Chem-Bio Informatics Journal, Vol. 4, No. 4, pp.142-147(2004) ***Communication***

High performance system for signal peptide prediction: SOSUIsignal

Masahiro Gomi[1], Masashi Sonoyama[2] and Shigeki Mitaku[1,2,*]

1 Tokyo University of Agriculture and Technology, Department of Biotechnology, Nakacho 2-24-16, Koganei, Tokyo 184-8588, JAPAN.
2 Nagoya University, School of Engineering, Department of Applied Physics, Furocho, Chikusa-ku, Nagoya 464-8603, JAPAN.

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

(Received September 9, 2004; accepted December 31, 2004; published online January 8, 2005)

Abstract

We describe a novel method for predicting a signal peptide of which three-domain (tripartite) structure is recognized by three modules of the software system. The first module numerates hydrophobic segment in N-terminal 100 residues, the second predicts signal sequences including both signal peptides and signal anchors, and the third discriminates signal peptides. Two novel indexes, SS- and SP-indexes, were developed for the discrimination of signal sequences and signal peptides, respectively, by calculating the relative propensities of amino acids at the carboxyl-terminal end of the hydrophobic region. The number of adjustable parameters in the whole system was only five. When three groups of data (917 signal peptides, 103 signal anchors and 544 non-signal sequences) were analyzed, signal peptides of eukaryotes could be discriminated with the Matthews correlation coefficient of 0.89. The signal peptide predictor SOSUIsignal is available at the web site: http://bp.nuap.nagoya-u.ac.jp/sosui/sosuisignal/sosuisignal_submit.html. This system has the advantage of very fast calculation.

Key Words: signal peptide, secretory protein, prediction, bioinformatics

Area of Interest: Bioinformatics and Bio computing

1. Introduction

Signal peptides are amino-terminal extensions of polypeptides which target them to the cytoplasmic membrane of prokaryotes or to the endoplasmic reticulum of eukaryotes. Because the secretion of proteins is closely related to the interaction of a cell with its environment, the prediction of signal peptides within a proteome will provide important information about the living strategy of a cell. It is known that a signal peptide has common features of sequences: a

hydrophobic segment, positively charged residues at the amino terminal end of the hydrophobic segment and the cleavage site at its carboxyl end. Because a hydrophobic segment with positively charged end is the common feature of segments which are translocated into membrane [2], it is difficult to discriminate a signal peptide from transmembrane helices.

Two approaches for the prediction of signal peptides have been reported [9]: window-based methods [1][10][11][12] and global structure-based methods [5][7][8]. In the former approach, a window of fixed length is examined at each position of the target sequences. The methods of the latter approach try to recognize the three-domain (tripartite) structure of signal peptides. The advantage of the former is considerably high accuracy, and that of the latter is the physicochemically interpretable rules with a small number of adjustable parameters.

In this work, we developed a method for the recognition of the tripartite structure of signal peptides, constituting the system with three modules. The accuracy is as good as the window-based methods, and the speed of calculation is very fast because of the simple algorithm.

2. Methods

2.1 Modules of prediction system

Most signal peptides interact with three different types of cellular apparatus: a shaperon-like protein (Sec B or SRP) for the recognition, a translocon for the translocation through membrane and a signal peptidase which cleaves a signal peptide [9]. Figure 1 shows the structure of a signal peptide which has a hydrophobic segment and a positively charged cluster at the amino terminal end of the hydrophobic segment. There is a vague motif of the cleavage site, the (-3,-1) rule, at the carboxyl terminal region of the hydrophobic segment [10]. However, this motif is not decisive

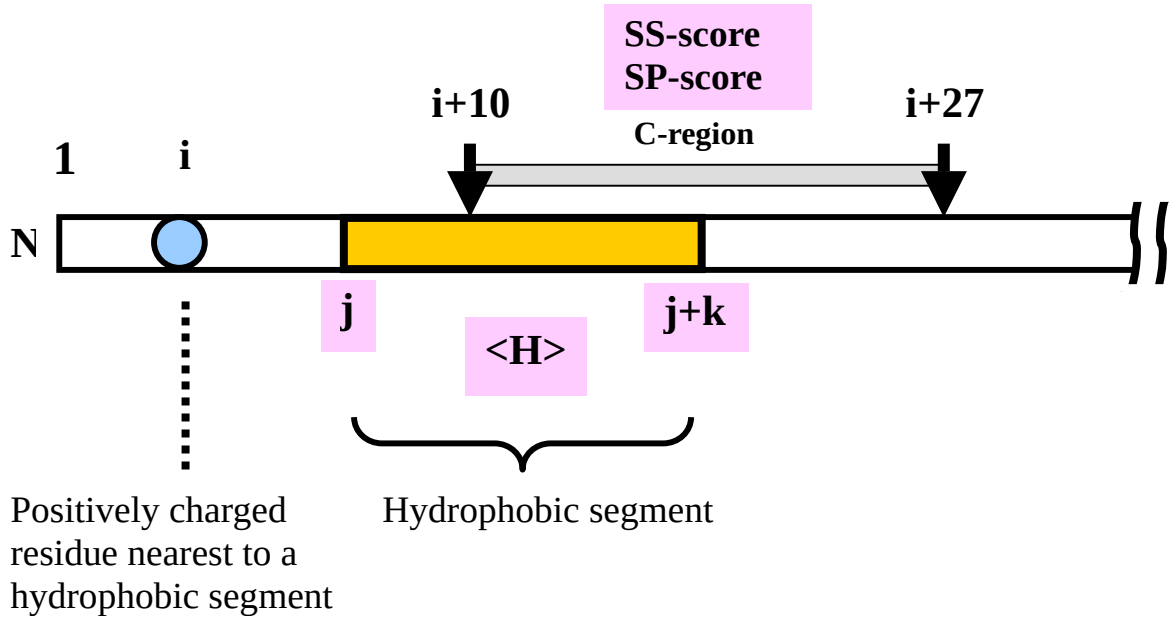

Figure 1. The structure of signal peptides and five parameters used for the prediction: the sequence number of the first residue of a hydrophobic segment j, the length of hydrophobic segment k, the average hydrophobicity <H> and SS- and SP-scores.

and it seems that the characteristics of longer segment at carboxyl end of the hydrophobic segment should be analyzed for more accurate prediction of signal peptides. We reexamined the amino acid sequences around the hydrophobic segments, comparing three kinds of datasets: signal peptides, signal anchors and soluble hydrophobic segments. The propensity of amino acids showed different profiles in the carboxyl end region (C-region) of the hydrophobic segments among the three kinds of datasets. Therefore, we calculated the propensity of amino acids in the region from 10-th to 27-th residues from the positive charge at the amino-terminal end of the hydrophobic segment (Figure 1).

We constructed three modules of the prediction system each of which corresponds to the characteristics of signal peptides. First module numerates a hydrophobic segment longer than 8 residues in the amino-terminal 100 residues. The threshold of the average hydrophobicity is zero of Kyte-Doolittle hydropathy [3]. Numerated segments by this module contain almost all signal peptides. The second module predicts signal sequences including both signal peptides and signal anchors, mainly using SS-score. Finally, the third module discriminates signal peptides from signal anchors, mainly using SP-score.

2.2 Indexes characterizing signal sequences

We carried out the cluster analysis for discriminating signal sequences (signal peptide + signal anchor) from other types of sequences, using various sets of parameters, including SS- and SP-scores. The best set of four parameters in the second module which discriminate signal sequences from non-signal sequences was the sequence position of the start point of a hydrophobic cluster, its length, the average hydrophobicity and a novel index, the SS-score, as shown in Figure 1. The SS-score was substituted by the SP-score in the third module for the discrimination between signal peptides from signal anchors.

The propensities of amino acids in the C-regions of signal sequences as well as signal peptides in each datasets were first calculated by the equations:

$$\boldsymbol{p}_{SS}(\boldsymbol{AA}) = \left(\sum_{i=1}^{m} \boldsymbol{N}_{AA}(\boldsymbol{i}) \right)_{SS} \Big/ (18\boldsymbol{m}) \qquad (1)$$

$$\boldsymbol{p}_{SP}(\boldsymbol{AA}) = \left(\sum_{i=1}^{m} \boldsymbol{N}_{AA}(\boldsymbol{i}) \right)_{SP} \Big/ (18\boldsymbol{m}) \qquad (2)$$

in which $\boldsymbol{p}_{SS}(\boldsymbol{AA})$ and $\boldsymbol{p}_{SP}(\boldsymbol{AA})$ are the propensity of an amino acid **AA** in datasets of signal sequences and signal peptides, respectively. The number of the amino acid in the C-region of i-th protein is represented by $\boldsymbol{N}_{AA}(\boldsymbol{i})$, and $\boldsymbol{m}$ is the number of proteins in the datasets. The propensity for total sequences was also calculated,

$$\boldsymbol{p}_{Total}(\boldsymbol{AA}) = \left(\sum_{i=1}^{m} \boldsymbol{N}_{AA}(\boldsymbol{i}) \right) \Big/ \left(\sum_{i=1}^{m} \boldsymbol{M}(\boldsymbol{i}) \right) \qquad (3)$$

in which $\boldsymbol{p}_{Total}(\boldsymbol{AA})$ represents the propensity of an amino acid $\boldsymbol{AA}$ for total sequences and $\boldsymbol{M}(\boldsymbol{i})$ and the size of i-th protein. Then, the relative propensity was calculated by the following equations,

$$\boldsymbol{x}_{SS}(\boldsymbol{AA}) = \boldsymbol{p}_{SS}(\boldsymbol{AA}) / \boldsymbol{p}_{Total}(\boldsymbol{AA}) \qquad (4)$$

$$\boldsymbol{x}_{SP}(\boldsymbol{AA}) = \boldsymbol{p}_{SP}(\boldsymbol{AA}) / \boldsymbol{p}_{Total}(\boldsymbol{AA}) \qquad (5)$$

in which $\boldsymbol{x}_{SS}(\boldsymbol{AA})$ and $\boldsymbol{x}_{SP}(\boldsymbol{AA})$ represent the SS- and SP-indexes for an amino acid $\boldsymbol{AA}$, respectively.

The parameters SS- and SP-scores for the discrimination were calculated from the SS- and SP-indexes by the following equations, respectively, averaging the values in the C-region:

$$\bar{x}_{SS} = \left(\sum_{j=i+10}^{i+27} x_{SS}(j) \right) \Big/ 18 \qquad (6)$$

$$\bar{x}_{SP} = \left(\sum_{j=i+10}^{i+27} x_{SP}(j) \right) \Big/ 18 \qquad (7)$$

The discrimination score for signal sequences or signal peptides was calculated by the cluster.

2.3 Datasets of signal sequences

We prepared three kinds of datasets (signal peptides, signal anchor type II and soluble sequences without signal sequences) from SWISS-PROT #40 for training and testing the system. Data of signal peptides were selected by the feature of "SIGNAL PEPTIDE". Data of signal anchors were selected by two conditions: One is the feature of "SIGNAL ANCHOR". The other is the feature of "TRANSMEM" together with the existence of the first transmembrane helix within 100 residues from the amino terminus. Redundancy of data was removed with the cutoff of 25 % homology. Merging the data of signal peptides and signal anchors, dataset of signal sequences was prepared. The numbers of data of signal peptides, signal anchors and soluble sequences were 917, 103 and 544 for eukaryotes, and 548, 0, 427 for prokaryotes, respectively. Since the

Table 1. SS- and SP-indexes of amino acids for eukaryotes and prokaryotes.

	Eukaryote		Prokaryote	
	SS-index	**SP-index**	**SS-index**	**SP-index**
Ala	2.66	2.04	3.73	2.63
Cys	2.39	1.26	1.70	3.04
Asp	0.76	7.78	0.78	4.13
Glu	0.74	5.82	0.62	2.33
Phe	0.69	0.30	0.80	0.36
Gly	1.33	1.28	0.76	0.58
His	0.87	2.09	0.85	1.62
Ile	0.69	0.30	0.32	0.16
Lys	0.46	3.08	0.46	1.25
Leu	0.90	0.40	0.54	0.30
Met	0.50	0.29	0.48	0.35
Asn	0.63	1.59	0.93	3.63
Pro	1.14	2.62	0.68	1.53
Gln	1.09	6.90	0.93	5.36
Arg	0.68	3.43	0.19	1.09
Ser	1.49	2.21	1.47	2.13
Thr	1.30	1.15	0.99	0.96
Val	1.27	0.58	0.89	0.60
Trp	1.14	0.60	0.72	0.39
Tyr	0.72	0.69	0.51	0.50
Xxx	0	0	0	0

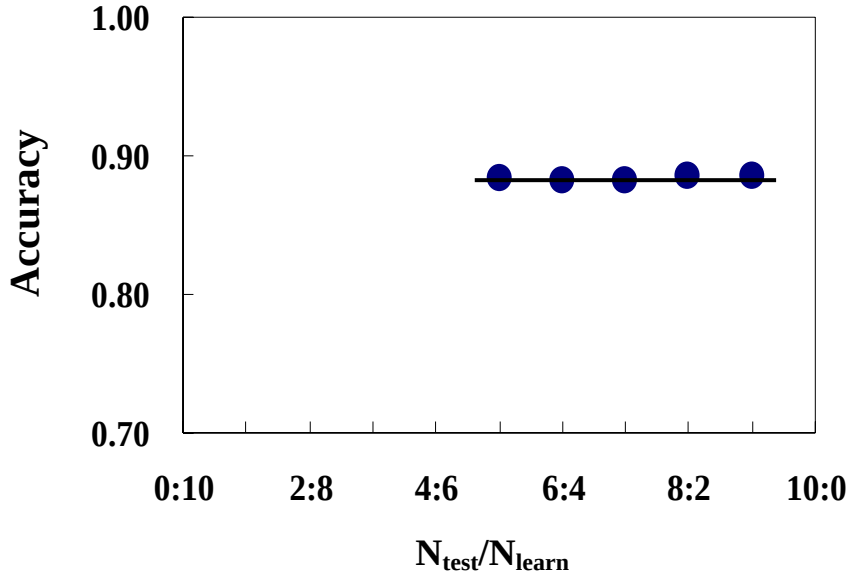

Figure 2. Result of Cross validation test of SOSUIsignal using three kinds of protein data of eukaryotic organisms, signal peptides, signal anchor type II and soluble sequences without signal sequences. Accuracy, the Matthews correlation coefficient, is plotted as a function of the ratio of the evaluation to learning data.

discrimination between signal peptides and signal anchors is essential for the analysis of a total proteome, we included more than 100 data of signal anchors in eukaryotes.

3. Results and Discussions

We defined the SS- and SP-indexes in this work for developing a high performance prediction system of signal peptides. Table 1 shows the values of the SS- and SP-indexes. The numbers for amino acids which are systematically larger than one are colored. The tendency of the SS- and SP-indexes indicates the characteristics of signal sequences and signal peptides. Small residues, Ala, Cys and Ser, are commonly abundant in the signal sequences of eukaryotes as well as prokaryotes. Other small residues such as Gly and Pro are also abundant in eukaryotes. This fact indicates that the clusters of small residues at the C-region of hydrophobic segments are required for signal sequences.

In addition to the clusters of small residues, signal peptides contain strongly polar residues: Asp, Glu, His, Asn, Gln, Lys and Arg. The cluster of this kind of amino acids in the C-region of signal peptides is probably for the direct interaction between signal peptidase and the C-region.

The present system of the signal peptide prediction is successful in discriminating signal peptides not only from non-signal sequences but also from signal anchors. As for eukaryotes, true positive prediction was obtained for 847 data which correspond to 92.4% of 917 signal peptides. The number of false positive prediction was only 20. Therefore, 97.7% of positive prediction was correct. The Matthews correlation coefficient [4] was as high as 0.89 by the self-consistency test. The prediction for prokaryotes was slightly worse than that of eukaryotes, and the Matthews correlation coefficient was 0.79. The reason why the accuracy for prokaryotes was not good

enough is probably due to the lack of data of signal anchors in prokaryotes.

The cross-validation test was also performed, and the average Matthews correlation coefficients were 0.88 for eukaryotes and 0.79 for prokaryotes. Figure 2 shows the results of the cross-validation tests for eukaryotes, in which the average Matthews correlation coefficient of 100 trials are plotted as a function of the ratio between the learning data and the evaluation data. The result shows that the accuracy of the prediction is not dependent on the ratio of data.

The apparent performance of prediction depends on the datasets in general. Our dataset for eukaryotes is larger than the dataset by Menne et al. [6], who tested various signal peptide prediction methods. The accuracy of SignalP v2-HMM was 0.86 in the Matthews correlation coefficient, and the accuracy of other systems (eg. SPScan and SigCleave) was not better than this value. The present system showed better accuracy than these previous methods. Finally, it is noted that this performance is attained by only 5 adjustable parameters and the speed of calculation is fast enough to analyze a proteome in reasonable time.

This work was partly supported by the Grant-in-Aid for Priority Area “Genome Science” from the Ministry of Education, Culture, Science and Sports of Japan and also by a Grant-in-Aid for the 21st Century COE “Frontiers of Computational Science” from the Ministry of Education, Culture, Sport, Science and Technology of Japan.

References

[1] Chou, K.-C., *Protein Eng.*, **14**, 75-79(2001).
[2] Hirokawa, T., Seah, B.-C. and Mitaku, S., *Bioinformatics Applications Note*, **14**, 378-379(1998).
[3] Kyte, J. and Doolittle, R. F., *J. Mol. Biol.*, **157**, 105-132(1982).
[4] Matthews, B. W., *Biochim. Biophys. Acta*, **405**, 442-451(1975).
[5] McGeoch, D. J. *Virus Res.*, **3**, 271-286(1985).
[6] Menne, K. M. L., Hermjakob, H. and Apweiler, R., *Bioinformatics Applications Note*, **16**, 741-742(2000).
[7] Nakai, K. and Kanehisa, M., *Genomics*, **14**, 897-911(1992).
[8] Nakai, K. and Horton, P., *Trends Biochem. Sci*., **24**, 34-36(1999).
[9] Nakai, K., Signal Peptides. In Cell-penetrating peptides. Processes and applications, CRC Press, 295-323(2002).
[10] Nielsen, H., Engelbrecht, J., Brunak S. and von Heijne, G.., *Protein Eng.*, **10**, 1-6(1997).
[11] Nielsen, H. and Krogh, A., *Intell. Syst. Mol. Biol.*, **6**, 122-130(1998).
[12] von Heijne, G., *Nucl. Acids Res.*, **14**, 4683-4690(1986).