Chem-Bio Informatics Journal, Vol. 4, No. 3, pp.110-120(2004)

Features of transmembrane helices useful for membrane protein prediction

Toshiyuki Tsuji*, and Shigeki Mitaku

Nagoya University, School of Engineering, Department of Applied Physics
Nagoya, Chikusa-ku, Furocho, 464-8606, Japan

**E-mail:tsuji@bp.nuap.nagoya-u.ac.jp*

(Received September 27, 2004; accepted October 28, 2004; published online November 10, 2004)

Abstract

For the development of a high-performance software system of membrane protein prediction, we analyzed the distribution of hydrophobicity and the amphiphilicity around transmembrane helices, using dataset of membrane proteins whose 3D structures or topologies are known. The moving average of 7 residues showed that there are a hydrophobicity peak in the center of a transmembrane helix and two amphiphilicity peaks at both ends of the hydrophobicity peak. The shapes of the peaks were asymmetric with respect to the center of the hydrophobicity peak. The membrane protein prediction system SOSUI (Hirokawa and Mitaku, Bioinformatics, 1998) was improved on the basis of the asymmetric profiles of hydrophobicity and amphiphilicity, resulting in the accuracy of 98% for positive dataset and 96% for negative one of prokaryotes, and the comparable accuracy was obtained also for eukaryotes.

Key Words: membrane protein, transmembrane helix, prediction, bioinformatics

Area of Interest: Bioinformatics and Bio computing

1. Introduction

The theoretical prediction of the cellular localization of proteins from amino acid sequences alone will be very helpful for screening orphan sequences because experiments of functional genomics are time consuming. The penetration of a protein into membrane is one of the most basic cellular localization, which is recognized by the translocation machinery in a cell. Because of the functional importance of membrane proteins, several groups have investigated the common features of membrane proteins as well as transmembrane regions [1][2][3][4][5][6][7] and developed software systems for the prediction [8][9][10][11][12][13][14][15]. The systems which are available on the internet can be categorized into two types: informatics methods and physicochemical approaches to the discrimination problem.

The former methods include the score matrix method, the neural network and the hidden Markov model. The score matrix and the neural network methods mainly use local properties of amino acid sequences [9][10][11], whereas the hidden Markov model can take the information of non-local sequences into account [12][13][14][15]. Anyhow, the main purpose of the informatics methods is to adjust parameters so that the highest accuracy is attained indifferent of the physical mechanisms.

In the physicochemical approaches, in contrast, physically well-defined parameters which give the best discrimination between soluble and membrane proteins are searched, taking care of the physical soundness of the combination. When a really good combination of such parameters is obtained, they lead not only to the information about the physical mechanism but also to a high performance software system of the prediction. The membrane protein prediction system SOSUI by Hirokawa and Mitaku [8] is one of those methods. Three parameters were combined in SOSUI to discriminate a membrane protein from other types of proteins: the high average hydropathy index of at least one segment, the high amphiphilicity index at the end regions of the hydrophobic segment, and the size of the protein [1][2][3][4]. This system attained high accuracy of the discrimination between membrane and soluble proteins. Analyses of amino acid sequences from total genomes have shown that approximately a quarter of ORFs code for membrane proteins [16][17].

However, preliminary analysis of organella membrane proteins indicated that they had intermediate properties between cytoplasmic membrane proteins and soluble ones [18]. Thus, three problems should be solved for the complete prediction of membrane proteins. The first problem is to improve the accuracy of the discrimination between membrane and soluble proteins, neglecting the dataset of proteins targeted to other organella than endoplasmic reticulum which is a pool of cytoplasmic membrane proteins. The second problem is the discrimination of organella membrane proteins by searching signal sequences that are recognized by the targeting machinery. The third is to predict all transmembrane helices including hydrophilic ones, determining the transmembrane regions together with the membrane topologies of helices.

In this work, we analyzed the detailed distribution of hydrophobicity and amphiphilicity indices of amino acids around transmembrane regions to help solve the first problem, which is improving the accuracy of the discrimination between membrane and soluble proteins. Then, we implemented the characteristic profiles of hydrophobicity and amphiphilicity indices into the algorithm of membrane protein prediction, making a new version of the SOSUI system.

2. Datasets

We used five sets of amino acid sequence data. The first set included the most reliable data: that of membrane proteins of known 3D structure, whose atomic coordinates are recorded in the Protein Data Bank [19][20][21][22][23][24][25][26][27][28][29][30][31][32][33][34][35]. Table 1 shows the names and PDB codes of membrane proteins, and the numbers of chains and helices they contain. The total number of membrane proteins in this category was 36. The topology of membrane proteins in these data is known. Redundancy of data was removed with the cutoff of 30% homology. The distribution of physical properties was analyzed by using transmembrane helices in this data set.

The second data set included 148 membrane proteins which were reported by Möller et al. [36]. Transmembrane regions and their topologies are experimentally confirmed for the proteins in this data set. Some of data, 22 proteins with 63 helices, were also included in the first data set and, therefore, omitted from this data set. Other data, 6 proteins with 22 helices, did not describe the information about membrane protein topology and omitted again from the data set. Redundancy

Table 1. Dataset of transmembrane proteins whose 3D-structure is known.

protein	ID	# of chains	# of helices
F1F0 ATPsynthase (E. coli)	1A91	1	2
Cyto. bc1 complex (Bovine)	1BGY(Chain C,D,G,J,K)	5	8,1,1,1,1
Bacteriorhodopsin (H. salinarium)	1AT9	1	7
Ca ATPase, SR (rabbit)	1EUL	1	10
Cyto. C oxidase (bovine)	1OCC (I,II,III,IV,VIa, VIc,VIIa,VIIb,VIIc,VIII)	10	12,2,7,1,1,1,1,1,1,1
Cyto. C oxidase (Thermus thermophilus)	1EHK (Chain I,II)	2	13,1
Fumarate Reductase (W. succinogenes 1)	1QLA (Chain C)	1	5
Fumarate Reductase (E. coli)	1FUM (15 kD anchor, 13kD anchor)	2	3,3
Glycophorin A (human)	1MSR	1	1
Halorhodopsin (H. salinarium)	1E12	1	7
KcsA potassium channel (S. lividans)	1BL8	1	2
Light Harvesting Complex (R. acidophila)	1KZU (Chain A,B)	2	1,1
Light Harvesting Complex (R. molischianum)	1LGH (Chain A)	1	1
MscL ion channel (M. tuberculosis)	1MSL	1	2
Reaction center (R. viridis)	1PRC (Chain M,L,H)	3	5,5,1
Rhodopsin (bovine)	1F88	1	7
Aquaporin (human)	1FQY	1	6
Glycerol Channel (E. coli)	1FX8	1	6
Total		36	129

of data was removed from the merged data set of 3D set and the Möller's set with the cutoff of 30% homology. 3 data (3 helices) in Möller set were eliminated. Finally, Möller's set included 117 proteins with 618 helices data (33 eukaryote proteins with 167 transmembrane helices and 84 prokaryote proteins with 451 transmembrane helices). The number of plasma membrane proteins with the information of topology consequently was 33 for eukaryotes and 101 for prokaryotes, which were used.

The third data set contained 85 membrane proteins with 267 transmembrane helices in intracellular organelles: mitochondria, nucleus, peroxisome and chloroplast. Combining these data with membrane proteins in organelle in the first and second dataset, the total number of organelle membrane proteins was 104. Membrane proteins in endoplasmic reticulum were not included in this data set because they are mainly targeted to the cytoplasmic membrane. Data was collected from Swissprot rel.40 by three conditions: 1) Keywords in "SUBCELLULAR LOCATION" of CC line are "MITOCHONDRIAL", "MITOCHONDRION", "NUCLEAR", "PEROXISOME" or "CHLOROPLAST". 2) A keyword "INTEGRAL MEMBRANE PROTEIN" is found in KW line and "TRANSMEM" in FT line. 3) There are not any keywords like "PROBABLE", "HYPOTHETICAL", "PUTATIVE", "THEORETICAL", "FRAGMENT" in "SUBCELLULAR LOCATION" of CC line. After the collection, the redundancy of data was removed with the cutoff of 30% homology.

The fourth data set is soluble protein set. This data set is very important as the negative control data against membrane proteins. This set is composed of two soluble protein data sets. One is obtained from PDBselect [37]. In order to obtain the information about the signal sequences at amino terminus, we checked the data by Swissprot database and obtained 350 soluble proteins (eukaryote protein: 152 and prokaryote one: 198). These data contained 1070 helices longer than 19 residues.

The fifth set also used as negative control data included 156 soluble proteins which directly extracted from Swissprot rel.40 by the conditions: 1) A keyword, "MITOCHONDRIAL" or "MITOCHONDRION", is found in "SUBCELLULAR LOCATION" of CC line. 2) There are no keywords, "INTEGRAL MEMBRANE PROTEIN" in KW line and "TRANSMEM" in FT line. 3) Any of keywords, "PROBABLE", "HYPOTHETICAL", "PUTATIVE", "THEORETICAL" or "FRAGMENT", is found in "SUBCELLULAR LOCATION" of CC line. This data set includes only soluble proteins in mitochondria. After extraction, redundancy of data was removed with the cutoff of 30% homology.

The forth and fifth set that included 198 prokaryote proteins and 308 eukaryote ones were used as negative control data for evaluation of prediction accuracy.

These five datasets can be found at http://bp.nuap.nagoya-u.ac.jp/~tsuji/paper/01/index.html.

3. Analysis of Transmembrane Helices

A previous work indicated that a transmembrane helix has a peak of hydrophobicity corresponding to the central region of membrane and two clusters of amphiphilic residues which probably stabilize the membrane-water interface [1][9][38]. However, this information could not exactly be implemented into the membrane protein prediction system SOSUI. For improving the system by using the landscape of the physical properties around transmembrane segments, we analyzed in more detail the hydrophobicity and the amphiphilicity of around transmembrane segments dividing them into seven regions: three regions within a helix and two regions in N- and C-terminal loops next to the helical segment (Figure1). The center region is defined 50% length of each transmembrane helix, end regions are defined 25% length each terminal of transmembrane helix. Loop1, 2 regions are five amino acid residues long. When a loop segment was shorter than 5 residues, all the amino acids in the loop segment were assumed to belong to Loop1 regions.

We found that not only the helical regions but also loop regions of about 10 residues next to the helix region had characteristic profiles of the hydrophobicity [4] and the amphiphilicity indices [1]. A peak of the hydrophobicity index was slightly asymmetric with the external end being more hydrophobic than the cytoplasmic end. We defined two kinds of amphiphilicity indices of amino acids, A- and A'- indices, which are different in the strength of polarity [4]. The amino acids of nonzero A-index were His, Lys, Arg, Glu, Gln and those for A'-index were Trp and Tyr. The profiles of A- and A'-indices had a common feature of twin peaks sandwiching the hydrophobicity peak. However, the profiles of A- and A'-indices were different in several points: (1) A-index was very low at the central region of transmembrane helices, whereas A'-index was considerably high at the same region. (2) A-index was much higher at the cytoplasmic side than the external side, while A'-index showed inversed profile. (3) A-index was higher at the loop region than the inside of the helix. In contrast, A'-index was larger at the end region of helix than the loop.

These features of three profiles in Figure 1 seem to be physically sound and suggest the mechanism of the stability of typical transmembrane helices. First, the high hydrophobicity of a sufficiently long segment is the basic requirement for the energy minimization of a transmembrane helix in the membrane. Second, amphiphilic residues, namely A- and A'-index residues, prefer the interface between the nonpolar and the aqueous environments, and the twin peaks of A- and A'-indices contribute to the location of the terminals of transmembrane helices at the membrane-water interface. Third, the asymmetric profile of the A-index seems reasonable, when we assume that the large peak of A-index at the membrane-water interface of the cytoplasmic side is a stop signal of the protein translocation. This role of A-index amino acids is physically

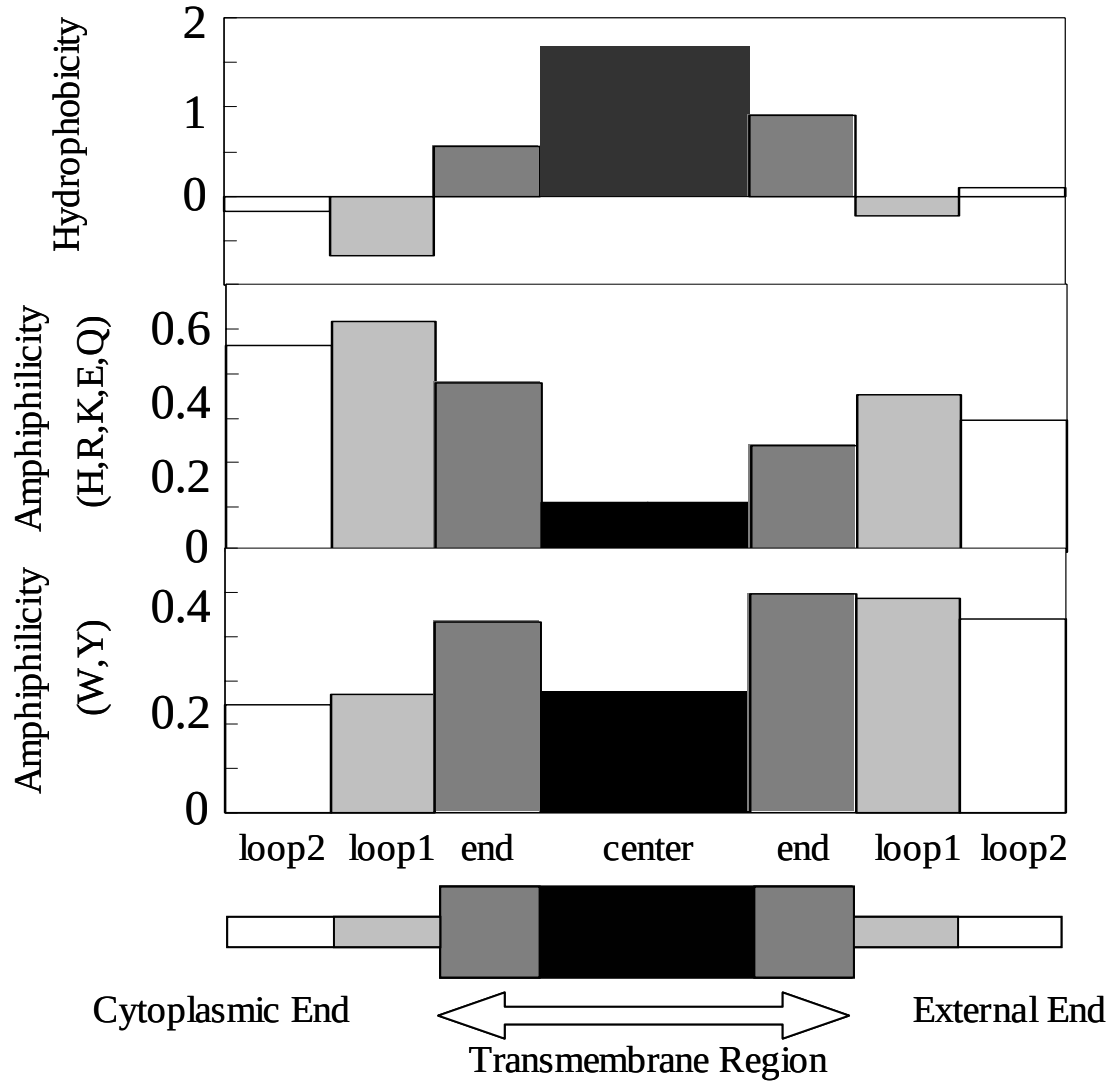

Figure 1. Distribution of hydrophobicity and amphiphilicity indices around transmembrane helices. Average values of H-, A- and A'-indices in each regions, a center, two ends, two loop1 and loop 2 regions are plotted as histograms.

reasonable because nonpolar hydrocarbon region of membrane with low dielectric constant forms the high energy barrier for strongly polar groups of A-index residues. Fourth, A'-index residues are not so polar and can easily be translocated to the other side of membrane. Therefore, residues of A'-index probably work as a stabilizing factor of transmembrane helices mainly at the external side.

Table 2. Weight factors for discriminating transmembrane segments.

	Cytoplasmic End					External End	
	loop2	loop1	end	center	end	loop1	loop2
H	-0.10	-0.40	0.33	1.00	0.54	-0.13	0.06
A	4.15	4.62	3.40	1.00	2.12	3.14	2.64
A'	0.90	0.97	1.58	1.00	1.81	1.77	1.60

Table 2 shows the average values of three parameters, H-, A-, A'-indices, for seven regions defined in Figure 1. In this table, all values are normalized by those at the central region of helices. These profiles are the standard distributions of the hydrophobicity and the amphiphilicity around transmembrane helices obtained by averaging many membrane proteins, but the distributions for individual transmembrane helices should fluctuate around the standard distributions. When the set of the H-, A- and A'-indices for seven regions are regarded as vectors, $\vec{H}$, $\vec{A}$ and $\vec{A}'$, respectively, the products, $\langle H \rangle$, $\langle A \rangle$ and $\langle A' \rangle$, of the vectors for an individual helix and the standard vectors can be used as good parameters, which indicate how an individual helix is similar to the standard one.

$$\langle H \rangle = \vec{H} \bullet \vec{H_S} \quad (1)$$

$$\langle A \rangle = \vec{A} \bullet \vec{A_S} \quad (2)$$

$$\langle A' \rangle = \vec{A}' \bullet \vec{A_S}' \quad (3)$$

Here, $\vec{H_s}$, $\vec{A_s}$ and $\vec{A'_s}$ represent the standard vectors whose elements are given by the numbers in Table 2.

Figure 2 shows the histograms of three parameters, $\langle H \rangle$, $\langle A \rangle$ and $\langle A' \rangle$, for the most hydrophobic helices in every proteins, comparing the datasets of soluble and membrane proteins. The difference between soluble and membrane proteins is significant for the histograms of $\langle H \rangle$, but all the pairs of $\langle H \rangle$, $\langle A \rangle$ and $\langle A' \rangle$ showed wide overlapping regions irrespective of prokaryotes and eukaryotes. Therefore, a single parameter is not effective for the discrimination of membrane proteins from soluble ones.

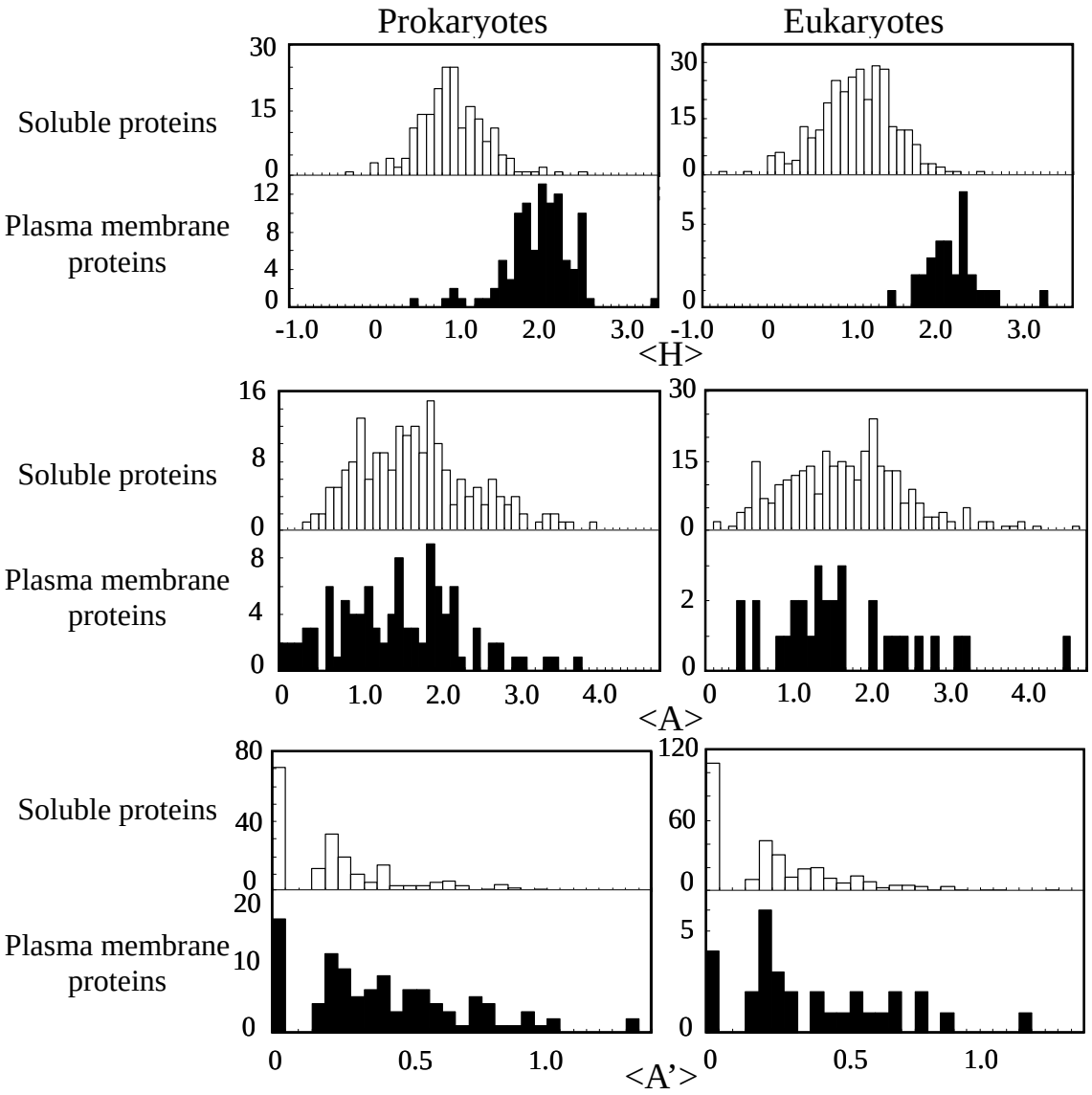

Figure 2. Histograms of three parameters, $\langle H \rangle$, $\langle A \rangle$ and $\langle A' \rangle$, for the most hydrophobic helices of soluble and membrane proteins in prokaryotes and eukaryotes.

4. Discrimination of membrane proteins

Principal component analysis by four parameters, $\langle H \rangle$, $\langle A \rangle$ and $\langle A' \rangle$ and protein size L, was performed for discriminating two kinds of proteins: soluble and membrane proteins. We omitted the data set of organelle membrane proteins in the present analysis, since the additional analysis of targeting signals is necessary for the discrimination of organelle membrane proteins. Principal component analysis involves a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called principal components. The uncorrelated variables are linear combinations of the original variables.

Table 3. Results of Principal component analysis for prokaryotes and eukaryotes

(1) Prokaryotes

Component	H	A	A’	L	Contribution
First	2.936	0.0499	2.481	-0.0018	99.99%
Second	0.505	1.137	-0.392	0.0019	0.01%

(2) Eukaryotes

Component	H	A	A’	L	Contribution
First	3.217	0.144	1.699	-0.0003	99.98%
Second	-0.5841	-0.1641	4.0118	0.0007	0.02%

Table 3 shows the first and the second components for prokaryotes and eukaryotes. The contribution of the first component is so large, 99.99 %, for both prokaryotes and eukaryotes that membrane proteins can be discriminated by a single component.

According to the principal component analysis, the discrimination lines for prokaryotes and eukaryotes were determined as Eqs. (5) and (6), respectively.

$$f_{pro}(\boldsymbol{H}, \boldsymbol{A}, \boldsymbol{A}', \boldsymbol{L}) = 7.49\boldsymbol{H} + 0.028\boldsymbol{A} + 6.17\boldsymbol{A}' - 0.005\boldsymbol{L} - 11.1 \qquad (5)$$

$$f_{euk}(\boldsymbol{H}, \boldsymbol{A}, \boldsymbol{A}', \boldsymbol{L}) = 6.55\boldsymbol{H} + 0.548\boldsymbol{A} + 4.52\boldsymbol{A}' - 0.001\boldsymbol{L} - 11.1 \qquad (6)$$

in which f_{pro} is the discrimination score for prokaryotes and f_{euk} is that for eukaryotes. Figure 3 shows the histograms of the discrimination score for the most hydrophobic segment in every soluble and membrane proteins. Two kinds of proteins are very well discriminated by this parameter.

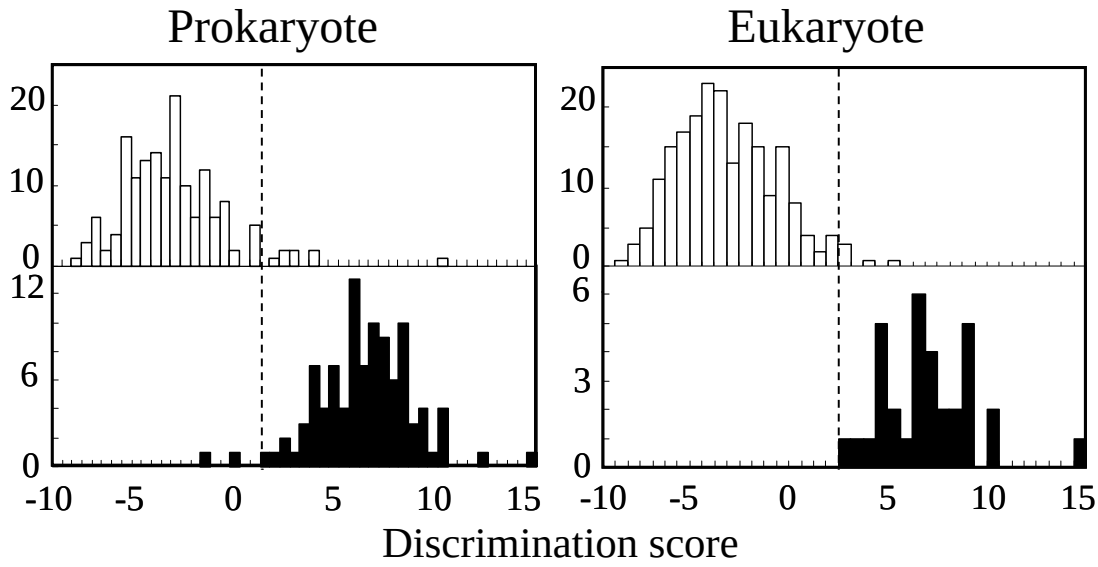

Figure 3. Histogram of discrimination score.
Histogram of the highest discrimination score in amino acid sequences.
Open bars are represented soluble proteins, solid ones are membrane proteins.

Table 4. Comparison of membrane protein predictors.

	Prokaryote					
	Plasma membrane proteins (101)			Soluble proteins (198)		
	TP	FN	Accuracy	TP	FP	Accuracy
This Research	99	2	98.02%	190	8	95.96%
SOSUI	98	3	97.03%	184	14	92.93%
TMHMM2	98	3	97.03%	177	21	89.39%
TopPred	101	0	100.00%	66	132	33.33%
MEMSAT2	101	0	100.00%	106	92	53.54%

	Eukaryote								
	Plasma membrane proteins (33)			Soluble proteins (308)			Other membrane proteins (104)		
	TP	FN	Accuracy	TP	FP	Accuracy	TP	FN	Accuracy
This Research	33	0	100.00%	303	5	98.38%	41	63	39.42%
SOSUI	32	1	96.97%	281	27	91.23%	63	41	60.58%
TMHMM2	33	0	100.00%	298	10	96.75%	72	32	69.23%
TopPred	33	0	100.00%	127	181	41.23%	104	0	100.00%
MEMSAT2	33	0	100.00%	206	102	66.88%	97	7	93.27%

The accuracy of the present system is compared with other membrane protein prediction systems in Table 4. The performances of membrane proteins prediction was evaluated using a jack knife procedure. The result showed that the accuracy of the prediction for membrane proteins was slightly lower than the values in the Tables, but the accuracy for soluble proteins were the same as the value in Tables. The accuracy of the true positive prediction is almost the same among various systems. However, the difference of the performance of the systems is well represented by the accuracy of the true negative prediction, that is the soluble protein prediction. Many of false positive data from soluble proteins had signal peptide: 8 false positive data for prokaryotes

and 5 for eukaryotes. Among those data 6 for prokaryotes and 3 for eukaryotes could be predicted as secretary proteins by SOSUIsignal, which will be reported else where. Many amino acid sequences of soluble proteins are incorrectly predicted as membrane proteins in TopPred and MEMSAT2. TMHMM2 is a considerably good system, but the prediction of soluble proteins has errors more than 10 %. On the other hand, the accuracy of the true negative prediction of the present system is as good as 96 %. The system is improved by about 3 % from the original SOSUI. Because about three quarters of proteins are soluble ones, the false negative prediction of 10 % means that about 30 % of membrane protein prediction is incorrect. This is a very important point in the classification of all amino acid sequences from genome information, which includes many orphan sequences whose biological meaning is unknown.

5. Discussion

The purpose of this work is to reveal the distributions of two physical parameters, the hydrophobicity and the amphiphilicity around typical transmembrane helices and to develop a high performance method of the membrane protein prediction, which is the basis of the prediction of cellular localization of proteins. Previously, we developed a membrane protein prediction SOSUI, which was one of the best systems in various methods, as shown in Table 4. The basic concept of SOSUI was that a membrane protein has at least one transmembrane helix whose central region is hydrophobic enough and sandwiched by the clusters of amphiphilic residues. In this work, we devised more systematic parameters that reflect the characteristic distribution of the hydrophobicity and the amphiphilicity indices of typical transmembrane helices. The improved system realized higher accuracy than not only the original SOSUI but also various informatics methods.

This result is important from two aspects: First, the improved system is practically very useful for the analysis of genome information. There are still many unknown sequences in total amino acid sequences in spite of extensive investigation of functional genomics. The improvement of the performance of the membrane protein prediction in this work provides more reliable data set of membrane proteins for those who want to carry out experiments of membrane proteins in genome scale. Second, the concept of this method in which to physical properties were coarse grained can be used for various other problems of protein classification. The membrane protein prediction in this work is based on the fact that clusters of hydrophobic residues and amphiphilic residues are substantial for stabilizing a transmembrane helix. Similarly, dumbbell-type proteins, for example, could be accurately predicted by the coarse graining approach [39].

This work is the first step of the development of a system which predicts membrane proteins, the targeting to various organella, and the location together with the topology of transmembrane helical regions. As shown in Table 4, the targeting to organella is still unsolved problem, and we will report in the next work a method for the targeting of proteins particularly to mitochondria.

References

[1] S. Mitaku, T. Hirokawa, T. Tsuji, *Bioinformatics*, **18**, 608-616 (2002).
[2] S. Mitaku, T. Hirokawa, *Protein Eng.*, **12**, 953-957 (1999).
[3] S. Mitaku, S. Hoshi, R. Kataoka, *J. Phys. Soc. Jpn.*, **54**, 2047-2054 (1985).
[4] J. Kyte, R. F. Doolittle, *J. Mol. Biol.*, **157**, 105-132 (1982).
[5] D. M. Engelman, A. Goldman, T. A. Steitz, *Meth. Enzymol.*, **88**, 81-88 (1982).
[6] D. Eisenberg, A. D. McLachlan, *Nature*, **319**, 199-203 (1986).

[7] D. Eisenberg, R. M. Weiss, T. C. Terwilliger, Proc. Natl Acad. Sci. USA, **81**, 140-144 (1984).
[8] T. Hirokawa, S. B. Chieng, S. Mitaku, *Bioinformatics*, **14**, 378-379, (1998).
[9] D. T. Jones, W. R. Taylor, J. M. Thornton, *Biochemistry*, **33**, 3038-3049 (1994).
[10] M. Cserzo, E. Wallin, I. Simon, von G. Heijne, A. Elofsson, *Protein Eng.*, **10**, 673-676 (1997).
[11] B. Rost, R. Casadio, P. Fariselli, Proc Int Conf Intell Syst Mol Biol. **4**, 192-200 (1996).
[12] G. E. Tusnady, I. Simon, *J Mol Biol.*, **283**, 489-506 (1998).
[13] G. E. Tusnady, I. Simon, *Bioinformatics.*, **17**, 849-850 (2001).
[14] E. L. Sonnhammer, von G. Heijne, A. Krogh, Proc Int Conf Intell Syst Mol Biol., **6**, 175-182 (1998).
[15] A. Krogh, B. Larsson, von G. Heijne, E. L. Sonnhammer, *J Mol Biol.*, **305**, 567-580 (2001).
[16] S. Mitaku., M. Ono, T. Hirokawa, B.-C. Seah, M. Sonoyama, *Biophys. Chem.*, **82**, 165–171 (1999).
[17] E. Wallin, von G. Heijne, *Protein Sci.*, **7**, 1029–1038 (1998).
[18] S. Kanaji, J. Iwahashi, Y. Kida, M. Sakaguchi, K. Mihara, *J. Cell Biol.*, **151**, 277-288 (2000).
[19] M. E. Girvin, V. K. Rastogi, F. Abildgaard, J. L. Markley, and R. H. Fillingame, *Biochemistry*, **37**, 8817-8824 (1998).
[20] S. Iwata, J. W. Lee, K. Okada, J. K. Lee, M. Iwata, B. Rasmussen, T. A. Link, S. Ramaswamy, and B. K. Jap, *Science*, **281**, 64-71 (1998).
[21] Y. Kimura, D. G. Vassylyev, A. Miyazawa, A. Kidera, M. Matsushima, K. Mitsuoka, K. Murata, T. Hirai, and Y. Fujiyoshi, *Nature*, **389**, 206-211 (1997).
[22] C. Toyoshima, M. Nakasako, H. Nomura, and H. Ogawa, *Nature*, **405**, 647-655 (2000).
[23] T. Tsukihara, H. Aoyama, E. Yamashita, T. Tomizaki, H. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaono, and S. Yoshikawa, *Science*, **272**, 1136-1144 (1996).
[24] T. Soulimane, G. Buse, G. P. Bourenkov, H. D. Bartunik, R. Huber, and M. E. Than, *EMBO J.*, **19**, 1766-1776 (2000).
[25] C. R. Lancaster, A. Kroger, M. Auer, and H. Michel, *Nature*, **402**, 377-385 (1999).
[26] T. M. Iverson, C. Luna-Chavez, G. Cecchini, and D. C. Rees, *Science*, **284**, 1961-1966 (1999).
[27] M. Kolbe, H. Besir, L.-O. Essen, and D. Oesterhelt, *Science*, **288**, 1390-1396 (2000).
[28] D. A. Doyle, C. J. Morais, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. MacKinnon, *Science*, **280**, 69-77 (1998).
[29] G. McDermott, S. M. Prince, A. A. Freer, A. M. Hawthornthwaite-Lawless, M. Z. Papiz, R. J. Cogdell, and N. W. Isaacs, *Nature*, **374**, 517-521 (1995).
[30] J. Koepke, X. Hu, C. Muenke, K. Schulten, and H. Michel, *Structure*, **4**, 581-597 (1996).
[31] G. Chang, R. H. Spencer, A. T. Lee, M. T. Barclay, and D. C. Rees, *Science*, **282**, 2220-2226 (1998).
[32] J. Deisenhofer, O. Epp, I. Sinning, and H. Michel, *Nature*, **318**, 618-624 (1985).
[33] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and M. Miyano, *Science*, **289**, 739-745 (2000).
[34] K. Murata, K. Mitsuoka, T. Hirai, T. Walz, P. Agre, J. B. Heymann, A. Engel, and Y. Fujiyoshi, *Nature*, **407**, 599-605 (2000).
[35] D. Fu, A. Libson, L. J. Miercke, C. Weitzman, P. Nollert, J. Krucinski, and R. M. Stroud *Science*, **290**, 481-486 (2000).
[36] S. Möller, E. V. Kriventseva, and R. Apweiler, *Bioinformatics*, **16**, 1159-1160 (2000).
[37] U. Hobohm, M. Scharf, and C. Sander, *Protein Science*, **1**, 409-417 (1992).
[38] M. Monne, I. Nilsson, M. Johansson, N. Elmhed, von G Heijne, *J. Mol. Biol.*, **284**, 1177-1183 (1998).
[39] N. Uchikoga, S. Mitaku, *Protein Sci.*, in press.