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

Yitzhak Pilpel¹*, Nir Ben-Tal² and Doron Lancet¹

¹Department of Molecular Genetics and the Crown Genome Center, The Weizmann Institute of Science, Rehovot 76100, Israel
²Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

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.

© 1999 Academic Press

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-
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 \( \alpha \)-helical periodicity. They used a power spectrum method to select TM sequences that display an \( \alpha \)-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 \( \alpha \)-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:

\[
\text{kPROT}_i = \ln \left( \frac{f_i^m}{f_i^s} \right)
\]

where \( f_i^m \) and \( f_i^s \) 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 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 (kPROT > 0) and membranophobic (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

| Residue | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Total |
| Numb   | 4888 | 45,179 | 1038 | 7310 | 214 | 4793 | 220 | 4875 | 3774 | 39,170 | 3402 | 36,316 | 289 | 3835 | 6216 | 14,984 | 678 | 5000 | 9456 | 73,628 | 1164 | 14,843 | 45,292 | 2307 | 2999 | 22,018 | 6128 | 45,966 | 891 | 9426 | 1383 | 14,657 |
| Frac   | 0.105 | 0.103 | 0.022 | 0.017 | 0.005 | 0.011 | 0.005 | 0.011 | 0.08 | 0.089 | 0.073 | 0.083 | 0.006 | 0.009 | 0.133 | 0.034 | 0.015 | 0.025 | 0.203 | 0.168 | 0.025 | 0.118 | 0.021 | 0.013 | 0.014 | 0.008 | 0.011 | 0.006 | 0.003 | 0.01 | 0.006 |
| Numb   | 532 | 1517 | 47 | 233 | 12 | 135 | 20 | 161 | 546 | 1533 | 370 | 1103 | 48 | 174 | 794 | 549 | 17 | 117 | 965 | 2523 | 141 | 1651 | 268 | 144 | 494 | 33 | 211 | 115 | 255 | 186 | 115 | 768 |
| Frac   | 0.103 | 0.1 | 0.009 | 0.015 | 0.002 | 0.009 | 0.002 | 0.011 | 0.067 | 0.010 | 0.072 | 0.073 | 0.009 | 0.011 | 0.154 | 0.036 | 0.003 | 0.008 | 0.187 | 0.167 | 0.027 | 0.109 | 0.014 | 0.007 | 0.028 | 0.006 | 0.008 | 0.008 | 0.028 | 0.002 |
| Numb   | 3038 | 26,134 | 388 | 4897 | 71 | 2002 | 73 | 2077 | 1907 | 21,221 | 2101 | 20,892 | 86 | 1621 | 3455 | 8211 | 947 | 1517 | 5653 | 40,961 | 591 | 28,959 | 4205 | 405 | 3739 | 131 | 2836 | 1292 | 1312 | 230 | 1292 | 3761 |
| Frac   | 0.12 | 0.109 | 0.023 | 0.012 | 0.003 | 0.008 | 0.003 | 0.009 | 0.075 | 0.088 | 0.083 | 0.087 | 0.003 | 0.007 | 0.136 | 0.034 | 0.004 | 0.006 | 0.223 | 0.171 | 0.023 | 0.121 | 0.018 | 0.016 | 0.025 | 0.005 | 0.008 | 0.005 | 0.012 | 0.001 |
| Numb   | 399 | 1403 | 213 | 338 | 27 | 270 | 23 | 254 | 436 | 1131 | 228 | 928 | 71 | 168 | 569 | 535 | 99 | 423 | 964 | 1395 | 132 | 1531 | 367 | 60 | 374 | 62 | 268 | 559 | 200 | 1526 | 752 |
| Frac   | 0.077 | 0.093 | 0.041 | 0.022 | 0.005 | 0.018 | 0.003 | 0.011 | 0.084 | 0.075 | 0.044 | 0.061 | 0.014 | 0.011 | 0.111 | 0.035 | 0.011 | 0.028 | 0.186 | 0.158 | 0.026 | 0.101 | 0.018 | 0.012 | 0.023 | 0.003 | 0.016 | 0.037 | 0.039 | 0.028 |
| Numb   | 1850 | 19,048 | 454 | 2933 | 143 | 2793 | 147 | 2798 | 1807 | 18,191 | 1301 | 15,426 | 203 | 2214 | 2761 | 6733 | 537 | 3483 | 3803 | 32,668 | 377 | 22,744 | 4295 | 496 | 6734 | 238 | 3111 | 3384 | 995 | 20,008 | 6735 |
| Frac   | 0.087 | 0.095 | 0.021 | 0.013 | 0.007 | 0.014 | 0.007 | 0.011 | 0.085 | 0.091 | 0.034 | 0.028 | 0.011 | 0.011 | 0.13 | 0.034 | 0.028 | 0.057 | 0.186 | 0.163 | 0.027 | 0.112 | 0.016 | 0.012 | 0.026 | 0.003 | 0.017 | 0.047 | 0.045 | 0.025 |

**A. Number and fraction of the residues in single and multi-set spans**

- **Total TM**: The total number of residues in the transmembrane region.
- **Extracellular**: The number of residues in the extracellular region.
- **Central**: The number of residues in the central portion.
- **Intracellular**: The number of residues in the intracellular region.
- **Both termini**: The number of residues at both termini.

**B. The kPROT scale**

- **Residue**: The amino acid residues.
- **Total TM**: The total number of residues in the transmembrane region.
- **Extracellular**: The number of residues in the extracellular region.
- **Central**: The number of residues in the central portion.
- **Intracellular**: The number of residues in the intracellular region.
- **Both termini**: The number of residues at both termini.

---

**Notes**

- The generation of the kPROT scale involves several steps, including the computation of residues for 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 the 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.
- 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.

---

**References**

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 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
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 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, Ser, 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.
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 favor residing in the electrostatically complex environment of the interface (Monne et al., 1998).

The derived k PROT 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 k PROT 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 k PROT 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 k PROT scales cluster with the water-membrane interface scale, k PROT for the TM-center is more akin to the other three scale classes. Within the latter branch of the dendrogram, the central k PROT 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 k PROT, which includes all of the k PROT segmental scales, reflects a balanced combination of physicochemical properties, so as

![Figure 3. The position-dependent k PROT 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.](image)

![Figure 4. A dendrogram depiction of mutual similarities between k PROT 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).](image)
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(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 $z$-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-
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 head-groups. 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 of membrane protein structures is highly complex (White & Wimley, 1994).

---

Figure 6. Benchmark test on (a) glycoporphin 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 glycoporphin A dimer are highlighted (a).
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.
et al., 1998). Thus, by combining kPROT analysis with experimental data on oligomerizing bitopic proteins (MacKenzie 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).

**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 archael 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 (Poptop & 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
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:

\[
\sigma(x \pm y) = \sqrt{(\sigma(x))^2 + (\sigma(y))^2}
\]

to the kPROT definition:

\[
kPROT = \log \left( \frac{f_s}{f_m} \right) = \log(f_s) - \log(f_m)
\]

and thus evaluate the standard deviation on kPROT as:

\[
\sqrt{(\sigma(\log(f_s)))^2 + (\sigma(\log(f_m)))^2}
\]

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

\[
\sigma(\log(f_s)) = \frac{\hat{\sigma}(f_s)}{f_s}
\]

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 oxi-dase, 1OCC (Tsukihara et al., 1996); potassium channel, 1BL8 (Doyle et al., 1998); mechanosensitive ion channel 1MSL (Chang et al., 1998); and glycoporphin A dimer.

---

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

### Table 2. Protein sequence statistics

<table>
<thead>
<tr>
<th>Protein set</th>
<th>All (single TM)</th>
<th>Single TM Polypeptide</th>
<th>Multiple TM Polypeptide</th>
<th>Single+Multiple TM Polypeptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original set</td>
<td>3682</td>
<td>2003</td>
<td>6634 (40,100)</td>
<td>2294 (14,333)</td>
</tr>
<tr>
<td>Non-redundant set</td>
<td>2164</td>
<td>1034</td>
<td>3322 (20,124)</td>
<td>498 (3041)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5486 (22,288)</td>
<td>1532 (4075)</td>
</tr>
</tbody>
</table>

Number of proteins (and number of TM segments in multi-span proteins) in the single and multi 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 as designated 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 determination 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 801; G-protein coupled receptors 145; oxidoreductases 110; symporters 82; ionic channels 74; electron transport 72; serine protease 24; metalloprotease 23; glycosidase 77; oxidoreductase 77; electron transport 72; 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.

---

Number 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:

\[
\sigma(x \pm y) = \sqrt{(\sigma(x))^2 + (\sigma(y))^2}
\]

to the kPROT definition:

\[
kPROT = \log \left( \frac{f_s}{f_m} \right) = \log(f_s) - \log(f_m)
\]

and thus evaluate the standard deviation on kPROT as:

\[
\sqrt{(\sigma(\log(f_s)))^2 + (\sigma(\log(f_m)))^2}
\]

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

\[
\sigma(\log(f_s)) = \frac{\hat{\sigma}(f_s)}{f_s}
\]

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 glycoporphin A dimer.

---

† http://blanco.biomol.uci.edu/
Membrane_Proteins_xtal.html
1AFO (MacKenzie et al., 1997). Two structures, of cytochrome bc1, (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) \( \mathbf{M} \) was computed for the different propensity scales as the moment length \(|\mathbf{M}|\) and the moment direction \(\Theta\) relative to the angular direction of the \( \alpha \)-carbon atom of the first amino acid residue in the TM segment, as follows:

\[
|\mathbf{M}| = \sqrt{\sum_{i=1}^{n}(M_i \cos \theta_i)^2 + \sum_{i=1}^{n}(M_i \sin \theta_i)^2} \quad (3)
\]

and:

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

where \( M_i \) 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 \( \alpha \)-carbon atoms of the TM-constituent amino acid residues.

Determination of angular orientations

For the purpose of benchmarking and for future \textit{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 \textit{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 \textit{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 proteins, and the mitochondrially 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 Piliusky 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


Edited by G. von Heijne

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