

# Prediction of Membrane Proteins in Post-Genomic Era

Daisuke Kihara<sup>1\*</sup> & Minoru Kanehisa<sup>2</sup>

1) Donald Danforth Plant Science Center  
893 North Warson Road  
St. Louis, Missouri 63141, USA  
kihara@danforthcenter.org  
Fax: 314-812-8075

2) Institute for Chemical Research, Kyoto University  
Gokasho, Uji, Kyoto, 611-0011, Japan  
kanehisa@kuicr.kyoto-u.ac.jp  
Fax: 81-774-38-3269

\*) To whom correspondence should be addressed

Running title: Membrane Proteins in Genome

Recent Res. Developments in Protein Engineering. 1: 179-196 (2001)

## ABSTRACT

The current status of the prediction methods of membrane proteins in the post-genomic era is described. In the first section, prediction methods of transmembrane segments in proteins are considered, including the TSEG program, which we have recently developed. Several prediction methods for the tertiary structure of membrane proteins are also mentioned. Finally, we show the unique feature of the distribution of membrane proteins in a genome. Our prediction of the function of several membrane proteins is also shown.

## INTRODUCTION

The beginning of the scientific research of this century is characterized by the report that the human genome sequencing project has been completed [1,2]. Beyond the human genome, a surge of world-wide genome projects in the last decade have been producing complete genome sequences of an increasing number of organisms. At this point (February 2001), complete genome sequences of 48 organisms are available in the KEGG database [3], which is a compilation of genome sequences and pathways we have been maintaining on the web. Complete genome sequences enable a comprehensive study of proteins or organisms through a catalog of proteomes, e.g., dynamic genome rearrangement of a pair of closely related organisms [4,5], the spread and evolution of a particular protein family among organisms [6,7], and physical principles of how related genes are encoded in the genome [8,9]. In this post-genomic era, every computational structure/function prediction method should take into account its application to genome sequences.

Throughout this manuscript, transmembrane (TM) protein is meant by membrane

protein. Membrane proteins have important roles in living cells, such as transport, energy production, cell signaling and cell adhesion. In the genomic context, membrane proteins have attracted more attention, because they discussed that in microbial genomes, the distribution of transporters, one of the major members of the membrane protein family, reflects the environment that each organism inhabits [10,11].

We begin this text by reviewing the prediction methods of TM segments in proteins, which are also used in detecting TM proteins in genome sequences. The prediction of TM segments is also a starting point of the tertiary structure prediction of a membrane protein. Here, we describe the TSEG program, which we have developed recently. The proper way to measure the performance of prediction methods is also mentioned. Next, methods for predicting the tertiary structure of TM proteins are reviewed. Currently, tertiary structures of only a few TM proteins are solved by experimental methods, and most of them have helical TM segments. Limiting the objects of prediction to those proteins with helical TM segments (exceptions at this point are porin, which has  $\beta$ -barrel structure [12] and bacterial toxin proteins which penetrate into membranes and form membrane pores [13-16]), this prediction procedure can be simplified to the assembly of semi-rigid helical rods. In a later section, we show a genome sequence analysis in terms of membrane proteins. The membrane protein content of a genome, and their physical distribution in a genome is the main issue here. We have also made predictions for the functions of membrane proteins based on the observations.

## 1. Prediction Methods of Transmembrane Segments

The TM segment prediction method originates from the hydrophathy plot by Kyte & Doolittle (1982) [17]. The idea is rather simple: recognizing highly hydrophobic stretches in the sequence as TM segments using a sliding window of a certain length (7-21 amino acids). Another important contribution of this paper is the derivation of the hydrophobicity index of amino acids, which is still commonly used. It was derived from water-vapor transfer free energies [18-20] and interior-exterior distribution of amino acids [21]. Various improvements have been performed on the hydrophathy analysis since then: taking amphiphilicity of TM helices into account [22, 23], using different or various hydrophobicity indices [24,25], employing discriminant analysis [26]. Hydrophathy-type methods can be generalized as follows: given a hydrophobicity index  $f_t$  for each amino acid type  $t$ , weights  $h_n$  for the position  $n$  in the sliding window and a threshold  $B$  which is constant:

$$D^{(j,m)} = \sum_n M^{(j,m-n)} h_n - B, 1 \leq m \leq L_j \quad (1)$$

where  $M^{(j,m-n)} = f_{p(j,m-n)}$ , and  $p^{(j,m)}$  denotes the residue type at position  $m$  in chain  $j$ .  $L_j$  is the length of the chain  $j$ . The position  $m$  in chain  $j$  is decided to be transmembrane if  $D^{(j,m)} > 0$ . Treating  $f_t$  and  $h_n$  as variables, Edelman [27] derived their optimal values by minimizing the following quadratic equation  $S$  (which corresponds to minimizing prediction errors):

$$S = \sum_{j,m} (D^{(j,m)} - q^{(j,m)})^2 \quad (2)$$

Here,  $q^{(j,m)} = 1$  when  $(j,m)$  in the training set is transmembrane, otherwise  $q^{(j,m)} = 0$ .

It has been observed that not only the inside of TM segments but also their flanking regions have preferable amino acids, e.g. aromatic residues [28,29]. One can utilize the

amino acid propensity to construct prediction methods [30-32]. Using the Neural network is another promising approach to these kinds of two-dimensional structure predictions [33-35]. Using multiple sequence alignments helps improving prediction accuracy [30,31,34,35]. Kroth *et al.* applied a hidden Markov model [36,37].

The topology of membrane proteins can be derived from some biochemical experimental evidence, even if their tertiary structures are not solved [38-40]. But it is often the case that contradictory results are suggested by other experiments. If one also considers that any prediction method has limited accuracy, it is optimal for a prediction method to output not a single prediction but a list of possibilities with certainty measures, so that further experiments can be designed to distinguish among several topology models. Jones *et al.* [32] elegantly applied a dynamic programming algorithm for this purpose. TopPred II by von Heijne [41] ranks several topology models according to the ‘positive-inside rule’ (see below, in Topology Prediction section). We also mention here that neural network-based and hidden Markov model-based methods can assign reliability index to the predictions made for each residue.

## TSEG program

In this section, we describe TSEG, which we have developed recently. When annotating biological functions of genes in a genome, the prediction of higher order structures can be used in order to compensate for the limitation of the conventional sequence similarity search [42-44]. This is especially true for membrane proteins, since the number of TM segments in a protein can be related to a functional subclass in some cases, such as seven-TM receptors or six-TM transporters. One of the objectives of developing the Transmembrane *SEG*ment prediction program [45] was to enhance the

functional identification of TM proteins. To capture detailed properties of TM segments, our method is based on a classification of TM segments in a database. In fact, not all TM segments are equally hydrophobic. For example, TM segments of single spanning TM proteins are known to be highly hydrophobic and have less amphiphilicity [22], whereas the last TM segments in seven-TM proteins are relatively less hydrophobic and often difficult to detect by prediction methods [30,32]. Thus, we have classified TM segments first by the total number of TM segments in a protein and the order in which they appear in the protein sequence, and at last merged similar ones into the same group. The second feature is that the TSEG enumerates possible models as ranked by their scores, where a model is distinguished by the number of TM segments in a protein and represented by the order of different groups (types) of TM segments. A model of globular proteins is included as well.

We have classified TM segments into five groups according to their average hydrophobicity and amphiphilicity (or AP value), using the Mahalanobis distance of the linear discriminant analysis [46]. The dataset of 2876 non-redundant TM protein sequences used here was extracted from Swiss-Prot *rel.* 34.0 [47]. The AP value is defined as follows:

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

where

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

$H_i$  is the hydrophobicity index [17] given to each amino acid,  $N$  is the length of the TM

segment, and  $\omega$  is the angle in degrees. Figure 1 is the model of the membrane protein used in the prediction, according to the locations of the five groups of TM segments. Here, membrane proteins with more than fourteen TM segments are excluded because there were not enough sequences in the database to execute statistical calculations. The most hydrophobic group, group1, appears only in the single spanning membrane proteins. Group2, which has relatively high hydrophobicity, appears at the N-terminal TM segments in most of the membrane protein classes (except for the 12TM proteins). It is possible that these TM segments are involved in the initiation of membrane insertion, which may correspond to what Eisenberg *et al.* [22] called ‘initiators’. The last segment of the seven TM proteins belongs to the least hydrophobic group 5, which also appears in the eighth segment of the nine TM proteins.

The prediction procedure is based on the detection of different TM segment groups using different discriminant functions, followed by matching with the 15 models shown in Figure 1. In the first stage, a query sequence is applied to each model and the best candidates for TM segments in the model are selected. This selection is done by a 17-residue-long sliding window, employing a discriminant function designed to distinguish the TM segments of the particular group from loop regions. Then, the models are compared by their scores. The score of a model is the sum of scores of the discriminant function given to each amino acid.

## Evaluation of Prediction Accuracy

The evaluation of the performance was made on 89 membrane proteins that were collected from the literatures which reported experimental evidence of TM topology (Table 1, 2). Homologous sequences are excluded from training sets used in the construction of

the discriminant functions. Results by two distributed programs, TopPred II [41] and MEMSAT [32] are also shown to clarify characteristics of performance of TSEG. Note that direct comparison between these distributed programs is difficult since there is a possibility that they used homologous sequences in the test set in training.  $Q_3$ , used in the table, is the overall percentage of correctly predicted residues in the sequence:

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

The symbol  $N_{pq}$  ( $p, q = T$  (transmembrane) or  $L$  (loop)) denotes the number of residues that are observed to be  $p$  and predicted as  $q$ .

A drawback of  $Q_3$  is that it does not take the difference of the amount of TM segments and loop region in a protein into account. For example, suppose one extreme prediction engine, which always predicts that there are no TM segments in a query protein (i.e. all loop region). Since there are fewer residues in TM segments than in non-TM segments, this engine could still get a  $Q_3$  value which gives an impression that the performance is not bad. The following measure  $Q_4$ ,  $Q_5$  and  $Q_7$  are to compensate for this difference.

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

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

$Q_7$  is Matthews' correlation coefficient [48]:

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

A tendency of TSEG is that it makes fewer overpredictions than underpredictions. The

other two methods score better in detecting TM segments (segment-based accuracy, Obs), but tend to overpredict them to reduce their protein-based accuracy, which is important as a clue of biological function. TSEG gains more improvements by considering the top 3 probable models than the others, which makes it reasonable to consider alternative models. As shown in Table 2, TSEG is superior in recognizing seven TM proteins. We have also tested the performance to distinguish globular proteins and membrane proteins. TSEG recognized 836 out of 928 (90.1%) globular protein sequences correctly, while seven out of 89 membrane protein sequences (7.9%) are falsely recognized as globular ones.

TSEG is a model-recognition approach to TM proteins. Models of TM proteins with more than 14 TM segments,  $\beta$ -type membrane proteins, membrane proteins with different topology can be added rather easily.

## Topology Prediction

It has been observed that basic residues are more abundant in loop regions in cytosol than in those in periplasma [49,50] (positive-inside rule), or more in N-terminal flanking regions of the first signal-anchor segment than in C-terminal flanking regions [51]. This fact can be readily implemented in prediction algorithms by counting positively charged residues [35,41] or more generally, by considering the amino acid composition [31,32]. It can also be built into the architecture of a hidden Markov model [36,37].

## 2. Prediction of Tertiary Structure

The large gap between the number of available sequences and the number of solved tertiary structures of membrane proteins makes attempts to predict structures from sequence information widespread. Since membrane

proteins are embedded in the two-dimensional membrane, the advantage of predicting tertiary structures of membrane proteins is that the procedure could be simplified to fit the arrangement of perpendicular TM helices followed by some modification. Baldwin made a model of rhodopsin using structural information for the arrangement of TM helices, which comes from careful investigation on the multiple sequence alignment of the family [52,53]. One of the basic ideas used is that the variation of amino acids would be found more frequently on the lipid-facing surface of helices rather than on the packing core sides. The location of polar residues is also taken into account. From the initial arrangement, the helices were inclined according to experimental evidence. Although this prediction procedure is based on a projection map by electron crystallography so that not all the possibilities of arrangements of TM helices are enumerated, the strategy shown by the author is reasonable and could be a foundation for subsequent studies. Donnelly proposed to use the AP value (3) for the arrangement of TM helices [54]. Suwa *et al.* used the polar interaction and the loop length between helices to allocate TM helices of bacteriorhodopsin on its projection map, in a fashion which is ready to be automated [55]. They also made an interesting attempt along this line by classifying of TM proteins by the pattern of polar energy surface calculated from their sequence [56]. The method by Du & Alkorta [57,58] is also an automated procedure, which uses the moment of the variability of hydrophobic/philic amino acids. Besides, the restriction used was the sequential packing of helices, not a projection map.

All methods discussed above use projection maps or other restrictions in the arrangement of TM helices. But ideally, it would be better to take all possible conformations into account without relying on

any other extra information, taking advantage of the limitation of conformations of TM proteins. One way to approach this is to use a hexagonal lattice for the arrangements of helices [59,60]. A drawback of the mentioned approaches would be in term of resolution, although the necessary resolution depends on the purpose. For example, they might not be able to differentiate structures of bacteriorhodopsin and rhodopsin, whose projection map shows some difference. Recently, Bowie presented an algorithm to construct a structure template set of more detailed conformations of TM proteins, which could be used by the threading approach [61]. Ideas of using threading for TM proteins have been presented previously by others [62,63]. Yet another approach is in the line of homology modeling [64]. Hu *et al.* predicted the structure of light-harvesting complex II, which is a complex of hetero dimers of helices. The TM regions are determined by prediction methods, and each helix is formed by homology modeling, followed by the energy minimization by molecular dynamics.

Although several interesting approaches have been presented, this field still awaits more tertiary structures of membrane proteins to come, not only for taking structure parameters from them but also for verifying the performance of the approaches.

### 3. Genome Analysis of Membrane Proteins

Previous studies of comparative genome analysis of membrane proteins discussed mostly the estimation of the number of membrane proteins in genome sequences [36,81-85]. Here, we investigated the distribution (location) of TM proteins in 16 complete genomes [65-80], and made function predictions of several TM proteins based on this observation together with structural information by TSEG [86].

## Detecting Membrane Proteins

One should be cautious in the detection of membrane proteins: One reason is due to the fact that some proteins have highly hydrophobic signal peptides at the amino-terminal region which will be cleaved off after translocation. Since the signal peptides could be easily misidentified as TM segments by predictive methods, these have to be properly removed beforehand. Another problem is that TM segments tend to bring about spurious hits in homology searches, because their amino acid composition is biased [87]. Therefore, the query membrane protein sequence should be preprocessed to mask out low-complexity regions by the SEG program [88]. We used both discriminant analysis and homology searches after the preprocess written above to collect candidates of TM proteins, followed by manual inspection to remove false positives.

## Number of Membrane Proteins

The proportion of membrane proteins in each organism ranges from 18% to 29% (Table 3). This estimate is smaller than those based only on a TM prediction method, which reports the values at around 35%. The distribution of membrane proteins grouped by the number of TM segments is shown in Figure 2. The representative organisms are selected from each organism category. In all organisms, single spanning membrane proteins dominate 30-50% of the entire membrane proteins, and those with more than twelve TM segments are rare. There are roughly two peaks in the distribution especially in bacterial genomes, one around 4-6 TM, and another one around 10-12 TM, which may be dominated by transporting proteins.

## Tandem Clusters of Membrane Proteins

Figure 3 shows the location of membrane proteins in *M. pneumoniae* genome. Surprisingly, in all 16 organisms, 43.9-60.1% of the membrane proteins were found to be located next to each other, namely, in tandem clusters, which is statistically significant. In most of the cases (70% in terms of membrane proteins in tandem clusters, except for *S. cerevisiae*, which is 59.1%) a tandem cluster is formed by the genes on the same strand, and furthermore, (again 70% in terms of membrane proteins in tandem clusters, except for *S. cerevisiae*, which is 22.1%) the gap between the genes is usually less than 300 bp, which implies that they are operons [89].

About 10-30% of the tandem clusters are conserved between organisms and/or within an organism. However, the conservation rate of *S. cerevisiae* is 2.2% and all of the conserved clusters are only within the organism, again showing its peculiarity. This is consistent with the fact that it doesn't have bacteria-like operons [90]. The majority (97.3%) of the conserved tandem clusters fall into one of the three functional categories: Membrane transporters, electron transport system, and cell motility. Membrane transporters include ABC (ATP-binding cassette) transporters, the PTS (phosphotransferase system), protein export membrane proteins, multidrug resistant proteins, glucose/hexose transporters of *S. cerevisiae*, and cobalt transport permease. The cluster size of this category is small, 2.6 on average. Compared with this, the category of the electron transport system has a larger cluster size, 5 on average. Another feature of this category is that typically one of the membrane proteins in a cluster contains more than ten TM segments. NADH dehydrogenase and proteins of cytochrome family belong to this category. The number of the conserved clusters in the third category, cell motility, is small (12 clusters in total). It includes proteins of flagellar motor and chemotaxis proteins.

If only clusters that have duplicated genes (i.e. a significantly similar pair of constituent genes, 150 in the Smith-Waterman score) are counted, the relationship among gene function and the cluster size and the number of TM segments in the genes becomes much clearer (Figure 4). More than 10% of the tandem clusters in each organism have duplicated genes. Using the difference of this distribution of TM proteins of transporters and the ones with the other functions, it is possible to make a chart which shows the likelihood of being a cluster of transporters rather than belonging to another category. Based on this empirical rule, we predict seven clusters of transporters out of 55 clusters of unknown function with gene duplication (Table 4).

## ABC Transporters

The ABC transporter is an active transport system of the cell, which utilizes the energy of ATP hydrolyzation [91,92]. They form the largest superfamily of paralogous proteins in bacterial and archaeal genomes [93]. Actually, in our analysis, ABC transporters dominate 42.3% of the entire conserved clusters and also contributes to making the characteristic distribution of transporters in Figure 4. A typical bacterial-type ABC transporter consists of two TM proteins, two ATP-binding subunits, and one periplasmic substrate-binding subunit. We searched components of ABC transporters by sequence similarity, starting with the reference set collected in our previous work [94] (Table 5). The proportion of the components is 2-7% in the genome, except for *S. cerevisiae*. Note here that the search used bacterial-type ABC transporters so that many eukaryotic ABC transporters [92] in *S. cerevisiae* were not detected. The ABC transporter is known to have well-conserved operon structure, but it turned out that there are some isolated components in each genome.

Table 5 indicates ATP-binding components are more likely to be isolated than membrane components. However, this may be simply due to the fact that membrane protein components are less conserved [94] so that they have not been detected by sequence similarity searches. To find partners of these isolated components, we examined conserved pairs of ATP-binding proteins (Table 3, the right-most column) and membrane proteins (Table 6). Those predicted to be ABC transporters in Table 6 could form new types of ABC transporters.

Recently, we have extended our study to eukaryotic ABC transporters in genome sequences of *C. elegans* and *D. melanogaster* (fruit fly) [95]. We have classified those using the hidden Markov Model.

## The Omnibus Hypothesis

We found that about half of the membrane protein genes in the genomes form tandem clusters, and only 10-30% of these were conserved among organisms. There are several possible explanations for this observation. First, the functional coupling of genes will be the most dominant biological constraint on such clustering in the genome [8,9,89]. The horizontal transfer may also result in gene clusters [96]. Second, clusters of functionally coupled non-membrane proteins may help form tandem clusters of membrane proteins as background clustering. However, since majority of non-membrane proteins do not form gene clusters as evidenced by the extensive shuffling of orthologous genes [97], we believe that this event alone cannot explain the statistically significant occurrence of tandem clusters of membrane proteins.

Third, it is possible that forming tandem clusters is favorable for the cellular mechanism of membrane protein expression. If this is true, it would be advantageous at the stage of protein translocation to the cell membrane. There are



two types of bacterial protein translocation machinery, *Sec* machinery dependent and signal recognition particle (SRP) dependent. In *Sec* machinery, SecB binds nascent proteins and delivers them to the SecY/E/G translocon, using energy from ATP hydrolysis by SecA and the proton motive force [98,99]. The two pathways seem to use a common translocon [100]. We speculate an implication of the tandem clusters for the SecB machinery. Considering the report that SecB forms a tetramer and can bind more than one polypeptide chain [101], and also the fact that bacterial mRNA is usually polycistronic, it may be favorable for the membrane protein genes to be positioned in tandem, so that SecB can deliver them all together like an omnibus. For *S. cerevisiae* this hypothesis could not be applied in the same way, because there is no SecB-like protein [102], though the translocon complex is similar to bacteria [103]. We name this story the *ominibus hypothesis* and await the experimental analysis of genome-scale translocation mechanisms to verify it.

## Concluding Remarks

In this manuscript, prediction methods of TM segments in proteins, including the TSEG program, and those of tertiary structures of membrane proteins are reviewed. In the latter section, the application of TSEG to genome sequences is described. A statistically significant number of membrane proteins form tandem clusters in a genome. We predicted seven clusters of transporters, according to an empirical rule concerning the cluster size and the number of TM segments. This was an attempt to use the information of structural features and that of the location of the coding region in the genome in functional annotation. In addition, we identified probable membrane protein partners of isolated ATP-binding protein components of ABC transporters by

searching for conserved pairs of a membrane protein and an ATP-binding protein. This approach can be further generalized to include other types of coupling, such as identifying sequence motifs that are known to be present on two interacting proteins. Generally speaking, a prediction method of TM segments is relatively accurate compared to other types of structure prediction methods, so it could compensate for the limitation of the sequence similarity search. Complete genome sequences bring us opportunities of comprehensive understanding of life through classified tables of genes. To classify transporters, Saier Jr. *et al.* proposed a transport commission (TC) number, which is a classification number of membrane transporting proteins [10]. In the next few years, our knowledge of membrane proteins will widen and become more organized through the analysis of genome sequences of higher multicellular organisms, which have membrane proteins with a large variety of functions, such as cell signaling and cell adhesion.

## ACKNOWLEDGEMENTS

This work was supported in part by a Grant-in-Aid for Scientific Research on the Priority Area 'Genome Science' from the Ministry of Education, Science, Sports, and Culture of Japan. We thank Katherine White and Dr. Adrian K. Arakaki for proofreading this manuscript.

## REFERENCES

- [1] International Human Genome Sequence Consortium. 2001. *Nature* 409: 860-921.
- [2] Venter, J.C., Adams, M.D., Myers, E.W. *et al.* 2001. *Science* 291: 1304-1351.
- [3] Kanehisa, M., and Goto, S. 2000. *Nucleic Acid Res.* 28: 27-30.

- <http://www.genome.ad.jp/kegg/>
- [4] Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., and Herrmann, R. 1997. *Nucleic Acid Res.* 25: 701-712.
- [5] de Rosa, R., and Labedan, B. 1998. *Mol. Biol. Evol.* 1998. 15: 17-27.
- [6] Aravind, L., Dixit, V.M., and Koonin, E.V. 2001. *Science* 291:1279-1284.
- [7] Leipe, D.D., Aravind, L., Grishin, N.V., and Koonin, E.V. 2000. *Genome Res.* 10: 5-16.
- [8] Tamates, J., Casari, G., Ouzounis, C., and Valencia, A. 1997. *J. Mol. Evol.* 44: 66-73.
- [9] Dandekar, T., Snel, B., Huynen, M., and Bork, P. 1998. *Trends Biochem. Sci.* 23: 324-328.
- [10] Paulsen, I.T., Sliwinski, M.K. and Saier Jr., H. 1998. *J. Mol. Biol.* 277: 573-592.
- [11] Clayton, R.A., White, O., Ketchum, A., and Venter, J.C. 1997. *Science* 387: 459-462.
- [12] Weiss, M.S., and Schulz, G.E. 1992. *J.Mol. Biol.* 227: 493-509.
- [13] Langs, D.A. *Science* 241: 188-191.
- [14] Luecke, H., Chang, B.T., Mailliard, W.S. et al. 1995. *Nature* 378: 512-515.
- [15] Parker, M.W., Postma, J.M., Pattus, F. et al. 1992. *J. Mol. Biol.* 224: 639-657.
- [16] Parker, M.W., Buckley, J.T., Postma, J.P. et al. 1994. *Nature* 367: 292-295.
- [17] Kyte, J., and Doolittle, R.F. 1982. *J.Mol.Biol.* 157: 105-132.
- [18] Hine, J., and Mookerjee, P.K. 1975. *J. Org. Chem.* 40: 292-298.
- [19] Wolfenden, R.V., Cullis, P.M., and Southgate, C.C.F. 1979. *Science* 206: 575-577.
- [20] Nozaki, Y., and Tanford, C. 1971. *J. Biol. Chem.* 246: 211-2217.
- [21] Chothia, C. 1976. *J. Mol. Biol.* 105: 1-14.
- [22] Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. 1984. *J. Mol. Biol.* 179: 125-142.
- [23] Cornette, J.L., Cease, K.B., Margalit, H. et al. 1987. *J. Mol. Biol.* 195: 659-685.
- [24] Rao, J.K.M., and Argos, P. 1986. *Biochim. Biophys. Acta* 869: 197-214.
- [25] Degli Esposti, M., Crimi M., and Venturoli, G. 1990. *Eur. J. Biochem.* 190: 207-219.
- [26] Klein, P., Kanehisa, M., and DeLisi, C. 1985. *Biochim. Biophys. Acta* 815: 468-476.
- [27] Edelman, J. 1993. *J. Mol. Biol.* 232: 165-191.
- [28] Landolt-Marticorena, C., Williams, K.A., Deber, C.M., and Reithmeier, A.F. 1993. *J. M. Biol.* 229: 602-608.
- [29] Nakashima, H., and Nishikawa, K. 1992. *FEBS Lett.* 303: 141-146.
- [30] Persson, B., and Argos, P. 1994. *J. Mol. Biol.* 237: 182-192.
- [31] Persson, B., and Argos, P. 1996. *J. Mol. Biol.* 5: 363-371.
- [32] Jones, D.T., Taylor, W.R., and Thornton, J. 1994. *Biochemistry* 33:3038-3049.
- [33] Dombi, G.W., and Lawrence, J. 1994. *Prot. Sci.* 3: 557-566.
- [34] Rost, B., Casadio, R., Fariselli, P., and Sander, C. 1995. *Prot. Sci.* 4: 521-533.
- [35] Rost, B., Fariselli, P., and Casadio, R. 1996. *Prot. Sci.* 5: 1704-1718.
- [36] Kroth, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L.L. 2001. *J. Mol. Biol.* 305: 567-580.
- [37] Tsunady, G.E., and Simon, I. 1998. *J. Mol. Biol.* 283: 489-506.
- [38] Manoil, C. 1991. *Methods Cell Biol.* 34: 61-75.
- [39] Cafiso, D.S. 1994. *Membrane Protein Structure*, White S.H. (ed.), Oxford Univ. Press, Newyork, 127-143.
- [40] Boyd, D. 1994. *Membrane Protein Structure*, White S.H. (ed.), Oxford Univ. Press, New York, 144-163.
- [41] von Heijne, G. 1992. *J. Mol. Biol.* 225: 487-494.
- [42] Pearson, W.R., and Lipman, D. 1988. *Proc. Natl. Acad. Sci. USA* 85: 2444-2448.

- [43] Altschul, S.F., Madden, T.L., Schaffer, A.A. *et al.* 1997. *Nucleic Acid Res.* 25: 3389-3402.
- [44] Fetrow, J.S., and Skolnick, J. 1998. *J. Mol. Biol.* 281: 949-968.
- [45] Kihara, D., Shimizu, T., and Kanehisa, M. 1998. *Protein Eng.* 11: 961-970.  
<http://www.genome.ad.jp/SIT/tseg.html>
- [46] Kendall, M., and Stuart, A. 1976. *The Advanced Theory of Statistics*, Vol. 3. Hafner Press, New York.
- [47] Bairoch, A., and Apweiler, R. 1997. *Nucleic Acid Res.* 25: 31-36.
- [48] Matthews, B.W. 1975. *Biochim. Biophys. Acta* 405: 442-451.
- [49] von Heijne, G. 1986. *EMBO J.* 5:3021-3027.
- [50] von Heijne, G., and Gavel, Y. *Eur. J. Biochem.* 174: 671-678.
- [51] Hartmann, E., Raport, T.A., and Lodish, H.F. 1989. *Proc. Natl. Acad. Sci. USA* 86: 5786-5790.
- [52] Baldwin, J.M. 1993. *EMBO J.* 12: 1693-1703.
- [53] Baldwin, J.M., Schertler, G.F.X., and Unger, V.M. 1997. *J. Mol. Biol.* 272: 144-164.
- [54] Donnelly, D., Overington, J.P., Ruffle, S.V. *et al.* 1993. *Prot. Sci.* 2: 55-70.
- [55] Suwa, M., Hirokawa, T., and Mitaku, S. 1995. *Proteins* 22: 363-377.
- [56] Suwa, M., Yudate, H.T., Masuho, Y., and Mitaku, S. 2000. *Proteins* 41: 504-517.
- [57] Du, P., and Alkorta, I. 1994. *Prot. Eng.* 7: 1221-1229.
- [58] Alkorta, I., and Du, P. 1994. *Prot. Eng.* 7: 1231-1238.
- [59] Taylor, W.R., Jones, D.T., and Green, N.M. 1994. *Proteins* 18: 281-294.
- [60] Hirokawa, T., Uechi, J., Sasamoto, H., Suwa, M., and Mitaku S. 2000. *Protein Eng.* 13: 771-778.
- [61] Bowie, J.U. 1999. *Prot. Sci.* 8: 2711-2719.
- [62] Efremov, R.G., and Vergoten, G. 1996. *Prot. Eng.* 9: 253-263.
- [63] Cronet, P., Sander, C., and Vriend, G. 1993. *Protein Eng.* 6: 59-64.
- [64] Hu, X., Xu, D., Hamer, K. *et al.* 1995. *Prot. Sci.* 4: 1670-1682.
- [65] Bult, C.J., White, O., Olsen, G.J. *et al.* 1996. *Science* 273: 1058-1073.
- [66] Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., *et al.* 1997. *J. Bacteriol.* 179: 7135-7155.
- [67] Klenk, H.P., Clayton, R.A., Tomb, J.-F., *et al.* 1997. *Nature* 390: 364-370.
- [68] Kawarabayasi, Y., Sawada, M., Horikawa, H. *et al.* 1998. *DNA Res.* 5: 55-76.
- [69] Blattner, F.R., Plunkett III, G., Bloch, C.A. *et al.* 1997. *Science* 277: 1453-1462.
- [70] Fleischmann, R.D., Adams, M.D., White, O. *et al.* 1995. *Science* 269: 496-512.
- [71] Tomb, J.-F., White, O., Kerlavage, A.R. *et al.* 1997. *Nature* 388: 539-547.
- [72] Kunst, F., Ogasawara, N., Moszer, I. *et al.* 1997. *Nature* 390: 249-256.
- [73] Fraser, C.M., Gocayne, J.D., White, O. *et al.* 1995. *Science* 270: 397-403.
- [74] Fraser, C.M., Casjens, S., Huang, W.M. *et al.* 1997. *Nature* 390: 580-586.
- [75] Fraser, C.M., Norris, S.J., Weinstock, G.M. *et al.* 1998. *Science* 281: 375-388.
- [76] Himmelreich, R., Hilbert, H., Plagens, H. *et al.* 1996. *Nucleic Acid Res.* 24: 4420-4449.
- [77] Cole, S.T., Brosch, R., Parkhill, J. *et al.* 1998. *Nature* 393: 537-544.
- [78] Deckert, G., Warren, P.V., Gaasterland, T. *et al.* 1998. *Nature* 392: 353-358.
- [79] Kaneko, T., Sato, H., Kotani, A. *et al.* 1996. *DNA Res.* 3: 109-136.
- [80] Goffeau, A., Aert, R., Agostini-Carbone, M.L. *et al.* 1997. *Nature* 387 (6632 Suppl.)
- [81] Arkin, I.T., Bruenger, A.T., and Engelman, D.M. 1997. *Proteins* 28: 465-466.
- [82] Boyd, D., Schierle, C., and Beckwith, J. 1998. *Prot. Sci.* 7: 201-205.

- [83] Jones, D.T. 1998. *FEBS Lett.* 423: 281-285.
- [84] Wallin, E., and von Heijne, G. 1998. *Prot. Sci.* 7: 1029-1038.
- [85] Kihara, D., and Kanehisa, M. 1997. *Genome Informatics 1997*. Universal Academy Press, Tokyo, Japan. 300-301.
- [86] Kihara, D., and Kanehisa, M. 2000. *Genome Res.* 10: 731-743.
- [87] Bork, P., and Koonin, E.V. 1998. *Nature Genet.* 18: 313-318.
- [88] Wooton, J.C., and Federhen, S. 1993. *Computers and Chem.* 17: 149-163.
- [89] Overbeek, R.M., Fonstein, M., D'Souza, G.D. *et al.* 1999. *Proc. Natl. Acad. Sci. USA* 96: 2896-2901.
- [90] Zhang, X., and Smith, T.F. 1998. *Microbial Comp. Gen.* 3: 133-140.
- [91] Higgins, C.F. 1992. *Annu. Rev. Cell Biol.* 8: 67-113.
- [92] Fath, M.J., and Kolter, R. 1993. *Microbial Rev.* 57: 995-1017.
- [93] Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. *Science* 278: 631-637.
- [94] Tomii, K. & Kanehisa, M. 1998. *Genome Res.* 8:1048-1059.
- [95] Igarashi, Y., Kihara, D., and Kanehisa, M. 2000. *Genome Informatics 2000*. Universal Academy Press, Tokyo, Japan. 274-275.
- [96] Xu, Y., Murray, B.E., and Weinstock, G.M. 1998. *Infect Immun.* 66: 4313-4323.
- [97] Watanabe, H., Mori, H., Itoh, T., and Gojobori, T. 1997. *J. Mol. Evol.* 44: 57-64.
- [98] Tokuda, H. 1994. *FEBS Lett.* 346: 65-68.
- [99] Randall, L.L., and Hardy, S.J.S 1995. *TIBS* 20: 65-69.
- [100] Valent, Q.A., Scotti, P.A., High, J.W. *et al.* 1998. *EMBO J.* 17: 2504-2512.
- [101] Randall, L.L., and Hardy, S.J.S., Topping, T.B. *et al.* 1998. *Prot. Sci.* 7: 2384-2390.
- [102] Lyman, S.K., and Schekman, R. 1996. *Experientia* 52: 1042-1049.
- [103] Jungnickel, B., Rapoport, T.A., and Hartmann, E. 1994. *FEBS Lett.* 346: 73-77.

**Table 1.** Prediction Accuracy

Method	Rank <sup>a</sup>	Protein -based <sup>b</sup> (%)	Segment-based <sup>c</sup>				Residue-based			
			Obs (%)	Prd (%)	Over	Under	Q <sub>3</sub> (%)	Q <sub>4</sub>	Q <sub>5</sub>	Q <sub>7</sub>
TSEG	top1	61.8	85.1	91.5	28	53	88.1	0.855	0.618	0.687
	top3	74.2	92.1	95.3	16	28	90.0	0.887	0.674	0.741
TopPredII	top1	57.3	90.4	83.3	64	34	86.8	0.845	0.610	0.681
	top3	60.7	92.1	86.0	54	29	87.9	0.858	0.642	0.714
MEMSAT	top1	52.8	86.2	85.7	51	49	89.3	0.844	0.626	0.701
	top3	66.3	87.0	92.5	25	46	90.8	0.856	0.663	0.740

<sup>a</sup> Up to 1, 3 predictions were considered.

<sup>b</sup> Counted if all the TM segments are predicted in their correct position (more than 11 residues overlap).

<sup>c</sup> Obs/Prd, observed/predicted segment overlaps, Over/Under, number of false positive/negative segments.

**Table 2.** The Number of Correctly Predicted Proteins <sup>a</sup>

TM	Number of Proteins	TSEG	TopPred II	MEMSAT
1	35	26	25	23
2	8	7	6	6
3	3	3	3	2
4	9	5	5	6
5	6	4	3	3
6	8	3	4	4
7	6	4	1	1
8	6	0	0	0
9	1	0	0	0
10	2	0	0	0
11	0	-	-	-
12	4	3	4	2
13	0	-	-	-
14	1	0	0	0

<sup>a</sup> Top1 predictions are counted.

**Table 3.** The Numbers of Predicted TM proteins and ATP-binding Proteins

Category	Organism	Total Genes	Membrane Proteins (%)	ATP-binding Proteins <sup>a</sup> (%)
Archaea	<i>M.jannaschii</i> (Mj)	1735	326 (18.8)	130 (7.5)
	<i>M.thermoautotrophicum</i> (Mt)	1871	395 (21.1)	127 (6.8)
	<i>A.fulgidus</i> (Af)	2407	499 (20.7)	158 (6.6)
	<i>P.horikoshii</i> (Ph)	1829	433 (23.7)	148 (8.1)
Bacteria	<i>E.coli</i> (Ec)	4289	1142 (26.6)	293 (6.8)
	<i>H.influenzae</i> (Hi)	1717	378 (22.0)	153 (8.9)
	<i>H.pylori</i> (Hp)	1566	334 (21.3)	118 (7.5)
	<i>B.subtilis</i> (Bs)	4100	1125 (27.4)	277 (6.8)
	<i>M.genitalium</i> (Mg)	467	90 (19.3)	61 (13.1)
	<i>M.pneumoniae</i> (Mp)	677	123 (18.2)	73 (10.8)
	<i>M.tuberculosis</i> (Mtu)	3918	949 (24.2)	213 (5.4)
	<i>B.burgdorferi</i> (Bb)	1256	297 (23.6)	85 (6.8)
	<i>T.pallidum</i> (Tp)	1031	228 (22.1)	95 (9.2)
	<i>Synechocystis sp.</i> (Ss)	3166	918 (29.0)	218 (6.9)
	<i>A. aeolicus</i> (Aa)	1522	304 (20.0)	130 (8.5)
Eukarya	<i>S.cerevisiae</i> (Sc)	6215	1652 (26.6)	414 (6.7)

<sup>a</sup> ATP-binding proteins are detected by P-loop ATP/GTP binding motif in PROSITE.

**Table 4.** Predicted Transporters According to the Number of TM segments and the Cluster Size

Organism	Duplicated Genes <sup>a</sup>	TM	Additional Protein	TM	Cluster Size
Mj	MJ0419 (+)	10			2
	MJ0420 (+)	9			
Ec	b0786 (+)	7	b0788 (-)	7	3
	b0787 (+)	7			
Bs	YybM (+)	5	YybL (+)	5	3
	YybK (+)	5			
Mg	MG225 (+)	10			2
	MG226 (+)	11			
Bb	BB0050 (+)	5			2
	BB0051 (+)	5			
Bb	BB0807 (+)	5	BB0806 (+)	1	3
	BB0808 (+)	6			
Tp	TP0883 (-)	6			2
	TP0884 (-)	5			

<sup>a</sup> The strand direction is shown in parenthesis.



**Table 5.** The number of Bacterial-type ABC Transporter Components

Organism	Total Number of ABC Transporter Components (% <sup>a</sup> )	Isolated Components <sup>b</sup>		
		Membrane Components	ATP-binding Components	Substrate-binding Components
Mj	39 (2.2)	1	5	-
Mt	36 (1.9)	-	6	-
Af	97 (4.0)	1	7	-
Ph	85 (4.6)	1	6	3
Ec	242 (5.6)	-	6	17
Hi	105 (6.1)	2	5	3
Hp	48 (3.1)	1	6	3
Bs	207 (5.0)	1	13	9
Mg	34 (7.3)	-	2	-
Mp	35 (5.2)	-	3	-
Mt	126 (3.2)	1	8	6
Bb	44 (3.5)	-	3	2
Tp	50 (4.8)	-	7	3
Ss	139 (4.4)	8	17	16
Aa	35 (2.3)	6	5	4
Sc	38 (0.61)	-	7	7

<sup>a</sup> The proportion to the total number of genes.

<sup>b</sup> Considered to be isolated if there is no other components within five gene positions on both sides.

**Table 6.** Predicted Membrane Protein Components of ABC Transporters

Cluster ID	Organism	ATP-binding Component <sup>a</sup>	Membrane Component	TM
1	Af	AF0004 (+)*	AF0008 (+)	10
	Hi	HI1252 (+)*	HI1242 (-)*	10
	Bs	ExpZ (-)*	YdgK (+)	11
	Bs	YfmM (-), YfmR (+)	YfmO (+)	11
2	Mg	MG065 (+)*	MG064 (+)	9
	Mg	MG468.1 (-), MG467 (-)	MG464 (-)	6
	Mp	R02_orf465 (-)	R02_orf1386V (-)	9
	Mp	K05_orf284 (+), K05_orf339 (+)	K05_orf385 (+), K05_orf1882 (+)	5,7
3	Bs	EcsA (+)*	EcsB (+)	8
	Bs	YthP (-)*	YthQ (-)	8
	Ss	s110489 (-)*	slr0096 (+)	11
4	Af	AF1136 (+), AF1139 (+)	AF1140 (+)	9
	Af	AF0393 (-)*	AF0392 (-)	12
	Ph	PH1230 (+)	PH1231 (+)	11
5	Tp	TP0881 (-)*	TP0880 (-)	6
6	Af	AF1170 (+)*	AF1169 (+)	4
	Ph	PH0157 (+)	PH0159 (-)	5

<sup>a</sup> Isolated components in Table 5 is shown with asterisks.

## Figure Legends:

### Figure 1

The models of the membrane proteins, which have one (1TM) to fourteen (14TM) TM segments, as well as for globular proteins (0TM). The N-terminus of the sequence is on the left. The larger black region corresponds to the higher average hydrophobicity.

### Figure 2

The distribution of membrane proteins grouped by the number of TM segments. *A.fulgidus* (Af) from archaea, *E. coli* (Ec) from bacteria, *S.cerevisiae* (Sc) from eukarya.

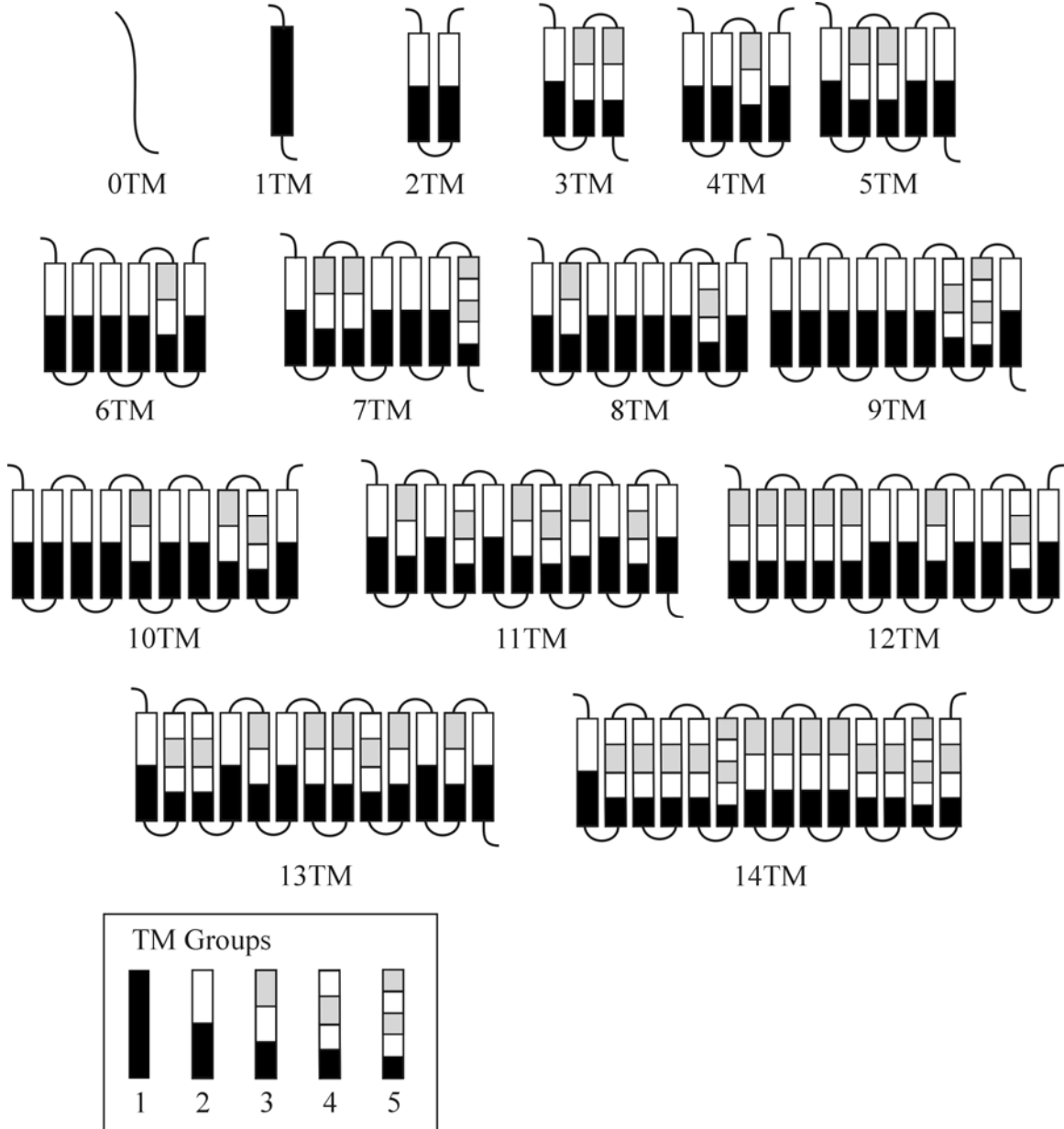
### Figure 3

Membrane proteins in the *M. pneumoniae* genome. Membrane proteins are indicated in solid boxes. The gray scale corresponds to the number of predicted TM segments: From left, 1-2TM, 3-4, 5-6, 7-8, 9-10, 11-12, 13-14, 15TM and more.

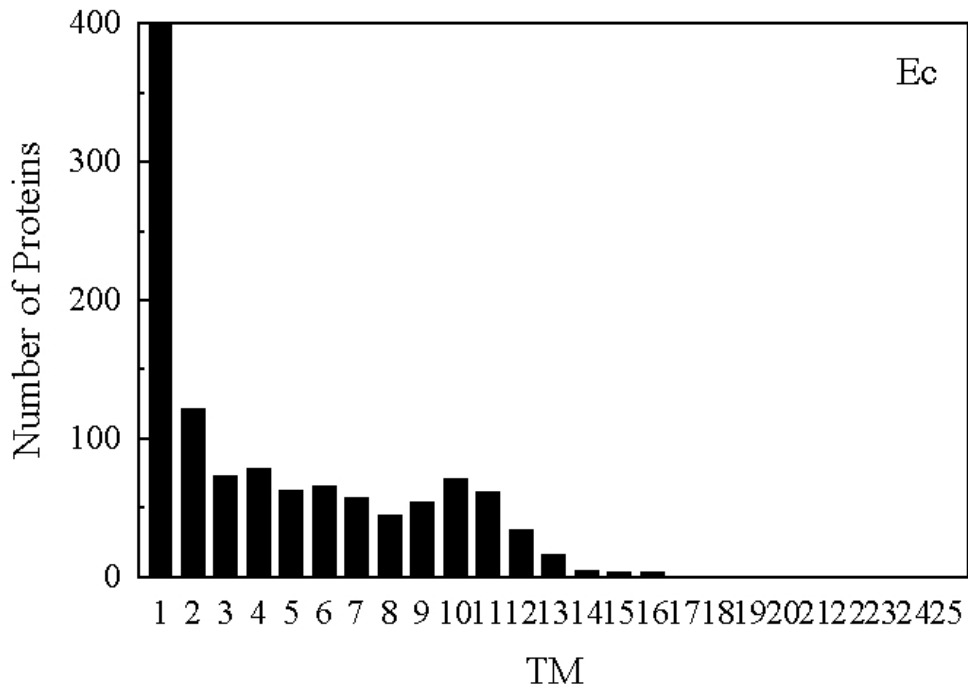
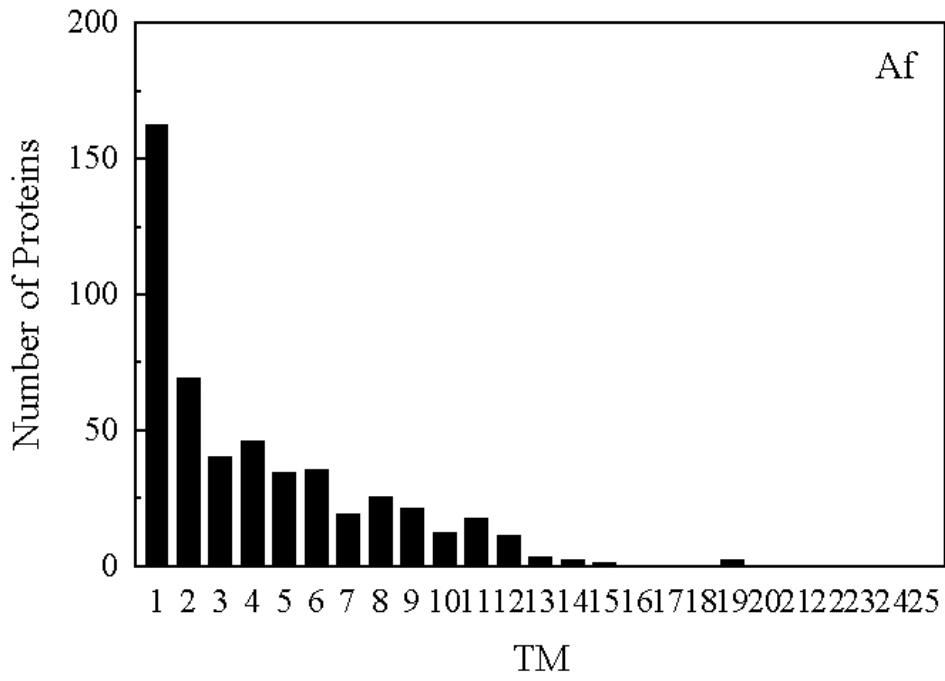
### Figure 4

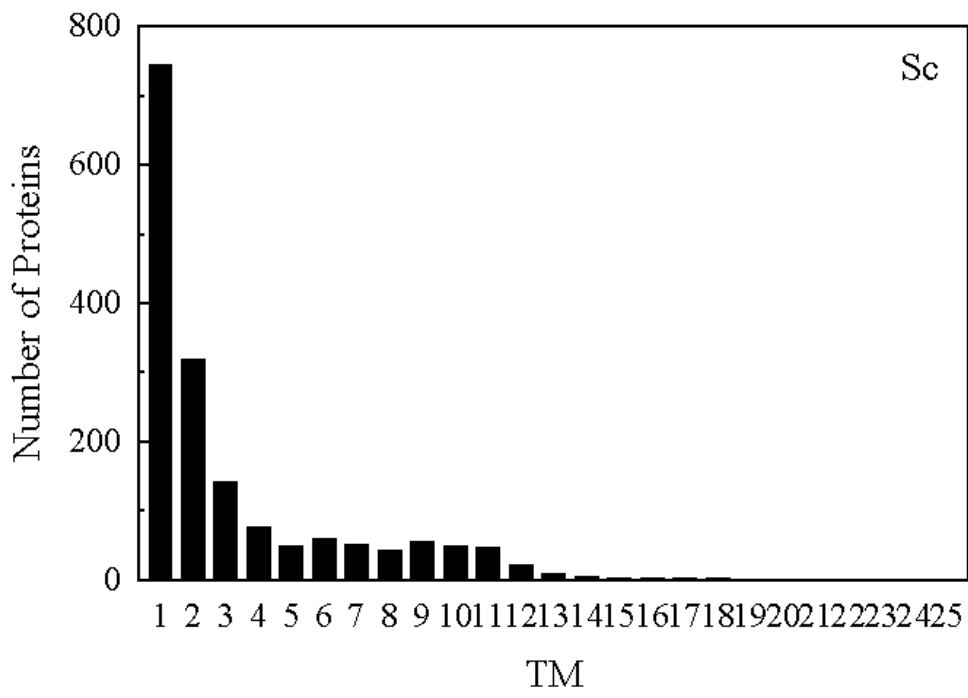
The frequency of membrane proteins that are paralogous within a tandem cluster, where each membrane proteins are classified according to the number of TM segments and the size of the cluster. (A), transporters; (B) membrane proteins with the other functions.

**Figure 1**



**Figure 2**





**Figure 3**

*Mycoplasma pneumoniae* M129 complete genome.

Page 1

