Predicting Signal Peptides and Their Cleavage Sites Using Support Vector Machines and Improved Position Weight Matrixes

Jingjing Sun
College of Information Engineering
Xiangtan University
Xiangtan, Hunan, China

Lipo Wang
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
E-mail: elpwang@ntu.edu.sg

Abstract

In this paper, we develop a method for predicting signal peptides and their cleavage sites. Unlike other published work, we divide proteins into two segments and calculate the amino acid compositions on both segments. After that, we hybridize the pseudo amino acid compositions (PseAAs) to the feature vectors. Using support vector machines (SVMs) to train the datasets, we get better results than those with the optimized evidence-theoretic K nearest neighbor (OET-KNN) classifier. The overall rate of correct prediction for signal peptides is over 97%. For identifying cleavage sites, we use the scaled window proposed by Chou to extract cleavable secretory segments and non-cleavable secretory segments and improve the position weight matrix (PWM) method proposed by Hiller et al.. By hybridizing the scaled window and PWM methods, the correct prediction for signal peptides cleavage sites is also better or comparable to other methods.

1. Introduction

Signal peptides have had an immense impact on modern cell biology. They are in charge of the access of almost proteins in eukaryotes and prokaryotes to the secretory pathway [1]. When a cell is divided, a large amount of proteins with various essential functions are being made and must be delivered the right positions. So the knowledge of how signal peptides work becomes very important when understanding the molecular mechanisms or studying new drugs. Signal sequence is often located in the N-terminal part of the protein and cleaved off while the protein is transferred through the membrane. If a sorting signal in a protein is cleaved at a wrong position, the protein could be delivered to a wrong cellular location, leading to varieties of diseases [2]. Therefore how to discriminate secretory proteins from non-secretory proteins and identify their cleavage sites becomes the first challengeable problem we must resolve. However, the length of signal sequences and the order of the residues vary obviously between different proteins so it makes the identification of the cleavage sites more difficult. Nevertheless, they still have some common features (See Figure 1). The most important feature of signal peptides is a series of hydrophobic amino acids called the h-region. It generally consists of seven to fifteen amino acids in length. The n-region, at the upstream of the h-region, has typically one to five amino acids generally carrying positive charges. Between the h-region and the cleavage site is the c-region, which consists of three to seven polar, but mostly uncharged, amino acids. Near to the cleavage site, a more specific subsite of amino acids exists: the residues at positions -3 and -1 (relative to the cleavage site) must be small and neutral for cleavage to occur correctly. For example, Ala often occurs at the last residue and the last third residue of the signal sequence [3]. In addition, since the number of protein sequences in data banks has been rapidly increasing, it is time-consuming and expensively for researchers to identify all of signal peptides entirely by experiments. So an automated method for predicting signal peptides has become a prerequisite.

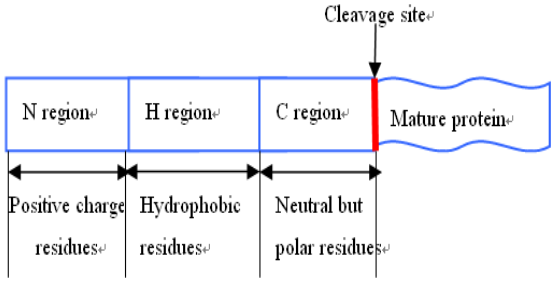

Figure 1. A schematic drawing of a signal peptide

Up to now, many efforts have actually been made to predict the presence of the signal peptides and their cleavage sites. The weight-matrix approach was developed by von Heijne [4], which was used to identify signal peptide cleavage sites. Folz and Gordon [5] implemented the prediction of signal peptides cleavage sites through two different algorithms. The widely used method, SignalP, developed by Nielsen et al. [6] [7] was based on neural network and hidden Markov model algorithms. And its new version, SignalP 3.0 [7], was enhanced greatly by adding other features into the network. PrediSi [8], based on a PWM approach, considered the amino acid bias present in proteins by using a frequency correction. The subsite coupling approach [9] [10] developed by Chou was also important, which was based on the sequence-encoded algorithm [11] and the scaled window approach [12].

This paper is organized as followings. In section 2, we give datasets in our experiment. In section 3, we introduce the methods used in our paper. In this section, we firstly resolve the problem of discriminating secretory proteins from non-secretory proteins by training SVM with protein sequences encoding with amino acid compositions and PseAAs [13]. Secondly, we briefly give an introduction about the SVM classifiers. Thirdly, we use the scaled window [12] and the PWM [8] methods to identify the cleavage sites of protein sequences confirmed as signal peptides by the SVM classifier. Different from the published work, we take the balance between the number of secretory-cleavable segments and non-secretory-cleavable ones into consideration. In the final section, we discuss our results using the methods in our paper and compare them with those from other published work.

2. Datasets

A reliable benchmark data set is very necessary for the assessment of the predictive performance. By comparing with the performance from other published work, we can easily tell the benefits of the methods. The datasets in our experiment are constructed by Chou [14], extracted from the most recent version of Swiss-Prot [15] database 50.7 (released on Sept-19-2006). There each protein contains the first N-terminal 100 amino acids. They consist of three benchmark datasets for eukaryotic, Gram-positive and Gram-negative proteins, respectively. The eukaryotic dataset contains 3302 secretory proteins and 3785 non-secretory proteins; the Gram-positive dataset, 269 secretory proteins and 306 non-secretory proteins; the Gram-negative dataset, 613 secretory proteins and 721 non-secretory proteins. We can get these data from the internet freely.

3. Methods

In our study, we take two steps to treat this problem. Firstly, we discriminate a query protein whether it is a secretory or non-secretory protein. If it is a secretory protein, we will deduce its cleavage site according to the second steps.

3.1. Discriminating secretory proteins from non-secretory proteins

According to the fact that most signal peptides are among the first 50 residues, we simplify the problem by using their first 50 residues as Chou did [14]. That is, given a protein P, we can express it as:

$$P = R_1 R_2 R_3 R_4 .... R_{49} R_{50} \quad (1)$$

where R_1 represents the 1st residue of the protein P, R_2 the 2nd residue, and so forth. To make sure whether the protein P is secretory or non-secretory, we attempt to construct a feature vector to represent it. In the feature vector, the information of amino acid compositions and PseAAs is included. The PseAAs are adopted to reveal the degree of the long distance interactions between residues along the whole sequence [13]. The feature vector is made up of two parts. Before constructing the first part, the sequence is divided into two same length segmentations. By this way, the rate of prediction is improved [16] [17]. Here the amino acid compositions which are the occurrence frequencies of different residues are calculated on both segmentations respectively. Suppose the numbers of different residues appearing in some segmentation are $n_1, n_2, ..., n_{20}$. And then the amino acid compositions vector of the segmentation is defined in equation (2).

$$V_{1i} = [p_1, p_2, ..., p_{20}] \quad where p_k = n_k / L, k = 1, ..., 20; i = 1, 2; \quad (2)$$

L is the length of the protein. After that, the two segmentations are merged together to form a whole sequence V_1,

$$V_1 = [V_{11} V_{12}] \quad (3)$$

where V_{11}, V_{12} mean the occurrence frequencies of the first and second segmentation of the first part. The physicochemical properties of the residues are considered in the second part of the feature vector in order to involve some information about long distance interactions between residues. We use two kinds of physicochemical properties in this paper (listed in Table 1). Because we have tried to add other properties

used by Li et al. [17] into the feature vector, but the rate of prediction is not improved. The values of these properties can be obtained from the amino acid index database (AAindex) [18]. In order to construct the feature vector, we substitute the amino acid residues with the normalized amino acid indexes, using the normalization procedure described by equation (4) to (6). More about the procedure can be found in Chou's hybridization space methods [19] [20] [21].

$$\bar{p}_i = \frac{1}{20}\sum_{k=1}^{20} p_i^{(k)} \quad (4)$$

$$Var(p_i) = \frac{1}{20}\sum_{k=1}^{20} (p_i^{(k)} - \bar{p}_i)^2 \quad (5)$$

$$p^{(k)}{}_{nomal\_i} = (p_i^{(k)} - \bar{p}_i) \Big/ (\sqrt{Var(p_i)}) \quad (6)$$

Table 1. The two physicochemical properties used in this work

Properties description	Reference
Hydrophilicity value	Hopp-Woods (1981)
Consensus normalized hydrophobicity	Eisenberg (1984)

where $p_k^{(i)}$, $1 \le k \le L$ is the i th normalized amino acid index of the k th residue in the sequence. Then using equation (7) (8), we calculate the set of values of auto correlation functions $V_{2j}(j = 1,2)$ for each property, where T is a constant.

$$V_{2j} = [R_j(1), R_j(2), ..., R_j(T)], (j = 1,2) \quad (7)$$

$$R_i(\tau) = 1/(L-\tau)\sum_{k=1}^{L-\tau} p_k^{(i)} p_{k+\tau}^{(i)} \quad (8)$$

where $R_i(\tau)(1 \le \tau \le T)$ is called auto correlation function. Now we get the second part of the feature vector V_2, expressed as equation (9).

$$V_2 = [V_{21} V_{22}] \quad (9)$$

Finally, the complete feature vector V in equation (10) has formed, with a $40 + 2T$ dimensions vector.

$$V = [V_1 V_2] \quad (10)$$

For each element p_u of the feature vector V, we also use the equation (11) proposed by Chou to normalize it. Finally, the new feature vector is used as the input of SVM.

$$p_u = \begin{cases} p_u \Big/ (\sum_{i=1}^{40} p_u + w\sum_{j=1}^{K} R_i(\tau)) & (1 \le u \le 40) \\ w * R_i(\tau) \Big/ (\sum_{i=1}^{40} p_u + w\sum_{j=1}^{K} R_i(\tau)) & (41 \le u \le k) \end{cases} \quad (11)$$

The values of T used in our paper for each group of dataset are listed in Table 2.

3.2. SVM classifiers

The SVM learning system, first proposed by Cortes and Vapnik [26], is based on statistical learning theory. Compared with other machine learning systems, a growing attraction for SVM in bioinformatics has emerged recently and resulted in better performance for many tasks [22][23][24]. Wide applications of them attribute to their many attractive features. They have an ability to jump out from the local minima; they have the acceptable speed and scalability; they can collect effectively the information included in the training set. This is the reason we adopt them here.

In this research, the publicly available LIBSVM software is used [25]. As the depiction above, we make the feature vectors of the proteins as inputs to LIBSVM for training. In the process of training, we choose the radial basis function (RBF) as the kernel function, given by the equation (12). And we cite a grid search approach to find a good parameter combination for C and γ, where C is the cost parameter of SVM and γ is the parameter in RBF kernel function.

$$K(\vec{u}, \vec{v}) = \exp(-\gamma \| \vec{u} - \vec{v} \|^2) \quad (12)$$

In order to save time, we use five-fold cross-validation method to find the best (C, γ). For each group of dataset, the best (C, γ) in our experiment is listed in Table 2. And the results in Table 3 are obtained by the jackknife method with the best (C, γ).

Table 2. The value of T and the best (C, γ) for dataset in the experiment

	T	(C, γ)
Gram-positive	15	(2^9,2^3)
Gram-negative	9	(2^10,2)
Eukaryotic	20	(2^11,2)

3.3. Identifying cleavage sites of signal peptides

In an N-terminal signal peptide, there is only one cleavage site locating at the position between the last residue of the signal peptide and the first residue of the mature protein. Thus once we find the cleavage site, the corresponding signal peptide is identified. The scaled window approach [12] and the improved PWM [8] are adopted for this study. Suppose a window with a scale of $-\xi_1, ..., -3, -2, -1, +1, +2, ..., +\xi_2$. Chou called this window as "scaled window" and symbolized as $[-\xi_1, +\xi_2]$. Segments, which are generated when sliding the scaled window $[-\xi_1, +\xi_2]$ along protein sequences, can be generally expressed as:

$$R_{-\xi_1} R_{-(\xi_1 - 1)} ... R_{-3} R_{-2} R_{-1} R_{+1} R_{+2} ... R_{+(\xi_2 - 1)} R_{+\xi_2} \quad (13)$$

where $R_{-\xi_1}$ represents the residue at the scale $-\xi_1$, R_{-1} the residue at the scale -1, and R_{+1} the residue at the scale $+1$, and so forth. For a sequence consisting of L residues, we can see $\phi = L-(\xi_1+\xi_2)+1$ different sequences by the rule depicted above. Among the ϕ sequence segments generated from the same protein, only the one is deemed as the secretion-cleavable segment, while all the other segments deemed as non-secretion-cleavable segments. All the secretion-cleavable segments form a positive set symbolized by S^+, and all the non-secretion-cleavable segments form a negative set S^-. Unlike what PrediSi [8] did, the set S^- here is derived from only secretory proteins, because we have discriminated them from non-secretory proteins in the first step. In addition, we have to note that the set S^- is far larger than S^+, following the procedure above. Considering the unbalance between S^+ and S^-, we divide S^- into ϕ parts. That is, all the segments which have the same distance from N-terminal of proteins are gathered in the same group. Therefore, we can get $(\phi+1)$ PWMs. Now the problem is, for a query protein, how to determine the one that is cleavable among its ϕ segments. To realize this, for each of the ϕ segments, we can define a score function similar to PrediSi [8], given by:

$$S_k = \sum_{j=-\xi_1}^{\xi_2} \log(P_{S^+}^j(R_j)\frac{P^0}{P_{S_k^-}^j(R_j)}) \quad (k=1,...,\phi) \qquad (14)$$

where $P_{S^+}^j(R_j)$ is the probability of amino acid R_j emerging at the position j $(j=-\xi_1,...,+\xi_2)$ for the secretion-cleavable segments and $P_{S_k^-}^j(R_j)$ the corresponding probability for the non-secretion-cleavable segment in S_k^-. The values of $P_{S^+}^j(R_j)$ can be calculated from a positive training dataset S^+, consisting of only secretion cleavable segments, and the values of $P_{S_k^-}^j(R_j)$ can be calculated from a negative training dataset $S_j^-(j=1,...,\phi)$, consisting of only non-cleavable segments. And P^0 is a frequency correction factor to adjust the amino acid bias proposed by PrediSi [8] (it is set to 0.05, the random frequency for each of the 20 native amino acids). This formulation in equation (13) is the same with the PWM [8] expect for the expression of $P_{S_k^-}^j(R_j)$. Thus according to the score of equation (13), the cleavage site of the query protein is predicted. It locates at the site with which its score is the highest. However, in the process of experiment, some protein sequences contain non-basic amino acids, because of failing to observe which amino acid residue occurs at this position. In this condition, we set all non-basic amino acids as the 21st amino acid. We also find some values of $P_{S^+}^j(R_j)$ or $P_{S_k^-}^j(R_j)$ equal to zero. Considering that they will affect the experiment results, so we set them equal to 0.05amino acid under this condition.

Table 3. Success rate comparison by both the self-consistency test and jackknife test on the datasets

Test method	Dataset	Success rate in discriminating secretory from non-secretory proteins			Success rate in identifying signal peptide cleavage site				
		PrediSi [8] (%)	Signal-CF [14] (%)	Ours (%)	PrediSi [8] (%)	Signal-CF [14] (%)	Ours (%)		
							(13,2)	(14,3)	(14,2)
Self-consistency	Gram-positive	97.9	99.3	99.0	74.4	82.2	77.3	80.7	-
	Gram-negative	97.2	99.6	99.5	84.2	89.6	87.9	88.1	87.4
	Eukaryotic	98.3	99.4	99.2	71.9	78.1	79.6	79.4	-
Jackknife	Gram-positive	94.6	98.1	98.6	60.2	67.3	67.3	66.2	-
	Gram-negative	91.2	94.4	97.2	80.3	84.2	84.0	83.4	84.7
	Eukaryotic	92.5	94.5	98.3	70.8	75.4	78.2	77.9	-

4. Results and discussion

To discuss the performance of the methods used in this paper, the comparison with PrediSi [8] and Signal-CF [14] is made. In order to compare with them, we also adopt two test methods, Self-consistency and Jackknife. In the self-consistency test, the training data and the testing data are the same one. In the jackknife test, each protein in the benchmark dataset is extracted out in turn as a "test protein" and all the remaining proteins are treated as the train proteins. In table 3, the results with the method of PrediSi [8] were calculated by Chou to compare with them under the same datasets. This is important in this area for prediction tasks. So we perform our methods on the same datasets. We can see our results are better than PrediSi [8] whether in

discrimination secretory proteins from non-secretory proteins or identification cleavage sites. For discriminating secretory proteins from non-secretory proteins, Signal-CF [14] used the PseAA model and the OET-KNN classifier. In our methods, before using the PseAA model, we firstly divide the sequences into two segments. By this way, we find the results are better than those from only treating the protein as the whole one and train the datasets with the SVM classifier. Our results are slightly lower than Signal-CF [14] when testing by the self-consistency, but higher than those when using the jackknife test on this problem. It is known that Matthews's correlation coefficient is a widely used and useful tool for measuring the performance of classification problems in bioinformatics. So we also calculate them on the given datasets respectively. And it is 0.94 for Gram-negative, while it reaches as high as 0.97 for Gram-positive and Eukaryotic. As for the identification signal peptide cleavage sites, our result is the best on the Eukaryotic dataset. The reason that our results are not high is our cutting all protein sequences length as 50 and the limitation width of windows, as a result some signal peptides are excluded and we consider them as wrong prediction. In spite of these limitations, our results can compare with the signal-CF [14] results. As other methods have found, we also find the best results are at where the window is (13, 2) or (14, 3). For gram-negative dataset, (14, 2) is the best window (See Table 3). During the process of experiment, we also find the differences between the cleavage sites that have higher scores are very small and most of cleavage sites are among these first ones. This encourages us to extract more physicochemical features to improve the rate of identifying the cleavage sites in the future work.

5. References

[1] L.M. Gierasch, "Signal sequences", Biochemistry 28, 1989, pp.923–930.

[2] Kuo-Chen Chou, "Prediction of Protein Signal Sequences", Current Protein and Peptide Science 3, 2003, pp.615-622.

[3] Gunnar von Heijne, "Signal sequences: The limits of variation", J. Mol. Biol., 1985, 184:pp.99–105.

[4] Gunnar von Heijne, "A new method for predicting signal sequence cleavage sites", Nucleic Research (14), 1986, pp.4683-4690.

[5] R. Folz and J. Gordon, "Computer-Assisted Predictions of Signal Peptidase Processing Sites", Biochem. Biophys. Res. Commun., vol. 146, 1987, pp.870-877.

[6] Nielsen et al., "Improved prediction of signal peptides--SignalP 3.0", J. Mol. Biol., 2004.

[7] Henrik Nielsen et al., "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites", Protein Engineering, vol.10 (1), 1997, pp.1–6.

[8] K. Hiller, A. Grote, M. Scheer, R. Munch, D. Jahn, "PrediSi: Prediction of signal peptides and their cleavage positions", Nucleic Acids Res. 32, 2004, W375–W379.

[9] Chou,K.C., "Using subsite coupling to predict signal peptides", Protein Engineering, 14, 2001, pp.75-79.

[10] Hui Liu, Jie Yang, Jian-Guo Ling, Kuo-Chen Chou, "Prediction of protein signal sequences and their cleavage sites by statistical rulers", Biochemical and Biophysical Research Communications 338, 2005, pp.1005–1011.

[11] K.C. Chou, "Prediction of protein signal sequences and their cleavage sites", PROTEINS: Structure, Function, and Genetics 42, 2001, pp. 136-139.

[12] K.C. Chou, "Prediction of signal peptides using scaled window peptides", peptides 22, 2001, pp.1973-1979.

[13] K.C. Chou, "Prediction of protein cellular attributes using pseudo amino acid composition", PROTEINS: Structure, Function, and Genetics 43, 2001, pp.246–255 (Erratum: 2001, vol. 44, 60).

[14] Kuo-Chen Chou, Hong-Bin Shen, "Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides", Biochemical and Biophysical Research Communications 357, 2007, pp.633–640.

[15] Bairoch, A., Apweiler, R., "The SWISS-PROT protein sequence data bank and its supplement TrEMBL", Nucleic Acids Res. 2000, 25, pp.31-36.

[16] J.-Y. Shi, S.-W.Zhang, Q. Pan, and G.-P. Zhou, "Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution", Amino Acids, 2008.

[17] Pufeng Du and Yanda Li, "Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence", BMC Bioinformatics, 2006, 7:518.

[18] Kawashima S, Ogata H, Kanehisa M: AAindex: amino acid index database. Nucleic Acids Research 2000, 28:374.

[19] Chou K-C, Cai Y-D, "Predicting of protease type in a hybridization space", Biochemical and Biophysical Research Communications 2006, 339:pp.1015-1020.

[20] Chou K-C, Cai Y-D, "Predicting protein-protein interactions from sequence in a hybridization space", Journal of Proteome Research 2006, 5:pp.316-322.

[21] Chou K-C, Cai Y-D, "Predicting enzyme family class in a hybridization space", Protein Science 2004, 13:pp.2857-2863.

[22] X.-D. Sun, R.-B., Huang, "Prediction of protein structural classes using support vector machines", Amino Acids, 2006, 30: pp.469–475.

[23] Chao Chen, Xibin Zhou, Yuanxin Tian, Xiaoyong Zou, Peixiang Cai, "Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network", Analytical Biochemistry 357, 2006, pp.116–121.

[24] Yu-Dong Cai, Shuo-liang Lin, Kuo-Chen Chou, "Support vector machines for prediction of protein signal sequences and their cleavage sites", Peptides 24, 2003, pp.159–161.

[25] C.C. Chang, C.J. Lin, “LIBSVM: A Library for Support Vector Machines” [software], 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[26] C. Cortes, V. Vapnik, Support-vector networks, Mach.Learn.20, 1995, pp.273–297.