Chem-Bio Informatics Journal, **Vol. 10, pp.61-73 (2010)**

SOSUImp1: high performance prediction system for single-spanning membrane proteins

T. Tsuji[1,2]*, F. Akazawa[1], R. Sawada[1] and S. Mitaku[1]

[1]Nagoya University, School of Engineering, Department of Applied Physics
Nagoya, Chikusa-ku, Furocho 464-8606, Japan
[2]University of Tokyo, Institute of Medical Science, Human Genome Center
4-6-1 Shirokane-dai Minato-ku Tokyo 108-8639, Japan

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

(Received September 29, 2009; accepted May 18, 2010; published online May 30, 2010)

Abstract

Single-spanning membrane proteins (MP1) occupy the largest component of membrane proteins in total open reading frames of organisms, having essential functions such as signal transduction, immunological reaction and cell adhesion. We developed a novel software system comprised of two filtering layers for predicting MP1 with or without a signal peptide region. In the first filtering layer, we selected membrane proteins with one or two transmembrane (TM) regions by the membrane protein prediction system SOSUI, which is accurate in predicting transmembrane regions but cannot identify signal peptide regions. The second filtering layer was comprised of several modules for distinguishing signal peptide regions. On the assumption that a signal peptide has two kinds of sequences at the N-terminus by which the signal peptide is embedded into membrane and cleaved at its C-terminal end, we calculated two discrimination scores by the canonical discriminant analysis, using averages of several physical properties around the first N-terminal hydrophobic cluster. This prediction system SOSUImp1 comprised of two filtering layers could discriminate very accurately among five types of proteins: cytoplasmic soluble proteins and secretory proteins, MP1 with and without a signal peptide, and multi spanning membrane proteins. The performance for MP1 with a signal peptide that is important in the cell-cell communication was particularly high compared with previous prediction systems.

The prediction system SOSUImp1 and the dataset of 5932 proteins used for developing the system are available at http://bp.nuap.nagoya-u.ac.jp/sosui/mp1/

Key Words: membrane protein, signal peptide, single spanning membrane protein, prediction, bioinformatics

Area of Interest: Bioinformatics and Bio computing

1. Introduction

Integral membrane proteins mediate a wide range of fundamental biological processes, such as cell signal transduction, molecular transport, immune system, and cell adhesion. Single-spanning membrane proteins (MP1) are the largest component of membrane proteins in the open reading frames of organisms [1][2]. Receptors of various growth factors [3], antibodies at cytoplasmic membranes [4][5] and activators of heterotrimeric G protein-coupled receptors [6] are typical MP1 with a signal peptide region. Despite the functional importance of MP1, high performance prediction of these membrane proteins has been difficult because of insufficient hydrophobicity of both their transmembrane (TM) and signal peptide (SP) regions. Such characteristics of MP1 are most likely to be related to their complex formation with other supplementary factors embedded in the membrane [7].

In this study, we compared the amino acid sequences of five types of proteins (cytoplasmic proteins, secretory proteins, MP1 with and without SP and multi-spanning membrane proteins) in terms of the position and physical properties around their hydrophobic regions. The position of TM region and various physical properties of the first hydrophobic segment were clearly different between MP1 with and without a signal peptide by which a high performance prediction system SOSUImp1 was developed.

2. Datasets of amino acid sequences

Eukaryotic proteins were extracted from the Swiss-Prot database release 54.6. Proteins with more than 25% homology and localized in the mitochondria, nucleus, peroxisome, and chloroplast were removed for unbiased discrimination. A total of 5932 amino acid sequences were used in this study.

Proteins were classified into the following five types based on the information in the CC and FT lines from the swissprot database: cytoplasmic soluble proteins, secretory proteins, MP1 with SP, MP1 without SP, multi-spanning membrane proteins (multi-MP). Proteins from all five types were randomly divided into training and test datasets. We generated ten different sets of training and complementary test datasets for the cross validation test. 80% of the data were used for developing the prediction system as the training data, and the remaining 20% was used for testing the system (Table 1). The ratio between the training and testing data was kept the same for all types of proteins.

Table 1. Number of proteins in the training and test datasets used for the development and evaluation of SOSUImp1

	total	training	testing
Cytoplasmic soluble proteins	1596	1276	320
Secretory proteins	1982	1585	397
MP1 without SP	348	278	70
MP1 with SP	758	606	152
Multi-MP	1248	998	250
Total	5932	4743	1189

3. Method for discriminating single-spanning membrane proteins

Our prediction system consisting of two filtering layers is outlined in Figure 1. The first layer is the crude classification of membrane proteins by the original SOSUI system, which has an overall accuracy of >95% [8][9]. Although the original SOSUI system is very useful for filtering membrane proteins, it was not designed for predicting SP. Therefore, the classification of proteins with SP by the original SOSUI system is incorrect. For example, the data of MP1 with SP will be predicted as either MP1 without SP or MP2 without SP. In the MP1 case, a SP was missed, whereas a SP was incorrectly predicted as a TM helix, in the MP2 case. Table 2 shows the result of the classification of the total datasets in terms of the number of predicted transmembrane regions by the SOSUI system. Almost all cytoplasmic soluble proteins were correctly predicted to be soluble proteins. However, each of four other types of proteins (secretory proteins, MP1 with or without SP, or multi MP) could not be predicted into a single class in terms of the number of transmembrane regions. About 95% of data for each type of proteins were classified into either of two classes of the number of transmembrane regions. For example, 94% of MP1 with SP were classified into membrane proteins having one or two transmembrane regions. Therefore, the problem of the classification of five types of proteins could be broken down to several smaller classification problems by using the SOSUI system as first filter.

The second filtering layer consists of several modules (Figure 1). Each module discriminates between two protein types, the combination of which varies according to the TM number (TM1, single helix; TM2, two helices; multi TM, more than or equal to two helices) predicted by SOSUI. The majority (96.5%) of predicted soluble proteins were comprised of two types of proteins, cytoplasmic (56.1%) and secretory (40.4%). Whereas, the predicted membrane proteins with TM2 were comprised of three types of proteins: MP1 with SP (48.0%), multi-MP (35.1%) and secretory proteins (14.1%). However, the secretory proteins in the last group have two SPs, one at both the N- and C-termini. Based on the positional distribution differences between the two predicted TMs, secretory proteins could be easily identified (data not shown), allowing for discrimination of predicted proteins with TM2 as either MP1 with SP or multi-MP.

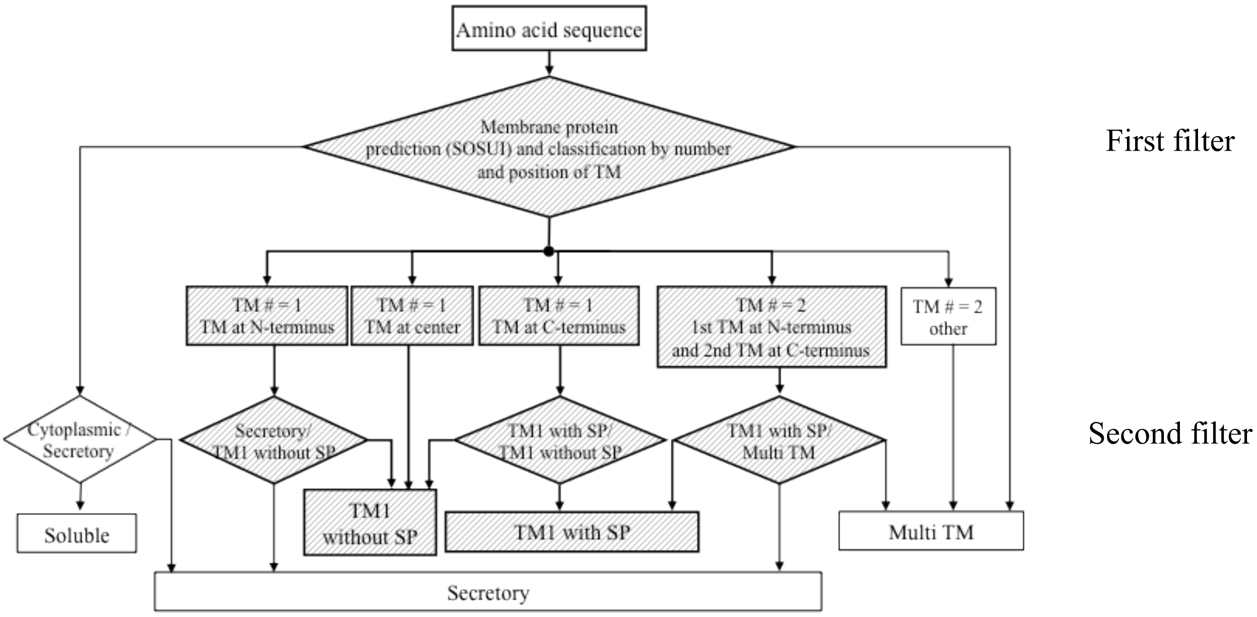
Amino acid sequence
Membrane protein prediction (SOSUI) and classification by number and position of TM
First filter
TM # = 1 TM at N-terminus
TM # = 1 TM at center
TM # = 1 TM at C-terminus
TM # = 2 1st TM at N-terminus and 2nd TM at C-terminus
TM # = 2 other
Second filter
Cytoplasmic / Secretory
Secretory/ TM1 without SP
TM1 with SP/ TM1 without SP
TM1 with SP/ Multi TM
Soluble
TM1 without SP
TM1 with SP
Multi TM
Secretory

Figure 1. Flowchart of SOSUI$_{MP1}$ for predicting MP1

Table 2. Classification of dataset based on TM number predicted by the membrane protein prediction system SOSUI, the first filter of SOSUImp1

Swiss-Protfeatures	Prediction by SOSUI				Total
	Soluble	TM1	TM2	Multi TM	
Cytoplasmic soluble proteins	1562	31	2	1	1596
Secretory proteins	1002	922	54	4	1982
MP1 without SP	31	304	11	2	348
MP1 with SP	14	542	197	5	758
Multi-MP	17	61	149	1021	1248
Total	2626	1860	413	1033	5932

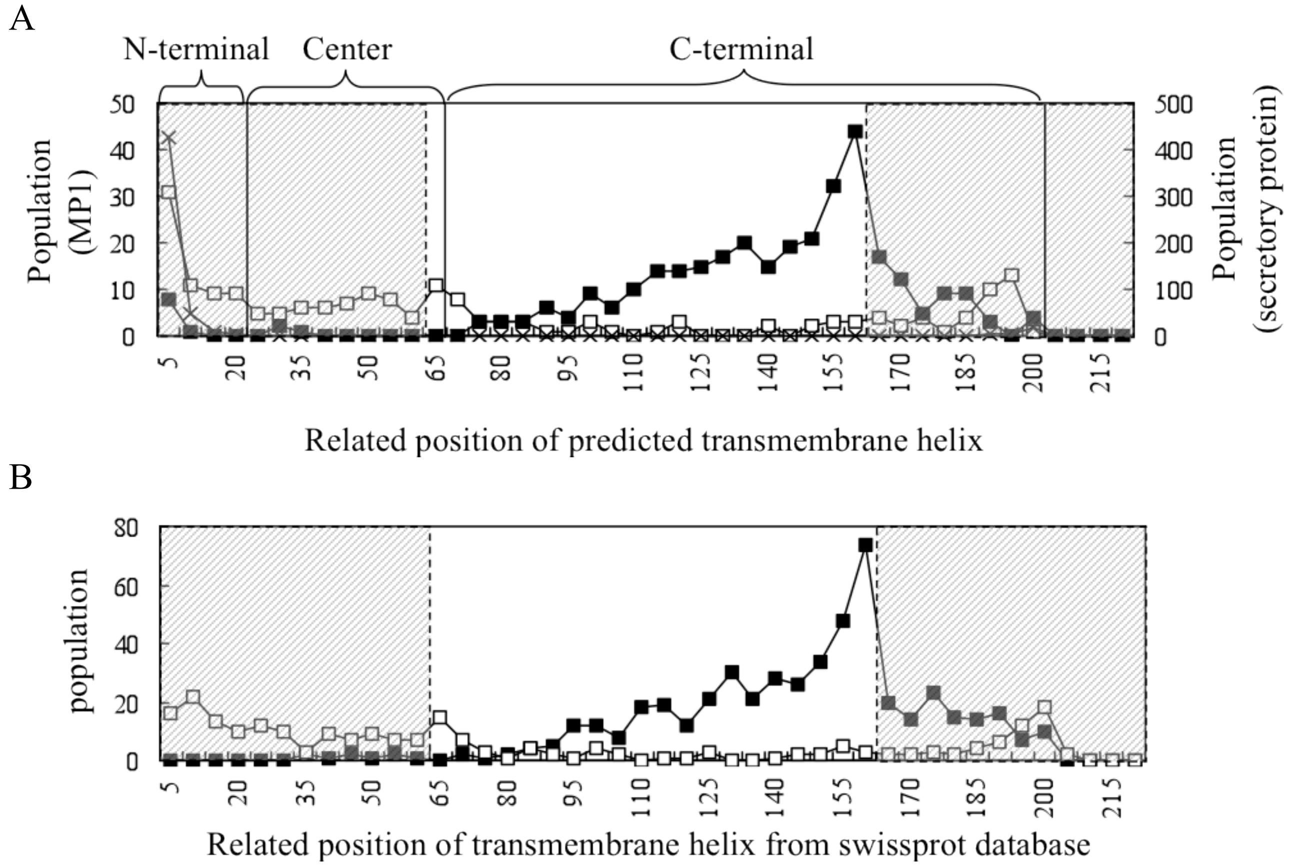

Figure 2. (A) Positional distribution of the starting point of TM1 regions predicted by the first filter for three protein types: MP1 with SP (solid square) and MP1 without SP (open square) and secretory proteins (X)
(B) The same plot for the positional distribution of the true single membrane region from swissprot annotation: MP1 with SP (solid square) and MP1 without SP (open square)

The analysis of the predicted membrane proteins with TM1 was more complicated than those for the predicted soluble proteins and the predicted membrane proteins with TM2. Figure 2A shows the position of predicted TM1 in amino acid sequences of three types of proteins, MP1 with and without SP and secretory proteins. It is difficult to compare the position of TM1 among various amino acid sequences of MP1 because the length of the amino acid sequences varies greatly among MP1. However, when the middle portion of total amino acid sequences was normalized to 100 residues, removing 60 residues from both N- and C-terminal ends (hatched areas in Figure 2), all the amino acid sequences could be compared in the same normalized sequence of 220 residues. The positional distribution of TM1 predicted by SOSUI differed among three protein types. When we divided the normalized amino acid sequence into three regions as indicated in Figure 2, N-terminal (residues 1-20), central (21-64), and C-terminal (65-200) regions, the problem of the discrimination of three types of proteins could be transformed to three simpler problems of the alternative discrimination,. The N-terminal region contained few MP1 with SP, but many MP1 without SP and secretory proteins; the central region contained only MP1 without SP; and the C-terminal region contained MP1 with and without SP but no secretory proteins. These results indicated that a protein with TM1 predicted by SOSUI in the N-terminal region can be assumed to be a MP1 without SP or secretory protein, while that in the C-terminal region can be assumed to be a MP1 with or without SP. In order to confirm the validity of the distribution of three types of proteins, we compared the positional distribution of the predicted TM1 (Figure 2A) with the graph of true transmembrane from the swissprot annotation (Figure 2B). The result indicated that the distribution of the predicted TM1 and that of the true TM1 is the same, leading to the conclusion that the classification of the three types of proteins by the positional distribution is reasonable.

In the second layer of filtering, the method is basically the same among all modules of the alternative discrimination. Therefore, we describe here the method for discriminating between MP1 with and without SP in the central region of Figure 2, as an example. In both types of MP1 (with/without SP), the position of TM1 is significantly different in the N-terminal region, at which SPs are located. We analyzed the physical properties of the N-terminal region for discrimination of these two types. SPs are located in the N-terminal region and usually possess two characteristics: they have a short cluster of hydrophobic amino acids which is translocated into the membrane and a segment at the C-terminal end of this hydrophobic cluster which is recognized and cleaved by signal peptidase. Considering these characteristics, we selected three segments around the hydrophobic cluster for SP prediction: 25 residues at the N-terminal end (region I), the hydrophobic segment in which the moving average of hydrophobicity of 21 residues is higher than 0.8 by the Kyte-Doolittle index [10] (region II), and the segment of 35 residues around the C-terminal end of the hydrophobic segment (region III) (Fig. 3A). The region III was further divided into three segments: a segment of 15 residues inside the hydrophobic region (S1 in Figure 3B) and two 10-residue segments outside of the hydrophobic segment (S2 and S3 in Figure 3B).

Figure 3B shows the averages of various physical properties in each region for training data of MP1 with and without SP, which were significantly different. The average hydrophobicity of region I was significantly higher for MP1 with SP than for MP1 without SP; this significant difference between MP1 with and without SP contributed to the accurate discrimination. Significant differences were also observed for three properties in region II: the starting point and width of the hydrophobic cluster, and the maximum hydrophobicity of the moving average; as well as for five properties in region III: the hydrophobicity, the positive and negative charges, the number density of small polar residues (Ser and Thr) and the SP index, which was previously defined [11].

All properties other than the start point and the width of the hydrophobicity peak were first smoothened by the moving average of seven residues:

$$\left\langle X_p(i) \right\rangle_7 = \left\{ \sum_{j=i-3}^{i+3} X_p(j) \right\} \Big/ 7 \qquad (1)$$

in which $X_p(j)$ the *p*-th property at the sequence number *j*. Then, the double average in the *k*-th region (the region I, II, S1, S2, or S3 in Fig. 3A) was calculated and used for the discrimination analysis.

$$\left\langle\left\langle X_p\right\rangle\right\rangle_k = \left\{ \sum_{i \in k\text{-th region}} \left\langle X_p(i)\right\rangle \right\} \Big/ l_k \tag{2}$$

in which l_k represents the length of the *k*-th region. Therefore, we obtained a parameter from the region I, three parameters from the region II and fifteen parameters from the region III from an amino acid sequence.

It is well known that there are two elementary processes in the signal peptide, the embedding into membrane and the cleavage at the external side of the signal peptide. Assuming that the parameters of the hydrophobic segment (region II) and its N-terminal side (region I) determine the embedding process, we determined a discrimination score D_1 from the analysis of four parameters of region I and II. Another discrimination score D_2, characterizing the cleavage of signal peptides at the C-terminal side of the hydrophobic segments, was determined by the analysis of fifteen parameters of region III.

The discrimination analysis was commonly applied to the determination of the score D_1 and D_2, so that the hydrophobic segments at the N-terminal end can be distinguished between the real signal peptide and the hydrophobic loops. The score D_1 is expressed by the linear combination of four parameters;

$$D_1 = a_1 C_{start} + a_2 C_{width} + a_3 \left\langle\left\langle H\right\rangle\right\rangle_{region\ II} + a_4 \left\langle\left\langle H\right\rangle\right\rangle_{region\ I} \tag{3}$$

in which C_{start} and C_{width} are the starting point and the width of the hydrophobic segment, respectively. The parameters $\left\langle\left\langle H\right\rangle\right\rangle_{region\ I}$ and $\left\langle\left\langle H\right\rangle\right\rangle_{region\ II}$ are the double average of the regions I and II, respectively. For determining the score D_2, we first calculated a parameter from three values of each physicochemical property at three sub-regions, S1, S2 and S3, of the region III by the following equation:

$$Z_p = \sum_{k \in S1,S2,S3} \left(\left\langle\left\langle X_p\right\rangle\right\rangle_k - \overline{\left\langle\left\langle N_p\right\rangle\right\rangle_k} \right) \times \left(\overline{\left\langle\left\langle P_p\right\rangle\right\rangle_k} - \overline{\left\langle\left\langle N_p\right\rangle\right\rangle_k} \right) \tag{4}$$

where $\overline{\left\langle\left\langle P_p\right\rangle\right\rangle_k}$ and $\overline{\left\langle\left\langle N_p\right\rangle\right\rangle_k}$ are the average values of $\left\langle\left\langle X_p\right\rangle\right\rangle_k$ for all of MP1 with SP and the hydrophobic loop segment in MP1 without SP, respectively. Finally, the score D_2 was determined by the discriminant analysis technique, using four parameters Z_p (*p* = 1~5):

$$D_2 = \sum_{p=1}^{5} \left(b_p Z_p \right) \tag{5}$$

The average of parameter, weight and contribution of various parameters in the discrimination analyses are shown in Table 3. Figure 4 shows a dispersion diagram of the D_1 versus D_2 space for the two types of proteins, MP1 with and without SP. The data of MP1 with SP which were predicted as TM1 by SOSUI were well discriminated from the data of MP1 without SP. This fact indicates that the physical properties of several regions around the first hydrophobic segment in an amino acid sequence determine the fate of the N-terminal hydrophobic segment, whether they would be embedded into membrane and cleaved by signal peptidase or not. According to the same approach, we developed discrimination tools for all problems simplified by the first filtering layer of the prediction system SOSUImp1.

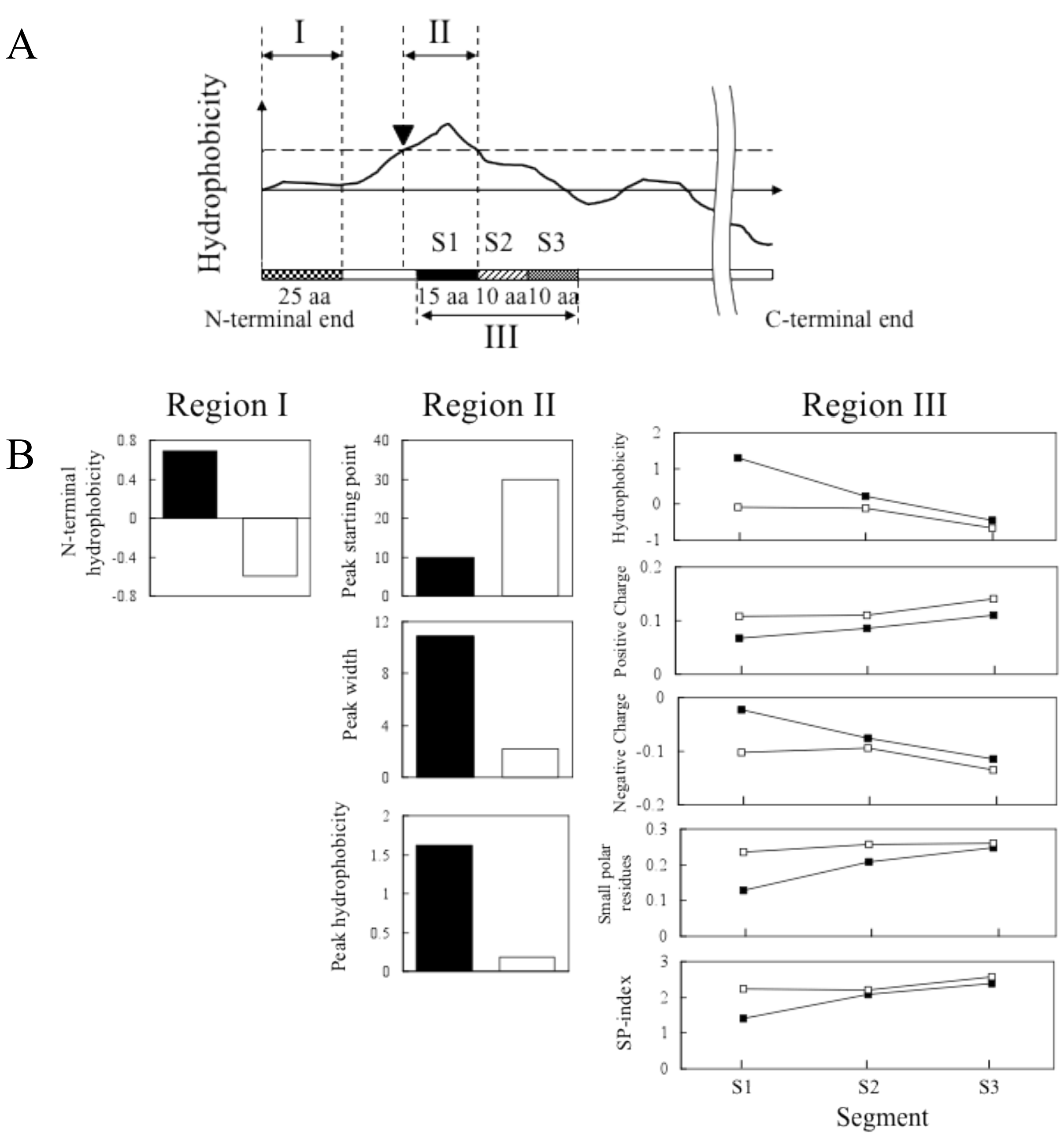

Figure 3. Physical parameters for discriminating between MP1 with or without SP which are predicted as TM1 in the C-terminal region by the first filtering layer of SOSUImp1

When the moving average of hydrophobicity (Kyte-Doolittle index [10]) of the 21-residue segment exceeds the threshold of 0.8 (dotted line), the peak is enumerated as a candidate of SP.

A. Three regions around the hydrophobicity peak were used for discriminating between MP1 with and without SP: the N-terminal segment of 25 residues (I), the segment exceeding the threshold (II) and the segments around the C-terminal end of the peak (III).

B. The average values were calculated for nine properties in three regions, I, II and III. Closed and open bars or squares represent MP1 with and without SP, respectively.

Table 3. Weight and contribution of the discrimination scores
(A) Score D1, (B) Score D2

A.

a1~a4	C_{start}	C_{width}	$\langle\langle H\rangle\rangle_{region\ II}$	$\langle\langle H\rangle\rangle_{region\ I}$
Weight	0.0178	-0.0349	-0.870	-0.652
Average of parameter	13.1	9.34	0.476	1.34
Contribution	-0.35	-0.31	-1.1	-0.92

B.

b1~b5	Hydrophobisity	Positive charge	Negetive charge	Small polar residues	SP-index
Weight	1.70	86.3	70.8	45.5	1.60
Average of parameter	-0.266	0.00217	0.00176	0.00714	0.520
Contribution	-1.65	-0.13	-0.20	-0.34	-0.54

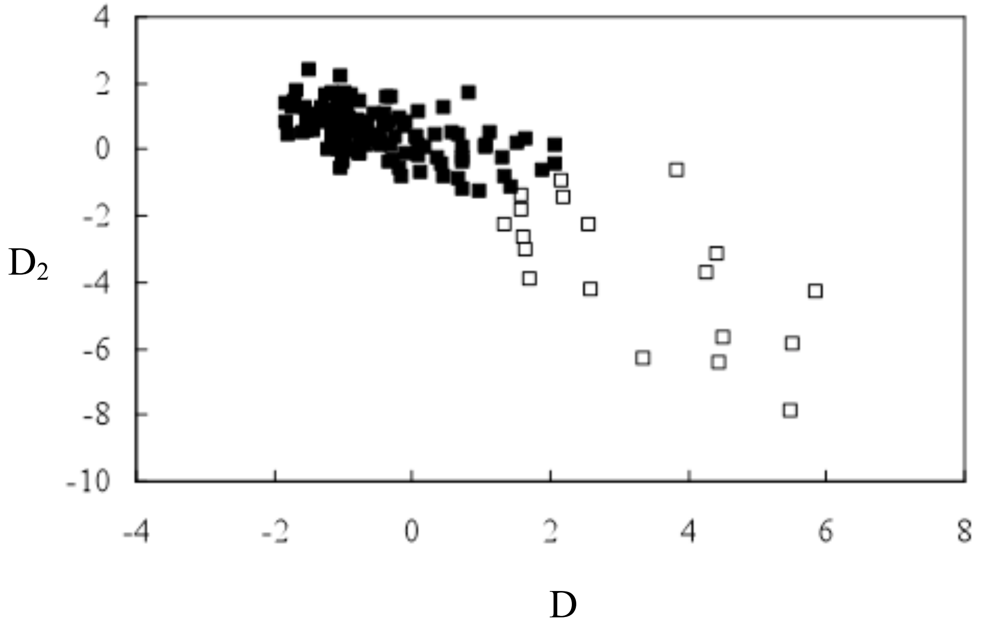

Figure 4. Dispersion diagram of discrimination scores D_1 versus D_2 for the N-terminal peaks of hydrophobicity in MP1 with (solid squares) and without (open squares) SP

3. Results and Discussion

We prepared ten pairs of training and test datasets for the cross validation of the prediction system for single spanning membrane proteins. Table 4 shows the result of the cross validation test. Both of the recall and the precision were larger than 90% for cytoplasmic, secretory and multi-spanning membrane proteins. The evaluation parameters for MP1 with SP were just below90%, whereas the recall was about 69% and the precision was 64% for MP1 without SP. The low performance of the prediction of MP1 without SP is due to the difficulty in the discrimination between MP1 without SP and secretory proteins. Since the number of test data for MP1 without SP was much smaller than that for secretory proteins, the same number of false prediction severely lowered the evaluation parameters of MP1 without SP.

Table 4. Result of prediction by cross validation test of the present system SOSUImp1

Dataset		Result of prediction					Recall
		Cytplasm	Secretory	MP1 without SP	MP1 with SP	Multi MP	
Cytplasm	321	302.5±5.0	11.8±5.9	4.8±1.0	0.7±0.6	1.0±0.0	94.2%
Secretory	399	17.1±2.6	362.1±7.7	11.3±3.9	6.9±2.3	2.0±0.7	90.8%
MP1 without SP	71	3.4±0.4	11.8±3.6	48.7±4.2	4.0±2.4	3.0±0.0	68.6%
MP1 with SP	153	1.2±0.8	8.1±2.8	4.9±2.5	135.3±4.8	3.5±1.5	88.8%
Multi MP	219	4.0±0.0	2.3±0.9	7.6±1.4	7.35±2.4	197.75±1.1	90.3%
Precision		92.9%	90.7%	64.2%	87.7%	95.4%	

In the Table 5, the performance of the present system SOSUImp1 was compared with several other prediction systems: the combination of the original SOSUI and SOSUIsignal, the combination of TMHMM2 [12] and SignalP ver.3 [13], and Phobius [14][15]. The performance of Phobius was comparable to our system, but the present system SOSUImp1 is completely different: In our system, amino acid sequences are transformed to the sequences of various physicochemical parameters, and this step of the analysis decreases the dataset dependence of the discrimination function, leading to the applicability to novel and unknown amino acid sequences. This advantage is particularly suitable for the analysis of all sequence from various biological genomes. Both of the recall and the precision of the prediction for MP1 with SP by our system were higher than other systems. It is well known that receptors of various growth factors, which are key proteins in the signal transduction of multi-cellular organisms including human beings, are single spanning membrane proteins with signal peptide. Therefore, the high performance of the present system in the prediction of MP1 with SP is very advantageous in the analysis of biological genomes.

The SOSUImp1 prediction system is a web-based application that can be used by inputting sequences having a minimum length of 60 amino acids. The SOSUImp1 system is available at http://bp.nuap.nagoya-u.ac.jp/sosui/mp1/.

This work was supported in part by the SENTAN, JST.

Table 5. Comparison of the performance of present system SOSUImp1 with other software tools for predicting single spanning membrane proteins

Protein type	Prediction System							
	SOSUImp1		SOSUI and SOSUIsignal		TMHMM2 and SignalP ver.3		Phobius	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)
Cytoplasmic	94.2	92.9	91.5	89.3	98.3	81.8	93.4	95.9
Secretory	90.8	90.7	87.6	78.3	74.7	95.7	88.6	94.7
MP1 without SP	68.6	64.2	53.0	60.9	79.1	42.1	77.4	60.5
MP1 with SP	88.4	87.7	69.5	67.6	60.3	88.4	87.3	84.3
Multi MP	90.3	95.4	79.2	95.5	96.6	92.2	98.9	94.9

References

[1] S. Mitaku, M. Ono, T. Hirokawa, S. Boon-Chieng, and M. Sonoyama, Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the SOSUI prediction system, *Biophys Chem.*, **82**, 165-171(1999).

[2] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J Mol Biol*, **305**, 567-580 (2001).

[3] X. Zhang, J. Gureasko, K. Shen, P. A. Cole, and J. Kuriyan, An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor, *Cell*, **125**, 1137-1149 (2006).

[4] J. Hennecke, and DC Wiley, T cell receptor-MHC interactions up close, *Cell*, **104**, 1-4 (2001).

[5] L. E. Samelson, Signal transduction mediated by the T cell antigen receptor: the role of adapter proteins, *Annu. Rev. Immunol.*, **20**, 371-394 (2002).

[6] T. B. Patel, Single Transmembrane Spanning Heterotrimeric G Protein-Coupled Receptors and Their Signaling, *Pharmacol Rev*, **56**, 371-385 (2004).

[7] C. P. Chen, A. Kernytsky and B. Rost, Transmembrane helix predictions revisited, *Protein Sci*, **11**, 2774-2791 (2002).

[8] T. Tsuji and S. Mitaku, Features of transmembrane helices useful for membrane protein prediction, *CBIJ*, **4**, 110-120(2004).

[9] T. Hirokawa, S. Boon-Chieng, and S. Mitaku, SOSUI: classification and secondary structure prediction system for membrane proteins, *Bioinformatics*, **14**, 378-379(1998).

[10] J. Kyte and R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, *J Mol Biol*, **157**, 105-132(1982).

[11] M. Gomi, M. Sonoyama, and S. Mitaku, High performance system for signal peptide prediction: SOSUIsignal, *CBIJ*, **4**, 142-147 (2004).

[12] E. L. Sonnhammer, G. von Heijne, A. Krogh, A hidden Markov model for predicting transmembrane helices in protein sequences, *Proc Int Conf IntellSyst Mol Biol.*, **6**, 175-82 (1998).

[13] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, Improved Prediction of Signal Peptides: SignalP 3.0, *J Mol Biol*, **340**, 783-795(2004).

[14] L. Kall, A. Krogh, and E. L. Sonnhammer, A Combined Transmembrane Topology and Signal Peptide Prediction Method, *J Mol Biol*, **338**, 1027-1036 (2004).

[15] L. Kall, A. Krogh, and E. L. Sonnhammer, Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server, *Nucleic Acids Res*, **35**, W429-W432 (2007).