proc. Natl Acad. Sci. USA,* **85**, 9012-9016.

Kuhner, M. K. & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459-468.

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

MacKenzie, K. R. & Engelman, D. M. (1998). Structure-based prediction of the stability of transmembrane helix-helix interactions: the sequence dependence of glycophorin A dimerization. *Proc. Natl Acad. Sci. USA,* **95**, 3583-3590.

MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science,* **276**, 131-133.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641-656.

Monne, M., Nilsson, I., Johansson, M., Elmhed, N. & von Heijne, G. (1998). Positively and negatively charged residues have different effects on the position in the membrane of a model transmembrane helix. *J. Mol. Biol.* **284**, 1177-1183.

Neuwald, A. F., Liu, J. S. & Lawrence, C. E. (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **4**, 1618-1632.

Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357-358.

Persson, B. & Argos, P. (1994). Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.* **237**, 182-192.

Pilpel, Y. & Lancet, D. (1999). The variable and conserved interfaces of modeled olfactory receptor proteins. *Protein Sci.* **8**, 969-977.

Popot, J. L. & Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry,* **29**, 4031-4037.

Preusch, P. C., Norvell, J. C., Cassatt, J. C. & Cassman, M. (1998). Progress away from 'no crystals, no grant'. *Nature Struct. Biol.* **5**, 12-14.

Prince, S. M., Papiz, M. Z., Freer, A. A., McDermott, G., Hawthornthwaite-Lawless, A. M., Cogdell, R. J. & Isaacs, N. W. (1997). Apoprotein structure in the LH2 complex from *Rhodopseudomonas acidophila* strain 10050: modular assembly and protein pigment interactions. *J. Mol. Biol.* **268**, 412-423.

Rees, D. C., DeAntonio, L. & Eisenberg, D. (1989). Hydrophobic organization of membrane proteins. *Science,* **245**, 510-513.

Rost, B., Casadio, R., Fariselli, P. & Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**, 521-533.

Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5**, 1704-1718.

Samatey, F. A., Xu, C. & Popot, J. L. (1995). On the distribution of amino acid residues in transmembrane alpha-helix bundles. *Proc. Natl Acad. Sci. USA,* **92**, 4577-4581.

Sato, H., Takeda, K., Tani, K., Hino, T., Okada, T., Nakasako, M., Kamiya, N. & Kouyama, T. (1999). Specific lipid-protein interactions in a novel honeycomb lattice structure of bacteriorhodopsin. *Acta Crystallog. sect. D,* **55**, 1251-1256.

Scherlter, G. F. X., Villa, C. & Henderson, R. (1993). Projection structure of rhodopsin. *Nature,* **362**, 770-772.

Stevens, T. J. & Arkin, I. T. (1999). Are membrane proteins ''inside-out'' proteins? *Proteins: Struct. Funct. Genet.* **36**, 135-143.

Taylor, W. R., Jones, D. T. & Green, N. M. (1994). A method for alpha-helical integral membrane protein fold prediction. *Proteins: Struct. Funct. Genet.* **18**, 281-294.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa, Itoh K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. *Science,* **272**, 1136-1144.

Tusnady, G. E. & Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **283**, 489-506.

Vajda, S., Sippl, M. & Novotny, J. (1997). Empirical potentials and functions for protein folding and binding. *Curr. Opin. Struct. Biol.* **7**, 222-228.

von Heijne, G. (1991). Proline kinks in transmembrane alpha-helices. *J. Mol. Biol.* **218**, 499-503.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487-494.

von Heijne, G. (1996). Principles of membrane protein assembly and structure. *Prog. Biophys. Mol. Biol.* **66**, 113-139.

Walz, T., Hirai, T., Murata, K., Heymann, J. B., Mitsuoka, K., Fujiyoshi, Y., Smith, B. L., Agre, P. & Engel, A. (1997). The three-dimensional structure of aquaporin-1. *Nature,* **387**, 624-627.

White, S. H. (1994). Global statistics of protein sequences: implications for the origin, evolution, and prediction of structure. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 407-439.

White, S. H. & Wimley, W. C. (1994). Peptides in lipid bilayers: structural and thermodynamic basis for partitioning and folding. *Curr. Opin. Struc. Biol.* **4**, 79-86.

Wimley, W. C. & White, S. H. (1996). Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Struct. Biol.* **3**, 842-848.

Yau, W. M., Wimley, W. C., Gawrisch, K. & White, S. H. (1998). The preference of tryptophan for membrane interfaces. *Biochemistry,* **37**, 14713-14718.

Yeates, T. O., Komiya, H., Rees, D. C., Allen, J. P. & Feher, G. (1987). Structure of the reaction center from *Rhodobacter sphaeroides* R-26: membrane-protein interactions. *Proc. Natl Acad. Sci. USA,* **84**, 6438-6442.

Zhang, L. & Skolnick, J. (1998). How do potentials derived from structural databases relate to ''true'' potentials? *Protein Sci.* **7**, 112-122.

Edited by G. von Heijne

(Received 23 March 1999; received in revised form 29 September 1999; accepted 30 September 1999)