

Prediction of membrane proteins based on classification of transmembrane segments

Daisuke Kihara, Toshio Shimizu¹ and Minoru Kanehisa²

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011 and
¹Department of Information Science, Faculty of Science, Hirosaki University, Hirosaki 036-8244, Japan

²To whom correspondence should be addressed

The number of transmembrane segments often corresponds to a structural or functional class of membrane proteins such as to seven-transmembrane receptors and six-transmembrane ion channels. We have developed a new prediction method to detect the membrane protein class that is defined by the number of transmembrane segments, as well as to locate the transmembrane segments in the amino acid sequence. Each membrane protein class is represented by a model of ordering different types of transmembrane segments. Specifically, we have classified the transmembrane segments in known membrane proteins into five groups (types) using the Mahalanobis distance with the average hydrophobicity and the periodicity of hydrophobicity as a measure of similarity. The discriminant functions derived for these groups were then used to detect transmembrane segments and to match with the models for one- to fourteen-spanning membrane proteins and for globular proteins. Using the test data set of 89 membrane proteins whose transmembrane positions are known by experimental evidence, 61.8% of the proteins and 85.1% of the transmembrane segments were correctly predicted. Because of the new feature to predict membrane protein classes, the method should be useful in the functional assignment of genomic sequences.

Keywords: discriminant analysis/Mahalanobis distance/membrane protein class/protein structure prediction/compatibility score

Introduction

A recent surge of whole genome sequencing is producing complete genome sequences of an increasing number of organisms. These sequence data have enabled genome-scale analysis of arrangements and compositions of genes in order to understand genome architecture, mechanisms of gene expression, and detailed evolutionary relationships among organisms (Tatusov *et al.*, 1997; Bono *et al.*, 1998). It is essential that the biological meaning (function and structure) is correctly assigned to each open reading frame (ORF) in the genome, but the current homology and motif search methods are unable to assign functions for at least one third of the ORFs in the genome. Even without any homology, predictive methods can still detect membrane proteins, which form a major class of proteins constituting about 30% of the entire ORFs in the microbial genomes thus far sequenced. Because the number of transmembrane segments in a membrane protein can be related to a functional subclass in some cases, such as seven-transmembrane receptors and six-transmembrane ion channels,

better prediction of transmembrane segments and membrane protein groups will help identify biological functions.

The prediction of transmembrane segments has been one of the main subjects in computational biology since Kyte and Doolittle (1982) presented the idea of hydrophathy analysis. Various improvements were performed on the hydrophathy analysis (Eisenberg *et al.*, 1984; Klein *et al.*, 1985; Rao and Argos, 1986). In addition, von Heijne (1992) used a 'positive-inside rule' to evaluate the validity of topology models derived from the hydrophathy analysis. The method by Jones *et al.* (1994) enumerates possible topology models with scores by the dynamic programming algorithm. Persson and Argos (1994) used amino acid propensity tables and multiple sequence alignments. Neural networks were applied by Rost *et al.* (1995).

These methods, as well as ours presented here, assume that in the majority of membrane proteins transmembrane segments are rich in hydrophobic amino acid residues and form helices in a direction almost normal to the membrane surface. This is based on the observations made in bacteriorhodopsin (Henderson *et al.*, 1990), photosynthetic reaction centre (Deisenhofer *et al.*, 1985), light-harvesting complex (Kühlbrandt *et al.*, 1994) and cytochrome c oxidase (Tsukihara *et al.*, 1996). Exceptions are β -barrel porins (Weiss and Schulz, 1992) and several types of bacterial toxins that form membrane pores, such as gramicidin A (Langs, 1988), annexin (Lücke *et al.*, 1995), colicin A (Parker *et al.*, 1992) and aerolysin (Parker *et al.*, 1994). Furthermore, helices parallel to the membrane surface were observed in prostaglandin H₂ synthase-1 (Picot *et al.*, 1994) and the transmembrane domains of nicotinic acetylcholine receptor might contain β -structures as well (Görne-Tschelnokow *et al.*, 1994). These exceptions cannot be predicted by any of the existing methods.

Our method is characterized by the following three features. The main feature is that different properties of different transmembrane segments are incorporated based on classification of transmembrane segments in a database. In fact, not all transmembrane segments are equally hydrophobic, but some of them have distinctive features. For example, transmembrane segments of single spanning membrane proteins are known to be highly hydrophobic and have small hydrophobic moments (Eisenberg *et al.*, 1984), whereas the last transmembrane segments in seven-transmembrane proteins are relatively less hydrophobic and are often difficult to detect by prediction methods (Jones *et al.*, 1994; Persson and Argos, 1994). Thus, we have classified transmembrane segments first by the total number of transmembrane segments in a protein and the order that they appear in the protein sequence, and then by merging similar ones into groups. Second, our method enumerates possible models ranked by their scores where a model is distinguished by the number of transmembrane segments in a membrane protein and represented by the ordering of different groups (types) of transmembrane segments. Even for a membrane protein whose topology is derived by some experimental evidence, it is often the case that contradictory results are

Table I. The training data set of membrane proteins taken from SWISS-PROT release 34.0

TM	Number of sequences in SWISS-PROT	Number of dissimilar sequences ^a	Number of sequences used in this work ^b
1	2823	1084	739
2	842	394	389
3	407	239	234
4	564	231	225
5	298	152	146
6	442	229	226
7	787	222	218
8	221	126	125
9	83	60	60
10	151	102	101
11	182	116	114
12	443	201	201
13	65	41	41
14	46	33	33
15	12	7	7
16	9	7	6
17	9	6	6
18	1	1	1
19	1	1	1
23	1	1	1
24	23	2	2
30	1	1	0
Total	7411	3256	2876

^aNo pair of sequences has more than 30% identity.

^bPrecursor entries with ambiguous signal sequence positions and fragment entries were excluded.

suggested by other experiments (Rost *et al.*, 1996). Therefore, it is desirable for a predictive method to output not a single prediction but a list of possibilities with certainty measures, so that further experiments can be designed to distinguish between several topology models. Third, the possibility that the query sequence is a globular protein is explicitly taken into consideration, which is not necessarily a feature incorporated in the existing methods.

We have collected from the literature 89 sequences of membrane proteins whose position of transmembrane segments are based on experimental evidence. The method has been applied to this test data set and to a set of globular protein sequences.

Materials and methods

Data set

For grouping of transmembrane segments and for construction of discriminant functions, the training data set of membrane proteins was prepared as follows. Starting from the 7411 membrane proteins taken from SWISS-PROT (Bairoch and Apweiler, 1997) *rel.* 34.0, similar sequences were first removed so that every pair of sequences had less than 30% of sequence identity. N-terminal signal sequences of precursor entries had been excised according to the annotation of SWISS-PROT. Next, fragment sequence entries as well as precursor entries with ambiguous signal sequence positions were removed from the set. The number of sequences grouped by the number of transmembrane segments is summarized in Table I. The training data set contained 14 353 transmembrane segments in 2876 proteins.

The test data set of 89 membrane proteins shown in Table II was collected from the literature reporting experimental

evidence of transmembrane topologies. Every pair of sequences had less than 35% of sequence identity. The data set contained 32 771 residues including 7817 residues in 355 transmembrane segments. This test set is available through the WWW (<http://web.kuicr.kyoto-u.ac.jp/~kihara/tmpdb.html>).

The test data set of globular proteins was prepared from the PDBSELECT database March 1997 version (Hobohm *et al.*, 1992). The 35% threshold list was used excluding the entries of membrane and lipid associated proteins (1AXN, 1COLA, 1HGGA, 1OCC, 1PRC, 1IDO, 2POR, 1ATY, 1SPF). This set consisted of 928 sequences.

Discriminant analysis

Our classification and prediction methods are based on the linear discriminant rule (Kendall and Stuart, 1976), which is briefly summarized below. The linear discriminant rule assumes that the distributions of two groups, G_1 and G_2 , against a given parameter set are Gaussian having the same covariance matrix Σ . Namely, the distribution functions are $N_1(\mu_1, \Sigma)$ and $N_2(\mu_2, \Sigma)$, where μ_1 and μ_2 are the mean vectors of parameters. Let x denote the vector of parameters for a transmembrane segment, and let $P(G_1/x)$ and $P(G_2/x)$ be, respectively, the conditional probabilities that the segment with the attribute x occurs in G_1 and G_2 . The linear discriminant rule assigns the segment with x to G_1 if:

$$P(G_1/x)/P(G_2/x) - 1 > 0 \quad (1)$$

We call the left-hand side of eqn (1) the score of the discriminant function. According to the Bayes theorem, the score can be rewritten using $N_n(\mu_n, \Sigma)$ ($n = 1, 2$) and the prior probabilities of the two groups, which we assumed to be the same, or 0.5. Thus, eqn (1) becomes a linear function for the vector x . The Mahalanobis distance D^2 , which we used as the measure of distance between groups of transmembrane segments, represents the difference between the means of the score distributions in the two groups:

$$D^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \quad (2)$$

Using the Mahalanobis distance, the score distributions of the two groups become $N(1/2D^2, D^2)$ for G_1 and $N(-1/2D^2, D^2)$ for G_2 .

Parameters for discriminant analysis

The parameters used in the discriminant analysis for the classification of transmembrane segments were the average hydrophobicity $\langle H \rangle$ and the AP value (Donnelly *et al.*, 1993) using the hydrophobicity index proposed by Kyte and Doolittle (1982). The AP value indicates helical periodicity of hydrophobicity defined by:

$$AP = \frac{\frac{1}{30} \int_{90}^{120} P(\omega) d\omega}{\frac{1}{180} \int_0^{180} P(\omega) d\omega} \quad (3)$$

where

$$P(\omega) = \left| \sum_{i=1}^N (H_i - \langle H \rangle) \sin(i\omega) \right|^2 + \left| \sum_{i=1}^N (H_i - \langle H \rangle) \cos(i\omega) \right|^2 \quad (4)$$

Table 2. The test data set of membrane proteins

TM	SWISS-PROT or PDB code
1	BCS1_YEAST CP5B_CANTR CYB5_RABIT DIVB_BACSU EXBD_ECOLI FTSL_ECOLI GEF_ECOLI MPRD_BOVIN NRAM_IAPUE OCH1_YEAST PBPB_ECOLI PGDR_MOUSE RCEH_RHOSH RCEH_RHOVI RIB1_HUMAN RIB2_HUMAN GHR_HUMAN TOLR_ECOLI TONB_SALTY TOXR_VIBCH THAS_HUMAN AMD2_XENLA VNB_INBLE OSTB_YEAST VG1_BPFD VS10_ROTHW 1KZU(A) 1KZU(B) 1OCC(D) 1OCC(G) 1OCC(I) 1OCC(J) 1OCC(K) 1OCC(L) 1OCC(M)
2	CPXA_ECOLI CYOA_ECOLI ENVZ_ECOLI FTSH_ECOLI IMM_BPT4 LEP3_ECOLI STS_HUMAN 1OCC(B)
3	CYOD_ECOLI EXBB_ECOLI KDGL_ECOLI
4	CXA1_RAT CXB1_RAT DSBB_ECOLI FIXL_RHIME IM17_YEAST IM23_YEAST IMMA_CITFR MYPR_HUMAN VRXB_LAMBD
5	CYOC_ECOLI DHG_ECOLI HISM_SALTY HISQ_SALTY RCEL_RHOSH RCEM_RHOSH
6	CNG1_BOVIN MALG_ECOLI MPCP_BOVIN NO26_SOYBN OPPB_SALTY OPPC_SALTY PTMA_ECOLI UCP_HUMAN
7	BACR_HALHA CYDA_ECOLI CYOE_ECOLI HOXN_ALCEU LSHR_RAT 1OCC
8	ATN1_PIG ATP6_ECOLI CPB5_RABIT CYB_RHOSH HMDH_CRIGR LCRD_YERPE
9	CITN_KLEPN
10	CAN1_YEAST CLC1_HUMAN
12	GLPT_ECOLI LYSF_ECOLI TCR1_ECOLI 1OCC(A)
14	NNTM_BOVIN

H_i is the hydrophobicity index given to each amino acid residue, N is the length of the segment, and ω is the angle in degrees.

In addition to the average hydrophobicity and the AP value, we used other parameters to optimize the accuracy of prediction. ‘Height’ is the highest point of the hydropathy plot (a seven residue subwindow was used) in the sliding window of 17 residues long. ‘Area’ is measured originally as the sum of values of the hydropathy plot in the sliding window (Degli Esposti *et al.*, 1989), but here it is normalized by the length of the sliding window, because the length of the sliding window in prediction and the lengths of segments in the training data were different. ‘Polarity’ is the largest average polarity of 11 residue subwindows in the sliding window. ‘Size’ is the largest average size of three residue long subwindows in the sliding window. The indices of polarity and size (molecular volume) were taken from Grantham (1974).

Evaluation of prediction accuracy

We have evaluated the accuracy of prediction in three ways: protein-based, segment-based and residue-based. The protein-based accuracy is the most strict one. Predicted proteins are counted to be correct if they have the correct number of transmembrane segments and all of the positions of the predicted transmembrane segments are accurate, which we consider when predicted segments overlap by 11 or more residues with the observed ones. This threshold length was chosen because 11 residues are more than half of typical transmembrane α -helices which are on average 21 residues long. The segment-based accuracy counts just the number of correct segment assignments according to the same criterion.

For the evaluation of the residue-based accuracy, the measures Q_3 , Q_4 , Q_5 and Q_7 were shown to be useful (Schulz and Schirmer, 1979). Q_3 is the overall percentage of correctly predicted residues in the sequence:

$$Q_3 = \frac{N_{TT} + N_{LL}}{L} \quad (5)$$

Q_4 is the average of the percentages of correctly predicted

residues in transmembrane segments and correctly predicted residues in loop regions:

$$Q_4 = \frac{1}{2} \left(\frac{N_{TT}}{N_{TT} + N_{TL}} + \frac{N_{LL}}{N_{LL} + N_{LT}} \right) \quad (6)$$

Q_5 is the measure which also takes into account the inequality of the amount of residues in transmembrane segments and in loop regions:

$$Q_5 = \frac{N_{TT}}{1 - N_{LL}} \quad (7)$$

and Q_7 is the Matthews’ correlation coefficient (Matthews, 1975):

$$Q_7 = \frac{N_{TT}N_{LL} - N_{LT}N_{TL}}{\sqrt{(N_{LL} + N_{TL})(N_{LL} + N_{LT})(N_{TT} + N_{TL})(N_{TT} + N_{LT})}} \quad (8)$$

The symbol N_{pq} [$p, q = T$ (denotes transmembrane) or L (denotes loop)] used in the formulae above is the number of residues which is observed to be p and predicted to be q .

Results

Classification of transmembrane segments

First, all the transmembrane segments in the training data set were classified into ‘subgroups’ according to the total number of transmembrane segments in a membrane protein and the order that they appear in the membrane protein sequence. The membrane proteins with more than fourteen transmembrane segments were excluded in the present analysis because the number of sequences was too small to execute statistical calculations. Thus, there were 105 ($14 \times 15/2$) subgroups. Next, similar subgroups were merged into a ‘group’. Since each transmembrane segment was characterized by the two parameters, the average hydrophobicity and the AP value, a subgroup is a distribution of these two parameters. We performed the clustering of subgroups based on the discriminant analysis. We used two clustering methods, single linkage and complete linkage, with several threshold values for two

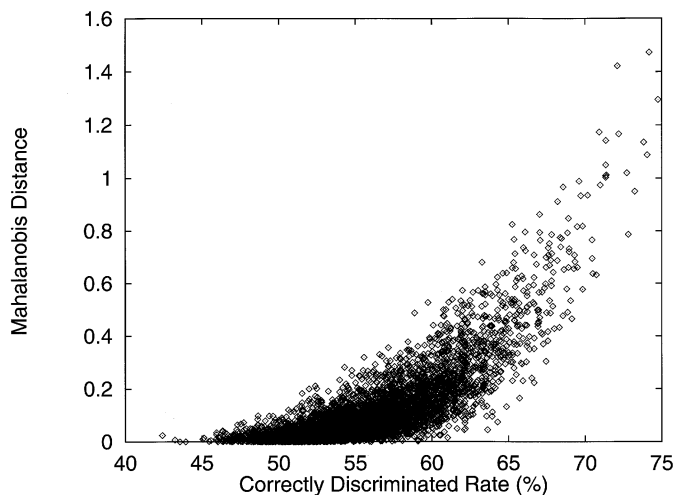


Fig. 1. The correlation between the correct discrimination rate and the Mahalanobis distance for the 5460 ($105 \times 104/2$) pairs of subgroups. The correlation coefficient was 0.79.

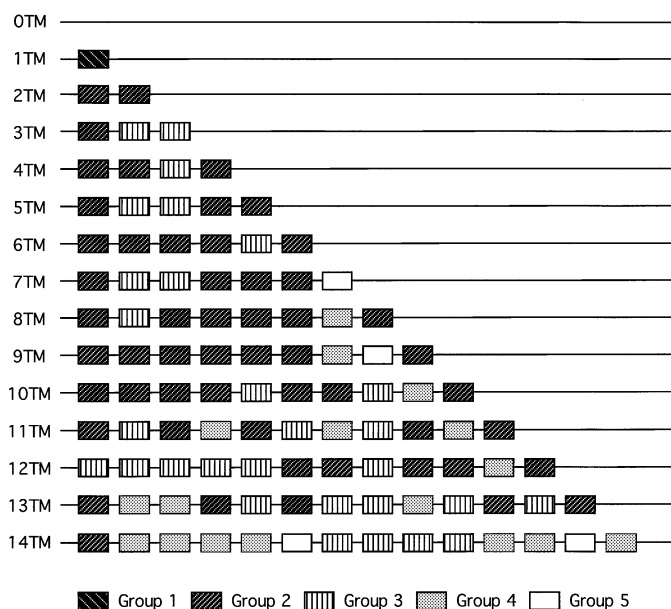


Fig. 2. The models for the membrane protein classes containing one (1TM) to 14 (14TM) transmembrane segments, as well as for globular proteins (0TM), according to the locations of the five groups of transmembrane segments. The N-terminal transmembrane segments are located on the left. The boxes with the same pattern belong to the same group, and the darker shading corresponds to higher average hydrophobicity with group 1 being most hydrophobic and group 5 least hydrophobic.

different measures, the Mahalanobis distance and the actual discrimination rate.

The Mahalanobis distance represents an ideal condition of Gaussian distributions with the same covariance matrix. To see if it is appropriate to use as a similarity measure of transmembrane segments, we checked the correlation with the actual rate of correct discrimination for all the 5460 ($105 \times 104/2$) pairs of subgroups. The result is shown in Figure 1, where the correlation coefficient was 0.79. We conclude that the Mahalanobis distance correlates well enough with the actual discrimination rate to be used as a measure of classifying transmembrane segments, especially when the Mahalanobis distance is above 0.2.

The result of classifying subgroups into groups is summar-

ized in Table III. The single linkage analysis generally tends to form larger clusters, and in most of the cases in Table III all of the subgroups were merged into one cluster. On the other hand, the complete linkage analysis will produce smaller, but uniform clusters, because the distances of all pairs of elements in a cluster are assured to be within the threshold value. We used the results of the complete linkage analysis, specifically the five groups formed using the Mahalanobis distance of 0.25, as the threshold value, because this clustering level agreed with our empirical knowledge in which the subgroup of the last spanning segment of seven transmembrane proteins is somewhat different. Note that from a practical point of view of better prediction, the choice of the clustering level should be made to achieve best accuracy. However, since the prediction accuracy depends on a number of other parameters as well, we did not perform a systematic search of the clustering level.

Models of membrane protein classes

One of the main features of our prediction method is that it distinguishes membrane protein classes according to the number of transmembrane segments. Each class is represented by a model that is a combination of different types of transmembrane segments. According to the above grouping of transmembrane segments, the 15 models that contain zero to 14 transmembrane segments are shown in Figure 2. The model that has no transmembrane segments corresponds to that for globular proteins. The average hydrophobicity of each transmembrane segment group is represented by the shading, darker shading represents greater hydrophobicity. The most hydrophobic group 1 appears only in the single spanning membrane proteins, which supports the validity of the method by Jones *et al.* (1994) who used special parameters for single spanning membrane proteins. Group 2, which has relatively high average hydrophobicity, appears at the first transmembrane segments in most of the membrane protein classes (except the 12 transmembrane proteins). It is possible that these N-terminal highly hydrophobic transmembrane segments are involved in the initiation of membrane insertion. These segments may correspond to what Eisenberg *et al.* (1984) called the 'initiators', although they did not mention the locations in the sequence. The last segment of the seven spanning membrane proteins belongs to the least hydrophobic group 5, again supporting previous suggestions, which also appears in the eighth segment of the nine spanning membrane proteins. Furthermore, groups 3 and 4 with relatively low average hydrophobicity appear more abundantly in membrane proteins with many transmembrane segments. Most of the segments in the 14 spanning membrane proteins seemed to be less hydrophobic, but the number of proteins in the data set was too small (Table I) to make further assessments.

Figure 3 presents a more quantitative picture of the grouping. It is a distance matrix of the 105 subgroups ordered along the axis according to the five groups identified. It shows again clearly that the subgroup of single spanning membrane proteins (group 1) is quite different from the rest of the subgroups. The five groups also stand out well; especially, groups 1 and 2 are well distinguished from groups 4 and 5. The distances between subgroups within each group are relatively small; most of them are less than 0.1 in the Mahalanobis distance.

Figure 4 illustrates the average hydrophobicity and the AP value of each subgroup in the five groups. The five groups seem to be distinguished mainly by the average hydrophobicity.

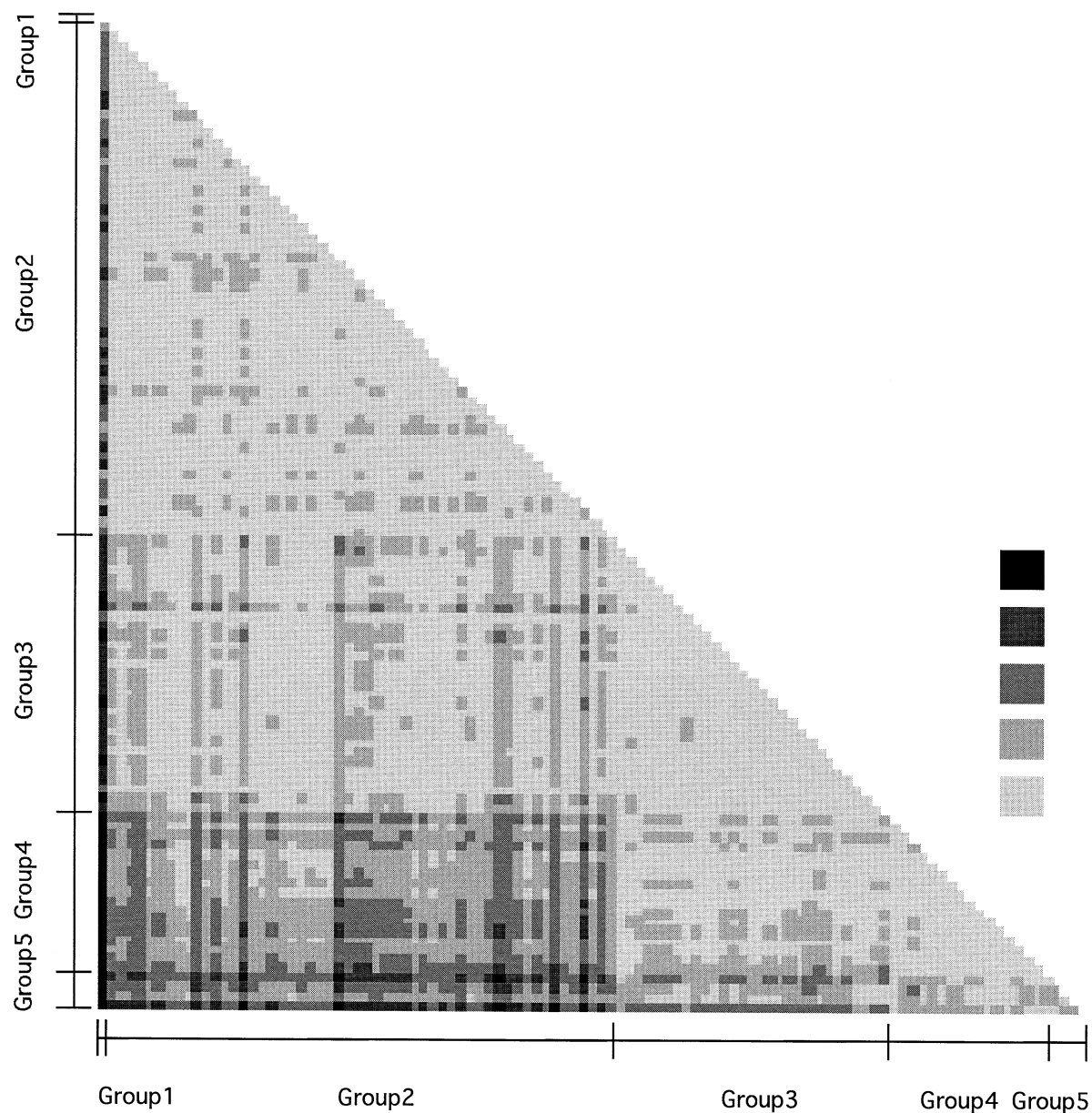


Fig. 3. The Mahalanobis distances between all pair of the transmembrane subgroups. The 105 transmembrane subgroups are aligned according to the five groups from top to bottom and left to right. Within each group the subgroups are aligned according to the total number of transmembrane segments and the order in the sequence (see Figure 2). The distance (d) is represented in five levels from darker to lighter shading, corresponding to the ranges of: $d \geq 0.75$, $0.75 > d \geq 0.5$, $0.5 > d \geq 0.25$, $0.25 > d \geq 0.1$ and $d < 0.1$.

Table III. The number of groups of transmembrane segments formed with two different measures and at various threshold values

Mahalanobis distance			Discrimination rate		
Threshold value	Complete linkage	Single linkage	Threshold value (%)	Complete linkage	Single linkage
0.20	7	1	55.0	21	2
0.25	5	1	57.5	14	2
0.30	4	1	60.0	9	1
0.40	4	1	62.5	6	1
0.60	3	1	65.0	4	1
0.80	2	1	67.5	3	1
1.00	2	1	70.0	3	1

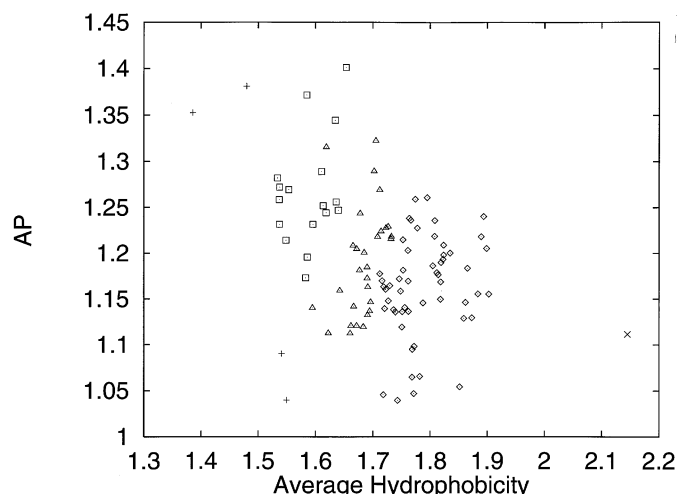


Fig. 4. The average hydrophobicity and the AP value in each of the 105 subgroups according to the classification into five groups: ×, group 1; ◇, group 2; △, group 3; □, group 4; and +, group 5.

However, the AP value takes effect on distinguishing groups 3 and 4. The last spanning segment of the seven transmembrane proteins and the eighth spanning segment of the nine transmembrane proteins in group 5 are plotted at (1.48, 1.38) and (1.54, 1.09), respectively. Both have low average hydrophobicity, but the AP value is quite different.

Prediction procedure

Our prediction procedure is based on the detection of different groups of transmembrane segments using multiple discriminant functions and matching with the 15 models of membrane proteins. The linear discriminant function was constructed for each of the groups of transmembrane segments against the group of loop segments. We define a loop as a segment that is not transmembrane. Thus, the group of loop regions consisted of all non-transmembrane segments that were longer than four residues, which was long enough to calculate the AP value, in the membrane protein sequences of the training data set. Another discriminant function was also constructed for distinguishing the combined group of all transmembrane segments against the group of loop regions.

Figure 5 illustrates the prediction procedure. In the first stage, a query sequence is applied to each model to see if it is compatible. When the model is represented only by a single discriminant function, transmembrane segments are selected in order of their scores. When the model contains multiple discriminant functions, each function first selects candidate transmembrane segments, then the combination that gives the highest score is adopted for the model avoiding overlapping segments. The selection of transmembrane segments in the query sequence is made by a window search, where the score for a window is the result of computing the parameter values and employing the discriminant function. The window length is set to 17 because it was the most successful in the previous work (Klein *et al.*, 1985). Those windows determined to be transmembrane are called 'cores' of transmembrane segments. Then the length of each core is adjusted as follows. The 17 residue window is moved from both ends of the core toward N- and C-terminal directions until the discriminant function gives a negative score, which determines outer boundaries of the transmembrane segment. The final predicted boundary is taken to be the halfway between the outer boundary and the

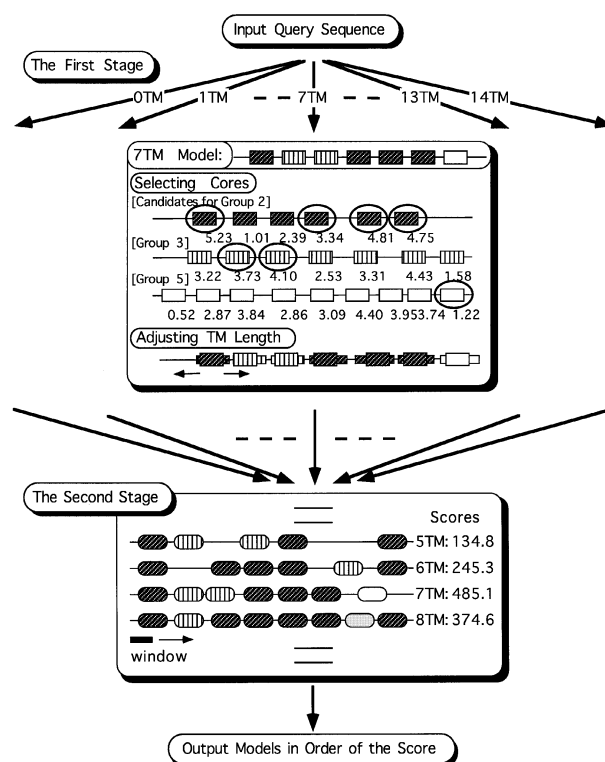


Fig. 5. A schematic diagram showing the prediction procedure in TSEG. In the first stage, the query sequence is applied to each model and the transmembrane segments are determined using multiple discriminant functions. Here an example of applying the sequence to the seven transmembrane protein model is shown. Because the seven transmembrane protein model consists of transmembrane segments of groups 2, 3 and 5 (Figure 2), three different discriminant functions first select candidates of fixed-length cores. Next, the combination of the cores that gives the highest score is adopted for the model (circled cores). Then, the lengths of the transmembrane segments are adjusted. In the second stage, the transmembrane structure selected by the models are compared by their scores that represent the compatibility to the models [see eqn (9)]. The prediction result is a list of probable structures enumerated in order of the scores.

core boundary. In case the outer boundaries of neighbouring transmembrane segments overlap, both boundaries are shortened as little as possible to eliminate the overlap. If the transmembrane segments obtained are longer than 35 residues, they are shortened not to exceed 35 residues.

In the second stage, the models are compared by their scores. The score given to a model is defined by:

$$\text{Score } i = \sum_{n=1}^{L-w+1} f_j(x_n) \quad (9)$$

where i is the model number ($i = 0 \sim 14$), L is the length of the query sequence, w is the length of the sliding window, x_n is the vector of parameters for the n th window, f_j is the discriminant function for one of the groups to which the window belongs. If the window is in the loop region, the discriminant function which was constructed using the loop regions against all-transmembrane segments is used. Note that the locations of possible transmembrane segments are already determined in the first stage for each model, and the score for each model is the sum of scores for all the windows in the sequence. After comparing all the models, the model with the highest score is considered to be most reliable, and the list of models with their scores is output. This is analogous to the protein fold recognition procedure where a query sequence is

Table IV. Prediction accuracy with different parameter sets

Parameter set	Protein-based		Segment-based				Residue-based					
	Q ₂ (%)	<L>	Obs sov (%)	Prd sov (%)	Nseg over	Nseg under	Q ₃ (%)	Obs TM (%)	Prd TM (%)	Q ₄	Q ₅	Q ₇
<H>	52.8	26.6	82.0	90.4	31	64	87.9	78.1	73.0	0.845	0.606	0.675
<H>,AP	53.9	26.6	82.5	90.2	32	62	87.8	78.3	72.6	0.845	0.605	0.673
<H>,AP,Height	52.8	26.9	84.5	86.6	46	55	87.0	81.0	69.5	0.849	0.597	0.664
<H>,AP,Polarity	61.8	26.7	85.1	91.5	28	53	88.1	80.5	72.7	0.855	0.618	0.687
<H>,AP,Size,Height	22.5	30.3	31.8	97.4	3	242	81.4	29.7	79.6	0.637	0.276	0.409
<H>,AP, Area,Height	51.7	26.3	81.1	89.1	35	67	87.8	77.3	73.0	0.842	0.601	0.670
<H>,AP, Height, Polarity	23.6	29.5	33.0	97.5	3	238	82.1	30.8	84.0	0.645	0.291	0.437

Abbreviations:

Q₂, fraction of proteins correctly predicted.

Q₃, Q₄, Q₅, Q₇, see Materials and methods for definition.

<L>, average length of predicted transmembrane segments.

Obs/Prd sov, observed/predicted segment overlaps.

Nseg over/under, number of false positive/negative segments.

Obs/Prd TM, correctly predicted residues in the observed/predicted transmembrane segments.

compared with each of the predefined models, namely, the three-dimensional profiles of representative folds.

The program TSEG (prediction of Transmembrane SEGment in proteins) is written in ANSI C and runs under UNIX. We will make the WWW version of TSEG publicly available as a part of the Japanese GenomeNet service (<http://www.genome.ad.jp/>).

Prediction accuracy

The evaluation of prediction accuracy was made against the test data set using the discriminant functions obtained by the training data set. For each sequence in the test set, sequences that had 30% or more sequence identity were excluded from the training set. Therefore, the discriminant functions were renewed every time the prediction of the sequences in the test set was performed.

Table IV shows the result of using different parameter sets for discriminant functions, and summarizes the protein-based, segment-based and residue-based prediction accuracies (see Materials and methods). More than 80% of the transmembrane segments were correctly predicted (Segment-based Obs in Table IV) in most parameter sets. The best result was achieved with the parameter set of <H>, AP value and polarity, which correctly predicted 61.8% of proteins, 85.1% of transmembrane segments, 88.1% of residues in the test set. Compared with the results using <H> and AP value only or using <H>, AP value and height, it is evident that polarity was effective. This is in agreement with the report that transmembrane segments are stabilized with each other by polar interactions (Mitaku *et al.*, 1995). In general, TSEG tended to predict transmembrane segments longer than the actual observations. Another tendency of TSEG was that it missed rather than overpredicted transmembrane segments. These aspects become clearer when compared with other methods (see below).

In the discriminant analysis, the results were very sensitive to the parameter sets used to represent transmembrane segments. If a bad parameter was included in the parameter set, the results became drastically worse; an example is the set of <H>, AP value and size. Or if the combination of parameters was not good, the results were also bad; an example is the set of <H>, AP value, height and polarity.

To represent the residue-based accuracy, we showed Q₄, Q₅ and Q₇ together with Q₃. The purpose of this is to make up

for the drawback of Q₃, i.e., Q₃ often indicates too high a value regardless of the actual prediction accuracy for residues in transmembrane segments, because usually loop regions are much longer than transmembrane segments in a membrane protein. The apparently too good results measured by Q₃ for the two parameter sets of <H>, AP value, size and <H>, AP value, height and polarity in Table IV would become more reasonable by using, especially, Q₅ when compared with the segment-based accuracy.

The training data of transmembrane segments and loop regions used to construct discriminant functions were various in length. We also tested discriminant functions that had been constructed from the most hydrophobic 17-residue segments in the training data of transmembrane segments and loop regions, but those functions did not perform well (data not shown). This is probably because some long loop regions contained highly hydrophobic segments, which made discrimination less effective.

Incorporation of alternative models

To examine the merit of enumerating possible models, we investigated if the correct model appeared within top three probable models. In all the results using different parameter sets shown in Table V, improvements were observed by considering the second probable models, which is indicated by the difference between top1 and top2, but improvements were minor by further considering the third probable models, which is indicated by the difference between top2 and top3. In the case of using the parameter set of <H>, AP value, polarity (5 groups), protein-based accuracy was improved by 11.2% (10 proteins) and segment-based accuracy was improved by 4.8% (17 transmembrane segments) in top2, whereas the improvements were 1.2% (1 protein) and 2.2% (eight transmembrane segments), respectively, in top3. Thus, it is worth while examining, at least, the second probable models.

Now the question is how to decide the likelihood that the second or other high-scoring models are correct rather than the best-scoring one. Using the parameter set of <H>, AP value and polarity (5 groups), 55 were correctly predicted and 34 were not correctly predicted in the best-scoring model out of 89 query membrane proteins (Table II). For the 55 correctly predicted proteins, the average score per residue in the best model was 2.41 and the score difference per residue between

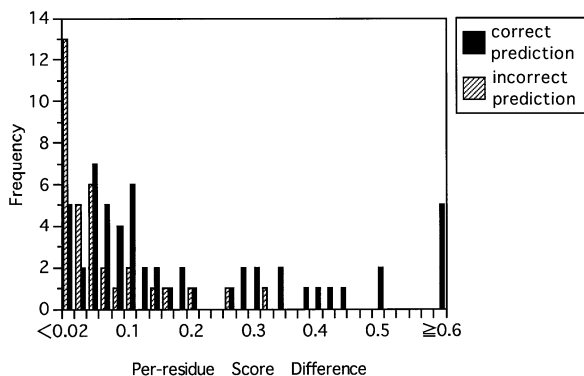


Fig. 6. The histogram showing the correctness of the best scoring models plotted against the difference of per-residue scores between the best and the second best models. The five discriminant functions were used with the parameters of average hydrophobicity, AP value and polarity. The correctness was based on the protein-based accuracy. There were 55 correct predictions and 34 incorrect predictions in total.

the best and the second best models was 0.23, while for the 34 incorrectly predicted proteins, they were 2.15 and 0.07, respectively. The average sequence lengths of the former and the latter groups were 314.3 and 455.4, respectively, which suggests that longer query sequences are more difficult to predict correctly.

Further to obtain a measure of evaluation whether to rely on the most probable model, we carried out the discriminant analysis using the score difference of the two groups. Figure 6 shows the two groups plotted against the difference of scores between the best and the second best models. In this case, we constructed a quadratic discriminant function, because the variances in the two groups were quite different, 0.084 for the group of correctly predicted proteins and 0.0064 for the group of incorrectly predicted proteins. The discriminant function obtained was:

$$72.2x^2 - 8.2x > 0 \quad (10)$$

where x denotes the score difference between the best and the second best models. If the score difference satisfies the condition above, the most probable model can be considered to be highly reliable. The TSEG program incorporates this post-processing of alternative models. In the example above, 27 out of 55 correctly predicted proteins and 6 out of 34 incorrectly predicted proteins satisfied the condition. If the score difference does not satisfy the condition, best and alternative models should be examined carefully, perhaps, by motif search and other methods.

Comparison with single discriminant function

To examine the merit of using the multiple different discriminant functions to detect transmembrane segments, we performed the same prediction procedure using the single discriminant function that had been constructed by the single combined group of all transmembrane segments and the group of loop segments. Compared with the results using the five discriminant functions, the accuracy was lower in all cases (Table V). It is remarkable that four out of six entries in the test set that have seven transmembrane segments were correctly predicted using the five discriminant functions, whereas only one entry (BAC-R_HALHA) was correct using the single discriminant function (Table VI). The five discriminant functions were successful in identifying the characteristics of unusual transmembrane segments in seven-transmembrane proteins. Although the five

discriminant functions were effective in general, there was one entry, 1OCC(J), which was not correctly predicted by the five discriminant functions, but correctly predicted by the single discriminant function. This entry is one of the subunits of cytochrome c oxidase which has one transmembrane segment, and the method with the five discriminant functions predicted this entry to be a globular protein.

Comparison with other published methods

We compared TSEG with two published prediction methods, TopPred II (Claros and von Heijne, 1994) and MEMSAT (Jones *et al.*, 1994), using the same test data set (Tables V and VI). We chose these methods because both output a list of predictions in a similar way as TSEG. However, direct comparison with these methods is somewhat difficult because of the following reasons. First, although they all output a list of predictions, the two published methods predict topology (orientation as to the membrane) whereas TSEG does not. Therefore, the list of predictions by the two published methods often contain two predictions with the same number of transmembrane segments with opposite topology (the locations of the corresponding transmembrane segments are identical in the TopPred II, but not identical in general in MEMSAT), while TSEG outputs one prediction for each model. Second, as we used the distributed programs of TopPred II and MEMSAT, strictly speaking, there is a possibility that some entries or closely related entries in our test set may have existed in the training data set of these programs.

As shown in Table VI, TSEG was superior in predicting seven transmembrane proteins. Another advantage of TSEG was that the protein-based accuracy was higher (Table V). TopPred II and MEMSAT predicted observed transmembrane segments (Segment-based Obs) more correctly than TSEG, but they tended to overpredict them reducing their protein-based accuracy. In TSEG more improvements were observed by considering up to the second or third probable models.

Prediction of globular proteins

In the genome-scale sequence analysis, it is an indispensable process to distinguish membrane proteins from others. One of the models TSEG examines is for globular proteins, and we tested the performance of discriminating this model and membrane protein models. In the test set of globular proteins, 836 out of 928 sequences (90.1%) were correctly predicted. In the test set of membrane proteins, seven out of 89 sequences (7.9%) were falsely predicted to be globular ones. Though more than 90% accuracy was achieved, further improvements will be needed to deal with genome sequences which can contain several thousand ORFs.

Discussion

The prediction of transmembrane segments has been important in the structural analysis of proteins. Predicted locations in a given sequence can be a starting point of three-dimensional structure prediction in a membrane protein (Taylor *et al.*, 1994; Suwa *et al.* 1995). In addition, the membrane protein prediction is becoming more important in the functional assignment and analysis of ORFs identified in the complete genomes. For example, Clayton *et al.* (1997) performed an inter-genome comparison of transport proteins and suggested that their composition may reflect the biosynthetic potential of the organisms and the environment in which they inhabit.

In view of many prediction methods that are already

Table V. Prediction accuracy in comparison with the other methods

Parameter set/ methods	^a Rank	Protein-based		Segment-based				Residue-based					
		Q ₂ (%)	<L>	Obs sov (%)	Prd sov (%)	Nseg over	Nseg under	Q ₃ (%)	Obs TM (%)	Prd TM (%)	Q ₄	Q ₅	Q ₇
<H>,AP,Polarity (5 groups ^b)	top1	61.8	26.7	85.1	91.5	28	53	88.1	80.5	72.7	0.855	0.618	0.687
	top2	73.0	26.1	89.9	93.8	21	36	89.5	84.7	74.6	0.878	0.657	0.726
	top3	74.2	26.1	92.1	95.3	16	28	90.0	86.3	75.5	0.887	0.674	0.741
<H>,AP,Polarity (1 group ^c)	top1	57.3	26.8	84.5	90.3	32	55	87.7	80.4	71.7	0.852	0.610	0.678
	top2	69.7	26.4	87.6	92.8	24	44	89.0	83.4	73.7	0.871	0.643	0.712
	top3	69.7	26.2	91.3	92.6	26	31	89.2	86.0	73.3	0.881	0.655	0.723
<H>,AP,Height (5 groups ^b)	top1	52.8	26.9	84.5	86.6	46	55	87.0	81.0	69.5	0.849	0.597	0.664
	top2	66.3	26.3	89.9	89.6	37	36	88.4	85.7	71.5	0.875	0.639	0.707
	top3	69.7	26.4	90.7	91.2	31	33	89.0	86.3	72.6	0.881	0.651	0.719
<H>,AP (5 groups ^b)	top1	53.9	26.6	82.5	90.2	32	62	87.8	78.3	72.6	0.845	0.605	0.673
	top2	70.8	25.6	89.0	92.9	24	39	89.4	83.5	74.9	0.874	0.652	0.721
	top3	71.9	25.6	91.3	94.7	18	31	90.1	85.3	76.0	0.884	0.672	0.740
TopPredII	top1	57.3	21.0	90.4	83.3	64	34	86.8	77.2	74.5	0.845	0.610	0.681
	top2	58.4	21.0	91.8	84.4	60	29	87.3	78.4	75.6	0.852	0.626	0.696
	top3	60.7	21.0	92.1	86.0	54	29	87.9	78.7	77.8	0.858	0.642	0.714
MEMSAT	top1	52.8	20.8	86.2	85.7	51	49	89.3	75.0	79.1	0.844	0.626	0.701
	top2	61.8	20.9	87.0	90.1	34	46	90.4	75.6	82.5	0.853	0.651	0.728
	top3	66.3	20.9	87.0	92.5	25	46	90.8	75.5	84.5	0.856	0.663	0.740

^aUp to 1, 2 or 3 predictions were considered.

^bThe multiple discriminant functions for the five groups were used.

^cThe single discriminant function for all transmembrane segments was used.

Table VI. The number of correctly predicted proteins by the three methods

TM	Number of proteins	TSEG ^a		TopPred II	MEMSAT
		5 Groups	1 Group		
1	35	26	26	25	23
2	8	7	7	6	6
3	3	3	3	3	2
4	9	5	5	5	6
5	6	4	3	3	3
6	8	3	3	4	4
7	6	4	1	1	1
8	6	0	0	0	0
9	1	0	0	0	0
10	2	0	0	0	0
11	0	—	—	—	—
12	4	3	3	4	2
13	0	—	—	—	—
14	1	0	0	0	0

^aOnly the most probable models were counted.

available, our objective of developing one more method is toward better functional identification of membrane proteins. We have proposed a method based on a new idea of classifying membrane proteins into different classes. The prediction is implemented by comparison against a library of membrane protein models as well as a globular protein model, and a list of compatible models is output with their scores. We note here that though the current version of TSEG holds 15 models in total, the methodology can be extended to consider multiple families, hence multiple models, for a group of membrane proteins defined by the number of transmembrane segments.

Not surprisingly, TSEG was relatively good at protein-based accuracy, namely, at distinguishing different membrane protein models, which is a unique feature among a number of existing prediction methods. Another feature of TSEG is its capability to provide alternative structures. As discussed by Jones *et al.* (1994), taking account of alternative structures is useful

considering that any current structure prediction method has limited accuracy. In the case of membrane proteins, this feature is more important because there is only a small number of three-dimensional structures available and, furthermore, experimental evidence for determination of topology is not necessarily very strong.

We discuss here some possible improvements of our method. As it became clear in the comparison with the other methods, TSEG tends to underpredict rather than overpredict transmembrane segments. To optimize the balance of over- and underpredictions is difficult, but it appears that the balance in TSEG is not well optimized yet. To reduce underpredictions is not simply to change the threshold values for detection of transmembrane segments, but to re-examine the discriminant functions themselves, such as the parameters and the training data set used to construct the discriminant functions.

Another weak point of TSEG is that it cannot predict membrane proteins that have more than 14 transmembrane segments. Because these membrane proteins are relatively scarce in the current database, e.g., 0.77% of the membrane protein sequences in SWISS-PROT *rel.* 34.0, we could not define models for them. A simple way to include them is to use the general (single) discriminant function which was constructed to distinguish all transmembrane segments from loop regions. Alternatively, since larger membrane proteins tend to be composed of smaller domains, models for domains can be repeatedly applied to detect larger proteins.

To predict β -type membrane proteins is of special interest. It can be implemented by adding a model for β -type membrane proteins, when more sequence data become available. In the case of typical β strands, the AP value has a distinctive peak around 180 degrees. Prediction of the topology of membrane proteins can also be implemented in TSEG by preparing two models which have the same number of transmembrane segments but have opposite topology and by applying, for example, the positive-inside rule. In addition, utilization of multiple aligned sequences will improve the prediction accu-

racy as discussed by many authors. A possible way is to use the average indices of aligned residues.

The accuracy of discriminating membrane proteins from globular proteins should also be improved. One possible practical way of improvement is to use a better filter for the discrimination before the prediction of transmembrane segments by TSEG. We reported that a discriminant function which had been constructed to discriminate the most hydrophobic 21 residue segment in the training sets of globular and membrane proteins performed better, namely, 93.5% of membrane proteins and 95.6% of globular proteins were correctly discriminated (Kihara and Kanehisa, 1997). Detecting hydrophobic signal sequences is also needed in distinguishing from single spanning membrane proteins. This can be implemented perhaps by using a pattern matching method as in PSORT (Nakai and Kanehisa, 1992).

The membrane protein prediction is an essential process for functional assignment of ORFs identified in the complete genomes, especially when no sequence similarity is found. It is expected to limit the structural possibilities and to give clues to further functional analysis. We have started working on the application of TSEG, together with further improvements, in the genome-scale sequence analysis.

Acknowledgements

We thank Dr Kenta Nakai for his help with collecting the test data set. This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Science', from the Ministry of Education, Science, Sports and Culture in Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

- Bairoch, A. and Apweiler, R. (1997) *Nucleic Acids Res.*, **25**, 31–36.
- Bono, H., Ogata, H., Goto, S. and Kanehisa, M. (1998) *Genome Res.*, **8**, 203–210.
- Claros, M.G. and von Heijne, G. (1994) *Comp. Appl. Biosci.*, **10**, 685–686.
- Clayton, R.A., White, O., Ketchum, A. and Venter, J.C. (1997) *Science*, **387**, 459–462.
- Degli Esposti, M., Ghelli, A., Luchetti, R., Crimi, M. and Lenaz, G. (1989) *Ital. J. Biochem.*, **38**, 1–22.
- Deisenhofer, J., Epp, O., Miki, K., Huber, R. and Michel, H. (1985) *Nature*, **318**, 618–624.
- Donnelly, D., Overington, J.P., Ruffe, S.V., Nugent, J.H.A. and Blundell, T.L. (1993) *Protein Sci.*, **2**, 55–70.
- Eisenberg, D., Schwartz, E., Komaromy, M. and Wall, R. (1984) *J. Mol. Biol.*, **179**, 125–142.
- Görne-Tschelnokow, U., Strecker, A., Kaduk, C., Naumann, D. and Hucho, F. (1994) *EMBO J.*, **13**, 338–341.
- Grantham, R. (1974) *Science*, **185**, 862–864.
- Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E. and Downing, K.H. (1990) *J. Mol. Biol.*, **213**, 899–929.
- Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) *Protein Sci.*, **1**, 409–417.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1994) *Biochemistry*, **33**, 3038–3049.
- Kendall, M. and Stuart, A. (1976) *The Advanced Theory of Statistics*, Vol. 3. Hafner Press, New York.
- Kihara, D. and Kanehisa, M. (1997) *Genome Informatics 1997*. Universal Academy Press, Tokyo, pp. 300–301.
- Klein, P., Kanehisa, M. and DeLisi, C. (1985) *Biochim. Biophys. Acta*, **815**, 468–476.
- Kühlbrandt, W., Wang, D.N. and Fujiyoshi, Y. (1994) *Nature*, **367**, 614–621.
- Kyte, J. and Doolittle, R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
- Langs, D.A. (1988) *Science*, **241**, 188–191.
- Lücke, H., Chang, B.T., Maillard, W.S., Schlaepfer, D.D. and Haigler, H.T. (1995) *Nature*, **378**, 512–515.
- Matthews, B.W. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.
- Mitaku, S., Suzuki, K., Odashima, S., Ikuta, K., Suwa, M., Kukita, F., Ishikawa, M. and Itoh, H. (1995) *Proteins Struct. Funct. Genet.*, **22**, 350–362.
- Nakai, K. and Kanehisa, M. (1992) *Genomics*, **14**, 897–911.
- Parker, M.W., Postma, J.P., Pattus, F., Tucker, A.D. and Tsernoglou, D. (1992) *J. Mol. Biol.*, **224**, 639–657.
- Parker, M.W., Buckley, J.T., Postma, J.P., Tucker, A.D., Leonard, K., Pattus, F. and Tsernoglou, D. (1994) *Nature*, **367**, 292–295.
- Persson, B. and Argos, P. (1994) *J. Mol. Biol.*, **237**, 182–192.
- Picot, D., Loll, P.J. and Garavito, R.M. (1994) *Nature*, **367**, 243–249.
- Rao, J.K.M. and Argos, P. (1986) *Biochim. Biophys. Acta*, **869**, 197–214.
- Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995) *Protein Sci.*, **4**, 521–533.
- Rost, B., Fariselli, P. and Casadio, R. (1996) *Protein Sci.*, **5**, 1704–1718.
- Schulz, G.E. and Shirmer, R.H. (1979) *Principles of Protein Structure*. Springer-Verlag.
- Suwa, M., Hirokawa, T. and Mitaku, S. (1995) *Proteins Struct. Funct. Genet.*, **22**, 363–377.
- Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) *Science*, **278**, 631–637.
- Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. and Yoshikawa, S. (1996) *Science*, **272**, 1136–1144.
- Taylor, W.R., Jones, D.T. and Green, N.M. (1994) *Proteins Struct. Funct. Genet.*, **18**, 281–294.
- von Heijne, G. (1992) *J. Mol. Biol.*, **225**, 487–494.
- Weiss, M.S. and Schulz, G.E. (1992) *J. Mol. Biol.*, **227**, 493–509.

Received April 4, 1998; revised June 10, 1998; accepted July 7, 1998