A Model Recognition Approach to the Prediction of All-Helical Membrane Protein Structure and Topology[†]

D. T. Jones,*[,‡,§] W. R. Taylor,[§] and J. M. Thornton[‡]

Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, United Kingdom, and Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, United Kingdom

Received July 13, 1993; Revised Manuscript Received January 4, 1994[®]

ABSTRACT: This paper describes a new method for the prediction of the secondary structure and topology of integral membrane proteins based on the recognition of topological models. The method employs a set of statistical tables (log likelihoods) compiled from well-characterized membrane protein data, and a novel dynamic programming algorithm to recognize membrane topology models by expectation maximization. The statistical tables show definite biases toward certain amino acid species on the inside, middle, and outside of a cellular membrane. Using a set of 83 integral membrane protein sequences taken from a variety of bacterial, plant, and animal species, and a strict jackknifing procedure, where each protein (along with any detectable homologues) is removed from the training set used to calculate the tables before prediction, the method successfully predicted 64 of the 83 topologies, and of the 37 complex multispanning topologies 34 were predicted correctly.

Integral membrane proteins represent an important and functionally diverse class of protein structure. From the observations made on the structure of bacteriorhodopsin (determined by electron diffraction; Henderson *et al.*, 1990) and the photosynthetic reaction center (determined by X-ray crystallography; Deisenhofer *et al.*, 1985), it has been concluded that transmembranal segments are typically apolar helices, 17–25 residues in length. Such an arrangement allows the apolar side chains to interact favorably with the lipid environment, while fully satisfying the hydrogen-bonding potential of the peptide units in a regular secondary structure. An alternative arrangement has been observed in the crystallographically determined structure of a bacterial outer membrane porin (Weiss *et al.*, 1992), where the transmembranal segments are β-strands arranged in a 16-stranded barrel. Typical methods for the prediction of membrane-spanning segments of integral membrane proteins (von Heijne, 1981, 1992; Argos *et al.*, 1982; Eisenberg, 1984; Engelman *et al.*, 1986) implicitly assume the predicted segments to be helices which are completely exposed to lipid, and in terms of predicting the tertiary structure, that these helices completely traverse the membrane in a direction normal to its surface. How reasonable this assumption is cannot be determined until more integral membrane protein structures are solved. Circular dichroism studies, however, at least support the notion in that they indicate that most transmembranal proteins have a very high helix content.

Methods for the prediction of transmembrane-spanning segments are typically based on hydropathy analysis (Kyte & Doolittle, 1982). The simplest scheme is to generate a hydrophobicity plot for a given sequence, with transmembrane segments being centered at the peaks of the plot. For an initial hydrophobicity plot a window of 17–22 residues is taken, and using a suitable hydrophobicity scale (Kyle & Doolittle, 1982; Engelman *et al.*, 1986; Cornette *et al.*, 1987) the average residue hydrophobicity calculated. Plots based on smaller windows (5–11 residues) are helpful in delineating the end points of the segments. Despite being generally apolar, transmembranal segments often exhibit a degree of amphipathicity, and this can be used in addition to the simple hydrophobicity plot to improve predictive accuracy (Eisenberg, 1984; Stirk *et al.*, 1992).

Recent studies (von Heijne & Gavel, 1988; Nakashima & Nishikawa, 1992; Landolt-Marticorena *et al.*, 1993) have indicated the presence of topogenic signals in integral membrane proteins, i.e., sequence patterns which correlate with the topology of the membrane-spanning segments. The most evident of these signals is the prevalence of positively charged residues in the interior (cytoplasmic) loops which is now familiarly known as the "positive inside rule" (von Heijne & Gavel, 1988). Such topogenic signals can be used to evaluate the plausibility of predicted integral membrane structures, a fact which has been very elegantly demonstrated by von Heijne (1992).

In this work a method is described that simultaneously takes into account the prediction of transmembrane secondary structure and the location of topogenic signals. For any given topology and scoring scheme, a mathematically optimal solution is found, which enables the likelihood of each suggested topology to be objectively assessed. The basic idea here is the idea of *expectation maximization*, a simple statistical method which is concerned with the generation and fitting of models to data. Traditional prediction schemes attempt to determine the most reasonable underlying model based on an analysis of one or more sequences. In contrast, expectation maximization attempts to search for the model which best explains the given data. Given a function which calculates the total probability for the match of a given model with a given sequence, the resulting model from expectation maximization should correspond to the maximum of this function.

MODEL DEFINITION

The first requirement for expectation maximization is the definition of a model (used here in the statistical sense, rather

[†] This work was supported by a SERC CASE studentship and a Wellcome Trust Biomathematics Fellowship awarded to D.T.J.

[‡] University College.

[§] National Institute for Medical Research.

[®] Abstract published in *Advance ACS Abstracts,* February 15, 1994.

0006-2960/94/0433-3038$04.50/0

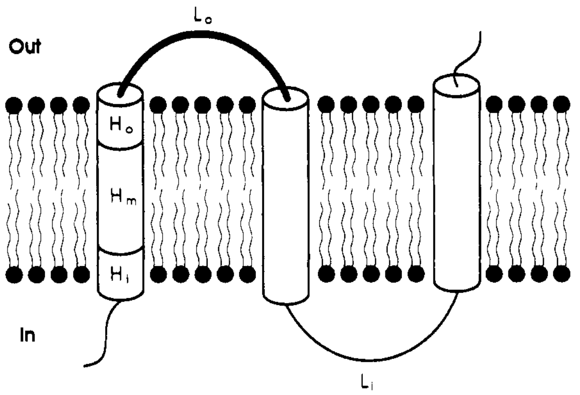


FIGURE 1: Structural states defined for a typical helical transmembrane protein.

Table 1: Multispanning Protein Sequences Used To Calculate Topogenic Parameters[a]

RCEL_CHLAU	REACTION CENTER PROTEIN L CHAIN. 5/92
RCEM_CHLAU	REACTION CENTER PROTEIN M CHAIN. 5/92
5HT2_CRIGR	5-HYDROXYTRYPTAMINE 2 RECEPTOR (5-HT-2). 5/92
5HT3_MOUSE	5-HYDROXYTRYPTAMINE 3 RECEPTOR PRECURSOR (5-HT-3). 3/92
5HTA_HUMAN	5-HYDROXYTRYPTAMINE 1A RECEPTOR (5-HT-1A). 5/92
A1AA_HUMAN	ALPHA-1A ADRENERGIC RECEPTOR. 5/92
A2AA_HUMAN	ALPHA-2A ADRENERGIC RECEPTOR (SUBTYPE C10). 5/92
EDG1_HUMAN	PROBABLE G PROTEIN-COUPLED RECEPTOR EDG-1. 8/92
MOTA_ECOLI	CHEMOTAXIS MOTA PROTEIN. 11/91
MALF_ECOLI	MALTOSE TRANSPORT INNER MEMBRANE MALF PROTEIN. 11/90
SECY_BACSU	SECY PROTEIN. 3/92
OPS1_CALVI	OPSIN RH1 (OUTER R1-R6 PHOTORECEPTOR CELLS OPSIN). 8/91
OPSB_HUMAN	BLUE-SENSITIVE OPSIN (BLUE CONE PHOTORECEPTOR PIGMENT). 8/91
OPSD_BOVIN	RHODOPSIN. 8/91
AA1R_CANFA	ADENOSINE A1 RECEPTOR. 8/92
AA2R_CANFA	ADENOSINE A2 RECEPTOR. 8/92
C561_BOVIN	CYTOCHROME B561. 7/89
ADT_RICPR	ADP,ATP CARRIER PROTEIN (ADP/ATP TRANSLOCASE). 8/91
CYOB_ECOLI	CYTOCHROME O UBIQUINOL OXIDASE SUBUNIT I (EC 1.10.3.-). 11/90
CYOC_ECOLI	CYTOCHROME O UBIQUINOL OXIDASE SUBUNIT III (EC 1.10.3.-). 11/90
CYOD_ECOLI	CYTOCHROME O UBIQUINOL OXIDASE OPERON PROTEIN CYOD. 11/90
SECE_ECOLI	INNER MEMBRANE PROTEIN SECE. 8/91
GAPB_HUMAN	GAP JUNCTION BETA-1 PROTEIN (CONNEXIN 32). 3/92
LEP_ECOLI	SIGNAL PEPTIDASE I (EC 3.4.-.-) (SPASE I). 8/92
PT2M_ECOLI	PHOSPHOTRANSFERASE ENZYME II, MANNITOL-SPECIFIC. 3/92
ATHP_NEUCR	PLASMA MEMBRANE ATPASE (EC 3.6.1.35). 12/92
LACY_ECOLI	LACTOSE PERMEASE (LACTOSE-PROTON SYMPORT). 12/92
OPPB_SALTY	OLIGOPEPTIDE PERMEASE PROTEIN OPPB. 12/92
TAPA_HUMAN	CELL SURFACE PROTEIN TAPA-1. 8/92
DHSC_BACSU	SUCCINATE DEHYDROGENASE CYTOCHROME B-558. 7/89
LSPA_ECOLI	LIPOPROTEIN SIGNAL PEPTIDASE. 5/91
IMM1_ECOLI	IMMUNITY PROTEIN FOR COLICIN E1. 11/90
IMMA_CITFR	IMMUNITY PROTEIN FOR COLICIN A. 8/91
TCR1_ECOLI	TETRACYCLINE RESISTANCE PROTEIN. 2/91
UHPT_ECOLI	HEXOSE PHOSPHATE TRANSPORT PROTEIN. 8/92

[a] Codes are from SWISS-PROT Release 23. To reduce bias in the calculated parameters, the full list was reduced so that no remaining pair of sequences is significantly more than 60% sequence identical.

than the biomolecular sense). In the case of transmembrane prediction such a model includes parameters for the number of membrane-spanning segments, *n*, the topology, *t* (N-terminus in or out), the length, *l*, and the location, *i*, in the sequence of each segment.

In the case of the work shown here, residues are classified as being in five structural states, as shown in Figure 1. The five states are as follows: L_i (inside loop), L_o (outside loop), H_i (inside helix end), H_m (helix middle), and H_o (outside helix end). The number of residues taken to be in the helix end caps was arbitrarily taken as being four. Other cap definitions could be used, and it is possible that a more rigorous definition of the cap/middle boundaries might improve results. The term "loop" is not particularly appropriate for single-spanning segments, but the term loop is here used rather loosely, in this case to indicate residues not in the *transmembrane* secondary structure. In this definition, therefore, an entire globular domain which happened to be anchored to a membrane would be classified as either an inside or outside loop.

The source data for this work were a set of documented transmembrane proteins extracted from Release 23.0 of SWISS-PROT (Bairoch & Boeckmann, 1991). The initial set of 1765 membrane sequences was split into two subsets, one containing the proteins with a single-membrane-spanning-segment, and the other with multiple-membrane-spanning segments. Both sets were further reduced to include just those sequences for which the membrane topology was given. In addition entries for which any transmembrane segment was listed as shorter than 17 or longer than 25 were eliminated. In view of the difficulty in assigning membrane-spanning segments for multispanning proteins, only the multispanning segments for which at least some experimental data were available were included (listed in Table 1). Some of this structure and topology information must be taken as being hypothetical, but it is to be hoped that the majority of the data are correct. Our experience with the described method has indicated that it is insensitive to the precise composition of the training set of proteins. Unfortunately, until more experimental data become available on membrane protein structure, there is no option but to make the most of the available mixture of experimental and hypothetical data and to rely on the judgements of the depositing authors. The final single-spanning and multispanning data sets comprised 285 and 35 sequences, respectively (Table 2).

Table 2: Composition of Data Sets Used To Calculate Topogenic Parameters[a]

	single-spanning data set	multispanning data set
sequences	285	35
transmembrane segments	285	174
inside loop residues	5699	2032
outside loop residues	830	1498
inside helix residues	1140	696
outside helix residues	1140	696
middle helix residues	4003	2037
total residues	179 037	8730

[a] Note that the total residue counts include residues not assigned to any structural state, i.e., oversized loops.

PROPENSITIES

For each of the 5 structural classes, log likelihood ratios for each of the 20 amino acids were calculated:

$$s_i = \ln(q_i/p_i)$$

where p_i is the relative frequency of occurrence (or fraction) of amino acid *i* in all the sequences in the data set, and q_i is the relative frequency of occurrence of amino acid *i* in a particular structural class. A positive score indicates a higher than expected frequency for a given amino acid to be found in a particular structural class and a negative score a lower than expected frequency, and a score close to zero indicates that the frequency of occurrence of the given amino acid in a particular class is no different from that expected from chance alone. To circumvent the previously mentioned problem of classifying globular domains as loops, loops longer than 100 residues are *not* classified as loops, and are ignored in the calculation of q_i values. These oversized loops are, however, included in the calculation of the overall relative frequencies of occurrence p_i. The scores calculated for the previously described set of sequences are shown for both single-spanning (Table I in the supplementary material and Figure 2) and multispanning segments (Table II in the supplementary material and Figure 3).

To test the statistical significance of the observed frequencies, χ^2 probabilities were calculated for each amino acid across every pair of structural locations (10 in total). The null hypothesis for this test is that the frequency of occurrence for

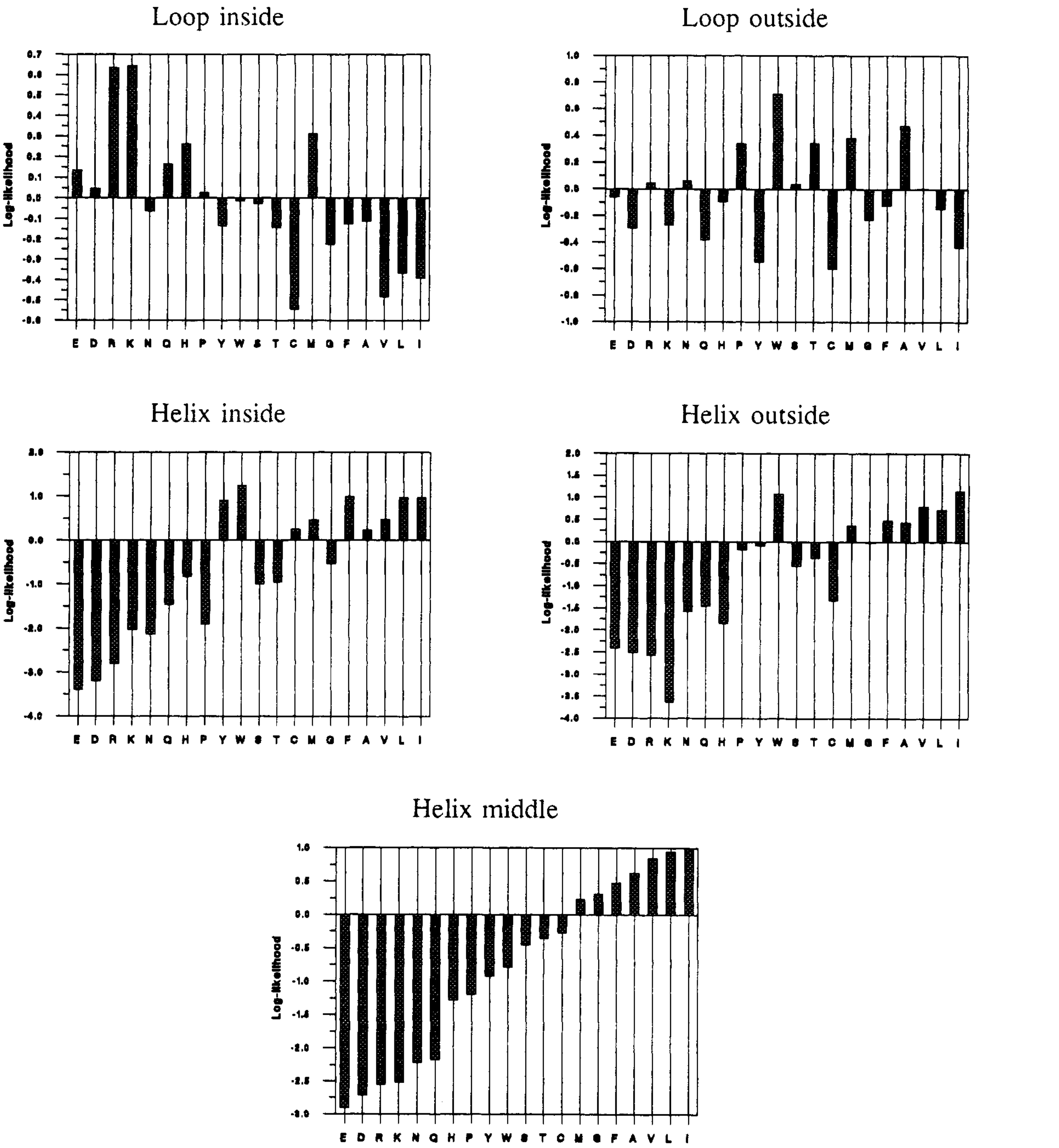


FIGURE 2: Topogenic parameters for single-spanning transmembrane segments.

residue *a* is identical in structural location *x* and location *y*. If *R* and *S* are the frequencies of occurrence for a particular residue in two structural locations, *x* and *y*, respectively, and N_x and N_y are the total frequencies of all residues in the two locations, then we estimate the χ^2 value using the expression

$$\chi^2 = \frac{[R(N_y/N_x)^{1/2} - S(N_x/N_y)^{1/2}]^2}{R+S}$$

which is based on the unequal sample size variant of the more familiar χ^2 calculation. The χ^2 probability (for 1 degree of freedom in this case) may be calculated using the appropriate incomplete γ function. The complete set of probabilities is shown in Tables III and IV in the supplementary material, each value denoting the probability of the particular null hypothesis holding or in other words the probability that the two frequencies of occurrence are not significantly different. As an example, the χ^2 probability for alanine in the inside/outside multispanning loop is 0.5649 which indicates that there is a greater than even chance that the *population* frequencies of occurrence for alanine in these two structural locations are identical on the basis of the observed *sample* frequencies. In contrast, the χ^2 probability for arginine in these locations is <0.0001, indicating that there is less than 1 chance in 10 000 that arginine occurs as frequently in inside loops as outside loops on the basis of the observed samples.

χ^2 calculations were also performed between each structural location as a whole to determine whether each of the five structural locations could be distinguished on the basis of amino acid frequencies of occurrence. Each of the five locations could be distinguished from the remaining four with a confidence of at least 98%, indicating that the arbitrary division of transmembrane structures into five regions is quite valid.

The log likelihood values clearly encode a variety of topogenic signals. The most significant components of the

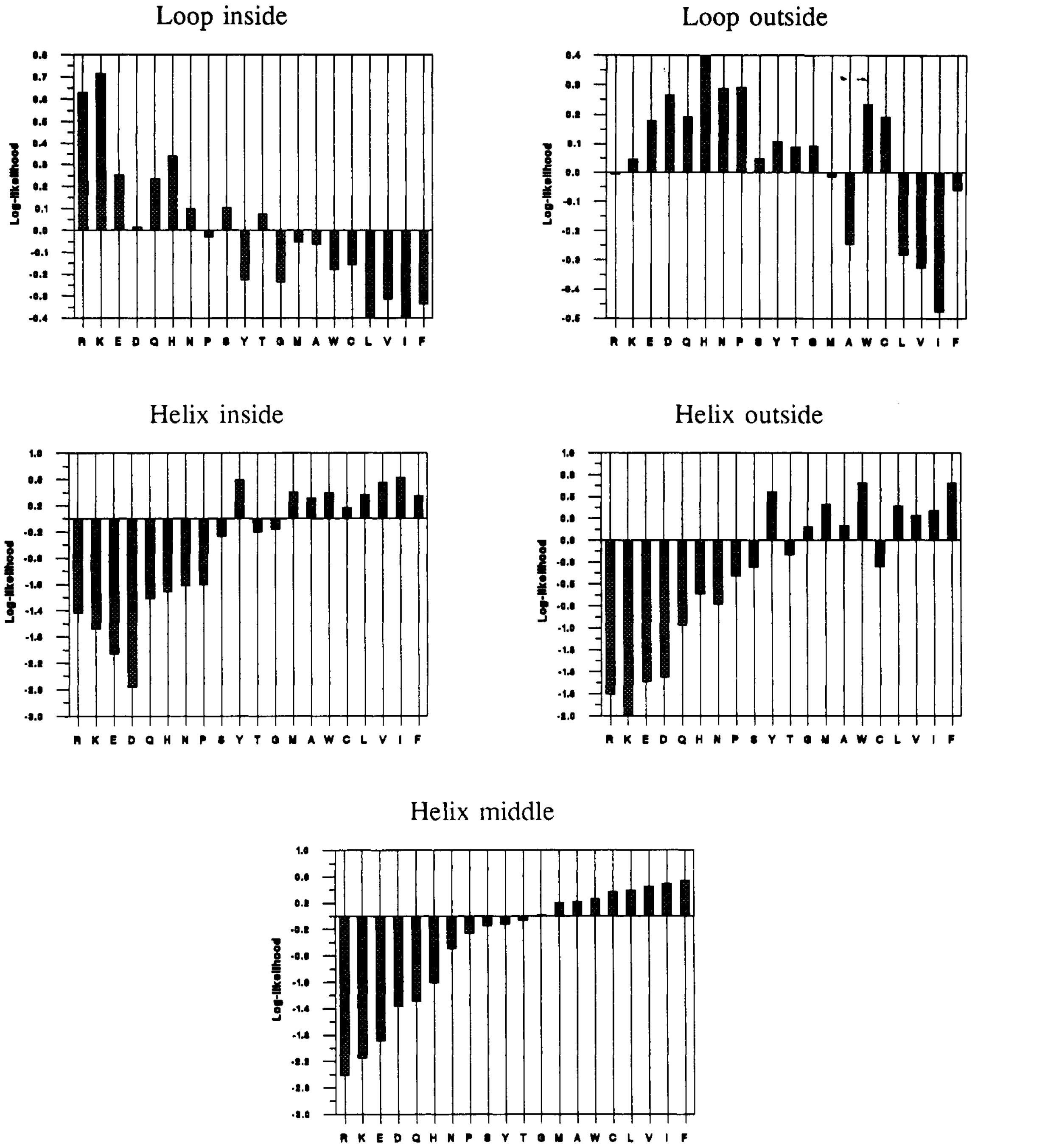


FIGURE 3: Topogenic parameters for multispanning transmembrane segments.

propensities merely determine the lipophilic preferences of amino acids, or in other words that hydrophobic residues occur more frequently in the helical segments than the flanking regions. The signals that cannot be explained away by hydrophobicity alone are perhaps of more interest. The preference for positively charged residues to be found in the inside loops is clearly seen. It is interesting to note that a similar effect is seen between the inside and outside helix caps, though this could be due to the indeterminate boundaries between the author-defined membrane-spanning segments and their flanking regions. For multispanning proteins, the most significant preferences for the inside/outside loop are found for Arg, Gly, His, Lys, and Pro, whereas for single-spanning proteins, Ala, Arg, Asp, Gln, Lys, Pro, Thr, Trp, and Val have the most significant propensities. For inside/outside helix caps, only Phe and Trp have highly significant topogenic propensities for multispanning helices whereas Cys, Gly, His, Leu, Lys, Phe, Pro, Ser, Thr, Tyr, and Val show clear inside/outside preferences for single-spanning helix caps. The unusual abundance of tryptophan residues in outside locations of the photosynthetic reaction center has been noted by Schiffer *et al.* (1992), and has also been seen in porin, but it would appear from the results presented here that this is a general feature of transmembrane proteins as a whole. Without further experimental evidence, the question of whether these residues help in the direction of membrane topology or merely act to stabilize the final topology remains open.

To test the effectiveness of this topogenic scoring system, the ability of these scores to predict the correct location of polar flanking regions was tested. Each documented polar flanking region was extracted from the sequence databank, and the total score calculated using both inside and outside loop log likelihood values. Where the total inside score exceeded the outside score, the region was predicted as being inside; otherwise it was predicted as being outside. The results of predicting the location of the flanking regions of a test set of proteins (see Table 5) are shown in Tables 3 (multispanning) and 4 (single-spanning). In all cases, the protein under test

Table 3: Results of Predicting the Location of a Set of Multispanning Loop Segments Using the Multispanning Segment Topogenic Parameters

loop length	total in data set	number correct	loop length	total in data set	number correct
0–9	34	25	60–69	3	2
10–19	81	60	70–79	7	5
20–29	32	25	80–89	4	2
30–39	29	21	90–99	0	
40–49	7	6	≥100	12	9
50–59	2	0			

Table 4: Results of Predicting the Location of a Set of Single-Spanning Loop Segments Using the Single-Spanning Segment Topogenic Parameters

loop length	total in data set	number correct	loop length	total in data set	number correct
0–9	2	2	60–69	2	2
10–19	4	4	70–79	0	
20–29	6	6	80–89	1	1
30–39	8	7	90–99	1	1
40–49	4	3	≥100	42	23
50–59	5	5			

was excluded from the calculation of the topogenic parameters, along with any related sequences (sequence identity >25%).

The average scores for the scheme proposed here are 73% for multispanning loops and 70% for single-spanning loops which compare favorably with the random expected score of 50%. Interestingly, in the case of the single-spanning loops, of the 31 loops shorter <70 residues in length, the locations of 29 (94%) are correctly predicted, and of the 42 loops ≥100 residues in length, only 23 (55%) are correctly predicted. In the case of single-spanning proteins, therefore, loops of 70 residues or more contain little information regarding their location with respect to the membrane. For multispanning segments, it is important to note that this scheme is clearly able to predict the location of long flanking regions as well as short regions. While it is not possible to use the positive-inside rule to directly predict the location of a given flanking region, it should be noted that previous studies have shown no significant positive bias for regions longer than 70 residues (von Heijne & Gavel, 1988). The results here clearly show little dependence on flanking region length for multispanning segments, yet a clear length-dependent effect for single-spanning segments. Why should the topogenic parameters be so sharply defined for single-spanning segments? The reason for this is no doubt due to the fact that, for a single-spanning protein, almost the entire responsibility for correctly orienting the protein in the membrane resides with the short N- or C-terminal flanking segment. For multispanning proteins, the location of a loop depends not only on the amino acids in the loop itself, but also on those in other loops. The method presented here makes use of this cooperativity, in that it is the score obtained for the whole protein with a given topology that is the basis of the prediction, rather than the score obtained for a segment taken in isolation.

While the analysis by Nakashima and Nishikawa (1992) on the amino acid composition of polar flanking regions produced some observations similar to those presented here, it is hard to directly compare results due to the fact that their analysis was based on compositional preferences, and was therefore inherently limited to regions at least 50 residues in length. The analysis of human type I single-span membrane proteins by Landolt-Marticorena *et al.* (1993) provided observations similar to those from this analysis for single-spanning segments, though some of the signals identified by Landolt-Marticorena *et al.* (in particular the Ser and Asn preferences for extracellular flanking regions, and the Trp preference for outside helix cap regions) did not appear statistically significant in our data set. It is possible that the "missing" signals are only significant for the class of sequences selected by Landolt-Marticorena *et al.*

For both multispanning and single-spanning helices, the transition between the helix cap and helix middle states was trapezoidally smoothed (von Heijne, 1992) as follows:

$$\text{score}_{\text{helix cap}} = \sum_{i=1}^{4}\left(\frac{iH_{\text{m}}}{5} + \frac{(5-i)H_{\text{i/o}}}{5}\right)$$

To prevent overprediction, a filter was applied to the final topologies by means of a score cutoff on helical segments. Predicted topologies including helical segments with trapezoidally smoothed scores less than a predetermined cutoff were rejected from consideration. A value of 0.1 for this cutoff produced acceptable results, and was used for all the results shown here.

TOPOLOGY PREDICTION

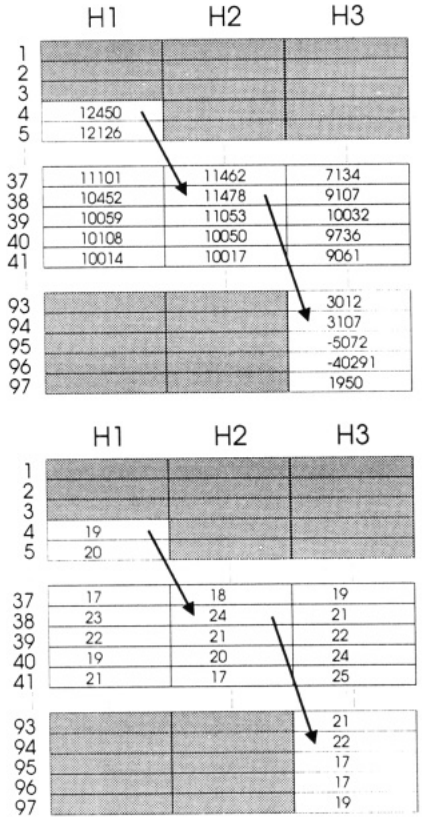
Using the propensities shown in Figures 2 and 3, it is possible to calculate a score relating to the compatibility of a given sequence with a given topology and secondary structure. This is analogous to the protein fold recognition approaches described for globular protein folds (Jones *et al.*, 1992; Bowie *et al.*, 1991), though the structural description used here is not at the full tertiary structure level. In common with the globular protein fold recognition methods, an algorithm is required that is capable of finding the optimum hatch between the given sequence and the given structural model. Given the simplicity of the structural models used here, at first sight it might appear feasible to use a brute-force search to identify the most likely match. For a sequence of length m, and a given transmembrane topology (n, t), there are approximately $9^n((M - 21n)/n)^n$ possible models. Taking as an example a typical case of a 7-helix transmembranal topology and a sequence of length 250, the total number of different models that could be generated for this sequence is $\sim 7 \times 10^{14}$. Clearly a brute-force approach is inappropriate.

Despite the apparent complexity of the problem, it should be noted, however, that the score for a particular residue depends solely on the identity of the residue, and its structural environment (L_i, L_o, H_i, H_m, or H_o). As a result of this single dimensionality, it is straightforward to formulate a dynamic programming solution to the problem, which will ensure that the global optimum model will be found every time.

The overall problem of determining the optimal position and length of n transmembranal helices in a sequence of length m is divided into n subproblems: namely, determining the optimal position and length of a single transmembranal helix along with its associated C-terminal coil segment. Let s^{il} be the score associated with a transmembranal helix of length l at position i in the given sequence. This score is calculated according to the diagram shown in Figure 1, where the helix is divided into three sections (two caps of length 4, and a center region of length l – 8). Whether the cap and its associated loop are inside or outside depends on the initially specified membrane topology. In order to find the best set of s_j^{il}, we use a recursive algorithm almost identical to the algorithms used for pairwise sequence alignment (Needleman & Wunsch, 1970; Sellers, 1974). A score matrix $\mathbf{S}_j^i(i{:}1...n,j{:}1...m)$ is thus defined (see Figure 4):

$$\mathbf{S}_j^i = \max_{l=17\to25}\{s_j^{il} + \max_{k=i+l+A\to n}\{\mathbf{S}_{j-1}^k\}\}$$

where A is the minimum length of a loop segment.



	H1	H2	H3
1			
2			
3			
4	12450		
5	12126		
37	11101	11462	7134
38	10452	11478	9107
39	10059	11053	10032
40	10108	10050	9736
41	10014	10017	9061
93			3012
94			3107
95			-5072
96			-40291
97			1950

	H1	H2	H3
1			
2			
3			
4	19		
5	20		
37	17	18	19
38	23	24	21
39	22	21	22
40	19	20	24
41	21	17	25
93			21
94			22
95			17
96			17
97			19

FIGURE 4: A hypothetical score matrix for three transmembrane helices. The upper matrix holds the highest achievable path score for each cell, and the lower matrix stores the helix length which permits this score.

Having computed the score matrix **S**, the highest value in the column $j = 1$ is the score for the best path through the matrix, which represents the optimal lengths and positions of m transmembranal helices in the given sequence. It should also be noted that the highest value in column 2 is the optimal path score for $m - 1$ helices, but with inverted topology, and this can be extended to the other columns. In this way, only two score matrices need to be calculated to evaluate all possible membrane topologies for a given case: one with helix 1 (column 1) defined with the N-terminus on the inside, and the other with the helix 1 N-terminus on the outside. If we calculated two matrices for $m = 7$, one matrix would therefore provide optimal paths for topologies +7, −6, +5, −4, +3, −2, and +1, and the other would provide paths for −7, +6, −5, +4, −3, +2, and −1 (where +ve indicates the N-terminus inside). A further point to note is that the raw values in **S** do not include the appropriate score for the N-terminal loop, and this must be added to the appropriate matrix values. For example, consider a path starting in column 1, row 5. This initial cell represents the position (j5) of the first helix, and by definition implicitly represents an N-terminal loop of length 4.

To illustrate the matrix representation, Figure 4 shows a portion of the final score matrix for a hypothetical protein sequence encoding three transmembrane segments (N-terminus inside). Gray cells indicate cells which cannot be traversed by a valid 3-segment path, with the condition that loops are of length 3 or greater. Unlike a traditional sequence alignment matrix path, deletions are not permitted in the horizontal "sequence", which would represent omitted helical segments in the predicted structure. Figure 5 shows the interpretation of the indicated path in terms of segment positions and lengths.

To enable a different scoring table to be used to evaluate single-spanning topologies, these topologies are treated separately. This requires an extra two calculations of the score matrix **S**, though only for the trivial cases of a single helix in both topologies.

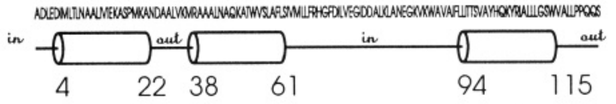

ADLEDMLTLNAALVEKASPMKANDAALVKMRAAALNAQKATWVSLAFLSVMLLFRHGFDLVEGIDDALKLANEGKVKWAVAFLLITSVAYHQKYRIALLLGSWVALLPPQQS
in
out
in
out
4
22
38
61
94
115

FIGURE 5: Predicted structure and topology relating to the optimal path shown in Figure 4.

PROGRAM IMPLEMENTATION

The algorithm described has been implemented in ANSI C on a number of Unix workstations and also on a PC compatible microcomputer. The results shown here were generated on both a Solbourne 5/602 and a Hewlett-Packard Apollo 9000/710 workstation. For computational efficiency, the propensities are scaled and converted to integers. Certain aspects of the final predicted structure and topology depend on a number of control parameters, as follows: Maxnhel is the maximum number of transmembrane helices that may be predicted, which also denotes the number of columns in the score matrix **S**. For a sequence of length n the maximum expected number of transmembrane helices is taken as $n/32$, with an upper limit of 20 helices. Minllen specifies the minimum length of the flanking regions (loops), and a value of 6 is generally used for this parameter. Minhlen and maxhlen specify the range of helix lengths that will be considered in the search, and values of 17 and 25, respectively, are used for these parameters. Those interested in obtaining the program (and machine–readable copies of the supplementary material) should send e-mail to jones@bsm.bioc.ucl.ac.uk.

RESULTS AND DISCUSSION

Perhaps the most important aspect of the method described here is that it provides not just a single final prediction, but a list of predictions for all achievable topologies. It is of course reasonable to pick the highest scoring prediction as the final result, but often it is important to take note of predictions whose scores are almost on a par with the optimum. Figure 6 shows the optimal models for all the achievable topologies of bacteriorhodopsin from *Halobacterium halobium* using the propensity values shown in Figures 3 and 4 (Tables I and II in the supplementary material). A low-resolution structure for this integral membrane protein has been determined by electron microscopic techniques (Henderson *et al.*, 1990), and has been found to comprise seven transmembrane helices, with the N-terminus located outside. The results in Figure 6 show that the native topology (seven helices, N-terminus outside) is clearly favored over all others.

Given that the correct topology of at least one integral membrane protein with a well-determined structure is favored over the possible alternatives, the next step is to evaluate the method over other examples. A significant problem again arises here in that very few integral membrane proteins have a known 3-D structure. Fortunately, the membrane topology of transmembrane proteins may be determined with a moderate degree of accuracy, and comparatively quickly by chemical, immunological, or genetic means (by the use of fusion proteins). The latter technique, whereby an alkaline phosphatase protein is genetically fused to the C-terminal end of part of the protein under study (Manoil & Beckwith, 1986), has provided topological information on many proteins over the past few years, and is the source of most of the topological information used here.

To evaluate the model recognition method, a set of 83 bacterial and eukaryotic integral membrane protein sequences

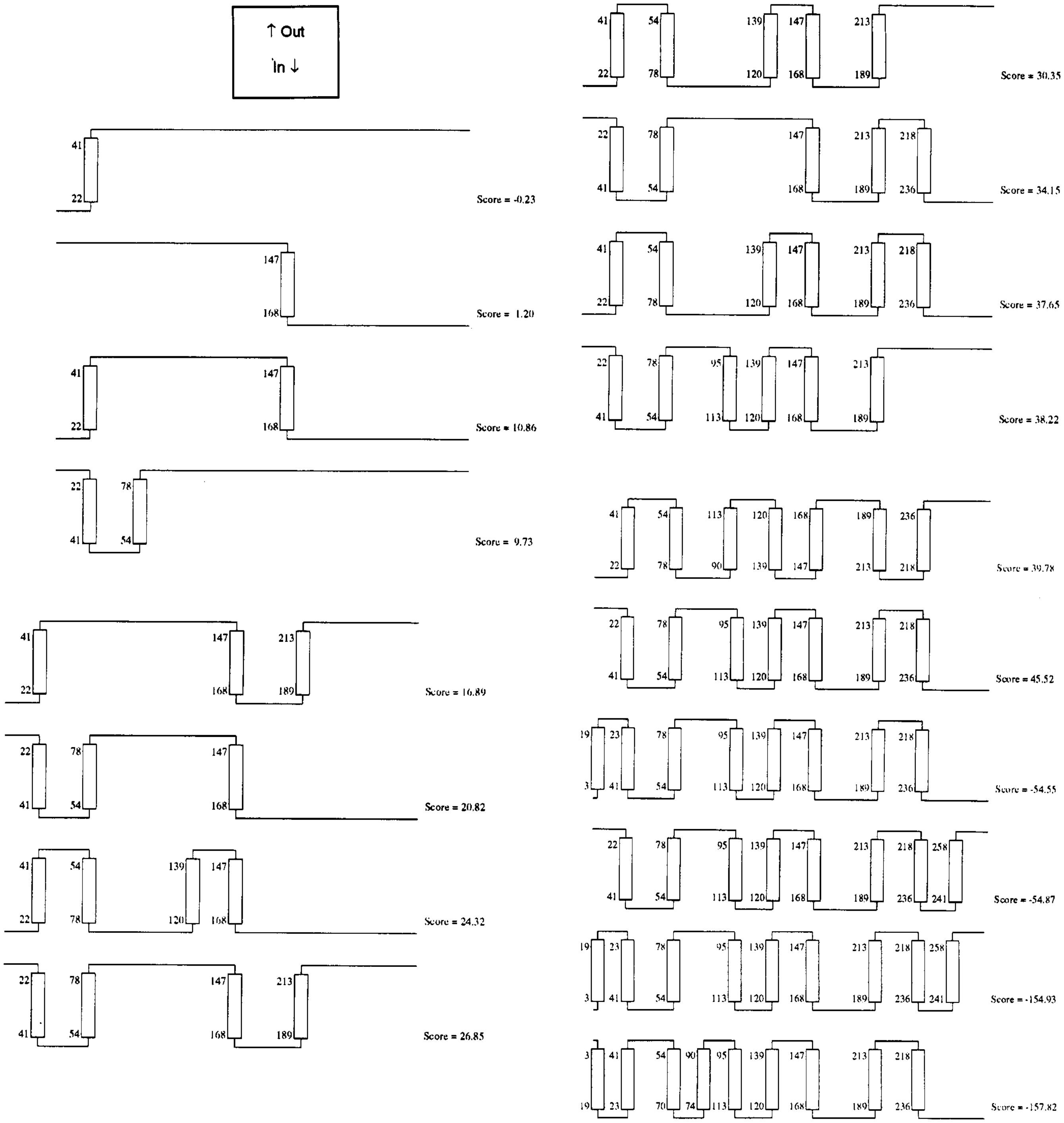

↑ Out
In ↓
41 22 Score = -0.23
147 168 Score = 1.20
41 22 147 168 Score = 10.86
22 41 78 54 Score = 9.73
41 22 147 168 213 189 Score = 16.89
22 41 78 54 147 168 Score = 20.82
41 22 54 78 139 120 147 168 Score = 24.32
22 41 78 54 147 168 213 189 Score = 26.85
41 22 54 78 139 120 147 168 213 189 Score = 30.35
22 41 78 54 147 168 213 189 218 236 Score = 34.15
41 22 54 78 139 120 147 168 213 189 218 236 Score = 37.65
22 41 78 54 95 113 139 120 147 168 213 189 Score = 38.22
41 22 54 78 113 90 120 139 168 147 189 213 236 218 Score = 39.78
22 41 78 54 95 113 139 120 147 168 213 189 218 236 Score = 45.52
19 3 23 41 78 54 95 113 139 120 147 168 213 189 218 236 Score = -54.55
22 41 78 54 95 113 139 120 147 168 213 189 218 236 258 241 Score = -54.87
19 3 23 41 78 54 95 113 139 120 147 168 213 189 218 236 258 241 Score = -154.93
3 19 41 23 54 70 90 74 95 113 139 120 147 168 213 189 218 236 Score = -157.82

FIGURE 6: Topology schematics showing the optimal achievable topologies for bacteriorhodopsin. The scores (nats) for each topology are shown along the side.

were used. These proteins were deemed to have a reliable experimentally determined topology either from topology records in the databank entry or from the literature (von Heijne & Gavel, 1988; von Heijne, 1992), though it must be pointed out that, without a full 3-D structure, this information is only preliminary. In cases where a sequence in the databank was specified as a precursor, and the location of the leader peptide given, the leader peptide was removed before proceeding with the prediction, or else the whole precursor sequence was used.

The results shown in Table 5 demonstrate that the proposed method for recognizing membrane topology models by a process of expectation maximization is highly successful, with 64 out of 83 being correctly predicted (34 out of 37 of the multispanning topologies were correct). Of the proteins with known structures, all 6 of the topologies have been correctly predicted, with success rates of 13/15 being achieved for the G-coupled receptors, 40/56 for proteins with experimentally well-determined topologies, and 5/6 for proteins with partial topological data. The locations of predicted helices were in close agreement with the experimental data, though due to the uncertainty in this data no attempt was made to quantify this accuracy. Most of the failures were due to overpredictions for large globular (eukaryotic) proteins with single membrane anchoring segments. Presumably in these cases, buried β-strands (or more rarely buried α-helices) are mistaken for membrane-spanning segments, and this is a recurrent problem in all membrane protein structure prediction methods. A few possible ways of detecting such mispredictions will be discussed later. Taking the human epidermal growth factor receptor prediction as an example, where a single-spanning membrane segment is located roughly halfway along the sequence at position 618, two extra helices are predicted. One of these

Table 5: Results of Predicting the Structure and Topology of 83 Proteins from a Mixture of Organism Classes[a]

(a) Results for Proteins with Known 3-D Structure or for Which a Homologous Protein Exists with Known 3-D Structure

protein	predicted topology	predicted segments	score	obsd topology	obsd segments	protein	predicted topology	predicted segments	score	obsd topology	obsd segments
archaerhodopsin (hals1)	out*	1: 17–35	5.17	out	1: 15–37	photosynthetic reaction center protein L chain (rhosh)	in*	1: 29–51	6.65	in	1: 34–57
		2: 48–72	4.71		2: 49–74			2: 85–102	3.11		2: 86–114
		3: 91–107	1.71		3: 89–106			3: 111–134	4.40		3: 117–142
		4: 114–133	2.75		4: 113–132			4: 174–198	3.85		4: 172–201
		5: 143–162	5.47		5: 143–163			5: 232–256	6.15		5: 227–253
		6: 180–197	3.92		6: 169–198	reaction center protein M chain (rhosh)	in*	1: 50–74	5.98	in	1: 54–81
		7: 212–230	1.23		7: 206–229			2: 114–130	4.12		2: 112–141
bacteriorhodopsin (halha) (see Figure 5.6)	out*	1: 22–41	4.96	out	1: 24–46			3: 147–171	3.62		3: 144–169
		2: 54–78	4.57		2: 52–76			4: 203–226	3.81		4: 199–227
		3: 95–113	1.92		3: 94–114			5: 268–291	4.84		5: 261–287
		4: 120–139	2.83		4: 122–141	reaction center protein H chain (rhosh)	out*	1: 12–31	4.44	out	1: 14–32
		5: 147–168	4.97		5: 151–171						
		6: 189–213	2.50		6: 181–205						
		7: 218–236	1.17		7: 217–239						
halorhodopsin (halsp)	out*	1: 7–28	4.41	out	1: 7–30						
		2: 40–63	3.98		2: 42–65						
		3: 88–106	3.77		3: 83–101						
		4: 113–132	2.88		4: 112–135						
		5: 141–159	5.31		5: 139–163						
		6: 180–203	2.80		6: 172–195						
		7: 211–229	1.12		7: 208–231						

(b) Results for G-Coupled Receptor Proteins

protein	predicted topology	predicted segments	score	expected topology	expected segments	protein	predicted topology	predicted segments	score	expected topology	expected segments
5-hydroxytryptamine 1A receptor (human)	out*	1: 41–63	5.17	out	1: 37–62	opsin RH3 (*Drosophila*)	out*	1: 59–83	4.79	out	1: 63–83
		2: 73–96	3.50		2: 74–98			2: 95–116	1.04		2: 96–115
		3: 111–132	3.06		3: 110–132			3: 132–153	3.26		3: 131–151
		4: 154–174	4.86		4: 152–177			4: 172–189	3.21		4: 172–192
		5: 196–214	4.09		5: 191–216			5: 220–241	5.36		5: 220–240
		6: 344–368	6.22		6: 345–366			6: 285–309	4.72		6: 289–309
		7: 378–401	1.43		7: 378–402			7: 321–340	0.53		7: 320–340
5-hydroxytryptamine 2 receptor (Chinese hamster)	out*	1: 76–99	6.24	out	1: 76–99	opsin RH2 (*Drosophila*)	out**	1: 59–82	5.37	out	1: 57–81
		2: 112–136	3.42		2: 111–132			2: 94–116	0.58		2: 94–119
		3: 147–171	1.95		3: 148–171			3: 132–153	4.09		3: 134–160
		4: 196–215	3.59		4: 192–215			4: 177–195	4.44		4: 173–196
		5: 234–258	6.13		5: 234–254			5: 227–250	4.37		5: 221–248
		6: 324–348	5.75		6: 325–346			6: 284–308	4.37		6: 284–307
		7: 357–380	2.91		7: 363–384						7: 315–339
adenosine A1 receptor (dog)	out*	1: 10–34	3.08	out	1: 11–33	opsin RH4 (*Drosophila*)	out*	1: 55–79	5.26	out	1: 59–79
		2: 45–69	4.64		2: 47–69			2: 91–110	0.60		2: 92–111
		3: 81–103	4.32		3: 80–102			3: 128–149	1.97		3: 127–147
		4: 124–146	5.60		4: 124–146			4: 168–185	3.12		4: 168–188
		5: 185–207	4.39		5: 177–201			5: 216–237	5.76		5: 216–236
		6: 236–259	4.60		6: 236–259			6: 281–305	4.76		6: 285–305
		7: 268–290	1.76		7: 268–292			7: 317–336	0.64		7: 316–336
adenosine A2 receptor (dog)	out**	1: 14–33	4.46	out	1: 8–30	opsin RH1 (blow fly)	out*	1: 50–73	4.37	out	1: 48–72
		2: 43–67	3.92		2: 44–66			2: 83–99	1.44		2: 85–110
		3: 78–100	4.75		3: 78–100			3: 123–144	3.47		3: 126–151
		4: 123–143	4.89		4: 121–143			4: 168–186	4.26		4: 164–187
		5: 182–204	4.28		5: 174–198			5: 215–239	5.09		5: 212–239
		6: 235–258	5.10		6: 235–258			6: 275–299	3.50		6: 275–298
					7: 267–290			7: 310–329	0.29		7: 306–330
α-1A adrenergic receptor (human)	out*	1: 57–81	5.51	out	1: 54–79	probable G protein-coupled receptor EDG-1 (human)	out*	1: 47–71	5.00	out	1: 47–71
		2: 91–114	2.52		2: 92–117			2: 81–104	1.76		2: 79–107
		3: 129–150	3.45		3: 128–150			3: 124–140	4.23		3: 122–140
		4: 171–193	5.84		4: 172–196			4: 160–182	6.13		4: 160–185
		5: 212–233	4.28		5: 210–233			5: 202–222	6.16		5: 202–222
		6: 308–332	6.35		6: 307–331			6: 257–277	6.15		6: 256–277
		7: 344–360	1.62		7: 339–363			7: 294–312	2.41		7: 294–314
α-2A adrenergic receptor (subtype C10, human)	out*	1: 34–58	5.72	out	1: 34–59	red-sensitive opsin (human)	out*	1: 58–77	4.22	out	1: 50–74
		2: 70–92	3.23		2: 71–96			2: 90–112	0.77		2: 87–112
		3: 108–129	2.31		3: 107–131			3: 130–149	3.58		3: 127–153
		4: 152–172	4.12		4: 150–173			4: 168–191	4.40		4: 166–189
		5: 195–217	6.14		5: 193–217			5: 219–240	7.02		5: 216–243
		6: 372–392	6.00		6: 375–399			6: 269–293	4.61		6: 266–289
		7: 410–429	1.95		7: 407–430			7: 302–324	0.12		7: 297–322
blue-sensitive opsin (human)	out*	1: 38–60	4.17	out	1: 34–57	rhodopsin (bovine)	out*	1: 37–61	5.17	out	1: 37–61
		2: 71–94	5.38		2: 71–96			2: 74–91	2.88		2: 74–98
		3: 111–130	3.82		3: 111–136			3: 114–133	2.69		3: 114–140
		4: 150–172	4.82		4: 150–173			4: 153–175	4.42		4: 153–173
		5: 202–220	6.34		5: 200–227			5: 203–223	6.00		5: 203–230
		6: 250–273	4.51		6: 250–273			6: 253–276	6.19		6: 253–276
		7: 284–305	0.12		7: 282–306			7: 286–307	0.58		7: 285–309
green-sensitive opsin (human)	out*	1: 58–77	4.78	out	1: 53–77						
		2: 90–112	0.71		2: 90–115						
		3: 130–149	3.58		3: 130–156						
		4: 168–191	4.66		4: 169–192						
		5: 219–240	6.41		5: 219–246						
		6: 269–293	5.16		6: 269–292						
		7: 302–324	0.65		7: 300–325						

Table 5 (Continued)

(c) Results for Proteins with Experimentally Determined Topologies

protein	predicted topology	predicted segments	score	expected topology	expected segments
4F2 cell-surface antigen heavy chain (human)	in*	1: 82–104	5.62	in	1: 82–104
ADP, ATP carrier protein (ricpr)	in*	1: 28–45 2: 62–82 3: 93–113 4: 146–168 5: 183–206 6: 219–237 7: 279–297 8: 324–341 9: 349–370 10: 378–396 11: 443–459 12: 466–482	2.45 2.43 5.55 2.27 4.19 5.05 3.88 2.22 4.96 2.98 2.02 5.92	in	1: 35–55 2: 69–89 3: 94–114 4: 149–169 5: 186–206 6: 220–240 7: 281–301 8: 322–342 9: 350–370 10: 381–401 11: 440–460 12: 467–487
Alzheimer's disease A4 protein precursor (human)	out*	1: 683–706	6.39	out	1: 683–706
asialoglycoprotein receptor 2 (mouse)	out**	1: 59–77 2: 106–125	5.18 0.42	in	1: 59–79
asialoglycoprotein receptor 1 (human)	in*	1: 40–59	5.51	in	1: 40–60
cation-dependent mannose-6-phosphate receptor precursor (human)	out*	1: 160–184	5.34	out	1: 160–184
epidermal growth factor receptor (*Drosophila*)	in*	1: 11–28 2: 1011–1035	0.92 6.02	out	1: 1011–1035
epidermal growth factor receptor (*Drosophila*)	in**	1: 26–49 2: 741–764	0.35 6.82	out	1: 733–764
epidermal growth factor receptor (human)	in**	1: 412–429 2: 619–643 3: 753–775	0.18 5.33 1.68	out	1: 622–644
fibronectin receptor α-subunit (mouse)	out*	1: 359–381	8.26	out	1: 356–381
chemotaxis motB protein (*E. coli*)	in*	1: 33–49	3.47	in	1: 28–49
cytochrome O ubiquinol oxidase subunit III (*E. coli*)	in*	1: 26–50 2: 68–91 3: 98–115 4: 138–162 5: 177–198	3.88 3.27 3.86 3.48 3.03	in	1: 33–51 2: 68–86 3: 103–121 4: 144–162 5: 186–204
cytochrome O ubiquinol oxidase operon protein CYOD (*E. coli*)	in*	1: 18–37 2: 44–66 3: 78–100	5.37 3.77 7.06	in	1: 19–37 2: 47–65 3: 82–100
cytochrome O ubiquinol oxidase operon protein CYOE (*E. coli*)	in*	1: 13–29 2: 38–56 3: 79–103 4: 111–127 5: 134–154 6: 161–183 7: 209–225 8: 233–249 9: 264–281	3.25 4.12 5.23 3.92 2.30 1.87 4.09 3.41 3.82	in	1: 11–31 2: 36–56 3: 85–105 4: 107–127 5: 133–153 6: 160–180 7: 208–228 8: 231–251 9: 264–284
cytochrome O ubiquinol oxidase subunit I (*E. coli*)	out*	1: 16–38 2: 57–79 3: 107–129 4: 136–160 5: 190–213 6: 230–253 7: 287–303 8: 310–332 9: 347–371 10: 380–403 11: 423–440 12: 456–479 13: 494–514 14: 588–604 15: 611–627	6.03 2.61 4.61 3.00 3.56 4.57 2.58 5.25 2.94 4.74 3.69 5.68 6.11 2.76 5.63	out	1: 18–36 2: 59–77 3: 103–122 4: 145–163 5: 196–214 6: 233–251 7: 278–297 8: 321–340 9: 349–367 10: 383–402 11: 411–430 12: 459–477 13: 496–514 14: 589–607 15: 615–633
cytochrome O ubiquinol oxidase subunit II (*E. coli*)	in*	1: 10–30 2: 43–67 3: 90–108	2.74 6.33 5.72	in	1: 10–30 2: 47–67 3: 89–109
glycophorin (pig)	out*	1: 63–85	6.86	out	1: 63–85
glycophorin A precursor (human)	out*	1: 73–95	5.41	out	1: 73–95
glycophorin C (human)	out*	1: 59–81	6.67	out	1: 58–81
granulocyte-macrophage colony-stimulating factor receptor precursor (human)	out*	1: 303–324	6.22	out	1: 299–324
hemagglutinin-neuraminidase (cdvo)	in*	1: 37–58	6.97	in	1: 35–54
hemagglutinin-neuraminidase (P14HA)	in**	1: 28–46 2: 299–321	5.85 0.90	in	1: 28–47
hemagglutinin-neuraminidase (measles virus)	in*	1: 36–58	6.80	in	1: 36–54
hexose phosphate transport protein (*E. coli*)	in**	1: 28–45 2: 56–80 3: 97–113 4: 120–136 5: 159–183 6: 191–210 7: 260–278 8: 327–344 9: 352–376 10: 422–446	1.53 2.15 3.46 2.77 1.47 5.27 2.64 5.17 4.53 5.38	in	1: 25–45 2: 59–79 3: 97–117 4: 118–138 5: 167–187 6: 190–210 7: 258–278 8: 299–319 9: 326–346 10: 351–371 11: 405–425 12: 427–447
HLA class II histocompatibility antigen, γ-chain precursor (human)	in*	1: 47–63	3.15	in	1: 47–72
immunity protein for colcicin A (*E. coli*)	in*	1: 17–37 2: 72–89 3: 107–123 4: 147–171	7.01 1.29 2.78 4.77	in	1: 17–37 2: 72–92 3: 104–124 4: 143–163
immunity protein for colicin E1 (*E. coli*)	in*	1: 9–25 2: 39–57 3: 84–104	2.62 1.87 3.91	in	1: 6–26 2: 37–57 3: 90–110
immunoglobulin G binding protein precursor (strsp)	out*	1: 391–409	3.91	out	1: 390–410
inner membrane protein secE (*E. coli*)	in*	1: 19–35 2: 45–61 3: 95–116	4.46 4.24 3.61	in	1: 20–37 2: 46–64 3: 94–112
insulin-like growth factor I receptor precursor (human)	out**	1: 903–927 2: 1157–1173	7.26 1.11	out	1: 906–929 4: 110–130
interleukin-2 receptor α-chain precursor (human)	out*	1: 220–240	5.02	out	1: 220–238
interleucin-2 receptor β-chain precursor (human)	out**	1: 50–66 2: 215–239	0.35 4.54	out	1: 215–239
lactose permease (*E. coli*)	in*	1: 10–34 2: 46–66 3: 75–96 4: 103–125 5: 145–162 6: 169–188 7: 222–239 8: 260–283 9: 291–313 10: 321–337 11: 349–370 12: 385–409	5.67 3.47 6.28 3.17 4.24 5.26 2.42 1.93 1.71 0.48 2.11 4.13	in	1: 8–28 2: 46–66 3: 79–99 4: 103–123 5: 146–166 6: 168–188 7: 220–240 8: 264–284 9: 292–312 10: 316–336 11: 350–370 12: 383–403
low-affinity nerve growth factor receptor precursor (human)	out**	1: 195–211 2: 223–244	1.57 4.85	out	1: 221–242
low-affinity immunoglobulin ϵ FC receptor (human)	in*	1: 24–45	6.16	in	1: 22–47
maltose transport inner membrane protein (*E. coli*)	in*	1: 17–34 2: 41–58 3: 67–91 4: 282–306 5: 319–336 6: 371–392 7: 417–436 8: 484–504	4.36 3.74 5.35 6.48 3.97 3.29 1.29 6.07	in	1: 17–35 2: 40–58 3: 73–91 4: 277–295 5: 319–337 6: 371–389 7: 418–436 8: 486–504
matrix (M2) protein (iann)	out*	1: 26–43	4.44	out	1: 25–42
melibiose carrier protein (*E. coli*)	in*	1: 6–24 2: 31–49 3: 74–95 4: 102–126 5: 145–162 6: 176–193 7: 228–252 8: 265–281	0.43 2.75 4.48 2.63 3.02 5.35 3.34 2.48	in	1: 16–36 2: 42–63 3: 74–95 4: 110–130 5: 146–166 6: 174–194 7: 236–256 8: 263–283

Table 5 (Continued)

(c) Results for Proteins with Experimentally Determined Topologies (Continued)

protein	predicted topology	predicted segments	score	expected topology	expected segments
melibiose carrier protein (*E. coli*) (continued)		9: 292–314 10: 321–345 11: 376–392 12: 401–425	5.42 4.50 4.51 4.89		9: 293–314 10: 326–346 11: 374–394 12: 401–422
myelin-associated glycoprotein, long form precursor (mouse)	out*	1: 495–514	5.39	out	1: 498–517
myelin p0 protein precursor (human)	out*	1: 125–149	7.30	out	1: 124–151
NB glycoprotein (inbbe)	out*	1: 21–44	4.17	out	1: 19–40
neprilysin (human)	in*	1: 28–49	5.60	in	1: 28–50
oligopeptide permease protein OPPB (salty)	in*	1: 9–27 2: 100–121 3: 134–158 4: 173–190 5: 228–250 6: 274–298	3.59 5.39 4.93 3.74 3.38 4.41	in	1: 10–29 2: 100–121 3: 138–158 4: 173–190 5: 227–250 6: 272–293
oligopeptide permease protein OPPC (salty)	in*	1: 38–59 2: 105–129 3: 140–157 4: 164–180 5: 214–236 6: 270–290	5.83 5.04 2.32 2.60 3.90 5.04	in	1: 38–59 2: 103–122 3: 140–160 4: 164–180 5: 216–236 6: 268–290
platelet glycoprotein IB β-chain precursor (human)	in**	1: 123–147	4.20	out	1: 122–146
polymeric-immunoglobulin receptor (human)	in**	1: 621–643	4.65	out	1: 621–643
ribophorin I precursor (rat)	out*	1: 416–433	5.42	out	1: 417–435
secY protein (bacsu)	in*	1: 18–39 2: 67–87 3: 115–132 4: 149–166 5: 174–191 6: 210–234 7: 269–291 8: 310–329 9: 367–386 10: 394–410	3.49 2.88 2.87 4.51 2.77 5.60 2.97 4.51 5.41 1.97	in	1: 18–39 2: 59–80 3: 115–132 4: 148–167 5: 174–192 6: 217–234 7: 268–291 8: 310–329 9: 367–386 10: 392–410
signal peptidase I (*E. coli*)	out**	1: 7–23 2: 59–76 3: 86–103	2.98 2.09 1.18	out	1: 4–22 2: 58–76
sucrase-isomaltase, intestinal (human)	in**	1: 11–32 2: 622–639	7.15 2.72	in	1: 13–32
T-cell receptor β-chain precursor (rabbit)	out*	1: 293–313	2.08	out	1: 292–313
tetracycline resistance protein tn10 (*E. coli*)	in*	1: 7–30 2: 43–62 3: 73–95 4: 102–119 5: 130–152 6: 161–179 7: 212–234 8: 244–266 9: 277–295 10: 302–324 11: 336–357 12: 367–388	3.81 2.78 3.70 2.80 4.20 5.00 3.09 2.85 2.92 3.13 3.64 6.82	in	1: 7–27 2: 42–62 3: 81–101 4: 102–122 5: 133–153 6: 159–179 7: 201–221 8: 240–260 9: 276–296 10: 298–318 11: 337–357 12: 369–389
thrombomodulin precursor (human)	out**	1: 162–178 2: 185–202 3: 495–518	1.45 1.10 5.73	out	1: 516–539
transferrin receptor protein (human)	in**	1: 66–86 2: 540–558 3: 735–751	5.96 2.50 0.24	in	1: 63–88
tyrosine kinase receptor CEK2 precursor (chick)	out**	1: 346–370 2: 652–668	6.82 1.28	out	1: 346–370
UDP-*N*-acetylglucosamine–dolichylphosphate *N*-acetylglucosaminephosphotransferase (crilo)	out*	1: 11–32 2: 59–79 3: 95–114 4: 126–142 5: 157–179 6: 186–208 7: 222–240 8: 248–268 9: 275–297 10: 379–397	3.62 6.06 5.96 2.45 3.06 0.94 4.19 2.69 0.92 3.03	out	1: 8–33 2: 59–80 3: 96–115 4: 127–146 5: 166–185 6: 196–212 7: 223–241 8: 254–270 9: 276–295 10: 380–298

(d) Results for Transmembrane Proteins with at Least Some Experimental Topological Data

protein	predicted topology	predicted segments	score	expected topology	expected segments
5-hydroxytryptamine 3 receptor precursor (mouse)	out*	1: 224–247 2: 259–276 3: 284–307 4: 438–458	4.44 2.52 4.40 3.20	out	1: 223–249 2: 255–273 3: 283–301 4: 442–461
cytochrome B561 (bovine)	in*	1: 37–59 2: 77–94 3: 107–128 4: 148–172 5: 183–201 6: 221–241	4.09 1.58 4.57 3.71 4.43 4.72	in	1: 38–60 2: 75–97 3: 107–129 4: 145–167 5: 185–207 6: 219–241
gap junction β-1 protein (human)	in*	1: 23–40 2: 76–96 3: 131–155 4: 190–214	2.89 2.49 1.78 4.02	in	1: 22–41 2: 75–94 3: 130–149 4: 188–207
gap junction β-1 protein (clawed frog)	in*	1: 23–40 2: 76–98 3: 132–156 4: 189–213	2.57 2.40 1.54 3.99	in	1: 22–41 2: 75–94 3: 130–149 4: 188–207
gap junction β-1 protein (rat)	in*	1: 23–40 2: 76–96 3: 131–155 4: 190–214	2.89 2.49 1.78 4.02	in	1: 22–41 2: 75–94 3: 130–149 4: 188–207
phosphotransferase enzyme II, mannitol specific (*E. coli*)	in**	1: 19–43 2: 51–67 3: 79–103 4: 133–154 5: 161–181 6: 243–260 7: 269–292 8: 311–334	4.41 2.74 2.90 5.51 1.17 0.66 5.43 4.74	in	1: 25–44 2: 51–69 3: 135–154 4: 166–184 5: 274–291 6: 314–333

[a] Again, in all cases, the protein under test was excluded from the calculation of the topogenic parameters, along with any related sequences (sequence identity >25%). Topology entries indicate the location of the N-terminus; the following segments thereafter alternate in/out. Correct predictions are indicated with an asterisk, and incorrect predictions with two asterisks. The locations of some of the helices in the melibiose carrier protein, including the first two, are not experimentally determined, and it would appear that the locations predicted here and more reasonable. Key to organisms: ricpr, ***Rickettsia prowazekii***; hals1, *Halobacterium* sp. (strain Aus-1); halsp, *Halobacterium* sp.; halha, *Halobacterium halobium*; cdvo, canine distemper virus; pih4a, human parainfluenza 4A virus; strsp, *Streptococcus* sp.; iaann, influenza a virus; inbbe, influenza b virus; salty, *Salmonella typhimurium*; rhosh, ***Rhodobacter sphaeroides***; bacsu, ***Bacillus subtilis***; crilo, ***Cricetulus longicaudatus***.

helices can be eliminated on the basis of a marginal helix score (0.18 nat), but the score for the other helix (1.68 nats) is reasonable.

Of the multispanning proteins, the general trend in misprediction is toward underprediction, though the success rate (34/37) was very high for these proteins. For example, the top-scoring topology for *Escherichia coli* hexose phosphate transport protein includes only 10 of the expected 12 helices. In the case of the 12 helix topologies, helix 11 only achieves a score of −0.164 nat, which is of course below the set cutoff of 0.1 nat. If this cutoff is not applied, then the highest scoring topology is found to correspond with the one which has been

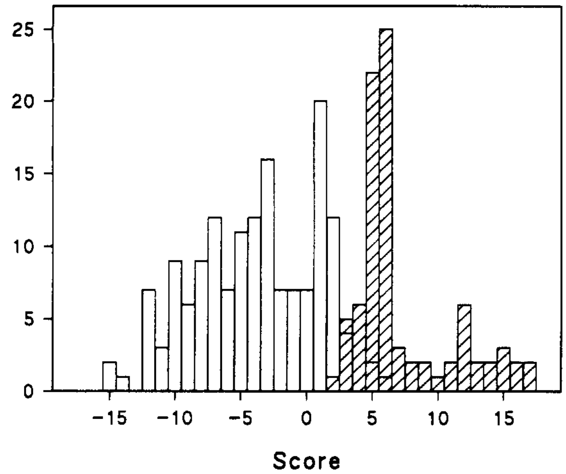


FIGURE 7: Distribution of maximum helix scores for transmembrane (hatched bars) and globular proteins.

experimentally determined. Low scoring helices appear to be common in the larger helix bundles: helix 10 of lactose permease, for example, only has a prediction score of 0.48 nat. This suggests, perhaps, that the helix score cutoff should not be applied to topologies involving more than six helices, though this is not yet verified, and is in any case hard to justify. Presumably, though, in large transmembrane helical bundles, fairly hydrophilic helices may be accommodated by means of shielding from the lipid environment by neighboring helices.

Individual helix scores also serve as useful indicators as to whether an unknown protein truly contains transmembrane segments. For the same reasons as those already offered for overpredicting transmembrane segments in large membrane-associated globular domains, methods for transmembrane helix prediction are prone to overpredicting transmembrane segments in entirely water-soluble globular proteins. Figure 7 shows the distribution of maximum helix scores (i.e., the score of the highest scoring predicted transmembrane segment) for the 83 integral membrane proteins listed in Table 5, and a representative set of 155 globular proteins of known 3-D structure. Using these distributions, a cutoff value can be estimated for which the overlap between the two sets is minimized. Minimum overlap is obtained for cutoff values in the range of 2.8–3.2, in which case one integral membrane protein (SECE_ECOLI) is assigned as globular and five globular proteins are assigned as integral membrane proteins. Edelman (1993) has determined an optimal set of parameters (hydrophobicity scale, window shape, and window size) for accurately predicting transmembrane segments using standard hydrophobicity profile analysis. The false positive rate for globular proteins was found to be 3 out of 14 (21%) chains, which can be compared against the rate demonstrated by the method described here (5 out of 155 chains (3%). The implication of these results is that integral membrane proteins require at least one strongly predicted (generally speaking strongly hydrophobic) segment. Perhaps this "major" helical segment forms part of the chain that is initially inserted into the membrane.

Encouragingly, the integral membrane proteins of known 3-D structure, or which have relatives of known structure, are correctly predicted by the method. In view of this, it is interesting to observe that the structure of the opsins is not predicted with great certainty. Despite confident prediction of bacteriorhodopsin, archaerhodopsin, rhodopsin, and most of the G-coupled receptors, most of the opsins and the adenosine A2 receptor, which are believed to have a seven-helix structure broadly similar to that observed in bacteriorhodopsin, have weakly predicted final helices, and in some cases the predicted topology misses this helix completely. In the absence of firm experimental data it is quite possible that these proteins only have six transmembrane segments, though this is not expected.

One interesting question that arises from looking at the enumerated segment arrangements (Figure 6, for example) is whether the helices which are located in the topologies involving the fewest helices (in particular the two-helix topologies), could be the helices which are assembled early on in the folding process itself. Unfortunately there is no experimental evidence available to adequately answer this question, but given the fact that the computer algorithm is keyed primarily on the hydrophobicity of sequence segments, and that this is the principal contribution to the stability of transmembrane protein structures, then this premise is at the very least tenable. In addition, the observation made earlier that integral membrane proteins incorporate at least one strongly predicted transmembrane segment meshes well with this suggestion.

CONCLUSION

The proposed method for membrane protein structure prediction appears to be very powerful. The most important point to note, however, is that it carefully considers all possible predictions in arriving at the final highest-expectation model, and ranks the alternative predictions alongside. As any prediction algorithm will have only a limited degree of accuracy (just over 77% in this case), it is vital that alternatives be considered when making use of the prediction results. Where, say, the top two topologies have almost indistinguishable scores, then the final prediction must be taken as being *either* topology, not just a single topology, which is the form of output from previous membrane structure prediction methods. It is important to realize that the method described here need not only make use of the topogenic parameters calculated here. Indeed any scoring system that can be encoded as a 1-D vector can be incorporated into the expectation maximization methodology. Despite the effectiveness of the parameters described, it is certain that improvements can be made, if only by an extension to the data set used in their compilation. As more experimental data become available, therefore, it is hoped that the predictive power of the parameters will increase.

Several improvements to the proposed method can be envisaged. The most immediate problem that needs solution is the problem of overpredicting large globular proteins with single-membrane-spanning segments. One possibility here is to calculate *directional* (i.e., pairwise) parameters for the scoring of helical segments, similar to those used in the GOR secondary structure prediction method (Garnier *et al.*, 1978), which would allow residue pair information to be incorporated. Including pair information will hopefully allow buried globular structure, in particular buried β-structure, to be discriminated from membrane-spanning helices, and preliminary experiments seem to bear this out.

Another important development would be to extend the simple in/out helical bundle model to encompass other structural elements observed in membrane-associated proteins. Ion-channel proteins, for example, include highly amphipathic helices in their overall topology, which will not score well with parameters biased toward strictly lipophilic helices. A rather more distant prospect is the consideration of membrane-spanning β-structure. As mentioned earlier, the sole example of an integral membrane protein containing β-structure is porin, and in this case the structure contains only β-strands, formed into a single β-barrel. It is not yet known whether it is possible for an $\alpha\beta$ protein to integrate into a membrane.

Without more information as to the likelihood of finding β-strands in an integral membrane protein structure, it is not possible to attempt to extend the recognition models along these lines. Clearly if it proves to be the case that both helices and strands can be found in almost any mixture in membrane proteins, then it will be necessary to alter the method such that only *observed* topologies are considered, as is the case for globular protein fold recognition.

The simplest and perhaps most powerful way to enhance the model recognition method is perhaps the use of *multiply aligned sequence families*. Rather than attempting to predict the optimal topology for a single sequence, it is clear that better discrimination will be achieved by summing the parameters over an aligned block of sequences. Early results along these lines are very promising.

ACKNOWLEDGMENT

We would like to thank Michael Green and Frances Richardson for useful discussion and comments.

SUPPLEMENTARY MATERIAL AVAILABLE

Tables of topogenic parameters and χ^2 probabilities for the observed frequencies of occurrence of the structural states for single-spanning and multispanning segments (4 pages). Ordering information is given on any current masthead page.

REFERENCES

Argos, P., Rao, J. K. M., & Hargrave, P. A. (1982) *Eur. J. Biochem. 128*, 565–575.

Bairoch, A., & Boeckmann, B. (1991) *Nucleic Acids Res. 19*, 2247–2249.

Bowie, J. U., Lüthy, R., & Eisenberg, D. (1991) *Science 253*, 164–170.

Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., & DeLisi, C. (1987) *J. Mol. Biol. 195*, 659–685.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., & Michel, H. (1985) *Nature 318*, 618–624.

Edelman, J. (1993) *J. Mol. Biol. 232*, 165–191.

Eisenberg, D., Schwarz, E., Komaromy, M., & Wall, R. (1984) *J. Mol. Biol. 179*, 125–142.

Engelman, D. M., Steitz, T. A., & Goldman, A. (1986) *Annu. Rev. Biophys. Biophys. Chem. 15*, 321–353.

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978) *J. Mol. Biol. 120*, 97–120.

Grantham, R. (1974) *Science 185*, 862–864.

Henderson, R., Baldwin, J. M. Ceska, T. A., Zemlin, F., Beckmann, E., & Downing, K. H. (1990) *J. Mol. Biol. 213*, 899–929.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992) *Nature 358*, 86–89.

Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105–132.

Landolt-Marticorena, C., Williams, K. A., Deber, C. M., & Reithmeier, R. A. F. (1993) *J. Mol. Biol. 229*, 602–608.

Manoil, C., & Beckwith, J. (1986) *Science 233*, 1403–1408.

Nakashima, H., & Nishikawa, K. (1992) *FEBS Lett. 303*, 141–146.

Needleman, S. B., & Wunsch, C. D. (1970) *J. Mol. Biol. 48*, 443–453.

Schiffer, M., Chang, C.-H., & Stevens, F. J. (1992) *Protein Eng. 5*, 213–214.

Sellers, P. H. (1974) *J. Comb. Theor. 16*, 253–258.

Stirk, H. J., Thornton, J. M., & Howard, C. R. (1992) *Intervirology 33*, 148–158.

von Heijne, G. (1981) *Eur. J. Biochem. 120*, 275–278.

von Heijne, G., & Gavel, Y. (1988) *Eur. J. Biochem. 174*, 671–678.

von Heijne, G. (1992) *J. Mol. Biol. 225*, 487–494.

Weiss, M. S., & Schulz, G. E. (1992) *J. Mol. Biol. 227*, 493–509.