Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 492174, 8 pages
doi:10.1155/2012/492174

Research Article

Signal-BNF: A Bayesian Network Fusing Approach to Predict Signal Peptides

Zhi Zheng,[1] Youying Chen,[1] Liping Chen,[1] Gongde Guo,[1] Yongxian Fan,[2] and Xiangzeng Kong[1,3]

[1] *Key Laboratory of Network Security and Cryptology, Fujian Normal University, Fuzhou 350007, China*
[2] *Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China*
[3] *School of Computing and Mathematics, University of Ulster at Jordanstown, Newtownabbey BT37 0QB, UK*

Correspondence should be addressed to Xiangzeng Kong, xzkongfjnu@sohu.com

Received 19 April 2012; Revised 9 September 2012; Accepted 9 September 2012

Academic Editor: George Perry

A signal peptide is a short peptide chain that directs the transport of a protein and has become the crucial vehicle in finding new drugs or reprogramming cells for gene therapy. As the avalanche of new protein sequences generated in the postgenomic era, the challenge of identifying new signal sequences has become even more urgent and critical in biomedical engineering. In this paper, we propose a novel predictor called Signal-BNF to predict the N-terminal signal peptide as well as its cleavage site based on Bayesian reasoning network. Signal-BNF is formed by fusing the results of different Bayesian classifiers which used different feature datasets as its input through weighted voting system. Experiment results show that Signal-BNF is superior to the popular online predictors such as Signal-3L and PrediSi. Signal-BNF is featured by high prediction accuracy that may serve as a useful tool for further investigating many unclear details regarding the molecular mechanism of the zip code protein-sorting system in cells.

1. Introduction

Signal peptides which are usually *N*-terminal extensions with 3–60 amino acids long direct proteins to their corresponding cellular and extracellular localizations. We treated its function as an "address tag" or "zip code." If the signal sequence in a nascent protein was changed, the protein could end up in a wrong cellular location causing various weird diseases.

The advent of signal peptides predictor has a significant impact on developing novel strategies for drug discovery, as well as for revealing the molecular mechanisms of some genetic diseases (refer a review [1]). Faced with the avalanche of new protein sequences emerging in the postgenomic era, to timely use them for basic research and drug discovery [2, 3], it is highly desirable to develop the fast and accurate algorithms to identify the signal sequences and predict their cleavage sites. Actually, many efforts have been made in this regard [4–17]. Based on different kinds of characteristics, several machine learning approaches have been proposed for this task, such as neural networks [8–11], hidden Markov models [12], and support vector machines [13–15]. Recently, Shen and Chou developed two algorithms based on evidence theory to predict the signal sequences and achieve favorable results [4, 5].

In this paper, we propose a novel predictor based on Bayesian learning algorithm to predict the *N*-terminal signal peptides and their cleavage sites. Bayesian learning algorithm has been previously applied in a number of other bioinformatics problems [18–23], such as protein-protein interactions. But their approaches are not designed to deal with the *N*-terminal signal peptide sequences and the amino acid preference at the cleavage sites [24]. Fundamentally differed from theirs, the Bayesian network is a method of statistical inference in which some kind of evidence or observations are used to calculate the probability if a hypothesis may be true, which is particularly suited for this task. Its advantage lies in that there are significant statistical preferences of different amino acids along the signal peptides mentioned in the previous studies [4, 5].

The integration system which was built by multiple base classifiers has a stronger generalization ability than a single

Table 1: Number of the secretory and nonsecretory proteins in each of the six different organism datasets.

Organism	Number of secretory proteins	Number of non-secretory proteins	Total
Human	894	1129	2203
Plant	338	559	897
Animal	1435	1762	3197
Eukaryotic	635	785	1420
Gram-positive	269	356	625
Gram-negative	613	721	1334

Table 2: Sampling proportion of S^- in each of the six different organism datasets.

Organism	$\lvert S^+\rvert$	$\lvert S^-\rvert$	$\lvert S^+\rvert : \lvert S^-\rvert$	Sampling proportion of S^-
Human	894	172047	1 : 192	1/20
Plant	338	76447	1 : 226	1/23
Animal	1435	268720	1 : 187	1/19
Eukaryotic	635	117089	1 : 184	1/19
Gram-positive	269	52940	1 : 196	1/20
Gram-negative	613	113314	1 : 184	1/19

good classifier. So we use integration system to improve the prediction accuracy. First, base classifier is built by using different feature datasets as Bayesian network input. Then, the ultimate result is fused by the results of different Bayesian classifiers through weighted voting system.

The experimental results show that Signal-BNF is superior to two other popular signal peptide predictors of Signal-3*L* [4] and PrediSi [25]. So, the approach we proposed is quite promising.

2. Materials and Methods

The datasets constructed in [4] were adopted in this paper, which contain the secretary proteins and the nonsecretary proteins from six different species. It was human, plant, animal, eukaryotic, Gram-positive, and Gram-negative (refer Table 1).

Signal peptide sequences are usually *N*-terminal extensions although they can also be located within a protein or at its *C*-terminal end. It will be cleaved off by a signal peptidase when the protein goes through a membrane. The cleavage site is the position between the last residue of the signal peptide sequence and the first residue of the mature protein. It is symbolized by $(-1, +1)$ (Figure 1). The signal peptide sequences of different secretory proteins are quite different in sequence components and orders. And they all have different sequence length. Figure 2 shows the length distribution of the signal peptides in the six species secretory proteins.

Since different proteins differ in the length of the signal peptide, we introduced the concept of scaled window to solve the difficulty in predicting the signal peptide for a general algorithm. The scaled window approach has been adopted for this study before [6].

The scaled window which is symbolized as $[-\xi_1, +\xi_2]$ is marked consecutively with $-\xi_1, \ldots, -2, -1, +1, +2, +\xi_2$ to define the corresponding position of amino acids of a protein sequence within the window. In this way, a segment can be used as a "benchmark window" to search the secretion-cleavable site along a protein sequence and can deduce its signal peptide accordingly. Only the one with the residue at the scale -1 being the very last residue of the signal sequence and the residue at the scale $+1$ being the first residue of the mature sequence are regarded as the secretion-cleavable segment (refer Figure 3(a)), while all the other segments are regarded as nonsecretion cleavable (refer Figures 3(b) and 3(c)).

For a $[-\xi_1, +\xi_2]$ protein segment sequence P can be generally expressed as

$$P = R_{-\xi_1} R_{-\xi_1+1} \cdots R_{-1} R_{+1} \cdots R_{\xi_2-1} R_{\xi_2}, \quad (1)$$

where $R_{-\xi 1}$ represents the amino acid residue at the position $-\xi_1$, R_{-1} represents the amino acid residue at the position -1, $R_{+\xi 2}$ represents the amino acid residue at the position $+\xi_2$, and so forth.

The whole prediction task is composed of two steps: (1) identifying whether a protein is secretory or not and (2) determining the signal peptide cleavage site for a secretory protein. In this study, we choose $\xi_1 = 13$, $\xi_2 = 2$ as the size of scaled window for predicting the cleavage site, which is demonstrated optimal in previous studies [6]. By sliding such a "window" along each of these protein sequences, we obtained 6 corresponding training datasets for the 6 species. It is important to point out that, for a secretory protein sequence of length $L1$, we can obtain $L1 - (\xi_1 + \xi_2) + 1$ different sequence segments. But in these segments only one secretion-cleavable segment, the others are nonsecretion cleavable segments. While a nonsecretory protein sequence of length $L2$ can obtain $L2 - (\xi_1 + \xi_2) + 1$ different sequence segments which are all nonsecretion cleavable segments. The one secretion-cleavable segment called positive sample and the other nonsecretion cleavable segments called negative sample. All the secretion-cleavable segments, namely, positive samples, denoted by S^+ and all the nonsecretion cleavable segments, namely, negative samples, denoted by S^-. Apparently, the scaled window approach causes the samples extreme imbalance. Hence, we take a random sampling process in the negative subset, which can relatively reduce the imbalance phenomena. The sampling proportion of S^- refers to Table 2.

As we known, most data classification techniques require the numeric discrete feature vectors as input. It means that the amino acid symbol should be replaced by the decimal integer, such as the local physicochemical properties. Due to that we need different feature datasets as different classifiers' input, we gain the different datasets through different coding schemes. In this paper, three different coding schemes (subsystems [26]) are adopted.

The first subsystem considers that each position in the scaled window has 21 possible values (20 amino acids and

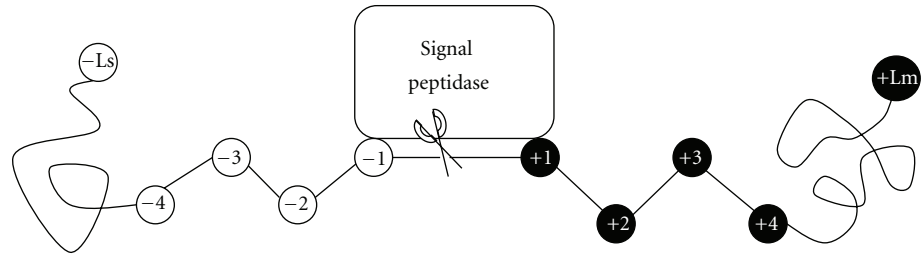

Figure 1: A schematic drawing shows the signal sequence of a protein and how it is cleaved by the signal peptidase. An amino acid in the signal part is depicted as a white circle with a black number to indicate its sequential position, while in the mature protein depicted as a black circle with a white number. The signal sequence contains Ls residues and the mature protein Lm residues. The cleavage site is at the position $(-1, +1)$, that is, between the last residue of the signal peptide sequence and the first residue of the mature protein.

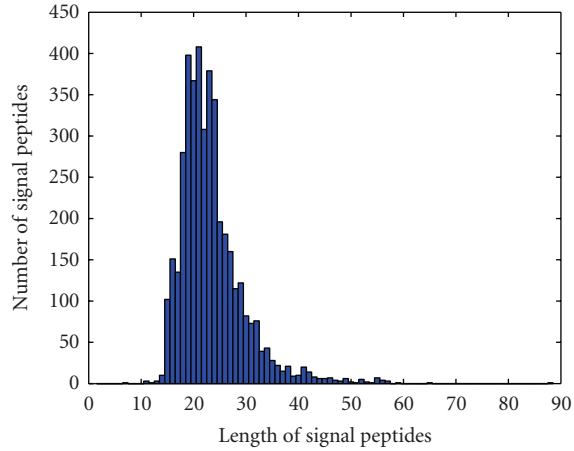

Figure 2: A histogram to show the distribution of signal peptides with their length in the 4,184 secretory proteins constructed in this study. Of the 4,184 proteins, 894 humans, 338 plants, 1,435 animals, 635 eukaryotic, 269 Gram-positives, and 613 gram-negatives.

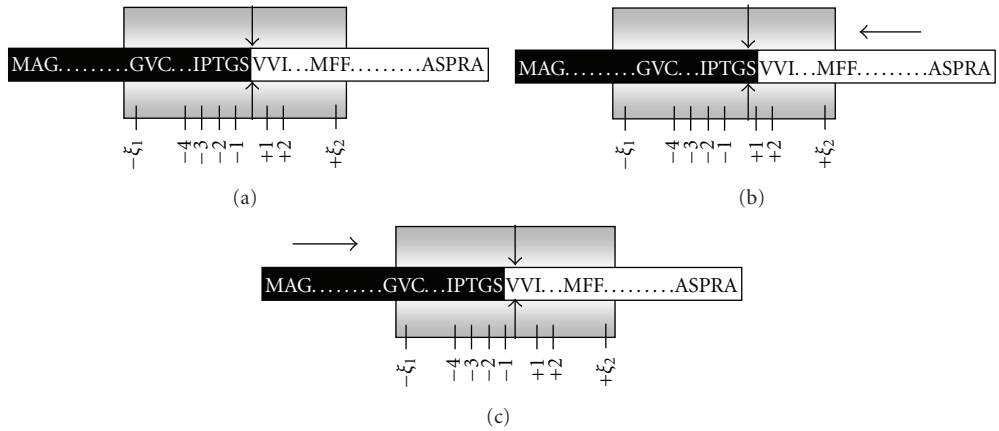

Figure 3: Illustration to show the sequence segments highlighted by sliding the scaled window $[-\xi_1, +\xi_2]$ along a protein sequence. During the sliding process, the scales on the window are aligned with different amino acids so as to define different peptide segments. When the scale -1 is aligned with the tail residue of the signal sequence and scale $+1$ aligned with the head residue of the mature protein as shown in (a), the peptide segment is seen within the window regarded as secretion cleavable. Peptide segments seen within the window for all the other cases, such as those shown in ((b) and (c)), are regarded as nonsecretion cleavable.

Table 3: Amino acid's integers encode ranging from 1 to 21 indicators of the first subsystem.

Amino acid title	Abbreviation title	Alphabet	Decimal code
Alanine	Ala	A	1
Cystine	Cys	C	2
Aspartic acid	Asp	D	3
Glutamic acid	Glu	E	4
Phenylalanine	Phe	F	5
Glycine	Gly	G	6
Histidine	His	H	7
Isoleucine	Ile	I	8
Lysine	Lys	K	9
Leucine	Leu	L	10
Methionine	Met	M	11
Asparagine	Asn	N	12
Proline	Pro	P	13
Glutarnine	Gln	Q	14
Arginine	Arg	R	15
Serine	Ser	S	16
Threonine	Thr	T	17
Valine	Val	V	18
Tryptophan	Trp	W	19
Tyrosine	Tyr	Y	20
Null	Null	Null	21

Table 4: Properties of amino acid residues of the second subsystem: 1: hydrophobic, 2: positive, 3: negative, 4: polar, 5: charged, 6: small, 7: tiny, 8: aliphatic, 9: aromatic, and 10: proline.

Amino acid alphabet	1	2	3	4	5	6	7	8	9	10	Decimal code
A	y					y	y				536
C	y					y					528
D			y	y	y	y					240
E			y	y	y						224
F	y								y		514
G	y					y	y				536
H	y	y		y	y				y		866
I	y							y			516
K	y	y		y	y						864
L	y							y			516
M	y										512
N				y		y					80
P						y				y	17
Q				y							64
R		y		y	y						352
S				y		y	y				88
T	y			y		y					592
V	y					y		y			532
W	y			y					y		578
Y	y			y					y		578

a null input). Hence, it uses an integer ranging from 1 to 21 indicators (refer Table 3), which is taken as the input of Signal-BNF, to present each amino acid.

The second subsystem deems that each amino acid is associated with 10-bit binary (i.e., value 0 or 1) indicators to represent its multiview properties. Each row in Table 4 shows that an amino acid can have multiple properties. And "*y*" means the amino acid has the property. If there is "*y*," the value is 1, otherwise 0. Then, the binary is converted to a decimal integer to represent each amino acid.

The last subsystem represents the relative hydrophobic value of amino acids with 3-bit binary indicators. In Table 5, each amino acid has been encoded into decimal integer.

Therefore, we received three different feature datasets according to the above subsystems.

2.1. Bayesian Networks. The term "Bayesian networks" was coined by Judea Pearl [18] in 1985, its theory, algorithms and applications can be found in [19–24]. A Bayesian network [27, 28], which is a kind of learning machine, encodes the joint probability distribution of a set of variables $\{x_1, \ldots, x_v\}$ as a directed acyclic graph and a set of conditional probability tables (CPTs). The probability of an arbitrary event $X = (x_1, \ldots, x_v)$ can be computed as

$$P(X) = \prod_{i=1}^{v} P(x_i \mid \pi_i) \tag{2}$$

Table 5: Relative hydrophobic value of amino acids of the third subsystem.

Amino acid alphabet	Polar	Neutral	Hydrophobic	Decimal code
A		y		2
C			y	1
D	y			4
E	y			4
F			y	1
G		y		2
H		y		2
I			y	1
K	y			4
L			y	1
M			y	1
N	y			4
P		y		2
Q	y			4
R	y			4
S		y		2
T		y		2
V			y	1
W			y	1
Y		y		2

Here π_i is the set of parents of x_i. Given a training set $D = \{X_1, \ldots, X_d, \ldots X_n\}$, where $X_d = (x_{d,1}, \ldots, x_{d,v})$, the goal of learning is to find the Bayesian network that best represents the joint distribution $P(x_{d,1}, \ldots, x_{d,v})$. In other words, when the Bayesian network is unknown we need to learn it by estimating the network structure and the parameters of the joint probability distribution from the training data and prior information.

We assume no missing data, then attention the problem on learning network structure. At present, there are mainly two kinds of Bayesian network learning methods [27]: conditional-independence-test-based method and search-based method. The conditional independence test is very sensitive of the error. And condition independence test times relative to the number of variables to increase exponentially in some cases. Search-based algorithm can search for the accurate and complete network structure, but the structure space is very large. Search the best Bayesian network structure from all possible network structure space is a *NP*-hard problem, so the commonly used method is heuristic algorithm. The widely used and the most representative heuristic algorithm is $K2$ algorithm which is a famous score-based algorithm.

Learning model structures from data is important for the construction of Signal-BNF. We have empirically compared the behavior of some Bayesian network classifiers base on Bayes Net in Weka [29] and base on Bayes Net Toolbox (BNT) in Matlab [30] over six datasets. In this paper, we use the $K2$ structure learning algorithm which performs relatively better than others. It maximizes the scoring measure of marginal likelihood. $K2$ is a greedy search algorithm which applies a known ordering of the nodes and the maximum limit on the number of parents for any node to constrain the search over network structure.

Followed by the network structure learning, the parameter learning is another important step, and we use the Bayesian estimation method for determining the related parameters. By doing this, we can get a Bayesian network that can be used to make inferences.

2.2. Classify the Secretory-Cleavable Peptides from Non-Secretory Cleavable Peptides by Base Classifier. Suppose a training set S of N samples $(S_1, S_2, \ldots, S_N)$ that can be separated into two subsets: S^+ consists of the secretion-cleavable peptides only and S^- the nonsecretion cleavable peptides only. We used Signal-BNF to distinguish secretion-cleavable peptides from nonsecretion cleavable. Through the Signal-BNF classifier the CPTs can be obtained, as formulated by

$$\rho\left(S_i, S^\theta\right) \quad (i = 1, 2, \ldots, N; \theta \in \{+, -\}), \tag{3}$$

where $\rho(S_i, S^\theta)$ mean the probability of the sample S_i belongs to the class S^θ. The criterion of predicting the secretion cleavability for a given peptide sequence can be formulated as follows:

$$\gamma(S_i) = \begin{cases} 1, & \rho(S_i, S^+) > \rho(S_i, S^-) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The sample is secretion-cleavable peptide if $\gamma(S_i) = 1$, otherwise is nonsecretion cleavable. If the sample is identified as secretion cleavable, it will be continued to predict the cleavage site.

2.3. Classify the Secretory-Cleavable Peptides from Non-Secretory Cleavable Peptides and Identify the Signal Peptide Cleavage Site of Secretory Proteins by Fusion Classifier System. The cleavage site is the position between the last residue of the signal peptide sequence and the first residue of the mature protein. Signal peptide can be automatically determined, while cleavage site is identified. We use Bayesian classifier to predict cleavage. But the result of Bayesian classifier may contain false result. To compensate for this error as much as possible, we consider to composite the base classifiers together. By the above three coding schemes, different feature datasets and Bayesian network constitute base classifiers. Then, the base classifiers fuse as Signal-BNF to predict the cleavage site.

The composite approach for classifying proteins has been used in previous study [31]. From the literature [32], we know that multiple classifier systems can be divided into three structures: cascade, parallel, and hierarchical (refer Figure 4). In cascade system, the result of base classifier directly depends on the success classification of the previous base classifier. The overall system error of this type classification system is the accumulation of each base classifier error. In other words, the error which previous classifier produced is unrecoverable. The parallel system, which each base classifier independently produces results, integrates the results of base classifier by decision logic. As long as the decision logic cleverly designed, you can get more satisfactory results. Hierarchical system is the combination of cascade system and parallel system. So we use the integrated classification model as shown in Figure 5 in the fusion stage.

Furthermore, we use voting as the decision-making method in integration of multiple classifier outputs. Generally, voting includes the weighted voting and the majority voting which has three decision methods: unanimity, simple majority and plurality. In this paper, we use weight voting which can obtain better accuracy to decide which candidate wins.

In fusion stage, discrimination of secretion-cleavable peptide from non-secretory cleavable can be formulated by

$$\delta(S_i) = \sum_{u=1}^{3} w_u \gamma_u(S_i) \quad (i = 1, 2, \ldots, N). \tag{5}$$

Here, $u \in \{1, 2, 3\}$ represent different classifiers. w_u is the weight of each base classifier. If $\gamma_2(S_i)$ equals to $\gamma_3(S_i)$, the weight is $w_1 = 0$, $w_2 = 1$, $w_3 = 0$, otherwise $w_1 = 1$, $w_2 = 0$, $w_3 = 0$. If $\sigma(S_i, S^\theta) = 1$, the sample is secretion-cleavable peptide, otherwise is nonsecretion cleavable peptide.

Then, we can continue to predict the cleavage site. As the protein has been cut into many segments, we have the starting position of each secretion-cleavable peptide in a protein, as formulated below:

$$\{\kappa_i\}, \quad (i = 1, 2, \ldots, N^+), \tag{6}$$

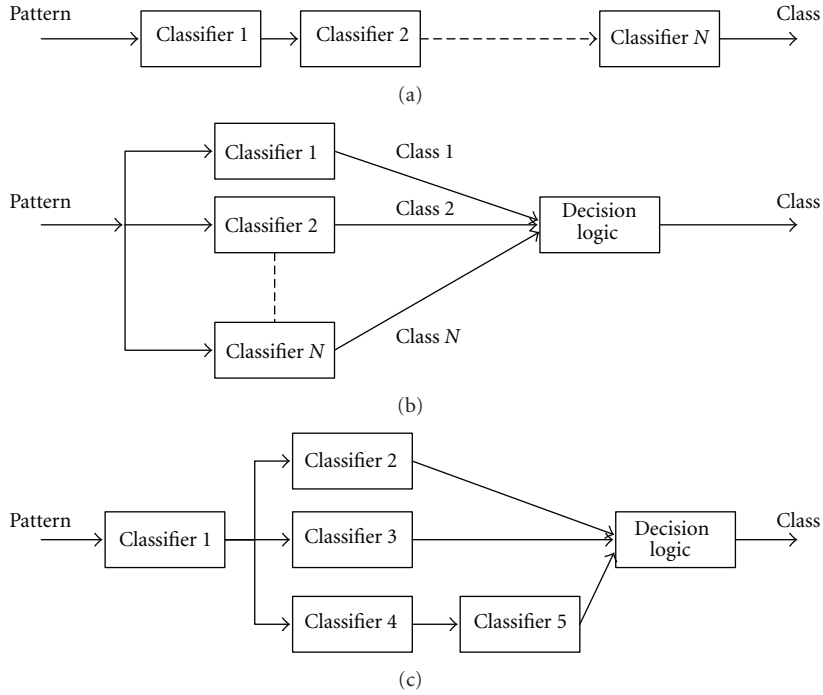

Figure 4: Structures of multiple classifier systems. (a) Cascading, (b) parallel, and (c) hierarchy.

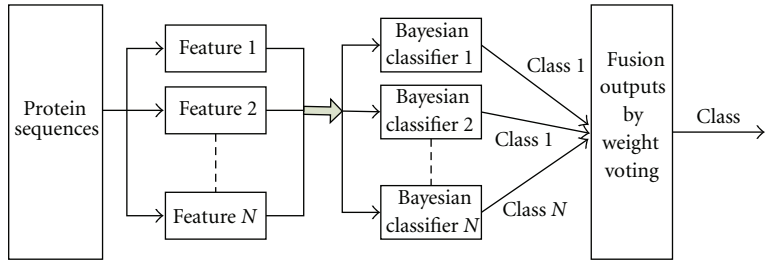

Figure 5: Integrated classification model used in this paper.

where N^+ represents the number of the set S^+, namely, the number of secretary proteins. The cleavage site position of a secretary protein is formulated as

$$\left\{\chi_i \mid \chi_i = \kappa_i + 12\right\}, \quad (i = 1, 2, \ldots, N^+) \tag{7}$$

3. Results and Discussion

The methods frequently used for cross-validating the accuracy of classifier in statistical prediction cover the single independent dataset test, sub-sampling test, and jackknife test. In this study, the 5-fold cross-validation test was performed on Signal-BNF.

Table 6 compares the accuracy of some Bayesian network classifiers in weka. Table 7 compares the accuracy of some Bayesian network classifiers in Matlab. From them we can conclude that $K2$ structure learning algorithm is performed relatively better than others.

Table 8 lists the prediction accuracy for secretory proteins from nonsecretory proteins by three subsystems and fusion system. Table 9 compares our approach's prediction accuracy for secretory proteins from nonsecretory proteins to other approaches. Table 10 lists the prediction accuracy for the cleavage sites by three subsystems and fusion system. Table 11 compares our approach's prediction accuracy for the cleavage sites to other approaches.

Table 6: The accuracy of some Bayesian network classifiers in weka by first subsystem.

Organism	$K2$ (%)	HC (%)	Simulated annealing (%)	Tabu search (%)
Human	95.91	95.91	95.91	95.91
Plant	95.77	95.58	95.58	95.58
Animal	96.34	96.34	96.34	96.34
Eukaryotic	94.87	94.87	94.87	94.87
Gram-positive	95.37	95.16	95.16	95.16
Gram-negative	97.26	97.26	97.26	97.26

Table 7: The accuracy of some Bayesian network classifiers in Matlab by first subsystem.

Organism	TAN (%)	*K*2 (%)	GS (%)
Human	94.20	96.78	96.88
Plant	94.01	96.85	96.72
Animal	95.75	97.12	97.16
Eukaryotic	94.19	95.92	95.98
Gram-positive	93.30	96.29	96.29
Gram-negative	96.30	97.99	97.99

Table 8: The prediction accuracy for secretory proteins from nonsecretory proteins by three subsystems and fusion system.

Organism	First subsystem (%)	Second subsystem (%)	Third subsystem (%)	Fusion (%)
Human	97.17	96.77	96.34	**97.73**
Plant	96.68	96.54	95.85	**97.46**
Animal	97.89	97.48	96.20	**98.18**
Eukaryotic	95.97	95.91	95.04	**96.80**
Gram-positive	95.53	90.49	92.70	**96.23**
Gram-negative	97.79	97.17	96.28	**98.11**

Table 9: Compare our approach's prediction accuracy for secretory proteins from nonsecretory proteins to other approaches.

Organism	PrediSi (%)	Signal-3*L* (%)	Signal-BNF (%)
Human	91.1	92.3	**97.73**
Plant	93.6	95.8	**97.46**
Animal	93.2	95.7	**98.18**
Eukaryotic	92.1	94.0	**96.80**
Gram-positive	94.6	**98.1**	96.23
Gram-negative	91.2	94.4	**98.11**

From Table 8, we can clearly conclude that the fusion system can complement the shortage of each base classifier to improve prediction accuracy. Similar results can also be observed from Table 10.

The comparison performances of the other two popular predictors of Signal-3*L* [4] and PrediSi [25] are listed in Tables 9 and 11. From Table 9, where the success rates of Signal-BNF is 1.63–6.91% higher than PrediSi [25] and 1.66–5.43% higher than Signal-3*L* except Gram-positive dataset. Signal-BNF achieves the best prediction accuracy when discriminating the cleavage sites which can be observed in Table 11. The success rates of Signal-BNF is 11.67–22.9% higher than PrediSi [25] and 3.84–17.5% higher than Signal-3*L*. These results indicate that the Signal-BNF can get a better prediction accuracy of the signal peptide sequences and their cleavage sites

Efficiently prediction of *N*-terminal signal peptides and their cleavage sites is important to both basically research and drug discovery. In this paper, we have proposed a novel

Table 10: The prediction accuracy for the cleavage sites by three subsystems and fusion system.

Organism	First subsystem (%)	Second subsystem (%)	Third subsystem (%)	Fusion (%)
Human	89.44	89.21	86.52	**90.90**
Plant	84.78	85.37	83.28	**87.16**
Animal	91.85	90.87	85.85	**92.47**
Eukaryotic	84.25	84.57	81.26	**86.14**
Gram-positive	80.75	71.70	73.58	**82.64**
Gram-negative	90.82	89.84	87.87	**91.97**

Table 11: Compare our approach's prediction accuracy for the cleavage sites to other approaches.

Organism	PrediSi (%)	Signal-3*L* (%)	Signal-BNF (%)
Human	68.0	73.4	**90.90**
Plant	70.1	82.8	**87.16**
Animal	71.9	77.7	**92.47**
Eukaryotic	65.7	76.2	**86.14**
Gram-positive	60.2	78.8	**82.64**
Gram-negative	80.3	88.1	**91.97**

Bayesian learning network approach named Signal-BNF to reach this goal. The experimental results also reveal that Signal-BNF can achieve the better prediction accuracy than other popular predictors. So we say that Bayesian networks can be a powerful computational tool for predicting signal peptide cleavage sites. The experiment also shows that fusing multiple predictors can provide effective complementarities among them for predicting *N*-terminal signal peptides since different algorithms have their own merits and shortcomings.

Acknowledgments

This work was supported by funds from China's Fujian Province Department of Education Category A Project no. JA10064 and Fujian Province Department of Science and Technology Category K Project no. JK2011007.

References

[1] K.-C. Chou, "Prediction of protein signal sequences," *Current Protein and Peptide Science*, vol. 3, no. 6, pp. 615–622, 2002.

[2] K.-C. Chou, "Structural bioinformatics and its impact to biomedical science," *Current Medicinal Chemistry*, vol. 11, no. 16, pp. 2105–2134, 2004.

[3] G. Lubec, L. Afjehi-Sadat, J. W. Yang, and J. P. P. John, "Searching for hypothetical proteins: theory and practice based upon original data and literature," *Progress in Neurobiology*, vol. 77, no. 1-2, pp. 90–127, 2005.

[4] H.-B. Shen and K.-C. Chou, "Signal-3L: a 3-layer approach for predicting signal peptides," *Biochemical and Biophysical Research Communications*, vol. 363, no. 2, pp. 297–303, 2007.

[5] K.-C. Chou and H.-B. Shen, "Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides,"

Biochemical and Biophysical Research Communications, vol. 357, no. 3, pp. 633–640, 2007.

[6] K.-C. Chou, "Prediction of signal peptides using scaled window," *Peptides*, vol. 22, no. 12, pp. 1973–1979, 2001.

[7] K.-C. Chou, "Using subsite coupling to predict signal peptides," *Protein Engineering*, vol. 14, no. 2, pp. 75–79, 2001.

[8] G. Schneider, S. Rohlk, and P. Wrede, "Analysis of cleavage-site patterns in protein precursor sequences with a perceptron-type neural network," *Biochemical and Biophysical Research Communications*, vol. 194, no. 2, pp. 951–959, 1993.

[9] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, "A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites," *International Journal of Neural Systems*, vol. 8, no. 5-6, pp. 581–599, 1997.

[10] J. D. Bendtsen, H. Nielsen, G. Von Heijne, and S. Brunak, "Improved prediction of signal peptides: signalP 3.0," *Journal of Molecular Biology*, vol. 340, no. 4, pp. 783–795, 2004.

[11] D. Plewczynski, L. Slabinski, K. Ginalski, and L. Rychlewski, "Prediction of signal peptides in protein sequences by neural networks," *Acta Biochimica Polonica*, vol. 55, no. 2, pp. 261–267, 2008.

[12] H. Nielsen and A. Krogh, "Prediction of signal peptides and signal anchors by a hidden Markov model," *Intelligent Systems for Molecular Biology*, vol. 1, no. 6, pp. 122–130, 1998.

[13] J. P. Vert, "Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings," in *Proceedings of Pacific Symposium on Biocomputing*, Kauai, Hawaii, USA, 2002.

[14] Y.-D. Cai, S.-L. Lin, and K.-C. Chou, "Support vector machines for prediction of protein signal sequences and their cleavage sites," *Peptides*, vol. 24, no. 1, pp. 159–161, 2003.

[15] C. Chen, X. Zhou, Y. Tian, X. Zou, and P. Cai, "Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network," *Analytical Biochemistry*, vol. 357, no. 1, pp. 116–121, 2006.

[16] P. G. Bagos, K. D. Tsirigos, S. K. Plessas, T. D. Liakopoulos, and S. J. Hamodrakas, "Prediction of signal peptides in archaea," *Protein Engineering, Design and Selection*, vol. 22, no. 1, pp. 27–35, 2009.

[17] P. P. Łabaj, G. G. Leparc, A. F. Bardet, G. Kreil, and D. P. Kreil, "Single amino acid repeats in signal peptides," *The FEBS Journal*, vol. 277, no. 15, pp. 3147–3157, 2010.

[18] J. Pearl, "Bayesian networks: a model of self-activated memory for evidential reasoning," in *Proceedings of the 7th Annual Conference of the Cognitive Science Society, Computer Science Department*, University of California, Los Angeles, Calif, USA, 1985.

[19] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[20] R. E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall Series in Artificial Intelligence, Chicago, Ill, USA, 2004.

[21] I. Ben-Gal, A. Shani, A. Gohr et al., "Identification of transcription factor binding sites with variable-order Bayesian networks," *Bioinformatics*, vol. 21, no. 11, pp. 2657–2666, 2005.

[22] K. Wang, J. Zhang, F. Shen, and L. Shi, "Adaptive learning of dynamic Bayesian networks with changing structures by detecting geometric structures of time series," *Knowledge and Information Systems*, vol. 17, no. 1, pp. 121–133, 2008.

[23] L. Sang, Y. Yang, Z. Wu, and W. Zhang, "Dynamic Bayesian network approach to speaker identification," *Electronics Letters*, vol. 39, no. 3, pp. 329–330, 2003.

[24] H. Liu, J. Yang, J. G. Ling, and K.-C. Chou, "Prediction of protein signal sequences and their cleavage sites by statistical rulers," *Biochemical and Biophysical Research Communications*, vol. 338, no. 2, pp. 1005–1011, 2005.

[25] K. Hiller, A. Grote, M. Scheer, R. Münch, and D. Jahn, "PrediSi: prediction of signal peptides and their cleavage positions," *Nucleic Acids Research*, vol. 32, pp. W375–W379, 2004.

[26] J.-Y. Wang, *Application of Support Vector Machines in Bioinformatics*, National Taiwan University, 2002.

[27] D. Grossman and P. Domingos, "Learning Bayesian Network classifiers by maximizing conditional likelihood," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 361–368, Department of Computer Science and Engineering, University of Washington, Seattle, Wash, USA, July 2004.

[28] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[29] R. R. Bouckaert, *Bayesian Network Classifiers in Weka*, University of Waikato, Hamilton, New Zealand, 2004.

[30] K. P. Murphy, "The Bayes Net Toolbox for Matlab," http://people.cs.ubc.ca/~murphyk/Papers/bnt.pdf.

[31] J. Lin, Y. Wang, and X. Xu, "A novel ensemble and composite approach for classifying proteins based on Chou's pseudo amino acid composition," *African Journal of Biotechnology*, vol. 10, no. 74, pp. 16963–16968, 2011.

[32] Y. Lu, "Knowledge integration in a multiple classifier system," *Applied Intelligence*, vol. 6, no. 2, pp. 75–86, 1996.