Article No. mb982107 *J. Mol. Biol.* (1998) **283**, 489–506

JMB

Principles Governing Amino Acid Composition of Integral Membrane Proteins: Application to Topology Prediction

Gábor E. Tusnády and István Simon*

Institute of Enzymology Biological Research Center Hungarian Academy of Sciences, H-1518 Budapest P.O. Box 7, Hungary

**Corresponding author*

A new method is suggested here for topology prediction of helical transmembrane proteins. The method is based on the hypothesis that the localizations of the transmembrane segments and the topology are determined by the difference in the amino acid distributions in various structural parts of these proteins rather than by specific amino acid compositions of these parts. A hidden Markov model with special architecture was developed to search transmembrane topology corresponding to the maximum likelihood among all the possible topologies of a given protein. The prediction accuracy was tested on 158 proteins and was found to be higher than that found using prediction methods already available. The method successfully predicted all the transmembrane segments in 143 proteins out of the 158, and for 135 of these proteins both the membrane spanning regions and the topologies were predicted correctly. The observed level of accuracy is a strong argument in favor of our hypothesis.

Keywords: hidden Markov model; transmembrane α-helices; topology prediction for helical transmembrane proteins; distribution of amino acids

Introduction

Integral membrane proteins play important and functionally diverse roles in living cells. So far, two basic classes are known, according to the structure of the membrane spanning segments. In the first class, all the transmembrane segments form an α-helical structure with lengths of 17 to 25 amino acid residues (von Heijne, 1994). Members of the second class are only known in the bacterial outer porins that have a 16-stranded β-barrel structure (Weiss & Schulz, 1992). While experimental structure determinations of globular proteins by means of X-ray crystallography are becoming more routine (Lattman, 1994), we cannot nurse such hopes for integral membrane proteins, due to the difficulties in crystallization of these proteins, though there are some new encouraging methods in sight (Gouaux, 1998).

However, it is commonly accepted that topology prediction of membrane proteins is easier, and results in higher accuracy than the prediction of the secondary structure of globular proteins. The number of known sequences is increasing rapidly, resulting in a large gap between that and the number of known structures. Since prediction methods are the most convenient and least expensive ways of determining proteins structures, there is a great demand for developing efficient prediction methods. In addition, comparison of prediction methods based on different ideas can help to reveal the principles governing the structure formation of proteins.

The development of prediction of transmembrane helices in integral membrane proteins proceeded *via* several steps. The first approaches were based on hydrophobicity analyses (Kyte & Doolittle, 1982; Eisenberg *et al.*, 1984; Engelman *et al.*, 1986; Cornette *et al.*, 1987; Esposti *et al.*, 1990; Ponnuswamy & Gromiha, 1993; Gromiha & Ponnuswamy, 1995), i.e. they used information only about the amino acids that contributed to the formation of transmembrane helices. Their accuracy could be increased by exploiting information not only from transmembrane segments: namely, by considering the different charge distribution between the inside and outside loops

Abbreviations used: HMM, hidden Markov model; NFL, non-fixed length state; FL, fixed length state.

E-mail address of the corresponding author: simon@enzim.hu

0022–2836/98/420489–18 $30.00/0

(Boyd *et al*., 1987; Hartmann *et al*., 1989; von Heijne, 1992; Sipos & von Heijne, 1993). As the number of experiments dealing with topology increased in the last few years, resulting in more reliable data, several statistical procedures were developed by applying whole amino acid distributions in various structural parts of proteins for the predictions (Jones *et al*., 1994). Using the advantages of neural network-based algorithms and combining prediction methods with multiple alignments (Persson & Argos, 1994, 1996; Lohmann *et al*., 1994; Rost *et al*., 1995, 1996; Casadio *et al*., 1996), the accuracy of the topology prediction reached the 70 to 80% level, while the accuracy of the prediction of the transmembrane helices reached the 90 to 95% level.

In a previous paper from our group a new method was used for sequence alignment of transmembrane proteins having a very low level of sequence similarities (Cserző *et al*., 1994). By this method we were able to locate the corresponding transmembrane segments and the method also give a high score for all pairs of transmembrane helices, indicating that certain transmembrane characteristics (namely the amino acid composition of these segments) are more relevant than the actual sequence similarity in the alignment. A prediction method based on this observation works well on a set of prokaryotic integral membrane proteins (Cserző *et al*., 1997). The application of the amino acid composition in distinguishing between the extracellular and intracellular proteins (Nakashima & Nishikawa, 1994) or in defining the folding class of proteins (Chou, 1995) shows that the amino acid composition of proteins contains enough information to predict their structure in ''large resolution''.

Studying amino acid similarity in a large database by means of independence divergence calculation indicates that from the viewpoint of structure formation amino acids may be classified into slightly different groups than one would expect on the basis of their physico-chemical parameters (Tusnády *et al*., 1995). Since there is a big difference between the physical environments of the membrane-spanning segments and the cytoplasmic or extracytoplasmic sides of the membrane proteins, it is not suprising that the amino acid compositions of these parts are different. Therefore, it seems reasonable to expect that a more accurate prediction can be developed when the amino acid compositions of these segments are considered instead of using physicochemical parameters like the hydrophobicity of the amino acids. Since integral membrane proteins have functionally diverse roles in cells and they are in different environments, these facts must be reflected in their amino acid compositions. Thus, enforcing some predetermined or common amino acid compositions of the structural parts of these proteins in topology prediction may produce false results.

Our method is based on the hypothesis that the differences between the amino acid distributions in the various structural parts are the main driving force in the folding of the membrane proteins, i.e. the topology of transmembrane proteins may be determined by the simple fact that the amino acid compositions of the various structural parts do show maximum differences rather than by enforcing specific compositions in these parts. The difference between two distributions can be characterized by the divergence function (Kullback, 1959; Gokhale & Kullback, 1978). Divergence calculation was demonstrated to be a useful tool in sequence database analyses in our earlier work (Tusnády *et al*., 1995). Here we use the sum of divergence values between the distribution of amino acids of the structural parts and the distribution of residues in the whole protein to measure differences in the amino acid distributions of the structural parts. This sum differs only in a constant from the log-likelihood, therefore the topology of membrane proteins can be determined if their amino acid sequences can be segmented to some part (e.g. inside, outside and membrane) in such a way that the product of the relative frequencies of the amino acids of these segments along the amino acid sequence should be maximal. Using more types of structural parts or enabling some controls on the length of the various segments may enhance the power of the method. We can solve this task with use of hidden Markov model (HMM).

HMM is widely used in bioinformatics. The most widespread use of this method is in aligning sequences and generating profiles for protein families (Baldi *et al*., 1994; Krogh *et al*., 1994a; Hughey & Krogh, 1996). The profile shows the common sequence motifs of biopolymers (Lawrence & Reilly, 1990) or can be used for database searching for finding new sequence homologs for a given family (White *et al*., 1993; Krogh *et al*., 1994b; Borodovsky *et al*., 1995). A special application of this alignment procedure is in protein topology prediction using secondary structure sequences (Francesco *et al*., 1997). Secondary structure predictions not based on alignment were also developed (Asai *et al*., 1993; Stultz *et al*., 1993; White *et al*., 1993), though their accuracies were modest.

In contrast with other prediction methods HMM can be suited to particular problems. Any actual structural knowledge may be incorporated into the model's architecture in order to increase its prediction power and to learn more about these proteins. Here, a special HMM is described showing that the maxima of the likelihood function on the space of all possible topologies of a given amino acid sequence correlate with the experimentally established topology. The accuracy of this method was tested in three different data sets. Prediction methods published earlier were compared with our method, to uncover the principles governing

the structure formation of integral membrane proteins.

Results and discussion

The hidden Markov model

Investigations of the transmembrane topology of proteins give the impression that transmembrane segments are not located randomly in the sequences. These segments tend to group. To test this hypothesis the length distribution of the segments between transmembrane helices or at the ends of polypeptide chains was checked. In a purely random case the distribution of these segments would be close to geometric distribution as shown in Figure 1, but segments of transmembrane proteins show a different distribution. Short loops with lengths between around five and 30 amino acid residues were observed significantly more often than would be expected, when transmembrane segments were placed into the sequences randomly. Building this distribution into a prediction method may increase its accuracy.

This particular loop length distribution may be the consequence of the structure of the membrane and its environment. The asymmetry of lipid composition between the two halves of the lipid bilayer in most membranes has been well known for a long time (Bergelson & Barsukov, 1977; Rothman & Lenard, 1977). While phospholipids are more abundant in the cytoplasmic part of membranes, glycolipids are found mostly in the extra-cytoplasmic part. It was shown that the orientation of membrane depends on the anionic phospholipid content of the membrane, which suggests that interactions between the negatively charged head of phospholipids and positively charged amino acid side-chains affect the orientation of membrane proteins (van Klompenburg *et al*., 1997). Keeping in mind this feature of membranes, one would expect characteristic length and amino acid distribution in short loops between transmembrane helices and in the polypeptide chains close to the helices.

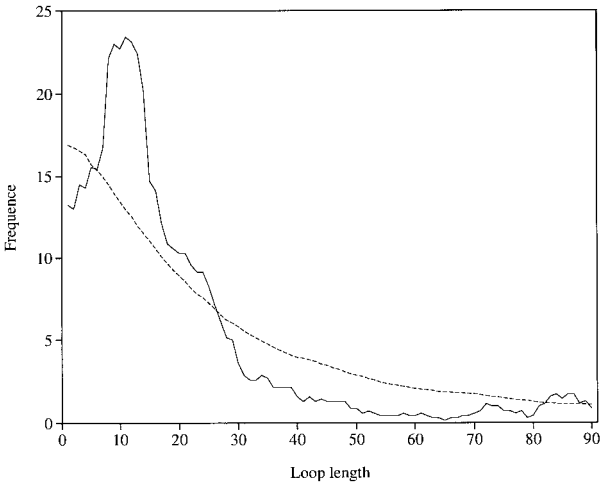

Figure 1. Distribution of loop lengths. The continuous line shows the length distribution of the non-membraneous part of the polypeptide chains in the reference data set. The broken line shows this distribution in the random sample in which transmembrane segments were shuffled for each protein in the reference data set; thus the number of membrane spanning segments remained the same, but their places were altered.

The architecture of HMM developed for topology was designed to exploit these particular properties of integral membrane proteins as well as the generally considered features. The model consists of five structural states, as shown in Figure 2. The five states are as follows: inside loop, inside helix tail, membrane helix, outside helix tail and outside loop. The helix parts are embedded in the membrane. The term loop means the longer part of a sequence outside the membrane, which can form a domain or a simpler structure. The tail is the elongation of the membrane helix, and it can be followed by a loop or another tail, forming a short loop interacting with the outside or inside part of the membrane. Note that this model is similar to that used by Jones *et al*. (1994); the differences are in the localizations and in the interpretation of helix tails, which were called helix ends in that study. While helix tails are not in the membrane, helix ends are the very ends of helices located in the membrane.

The power of the model lies in the architecture of possible transitions between states. According to the observations that the length distribution of the long loops (lengths above 30 residues) is close to geometric distribution, but the length of the short loops (about 5 to 30 residues) between helices follows a special distribution, two types of states were defined. These two types are the non-fixed length (NFL) and the fixed length (FL) states. From an NFL state there are only two possible transitions: one to the same state, which increases the length of this state and the other to the next state. This simple architecture of the NFL type transition matrix ensures that the length of this state can be arbitrary and the distribution of the lengths is geometric. The structure of the FL state is more complex. This state is split into MAXL substates in order to limit its length to between a minimum and a maximum (MINL and MAXL, respectively). There is only one possible transition from each of the first MINL substates and it is to the next substate. In each substate between MINL and MAXL there is another possible transition, which is to jump from the current state to the next state. The observation-symbol probabilities of substates in an FL state are the same, while transition probabilities are different between substates MINL and MAXL. The type of loops is defined as NFL, while tail and helix states are defined as FL. The next states are determined by the natural structure of the membrane proteins; for example, after an inside loop, the next state is the inside helix tail, then the helix, then the outside tail etc. The tail state on both sides

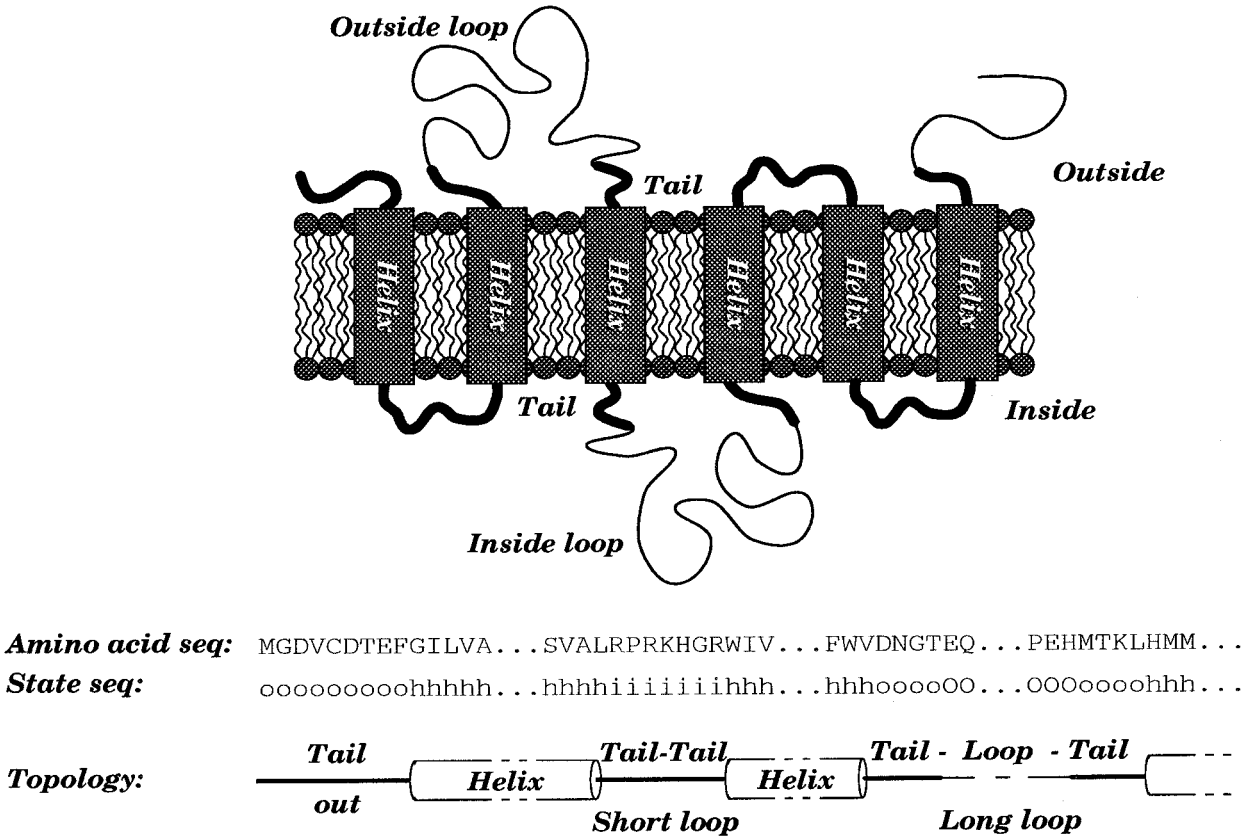

Figure 2. Structural states defined for a typical helical transmembrane protein. The five states are: inside loop (I), inside tail (i), membrane helix (h), outside tail (o) and outside loop (O). Tails (thick lines) are thought to interact with the inside or outside parts of the membrane, while loops (thin lines) do not. Two tails between helices can form a short loop, but longer loops are formed by tail-loop-tail sequences.

of the membrane, coming after the helix state, can be followed by another tail or by a loop; that is, the state sequence between two helix states can be a tail-tail forming a short loop or a tail-loop-tail resulting in a long loop (see Figure 2). Short loops are thought to be associated with the heads of phospholipids, while long loops form a well-defined structure in the cytosol or in the other side of the membrane, but their very ends interact with the membrane. The architecture of the possible transitions is shown in Figure 3.

The prediction method based on this model has three steps. First, the initial estimates of HMM parameters (the initial state, the observation symbol and the state transition probabilities) have to be set. The parameters can be chosen by random values, or by predetermined values. The next step is the optimization of these parameters for the amino acid sequence studied or for homolog sequences. The third step is to find the best state sequence by the so-called Viterbi algorithm, given the model and the parameters. Elements of the state sequence show the localization of each amino acid in the query sequence. The mathematical details of these procedures are given in Materials and Methods. An excellent tutorial for using HMMs was written by Rabiner (1989). By applying random values in parameter settings optimization produced various results, since the likelihood function over the sequence has many local optima. To avoid this problem iteration was started from a predetermined parameter set and the pseudocount method was used during the iteration process. Values of the parameters and the pseudocount array were derived from the amino acid sequences of transmembrane proteins whose topologies are experimentally well defined (see Materials and Methods). Since optimization of the parameters can work for multiple sequences (multiple observations), prediction can be made using multiple sequence information. One of the advantages of HMM is that related proteins do not have to be aligned before the prediction.

Prediction efficiency on various data sets

The prediction power of the newly developed HMM was tested on three different data sets, collected earlier for transmembrane prediction methods: 83TMP (Jones *et al.*, 1994), 48TMP (Rost *et al.*, 1996) and prokTMP (Cserző *et al.*, 1997), respectively (see Materials and Methods). The results on multiple sequences shown in Table 1 demonstrate the accuracy of the proposed prediction method for recognizing transmembrane topology. For the three data sets the transmembrane

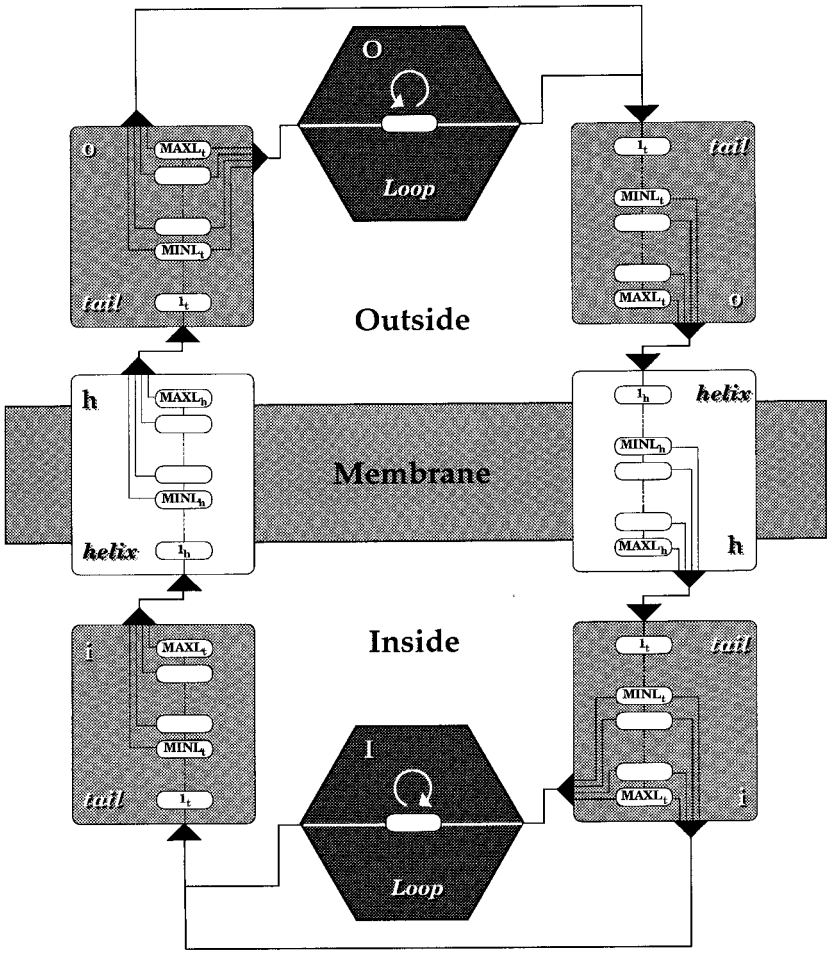

Figure 3. Architecture of HMM used for topology prediction. States with the same transition matrices are colored in the same way: white, helix states; light gray, tail states; dark gray, loop states. Rectangular areas FL type states; hexagonal ones, NFL type states. The observation-symbol probabilities used by states are marked in each state. The structure of substates in the case of the FL type is drawn within states. Lines and arrows show the possible transition between states or substates.

helix prediction accuracy is over 98% in each case (altogether, for the 698 transmembrane segments 709 were predicted, of which 694 were predicted correctly), all transmembrane segments were predicted correctly in 74 out of 83 proteins on the 83TMP set (89%), 45/47 (96%) on the 48TMP set and 38/44 (86%) on the prokTMP set. The number of proteins with correctly predicted orientation and transmembrane segments reached a high level as well: 72/83 (87%), 43/47 (91%) and 32/44 (73%) on the three data sets, overall 135 out of 158 proteins (85%). The published and predicted transmembrane segments and topologies can be found in the Appendix or see Tusnády (1998).

Prediction based on single sequence information was naturally less accurate. Nevertheless, comparing these results with previously published methods (see the next section) using single sequence information, the prediction power of this method is astonishing. From the three data sets, 714 helices were predicted, of which 689 were correct. This value is less by only five transmembrane helices than the value in the case of multiple sequences. The number of proteins with correctly predicted membrane spanning segments was 131, while the topology and the transmembrane segments were predicted correctly in 124 cases (78%). This is much higher than the accuracy of predictions based on hydrophobicity plot analysis, the

Table 1. Results on various data sets using HMM for topology prediction

Data set	No. of transmembrane helices					No. of correct proteins			
	N_{obs}	N_{prd}	N_{cor}	Q_P (%)	Q_2	N_{TOT}	N_{TM}	N_{TT}	Q_T (%)
83TMP	346	353	344	98.4	94.9	83	74	72	87
48TMP	194	197	194	99.2	94.6	47	45	43	91
prokTMP	262	264	259	98.5	89.8	44	38	32	73
Total	698	709	694	98.7	94.2	158	143	135	85

N_{obs}, N_{prd} and N_{cor} are the number of observed, predicted and correctly predicted transmembrane helices, respectively; $Q_P = 100 \cdot \sqrt{(N_{cor}/N_{obs}) \cdot (N_{cor}/N_{prd})} \cdot N_{TOT}$, N_{TM} and N_{TT} are the number of proteins in the data sets, the number of proteins for which all transmembrane segments were predicted correctly, and the number of proteins for which both the topology and the transmembrane segments were predicted correctly, respectively. Q_2 is the per residue accuracy.

accuracies of which are about 60% on these data sets.

The prediction accuracy on the prokTMP set is not as good as the other two. In this data set, however, there are a few proteins without experimentally well-defined topology. In the case of cytochrome *d* terminal oxidase subunits I and II (CYDA_ECOLI and CYDB_ECOLI, respectively), according to the original article (Georgiou *et al*., 1988), an independent experimental approach is required to determine the actual topology, and from other experiments (Dueweke & Gennis, 1990, 1991) it is only acceptable that the loop between residues 239 and 393 in subunit I is located in the periplasma. Since there is no evidence that this protein contains seven transmembrane segments, the prediction made using HMM does not contradict the results of the experiments. The experimental results obtained on subunit II of this protein do not distinguish between the former prediction based on hydrophobicity analysis and the recent prediction made by HMM. The cytochrome *o* terminal complex is the other component of the aerobic respiratory chain of *Escherichia coli*. This complex consists of five subunits, whose topologies were determined by Chepuri & Gennis (1990). According to their work there is no evidence that the last putative transmembrane segment of the subunit I (CYOB_ECOLI) really crosses the membrane. In case of the E subunit of this complex (CYOE_ECOLI), the results of experiments do not exclude the possibility that the polypeptide chain crosses the membrane twice between the fourth and fifth putative transmembrane segments in the original work, as the HMM predicted.

Comparison with other methods

To disclose the principles governing the structure formation of membrane proteins, it is interesting to compare the results of prediction methods based on different ideas. Three other prediction methods were taken into consideration. TOPPRED (von Heijne, 1992) uses the hydrophobic profile of transmembrane proteins and the observation that positively charged residues are more abundant in cytoplasmic regions (''positive-inside'' rule). MEMSAT (Jones *et al*., 1994) employs the amino acid log likelihood ratios in five structural classes of membrane proteins (inside loop, outside loop, inside helix end, helix middle and outside helix end) and uses a dynamic programming algorithm to maximize the sum of these log likelihoods over the sequences. In fact this method is one step of the HMM, i.e. finding the best state sequence for the amino acid sequence if the parameters and the model are given. The third method, PHDhtm_ref (Rost *et al*., 1996), introduced a refined neural network system to predict localization of transmembrane helices combined with the positive-inside rule.

The accuracies of these three prediction methods on the three data sets are listed in Table 2. Note that segment prediction accuracies (Q_p) are high for all of these methods (above 94%), thus to distinguish among them the number of proteins for which all the transmembrane segments and the topology are correctly predicted (N_{TT} or Q_T should be considered. TOPPRED, elaborated first, is the least accurate. The overprediction of this algorithm is remarkable, which may be the result of marking the apolar cores of the non-transmembrane

Table 2. Prediction accuracy of various algorithms on various data sets

		No. of transmembrane helices				No. of correct proteins			
Data set	Method	N_{obs}	N_{prd}	N_{cor}	Q_P (%)	N_{TOT}	N_{TM}	N_{TT}	Q_T (%)
83TMP	TOPPRED	346	381	336	92.5	83	54	54	**65**
	MEMSAT		351	336	96.4		69	65	**78**
	HMM_1		358	342	97.2		68	66	**80**
	PHDhtm_ref		351	342	98.1		75	73	**88**
	HMM_{multi}		353	344	98.4		74	72	**87**
48TMP	MEMSAT	194	174	165	89.8	47	26	23	**49**
	TOPPRED		200	193	98.0		40	25	**53**
	HMM_1		198	192	98.0		40	39	**83**
	HMM_{multi}		197	194	99.2		45	43	**91**
	PHDhtm_ref		192	192	99.5		45	42	**89**
prokTMP	PHDhtm_ref	262	259	254	97.5	44	31	28	**64**
	MEMSAT		255	250	96.7		33	29	**66**
	TOPPRED		264	255	97.0		32	30	**68**
	HMM_1		264	258	98.1		36	30	**68**
	HMM_{multi}		264	259	98.5		38	32	**73**
Total	TOPPRED	698	740	681	94.8	158	112	95	**60**
	MEMSAT		673	647	94.4		114	103	**65**
	HMM_1		714	689	97.6		131	124	**78**
	PHDhtm_ref		699	685	98.1		136	128	**81**
	HMM_{multi}		**709**	**694**	**98.7**		**143**	**135**	**85**

References for methods are as follows: TOPPRED (von Heijne, 1992), MEMSAT (Jones *et al*., 1994), PHDhtm_ref (Rost *et al*., 1996), HMM_1 hidden Markov model used on single sequence information in this article, HMM_{multi} hidden Markov model used on multiple sequence information in this article. The meanings of the columns are the same as in Table 1.

domains as transmembrane regions. The positive-inside rule improves the accuracy, but the inclusion of only the positive to the charged residues in the prediction is not as efficient as taking into account the distribution of all amino acids in the structural units of membrane proteins. This observation explains why MEMSAT works better than TOPPRED, since in this procedure the distribution of all amino acids of five structural classes are built into the prediction. The difference in accuracy between MEMSAT and our method might reflect the fact that the various membrane proteins have different amino acid frequencies in their structural units, so that forcing the same distribution for each protein may result in underprediction of transmembrane regions. If the amino acid composition of transmembrane segments differs from the general composition, then these segments will not be predicted using the model recognition approach, but will be accepted as transmembrane segments if changes in the composition are considered. To understand this, a spectacular example is shown. Multiple polar residues were gradually introduced into helix D of bacteriorhodopsin, from one glutamine to as many as five hydrophilic residues (four glutamine and one aspartic acid) (Chen & Gouaux, 1997). All of the mutants refold and show properties similar to wild-type protein, demonstrating that micelle-solubilized bacteriorhodopsin can tolerate multiple non-conservative substitution of amino acids. Application of the MEMSAT method to these mutants yielded correct results in the case of mutants containing one and two glutamine residues but failed for mutants containing more hydrophilic residues. On the contrary, the method presented here failed only for mutants containing five additional hydrophilic residues.

The lower accuracy of the MEMSAT and TOPPRED methods on the 48TMP set may be due to the homolog proteins in this set (e.g. GABA-receptor subunits or members of TM4 superfamily). One should also note that MEMSAT uses another definition of structural units. Though the distributions of apolar amino acids are different in helix middle and ends, as first shown by Sipos & von Heijne (1993), this observation provides similar information about localization and orientation of transmembrane helices. Choosing the parameters in HMM used in MEMSAT resulted in a prediction accuracy similar to theirs. In the total data set, using only a single sequence for prediction, setting the helix's minimum and maximum lengths to 9 and 17, respectively, and the tail's minimum and maximum length to 4, as used in MEMSAT, the method predicted 658 helices, of which 644 were correct (673 and 647 were the values predicted by MEMSAT, respectively; Table 2). The number of perfectly predicted proteins decreased from 124 to 100, which is in good agreement with the prediction accuracy reached by MEMSAT (103). This result suggests that amino acids locate near the membrane have special roles in determining the protein topology, so application of their distribution increases the efficiency of the prediction methods.

As it is known the prediction accuracy reaches a higher level when using more sequences, PHDhtm_ref reached the highest efficiency compared to the previous methods, since if includes sequence alignments. However, aligning the transmembrane segments can lead to wrong prediction, since the sequence identity is very low on these segments. For this reason, the way the HMM handles multiple observations can be a great advantage and may result in better prediction. Naturally, the neural network algorithm by which the markings of transmembrane segments was learned also increases the efficiency. However, we do not gain more knowledge about proteins, for the neural network algorithm is a black box. This algorithm cannot handle the length of the transmembrane segments properly and does not employ the charge bias between inside and outside loops, so the authors had to include them in the method as an inside filter.

Validation of the hypothesis

Here we suggest that the topology depends on the maximum divergence of the amino acid distributions of the various structural parts rather than on the absolute frequencies of amino acids in these parts. First, it is algebraic evidence that the maximum divergence (the sum of divergences between the amino acid distributions of these parts and the distribution of the whole proteins) can be obtained as the sum of the logarithm of the relative frequencies of residues in these parts along the given sequence (log likelihood), or without using a logarithm, the product of these frequencies (likelihood). When searching for the correct topology as the maximum of the likelihood function, the biological constraints have to be considered; for example, the length of a transmembrane helix cannot be arbitrary, or a helix after an inside loop can only be followed by an outside loop. However, even if the amino acid distributions in the five structural parts were known, the most likely topology could not be searched for by a direct searching method due to a combinatorical problem (Jones *et al.*, 1994). In addition, according to our hypothesis, amino acid distributions in these parts have also to be searched for, thus finding the most likely topology becomes even harder by ''brute-force searching''. Fortunately, this optimization problem can be solved by using HMM associated with the Baum-Welch algorithm. Thus HMM in this study is only a tool for searching for the topology corresponding to the maximum divergence, and therefore a high level of prediction accuracy, i.e. the observation that the most likely topology correlates with the natural topology of the proteins is a strong argument in favor of our hypothesis.

Starting the optimization from the amino acid distributions corresponding to the natural topology of proteins results in the same topology with only

a small alteration in the ends of transmembrane segments (data not shown). This observation shows that the likelihood functions have local optima at these distributions, which also supports our hypothesis. Using the pseudocount vector in the optimization process can be interpreted as a search for the topology of the query protein in a restricted space of the amino acid distributions. However, optimizations without pseudocount vector result in only a little lower accuracy (number of proteins in which all transmembrane segments are correctly predicted is 132, with correct topology is 123 (79%)), showing the high level of validity of the hypothesis.

Conclusion

The accuracy of the prediction method described here indicates that the topology is determined by the maximum divergences of the amino acid distribution of the different structural parts in the membrane proteins rather than by the absolute composition of these parts.

This work is a wide generalization of the work of Jones *et al*. (1994). Improvements proposed by them are included in HMM automatically; for example, usage of multiple sequence information. The other advantage of HMM is that there is no need to make alignments before prediction. Since the actual topology is determined by the principles mentioned above, the effect of the parameters originating from the experimental results is much weaker. Thus the experimental errors do not affect the prediction accuracy. Moreover, the proposed method can work without any external parameters, with very high success.

It is worth mentioning that the various segment distributions of the membrane proteins can be stimulated in HMM by its special architecture. This architecture ensures the distinction between the short and long loops connecting helices.

Naturally, there are several weak points in this model originating from the methodology of HMM. One of them is that, using multiple sequences, the same predicted topology for each sequence is not guaranteed. The next point is related to the multiple optima problem in the optimization process. Since the Baum-Welch algorithm cannot find the global optimum of the likelihood function, the correct way to handle this problem may be by an exhaustive search for the optimum. Because of the huge computational demand for searching, each iteration was started from the same point.

Materials and Methods

The hidden Markov model

To apply the hidden Markov model, the model architecture has first to be defined; namely, the number of states, the possible transitions between states and the observation symbols of each state. The model described here consists of five states: loops (inside and outside, I and O, respectively), tails (inside and outside, i and o, respectively) and helices (h). The model is presented in Figure 2; our notation is given in Table 3. For defining the possible transitions between these states: first, two types were defined. In the first one, called non-fixed length (NFL) type state, there are only two possible transitions: from current state to current state with $\mathcal{J}_{Curr,Curr}$ probability and from current state to the next one with $\mathcal{J}_{Curr,Next}$ probability. By definition: $\mathcal{J}_{Curr,Curr} + \mathcal{J}_{Curr,Next} = 1$ (for definition of the term ''Next'' state, see below). In the second case, called fixed length (FL) type state, the minimum and maximum lengths of the state are fixed (MINL and MAXL, respectively; they are different for various FL type states). This can be ensured by introducing maximum length number substates. The values of the transition probabilities in this case are as follows: for the first MINL substates transition probabilities are unity to the next substate, and zero to any other substates and states ($\mathcal{J}_{Curr(j),Curr(j+1)}: = 1, j = 1 \ldots \mathrm{MINL}_{Curr} - 1$). Between MINL and MAXL transition probabilities are $\mathcal{J}_{Curr(j),Curr(j+1)}$ to the next substate, $\mathcal{J}_{Curr(j),Next}$ ($j = \mathrm{MINL}_{Curr} \ldots \mathrm{MAXL}_{Curr} - 1$) to the first element of the nest state and zero to any other substates or to other states. Naturally, from the last substrate of the current state transition is only possible to the next state. When a state is followed by two other states (see below) the transition probabilities are split into two parts $\mathcal{J}_{Curr(j),Next}$ and $\mathcal{J}_{Curr(j),Other}$ (see Figure 3). Tail and helix states are defined as FL type states, while loop states are NFL type states.

Table 3. Notation

A	Amino acids ($A = a_1 \ldots a_{20}$)
B	States ($B = b_1 \ldots b_5$)
O,o,h,i,I	States (outside loop, outside tail, membrane helix, inside tail and inside loop)
N	Length of sequence
S	Sequence of amino acids ($S = s_1 \ldots s_N$)
Q	Sequence of states ($Q = q_1 \ldots q_N$)
α	Pseudocount vector ($\alpha = \alpha_{1,1} \ldots \alpha_{5,20}$)
$\mathcal{I}$	Initial state distribution $\mathcal{I}_j = P(q_1 = b_j), j = 1 \ldots 5$
$\mathcal{P}$	Emission probability distribution $\mathcal{P}_{ij} = P(a_i \mid b_j)$, $i = 1 \ldots 20$, $j = 1 \ldots 5$
$\mathcal{J}$	Transition probability distribution $\mathcal{J}_{ij} = P(q_k = b_j \mid q_{k-1} = b_i)$ $i = 1 \ldots 5$, $j = 1 \ldots 5$, $k = 2 \ldots N$

The sequence of states and the corresponding transition matrix follows the natural structure of transmembrane proteins, i.e. inside loop is followed by helix, helix is followed by outside loop and outside loop is followed again by helix. More exactly, a tail, which comes after a helix, can be followed by another tail or by a loop, thus a linker region between two helices can be formed by two tail states or by a tail-loop-tail state sequence.

The observation-symbol probabilities of substates were the same. So were the two kind of tails, which are on the same side of the membrane before and after helices. In this way the observation-symbol probability matrix ($\mathcal{P}$) contains five rows corresponding to five structural parts of the membrane proteins; each contains 20 observation-symbol probabilities for the 20 kinds of amino acids. For HMM the initial state probabilities ($\mathcal{I}$) have to be defined as well. They are zero for helix state and tail state, which are located after a helix. For the

other states they can be any value. This model corresponds to the natural structure of transmembrane proteins, containing a shorter or longer sequence before the first membrane spanning segment.

A prediction (i.e. sequence of states) for a given amino acid sequence can be generated by a ''random walk'' through the model. The first element of the state sequence (q_1) is chosen randomly according to the initial probability matrix ($\mathcal{I}$). The second one (q_2) is selected randomly according to the transition probabilities $\mathcal{J}(x|q_1)$, where x indicates any possible next state. The ith element of the state sequence is generated from transition probabilities $\mathcal{J}(x|q_i)$. The probability of this prediction (q_1, $q_2 \ldots q_N$) for a given amino acid sequence (s_1,$s_2 \ldots s_N$), if the model ($\mathcal{I}$, $\mathcal{P}$, $\mathcal{J}$) is given is:

$$P(q_1 \cdots q_N, s_1 \cdots s_N|model) = \mathcal{I}(q_1) \cdot \mathcal{P}(s_1|q_1) \cdot \prod_{i=2}^{N} \mathcal{J}(q_i|q_{i-1}) \cdot \mathcal{P}(s_i|q_i) \quad (1)$$

The probability of an amino acid sequence associated with a given model can be calculated by summing these probabilities over all possible state sequences:

$$P(s_1 \cdots s_N|model) = \sum_{All\ q_1 \cdots q_N} P(q_1 \cdots q_N, s_1 \cdots s_N|model) \quad (2)$$

Given a set of homolog proteins (S(1), S(2)…S(M) derived from the same model, the probability of the model is simply the product of the probabilities calculated for each sequence:

$$P(sequences|model) = \prod_{j=1}^{M} P(S(j)|model) \quad (3)$$

where M is the number of sequences, and each term $P(S(j)|\text{model})$ is calculated by substituting $s_1 \ldots s_N = S(j)$ in equation (2). In this way a probability distribution on the space of sequences is defined. The goal is to find a model (i.e. values of the observation-symbol and transition probabilities) that accurately describes the topology of a given protein (or proteins) by assigning a maximal probability to the sequence(s).

The original Baum-Welch (or forward-backward) algorithm was used to find this best model. The detailed description of the HMM and the Baum-Welch algorithm can be found in Rabiner's excellent tutorial (Rabiner, 1989). To ensure the correct sequence of states and avoid the incorrect ones (i → h → i or o → h → o, we used a special matrix in the forward-backward algorithm, where the two types of transmembrane helices (i → h → o and o → h → i) were distinguished but the same transition and observation symbol probabilities were used for them.

Many authors pointed out the weakness of the Baum-Welch algorithm, i.e. it finds only a local optimum, not a global one. There are two suggested solutions to this problem in the literature, the ''noise injection'' heuristic procedure used by Krogh *et al*. (1994b) and a simulated annealing variant proposed by Eddy (1995). We have found the latter one to be unsatisfactory, due to the changing of the optimum place during the temperature change (data not shown). Obviously, using more sequences and doing many optimizations from various probability distributions, proposed by Krogh *et al*. (1994a), can help to solve this problem. We found that introducing the Dirichlet mixture to the HMM (Brown *et al*., 1995; Sjölander *et al*., 1996), or its simpler variant, the pseudocount method, the number of the local optimum places decreased drastically. We used the pseudocount method, where the prior distribution (α) was given by the relative frequencies of the amino acids in the reference data set (see below). In the likelihood function the count vector has to be considered in the following way:

$$Prob(sequences,\ model|\alpha) = Prob(sequences|model) \cdot \prod_{j=1}^{5} \prod_{i=1}^{20} \mathcal{P}(a_i|b_j)^{\alpha_{ij}} \quad (4)$$

where the probability of sequences for a given model *Prob* (*sequences*|*model*) was calculated as in equation (3).

Rabiner (1989) emphasized the importance of the initial estimates of HMM parameters. If a good estimation is given as a starting point to the Baum-Welch algorithm, the multiple optima problem can be avoided. For this reason, besides using the pseudocount method each iteration was started from the same point located by the count arrays.

Data sets, measure the prediction accuracy

Three data sets, collected earlier for transmembrane prediction methods, were used to measure the prediction accuracy. The first data set was originally collected by Jones *et al*. (1994), and was also used by Rost *et al*. (1996) (83TMP set). The second one is an extension of it by Rost *et al*. (1996) (48TMP set). The third data set contains prokariotic transmembrane proteins collected by Cserző *et al*. (1997) (prokTMP set). These data sets contain transmembrane proteins whose topologies are established by two kinds of approaches. Indirect experiments, providing information about certain parts of a protein (for example, an amino acid in a given position is inside or outside), were combined with hydrophobicity plot analyses resulting in the most probable topology. Thus the uncertainty of the termini of transmembrane segments has to be kept in mind in measuring the accuracy of the method (see below) and in the interpretation of the results. Apparently, the following entries were missing from the Swissprot database release 34.0 (Bairoch & Boeckmann, 1991): EGFR_DROME, GP1B_HUMAN, PT2M_ECOLI and IGGB_STRSP in the 83TMP set. These entries were replaced by the corresponding files in the current Swissprot release, i.e. TOP_DROME, GPBB_HUMAN, PTMA_ECOLI and IG1B_STRSP, respectively. In the 48TMP set there were also some missing files, AD1_RAT and COX1_PARDE. AD1_RAT was eliminated because it is the same as CD63_RAT, which originally belonged to the 48TMP set. COX1_PARDE was replaced by CX1B_PARDE.

We have found some data in the data sets studied which were in contradiction with the original article or with other experimental results. According to van Beilen *et al*. (1992), the transmembrane segments of ALKB_PSEOL are as follows: 22-40, 41-69, 88-110, 114-137, 227-247 and 250-270. In COX2_PARDE the lengths of the two transmembrane segments were too long, so we shortened them according to the results of Iwata *et al*. (1995) to 66-88 and 108-128 instead of 56-88 and 103-134, respectively. The annotations of transmembrane segments were missing for UHPT_ECOLI and were added according to the results of Yan & Maloney (1993). Where the annotation of the topology, i.e. the localization of the first loop, was missing it was taken from Jones *et al*.

(1994) or Rost *et al*. (1996). Finally, annotation errors mentioned by Cserző *et al*. (1997) were corrected as well.

After these corrections 83TMP, 48TMP and prokTMP contain 83, 47 and 44 proteins, with 346, 194 and 262 transmembrane segments, respectively. Because of the overlapping proteins, the three data sets contain, altogether, 158 proteins and 698 transmembrane helices.

To measure the prediction accuracy we followed the method described by Cserző *et al*. (1997), with the following slight modification. The overlapping predicted and observed transmembrane segments were counted (N_{cor}). The total numbers of predicted (N_{prd}) and observed (N_{obs}) segments were also counted. If N_{cor} was higher than N_{prd} (which can happen if the observed helix overlaps two predicted helices), then it was reduced to N_{prd}. The efficiency of the transmembrane helix prediction was measured in terms of the following ratios: $M = N_{cor}/N_{obs}$ and $C = N_{cor}/N_{prd}$. The overall prediction power can be measured as the geometric mean of these ratios ($Q_P = 100 \cdot \sqrt{M \cdot C}$). As this value is very high for many prediction methods (above 90%), two other values were used to measure the prediction accuracy proposed by Rost *et al*., (1996): the number of proteins for which all the transmembrane segments were predicted correctly (N_{TM}), and the number of proteins for which both the transmembrane segments and the topology were predicted correctly (N_{TT}).

For using multiple observation sequences in the prediction method, homolog sequences of the query protein were searched by the BLAST automatic server (Altschul *et al*., 1990). Sequences above 25% identity with the query protein were applied. Because of the limiting factor of the computer hardware, a maximum of 50 related proteins was used in the prediction.

The sequence processing before prediction was the same as found by Jones *et al* (1994), i.e. if the localization of the signal peptide was given in the databank then it was removed. If the precursor protein was marked in the database then only the precursor sequence was applied in the prediction.

Parameters

The control parameters of the algorithm described here are $MINL_s$ and $MAXL_s$, the minimum and maximum lengths of the FL type state s, respectively. They are 1 and 15 for tails (inside and outside), 17 and 25 for helix states.

As described above, the pseudocount method was used to eliminate the local optima problems. Proteins containing one transmembrane segment and sequences longer than 500 residues were eliminated from the 83TMP data set. Proteins that have no well-confirmed topology, by experimental results, were omitted from the 83TMP set as well. After this filtration 63 proteins remained (marked by asterisks in the Appendix, and see Tusnády, 1998), which were used to create the initial estimate of parameters and also the pseudocount array. From this set, proteins which have a higher sequence identity than 25% to proteins under prediction were also omitted (jack-knife method). The amino acid frequencies after elimination from the five states were counted. The initial observation-symbol probabilities were their normalized arrays for each state. The pseudocount array (α) was calculated as follows: the amino acid frequencies from the selected proteins were counted in each state and were normalized to a given size ($T = |\bar{\alpha}|$, $\alpha_{i,j} = T \cdot \alpha_{ij}/\bar{\beta}$, where $\beta_{i,j}$ is the frequency of the jth amino acid in the ith state and $\bar{\beta} = \Sigma_{i=1}^{5}\Sigma_{j=1}^{20}\beta_{ij}$). The highest prediction accuracy was reached at $T = 10{,}000$.

The initial transition probabilities of the FL type states were also derived for the 63 selected proteins. Tail regions in this case were defined as follows: let l be the length of a linker region between two helices. If $l \geqslant 2 \cdot MAXL_t$ (i.e. 30 residues), then two $MAXL_t$ length (15 residues) tail regions were marked else two symmetrical $l/2$ length ones. The frequencies of the various lengths of loops and membrane helices were counted. The initial transition probabilities were set in a manner by which they could generate these distributions of length: let λ_{ij} be the frequency of the segments of j length in the ith state. The let $\Lambda_{ij} = \lambda_{ij}/\Sigma_{k=j}^{MAXL_i}\lambda_{ik}$ and $\tau_{ij0} = 1 - \Lambda_{ij}$ (the initial probability of the elongation of the j length segment in the ith state), $\tau_{ij1} = \Lambda_{ij}$ (the initial probability of the termination of the j length segment in the ith state), which is the transition to the next state if the ith state is followed by only one state (loops, tails before membrane helix and membrane helices). If the ith state may be followed by two states (tails coming after membrane helices, see Figure 2 and Figure 3), Λ_{ij} is split into two parts according to the relative frequencies of the two states after the j length segments in the ith state (ω_{ij} and $1 - \omega_{ij}$), so $\tau_{ij1} = \omega_{ij} \cdot \Lambda_{ij}$ and $\tau_{ij2} = (1 - \omega_{ij}) \cdot \Lambda_{ij}$.

The initial parameter set of the model was determined using the 83TMP data set and for each protein the sequences which have higher sequence identity than 25% to proteins under investigation were omitted (jack-knife method).

Programs

The hidden Markov model described here has been implemented in ANSI C language on a Unix workstation (Silicon Graphics, Indigo2). The prediction method is available *via* an automatic server on the World-Wide Web site http://www.enzim.hu/hmmtop.

Methods for comparison with our method were used *via* an Internet site or the source codes were purchased. TOPPRED predictions (von Heijne, 1992) for the three data sets were generated using its automatic prediction server (http://www.biokemi.su.se/~server/toppred2) with default parameters (upper cutoff, 1.0; lower cutoff, 0.6; window size; top, 11; bottom, 21). The source code of MEMSAT program (Jones *et al*., 1994) was obtained from the authors, and was implemented on our workstation. The method developed by Rost *et al*. (1996) was used *via* their automatic server (http://www.embl-heidelberg.de/predictprotein). In case of dubious prediction results we consulted Dr Rost. The results obtained with these programs on data sets were different from the original ones, due to the annotation errors mentioned above.

Acknowledgements

We thank Gábor Tusnády for very useful discussion and comments, and Zsuzsanna Dosztányi and Gábor Szirtes for their critical comments on the manuscript. We thank Burhard Rost for helping with discussion of results from the PHDhtm_ref method. This work was supported by research grants OTKA T017652, F019008 and F022051.

References

Altschul, F. S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Asai, K., Hayamizu, S. & Handa, K. (1993). Prediction of proteins secondary structure by the hidden Markov model. *Comp. Appl. Biosci.* **9**, 141–146.

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

Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA,* **91**, 1059–1063.

Bergelson, L. & Barsukov, L. I. (1977). Topological asymmetry of phospholipids in membranes. *Science,* **197**, 224–230.

Borodovsky, M., McIninch, J. D., Koonin, E. V., Rudd, K. E., Médigue, C. & Danchin, A. (1995). Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucl. Acids Res.* **23**, 3554–3562.

Boyd, D., Manoil, C. & Beckwith, J. (1987). Determinants of membrane proteins topology. *Proc. Natl Acad. Sci. USA,* **84**, 8525–8529.

Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K. & Haussler, D. (1995). Using Dirichlet mixture priors to derive hidden Markov models for protein families. In *Proceeding of First International Conference on Intelligent Systems for Molecular Biology* (Rawlings, C., ed.), pp. 47–55, AAAI/MIT Press, Menlo Park, CA, USA.

Casadio, R., Fariselli, P., Taroni, C. & Compiani, M. (1996). A predictor of transmembrane alpha-helix domains of proteins based on neural networks. *Eur. Biophys. J.* **24**, 165–178.

Chen, G.-Q. & Gouaux, E. (1997). Reduction of membrane proteins hydrophobicity by site-directed mutagenesis: introduction of multiple polar residues in helix d of bacteriorhodopsin. *Protein Eng.* **10**, 1061–1066.

Chepuri, V. & Gennis, R. B. (1990). The use of gene fusions to determine the topology of all of the subunits of the cytochrome o terminal oxidase complex of Escherichia coli. *J. Biol. Chem.* **265**, 12978–12986.

Chou, K. C. (1995). A novel approach to predicting protein structural classes in a (20–1)-d amino acid composition space. *Proteins: Struct. Funct. Genet.* **21**, 319–344.

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

Cserző, M., Bernassau, J., Simon, I. & Maigret, B. (1994). New alignment strategy for transmembrane proteins. *J. Mol Biol.* **243**, 388–396.

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

Dueweke, T. J. & Gennis, R. B. (1990). Epitopes of monoclonal antibodies which inhibit ubiquinol oxidase activity of Escherichia coli cytochrome d complex localize functional domain. *J. Biol. Chem.* **265**, 4273–4277.

Dueweke, T. J. & Gennis, R. B. (1991). Proteolysis of the cytochrome d complex with trypsin localizes a quinol oxidase domain. *Biochemistry,* **30**, 3401–3406.

Eddy, S. R. (1995). Multiple alignment using hidden Markov models. In *Proceedings of Third International Conference on Intelligent Systems for Molecular Biology* (Rawlings, C., ed.), pp. 114–120, AAAI Press, Menlo Park, CA, USA.

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

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

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

Francesco, V. D., Garnier, J. & Munson, P. J. (1997). Proteins topology recognition from secondary structure sequences: application of the hidden Markov models to the alpha class proteins. *J. Mol. Biol.* **267**, 446–463.

Georgiou, C. D., Dueweke, T. J. & Gennis, R. B. (1988). β-galactoside gene fusions as probes for the cytoplasmic regions of subunits I and II of the membrane-bound cytochrome d terminal oxidase from Escherichia coli. *J. Biol. Chem.* **263**, 13130–13137.

Gokhale, D. V. & Kullback, S. (1978). *The Information in Contingency Tables*, Marcel Dekker Inc., New York.

Gouaux, E. (1998). It's not just a phase: crystallization and x-ray structure determination of bacteriorhodopsin in lipidic cubic phases. *Structure,* **15**, 5–10.

Gromiha, M. M. & Ponnuswamy, P. K. (1995). Prediction of protein secondary structures from their hydrophobic characteristics. *Int. J. Peptide Proteins Res.* **45**, 225–240.

Hartmann, E., Rapoport, T. A. & Lodish, H. F. (1989). Prediction the orientation of eukaryotic membrane proteins. *Proc. Natl Acad. Sci. USA,* **86**, 5786–5790.

Hughey, R. & Krogh, A. (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comp. Appl. Biosci,* **12**, 95–107.

Iwata, S., Ostermeier, C., Ludwig, B. & Michel, H. (1995). Structure at 2.8 Å resolution of cytochrome c oxidase from *Paracoccus denitrificans*. *Nature,* **376**, 660–669.

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

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994a). Hidden Markov models in computational biology. *J. Mol. Biol.* **235**, 1501–1531.

Krogh, A., Mian, I. S. & Haussler, D. (1994b). A hidden Markov model that finds genes in *E. coli* DNA. *Nucl. Acids Res.* **22**, 4768–4778.

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York.

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

Lattman, E. E. (1994). Protein crystallography for all. *Proteins: Struct. Funct. Genet.* **18**, 103–106.

Lawrence, C. E. & Reilly, A. A. (1990). An Expectation Maximization (EM) algorithm for the identification and characterization of common sites in unaligned

biopolymer sequences. *Proteins: Struct. Funct. Genet.* **7**, 41–51.
Lohmann, R., Schneider, G., Behrens, D. & Wrede, P. (1994). A neural network model for the prediction of membrane-spanning amino acid sequences. *Protein Sci.* **3**, 1597–1601.
Nakashima, H. & Nishikawa, K. (1994). Discrimination of intercellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* **238**, 54–61.
Persson, B. & Argos, P. (1994). Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.* **237**, 182–192.
Persson, B. & Argos, P. (1996). Topology prediction of membrane proteins. *Protein Sci.* **5**, 363–371.
Ponnuswamy, P. K. & Gromiha, M. M. (1993). Prediction of transmembrane helices from hydrophobic characteristics of protein. *Int. Peptide Protein Res.* **42**, 326–341.
Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* **77**, 257–286.
Rost, B., Casadio, R., Fariselli, P. & Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**, 521–533.
Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5**, 1704–1718.
Rothman, J. & Lenard, J. (1977). Membrane asymmetry. *Science,* **195**, 743–753.
Sipos, L. & von Heijne, G. (1993). Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* **213**, 1333–1340.
Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S. & Haussler, D. (1996). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comp. Appl. Biosci.* **12**, 327–345.
Stultz, C. M., White, J. V. & Smith, T. F. (1993). Structural analysis based on state-space modeling. *Protein Sci,* **2**, 305–314.
Tusnády, G. E. (1998). Appendix to ''Principles governing amino acid compositions of integral membrain proteins: application to topology prediction''. WWW document, http://www.enzim.hu/hmmtop/appendix.html.
Tusnády, G. E., Tusnády, G. & Simon, I. (1995). Independence divergence-generated binary trees of amino acids. *Protein Eng.* **8**, 417–423.
van Beilen, J., Penninga, D. & Witholt, B. (1992). Topology of the membrane-bound alkane hydroxylase of pseudomonas oleovorans. *J. Biol. Chem.* **267**, 9194–9201.
van Klompenburg, W., Nilsson, I., von Heijne, G. & de Kruijff, B. (1997). Anionic phospholipids are determinants of membrane protein topology. *EMBO J.* **16**, 4261–4266.
von Heijne, G. (1992). Membrane protein structure prediction. *J. Mol. Biol.* **225**, 487–494.
von Heijne, G. (1994). Membrane protein assembly: rules of the game. *BioEssays,* **17**, 25–30.
Weiss, M. & Schulz, G. (1992). Structure of porin refined at 1.8 Å resolution. *J. Mol. Biol.* **227**, 493–509.
White, J. V., Stultz, C. M. & Smith, T. F. (1993). Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Math. Biosci,* **119**, 35–75.
Yan, R. & Maloney, P. (1993). Identification of a residue in the translocation pathway of a membrane carrier. *Cell,* **75**, 37–44.

Appendix

Predicted and observed topologies and transmembrane segments

Swissprot ID	Topology and TM Helices Predicted		Observed	
4f2_human	IN		IN	
	83	104	82	104
5h1a_human*	OUT		OUT	
	37	61	37	62
	74	95	74	98
	108	132	110	132
	153	174	153	178
	195	214	192	217
	347	371	346	367
	384	403	379	403
5h2a_crigr*	OUT		OUT	
	77	101	76	99
	112	136	111	132
	147	171	148	171
	192	213	192	215
	234	258	234	254
	324	348	325	346
	359	383	363	384
5ht3_mouse*	OUT		OUT	
	224	247	223	248
	259	281	255	273
	285	307	283	301
	438	460	442	461
a15_human	IN		IN	
	13	34	12	35
	55	76	52	70
	85	107	82	107
	210	232	209	229
a1aa_human*	OUT		OUT	
	99	120	96	121
	135	156	134	159
	171	192	170	192
	213	234	214	238
	255	274	252	275
	350	371	349	373
	386	405	381	405
a2aa_human*	OUT		OUT	
	34	55	34	59
	71	92	71	96
	108	129	107	129
	152	172	150	173
	195	214	193	217
	373	394	375	399
	410	429	407	430
a4_human	OUT		OUT	
	685	706	683	706
aa1r_canfa*	OUT		OUT	
	10	34	11	33
	45	69	47	69
	79	103	81	102
	126	146	124	146
	180	200	177	201
	235	259	236	259
	269	289	268	292
aa2a_canfa*	OUT		OUT	
	10	34	8	30
	43	67	44	66
	77	101	78	100
	123	143	121	143
	176	196	174	198
	235	259	235	258
	269	289	267	290
adt_ricpr*	IN		IN	
	28	48	34	54
	61	81	68	88
	94	112	92	113
	147	170	148	168
	183	206	185	205
	219	237	219	239

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
	274	297	280	300
	317	337	321	341
	350	370	349	369
	383	401	380	400
	442	461	439	459
	467	485	466	486
alkb_pseol	IN		IN	
	20	39	21	39
	46	66	42	67
	89	109	88	110
	119	137	114	137
	232	256	227	247
	329	348	250	270
atpl_ecoli	OUT		OUT	
	8	32	11	31
	53	77	53	73
bac1_hals1*	OUT		OUT	
	16	40	17	38
	48	72	50	75
	81	105	90	170
	114	133	114	133
	141	162	144	164
	173	197	180	199
	205	229	207	230
bach_halss*	OUT		OUT	
	33	56	33	56
	66	89	68	91
	115	134	109	127
	139	158	138	161
	163	186	165	189
	207	226	198	221
	231	254	233	256
bacr_halha*	OUT		OUT	
	10	34	10	29
	42	66	44	63
	77	101	82	101
	108	127	108	127
	135	156	135	154
	177	198	178	197
	206	224	204	223
c561_bovin*	IN		IN	
	35	56	38	60
	76	94	75	97
	107	128	107	129
	148	169	145	167
	183	201	185	207
	221	242	219	241
cb12_pea	IN		IN	
	67	88	62	81
	114	133	114	134
	182	201	182	199
cd37_human	IN		IN	
	14	38	13	36
	56	80	60	77
	85	109	86	111
	242	266	242	266
cd53_human	IN		IN	
	12	36	11	36
	51	75	55	72
	81	105	81	106
	182	206	182	206
cd63_human	IN		IN	
	11	34	11	34
	51	74	51	69
	81	105	81	106
	203	227	203	223
cd63_rat	IN		IN	
	11	34	11	34
	51	74	51	69
	81	105	81	106
	202	226	203	223
cd81_human	IN		IN	
	12	35	12	35
	60	83	58	78
	89	113	90	115
	202	226	202	226
cd82_human	IN		IN	
	12	36	11	34
	54	78	54	72
	83	107	84	109
	228	252	229	250
cd9_cerae	IN		IN	
	12	34	11	34
	59	81	55	75
	87	111	87	110
	194	218	195	220
cd9_felca	IN		IN	
	12	34	11	34
	56	78	53	73
	84	108	85	108
	194	218	193	218
cd9_human	IN		IN	
	12	34	11	34
	59	81	55	75
	87	111	87	110
	194	218	195	220
cek2_chick	OUT		OUT	
	350	370	346	370
	516	534		
co02_human	IN		IN	
	11	33	10	33
	54	78	58	75
	84	108	84	109
	206	230	206	230
cox2_parde	OUT		OUT	
	38	58	37	59
	79	103	79	99
cox3_parde	IN		IN	
	14	33	15	35
	51	70	48	73
	88	107	79	104
	139	158	139	164
	171	190	168	193
	208	232	203	228
	250	271	244	269
cx1b_parde	IN		IN	
	31	50	29	54
	95	119	84	109
	133	154	130	151
	181	205	178	203
	219	243	218	243
	270	294	263	288
	302	326	304	322
	340	364	334	359
	372	394	370	395
	408	432	404	429
	443	467	441	466
	488	512	483	508
cxb1_human*	IN		IN	
	16	40	20	40
	71	95	76	96
	133	157	143	163
	188	212	189	209
cxb1_rat*	IN		IN	
	16	40	20	40
	71	95	76	96
	133	157	143	163
	188	212	189	209
cxb1_xenla*	IN		IN	
	16	40	20	40
	71	95	76	96
	133	157	143	163
	189	213	189	209
cyda_ecoli	OUT		IN	
	12	36	23	42
	54	73	95	114

continued overleaf

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
	91	115	130	149
	127	151	188	207
	182	206	220	239
	219	237	393	412
	388	412	471	490
	425	444		
	471	494		
cydb_ecoli	OUT		IN	
	9	28	9	28
	50	74		
	80	99	80	99
	118	142	123	142
	164	188	165	184
	207	231	206	225
	264	283	263	282
	289	313	293	312
	335	359	337	356
cyoa_ecoli*	OUT		OUT	
	16	40	27	45
	66	84	69	87
cyob_ecoli*	OUT		OUT	
	15	39	17	35
	57	79	58	76
	105	129	102	121
	139	163	144	162
	189	213	195	213
	231	255	232	250
	273	297	277	296
	312	336	320	339
	346	370	348	366
	380	404	382	401
	414	438	410	429
	456	480	457	476
	490	514	494	513
	602	626	588	607
			613	633
cyoc_ecoli*	IN		IN	
	26	50	32	50
	68	88	67	85
	98	118	102	120
	136	160	143	161
	178	202	185	203
cyod_ecoli*	OUT		IN	
	15	39	18	36
	45	66	46	64
	78	102	81	99
cyoe_ecoli	IN		IN	
	21	30	10	28
	38	56	38	56
	84	102	79	97
	110	128	108	126
	132	150		
	162	180		
	208	226	198	216
	234	252	229	247
	164	283	269	287
dhg_ecoli	IN		IN	
	11	31	11	36
	36	57	41	58
	63	81	63	81
	87	108	96	113
	113	131	119	141
dmsc_ecoli	OUT		OUT	
	10	31	10	32
	44	64	44	66
	79	103	88	107
	116	136	113	134
	149	173	153	176
	178	201	183	203
	216	240	223	243
	253	277	255	280
dsbb_ecoli	IN		IN	
	13	32	15	32
	42	61	50	67
	71	89	72	89
	145	163	145	162
edg1_human*	OUT		OUT	
	47	71	47	71
	80	104	79	104
	122	140	122	140
	160	184	160	185
	203	221	202	222
	252	276	256	277
	295	313	294	314
egfr_chick	OUT		OUT	
	625	647	625	642
egfr_human	OUT		OUT	
	622	644	622	644
envz_ecoli	IN		IN	
	16	40	16	35
	160	179	162	182
exbb_ecoli	OUT		OUT	
	21	42	25	42
	133	157	132	150
	174	195	178	195
exbd_ecoli	IN		IN	
	15	36	26	43
fce2_human*	IN		IN	
	24	44	22	47
ftsh_ecoli	IN		IN	
	5	23	5	24
	102	120	96	120
ftsl_ecoli	OUT		IN	
	38	57	38	57
fucp_ecoli	IN		IN	
	23	41	24	44
	65	84	65	85
	93	112	92	112
	121	145	115	134
	161	179	161	181
	211	229	214	234
	262	282	262	282
	292	311	291	311
	327	346	327	347
	354	373	350	369
	382	401	385	405
	411	429	408	428
gaa1_chick	OUT		OUT	
	224	245	225	246
	256	277	252	273
	288	311	286	307
	397	415	394	414
gaa1_human	OUT		OUT	
	224	245	225	246
	256	277	252	273
	288	311	286	307
	398	416	395	416
gaa2_human	OUT		OUT	
	223	244	224	245
	255	276	251	272
	287	310	285	306
	395	413	392	417
gaa3_human	OUT		OUT	
	250	271	249	270
	276	297	276	297
	313	331	310	331
	433	451	430	451
gaa4_bovin	OUT		OUT	
	224	245	224	245
	255	276	250	270
	287	306	284	306
	491	509	491	510
gaa5_human	OUT		OUT	
	227	248	229	250
	259	280	255	276
	291	314	288	310

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
	400	418	397	418
gaa6_mouse	OUT		OUT	
	212	233	214	235
	244	265	240	261
	276	299	272	295
	394	412	391	412
gab1_human	OUT		OUT	
	219	240	221	242
	251	272	246	268
	283	306	280	302
	430	448	427	448
gab2_human	OUT		OUT	
	219	240	221	242
	251	272	246	268
	283	306	280	302
	431	449	428	449
gab3_human	OUT		OUT	
	219	240	221	242
	251	272	246	268
	283	306	280	302
	429	447	426	447
gab4_chick	OUT		OUT	
	218	239	220	241
	250	271	246	267
	282	305	279	301
	444	462	441	462
gab_lymst	OUT		OUT	
	227	248	229	250
	259	280	255	276
	291	314	288	310
	456	474	453	476
gac1_rat	OUT		OUT	
	236	257	238	259
	268	289	264	285
	300	323	297	319
	406	429	410	430
gac3_mouse	OUT		OUT	
	237	258	238	260
	269	290	264	286
	301	324	298	320
	426	449	427	450
gca4_chick	OUT		OUT	
	235	256	236	258
	267	288	262	284
	299	322	296	318
	412	435	413	436
gad_mouse	OUT		OUT	
	232	253	233	255
	264	285	259	281
	296	314	293	315
	414	432	411	433
gar1_human	OUT		OUT	
	261	282	261	284
	293	314	288	310
	325	349	322	344
	439	457	437	458
gar2_human	OUT		OUT	
	242	263	242	265
	274	295	269	291
	306	330	303	325
	426	444	424	445
glp_pig*	OUT		OUT	
	63	85	63	85
glpa_human*	OUT		OUT	
	73	95	73	95
glpc_human*	OUT		OUT	
	55	79	58	81
glpt_ecoli	IN		IN	
	31	55	28	45
	65	86	65	87
	94	112	98	115
	120	139	120	138
	155	179	167	184

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
	188	207	187	208
	254	278	254	273
	294	312	293	310
	322	341	322	341
	351	375	351	373
	384	408	382	406
	417	436	415	437
glr1_rat	OUT		OUT	
	470	488	521	540
	519	539	567	585
	569	587	596	614
	596	620	788	808
	788	812		
gmcr_human*	OUT		OUT	
	305	324	299	324
gpbb_human*	OUT		OUT	
	123	147	122	146
gpt_crilo*	OUT		OUT	
	11	29	7	32
	59	83	58	79
	95	114	95	114
	126	145	126	145
	157	181	165	184
	189	208	195	212
	222	241	222	240
	249	268	253	270
	276	294	275	294
	327	345		
	379	397	379	397
gra1_human	OUT		OUT	
	221	242	220	245
	253	274	253	270
	285	308	285	308
	392	410	393	410
gra2_human	OUT		OUT	
	228	249	227	252
	260	281	260	277
	292	315	292	315
	393	411	397	414
gra3_rat	OUT		OUT	
	221	242	220	245
	253	274	253	270
	285	308	282	307
	397	415	401	418
grb_rat	OUT		OUT	
	245	266	244	268
	277	298	277	294
	309	332	309	332
	453	472	456	473
hema_cdvo	IN		IN	
	37	58	35	55
hema_measi	IN		IN	
	22	43	35	58
hema_pi4ha	IN		IN	
	24	45	28	47
hg2a_human*	IN		IN	
	47	71	46	71
hism_salty	OUT		OUT	
	23	47	27	47
	61	85	59	79
	104	125	105	125
	158	182	158	178
	202	221	200	220
hisq_salty	OUT		OUT	
	13	37	13	33
	55	79	59	79
	93	117	88	108
	149	173	153	173
	191	215	195	215
hoxn_alceu	IN		IN	
	18	42	20	40

continued overleaf

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
	48	72	52	72
	88	112	95	115
	128	149	129	149
	198	222	200	220
	238	262	244	264
	278	302	270	290
	318	339	317	337
ig1r_human	OUT		OUT	
	904	928	906	929
il2a_human*	OUT		OUT	
	220	239	220	238
il2b_human*	OUT		OUT	
	221	239	215	239
im23_schja	IN		IN	
	13	36	13	36
	53	76	56	73
	83	107	83	108
	184	208	184	205
im23_schma	IN		IN	
	13	36	13	36
	53	76	56	73
	83	107	83	108
	184	208	184	205
imm1_ecoli*	IN		IN	
	5	24	9	26
	39	62	39	57
	86	110	84	104
imma_citfr*	IN		IN	
	16	37	14	37
	69	89	69	89
	103	123	107	124
	145	166	143	165
ita5_mouse	OUT		OUT	
	957	981	956	981
kdpd_ecoli	IN		OUT	
	403	421	403	422
	426	444	425	444
	449	473	447	466
	478	498	478	498
kgtp_ecoli	IN		IN	
	25	49	26	51
	59	83	62	80
	95	115	96	116
	125	149	120	137
	159	183	163	185
	195	213	196	214
	245	266	244	261
	278	302	275	300
	311	330	312	330
	339	360	337	360
	372	396	369	392
	405	423	403	423
lacy_ecoli*	IN		IN	
	9	27	11	33
	47	66	47	67
	75	94	75	99
	103	125	103	125
	145	164	145	163
	168	186	168	187
	222	240	212	234
	260	283	260	281
	292	311	291	310
	315	334	315	334
	347	369	347	366
	380	398	380	399
lech_human*	IN		IN	
	40	58	40	60
leci_mouse*	IN		IN	
	59	77	59	79
lep_ecoli*	OUT		OUT	
	4	28	4	22
	58	76	58	76
lha4_rhoac	IN		IN	
	14	35	15	35

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
lhb5_rhoac	IN		IN	
	21	40	14	36
lspa_ecoli	IN		IN	
	12	36	12	29
	70	88	70	88
	99	117	96	113
	134	157	139	156
magl_mouse	IN		OUT	
	3	21		
	490	514	498	517
malf_ecoli*	IN		IN	
	15	36	17	35
	40	58	40	58
	71	92	73	91
	285	306	277	295
	319	340	319	337
	371	392	371	389
	426	447	418	436
	485	506	486	504
malg_ecoli	IN		IN	
	13	37	19	39
	91	111	82	102
	124	144	124	144
	153	177	151	171
	207	227	205	225
	258	280	260	280
melb_ecoli*	IN		IN	
	9	30	8	28
	40	64	33	53
	74	96	76	96
	106	130	103	123
	140	163	146	166
	173	194	172	192
	229	253	231	251
	263	283	263	283
	293	312	293	313
	322	346	320	340
	374	393	370	390
	403	426	408	428
mota_ecoli	IN		IN	
	4	28	4	21
	34	54	34	51
	165	189	171	191
	195	219	201	222
motb_ecoli*	OUT		IN	
	28	49	28	49
mprd_human*	OUT		OUT	
	160	184	160	184
mtr_ecoli	IN		OUT	
	10	30	16	36
	34	56	42	62
	89	109	89	109
	129	148	129	149
	152	171	152	172
	190	209	191	211
	228	251	229	249
			256	276
	284	305	286	306
	325	344	326	345
	348	367	348	367
	387	410	386	406
myp0_human*	OUT		OUT	
	126	150	125	150
nep_human	IN		IN	
	28	47	28	50
ngfr_human*	OUT		OUT	
	223	244	223	244
oppb_salty*	IN		IN	
	9	27	10	30
	100	121	100	121
	135	156	138	158
	169	190	173	190
	229	250	227	250
	275	296	272	293

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
oppc_salty*	IN		IN	
	40	59	38	59
	104	128	103	122
	136	160	140	160
	164	183	164	181
	218	236	216	236
	271	290	268	290
ops1_calvi*	OUT		OUT	
	50	74	48	72
	85	109	85	110
	119	142	125	144
	163	187	164	187
	218	239	212	237
	275	299	275	298
	310	330	306	330
ops2_drome*	OUT		OUT	
	59	83	57	81
	94	118	94	119
	129	153	134	153
	173	197	173	196
	228	249	221	246
	284	308	384	307
	318	339	315	339
ops3_drome*	OUT		OUT	
	60	84	58	82
	95	119	95	119
	130	154	134	153
	165	189	172	196
	220	241	221	246
	285	309	285	308
	319	340	317	341
ops4_drome*	OUT		OUT	
	56	80	54	78
	90	114	91	113
	125	149	130	149
	162	186	168	192
	217	238	217	242
	281	305	281	304
	315	336	313	337
opsb_human*	OUT		OUT	
	39	60	34	58
	69	93	71	96
	106	130	111	130
	150	171	150	173
	200	221	200	225
	250	274	250	273
	284	306	282	306
opsd_bovin*	OUT		OUT	
	37	61	37	61
	71	95	74	99
	115	133	114	133
	153	174	153	176
	203	224	203	228
	253	277	253	276
	286	308	285	309
opsg_human*	OUT		OUT	
	55	79	53	77
	88	112	90	115
	125	149	130	149
	169	190	169	192
	219	240	219	244
	269	293	269	292
	302	325	301	325
opsr_human*	OUT		OUT	
	55	79	53	77
	88	112	90	115
	125	149	130	149
	169	190	169	192
	219	240	219	244
	269	293	269	292
	302	325	301	325
phor_ecoli	IN		IN	
	14	38	14	34
			38	58

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
pigr_human	OUT		OUT	
	621	642	621	643
ptma_ecoli	IN		IN	
	15	38	25	44
	48	69	51	69
	79	103		
	134	158	135	154
	168	189	166	184
	213	234		
	258	282	274	291
	313	334	314	333
rech_rhovi*	OUT		OUT	
	12	31	12	32
rcel_rhovi*	IN		IN	
	22	46	32	55
	78	102	84	109
	118	142	115	140
	174	198	170	195
	232	256	225	250
rcem_rhovi*	IN		IN	
	48	71	52	77
	111	129	110	135
	144	167	142	167
	202	225	197	222
	266	289	259	284
rfbp_salty	IN		OUT	
	15	39	15	32
	56	74	56	73
	90	108	90	107
	116	134	115	132
	160	179		
	240	258		
	284	303	285	302
rhat_ecoli	OUT		OUT	
	4	24	4	24
	36	57	38	58
	74	93	74	94
	97	117	101	121
	136	157	137	157
	175	195	175	195
	214	234	214	234
	253	274	259	279
	290	310	290	310
	322	343	323	343
rib1_rat	IN		OUT	
	88	107	416	433
	416	435		
secd_ecoli	IN		OUT	
	10	29	10	30
	454	472	452	472
	476	497	476	497
	502	524	504	524
	549	573	564	584
	578	597	587	605
sece_ecoli*	IN		IN	
	13	32	19	36
	39	63	45	63
	95	119	93	111
secy_bascu*	IN		IN	
	18	36	18	39
	68	87	59	80
	119	140	115	132
	148	166	148	167
	174	192	174	192
	214	233	217	234
	266	284	268	291
	310	329	310	329
	368	386	367	386
	394	413	392	410
secy_ecoli	IN		IN	
	24	43	23	42
	75	96	75	95

continued overleaf

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
	122	140	122	139
	154	175	154	174
	183	204	183	203
	218	237	217	237
	271	292	274	294
	318	337	316	335
	369	387	376	395
	407	426	399	416
spg1_strsp*	IN		OUT	
	–	–	389	409
ssrg_rat	IN		OUT	
	30	51	38	48
	59	77	55	76
	135	156	136	157
	164	182	164	184
suis_human	IN		IN	
	11	32	13	32
tal6_human	OUT		IN	
	10	30	10	30
	46	70	46	70
	91	115	89	114
	165	189	162	187
tcbl_rabit*	OUT		OUT	
	289	313	292	313
tcr1_ecoli*	IN		IN	
	7	31	8	28
	41	65	44	64
	75	99	76	96
	104	123	104	124
	133	157	133	153
	162	180	161	181
	212	236	216	236
	245	269	246	267
	279	297	279	298
	302	325	301	320
	334	358	338	358
	367	385	365	385

Swissprot ID	Topology and TM Helices			
	Predicted		Observed	
tolq_ecoli	OUT		IN	
	13	36	23	43
	127	150	127	152
	167	191	162	187
tolr_ecoli	IN		IN	
	16	40	16	40
top_drome	OUT		OUT	
	838	861	838	858
	973	994		
trbm_human	OUT		OUT	
	496	516	495	518
trsr_human	IN		IN	
	62	83	63	88
uhpt_ecoli*	IN		IN	
	27	45	26	46
	62	80	70	90
	97	115	99	120
	124	148	124	144
	158	182	159	179
	191	210	190	210
	260	278	258	279
	295	319	295	315
	327	345	327	348
	354	378	356	376
	394	418	382	402
	427	446	409	430
upkb_bovin	IN		IN	
	12	36	12	37
	56	80	60	80
	86	109	86	111
	229	253	230	253
vmt2_iaann*	OUT		OUT	
	25	43	25	42
vnb_inbbe*	OUT		OUT	
	16	40	19	40

*Entries which are included in the initial parameter and pseudocount settings from the 83TMP set.

Edited by J. Thornton

(Received 27 April 1998; received in revised form 9 July 1998; accepted 21 July 1998)