Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2013, 5(11):212-218

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A novel feature extraction method by compressive sensing for signal peptide

Cui-Fang Gao[1], Qiang Guan[1], Hao Zhang[2], Wei Chen[2] and Feng-Wei Tian[2*]

[1]School of Science, Jiangnan University, Wuxi, China
[2]School of Food Science and Technology, Jiangnan University, Wuxi, China

---

ABSTRACT

In this paper, we propose a novel mathematical expression of signal peptide that can truly reflect the intrinsic attributes of the sequence. To deal with every signal peptide that displays diverse length, we first transformed the original sequence into Markov transition matrix with a unified size. Next we obtained the expansion sparse vector on a unit orthogonal basis, and finally utilized random projection of compressive sensing (CS) to obtain an accurate feature expression. Analyzing from the perspective of traditional mathematical theory, we determined that the feature vector abstracted by CS was the optimal combination of the original information (amino acid composition, sequence order, and so on). Thus, the new method can be regarded as a comprehensive development of high-density discriminative information extraction. The experimental results suggested that the CS feature expression has the approving determinant and has the potential for future research and applications in the development of feature extraction.

Key words: Compressive sensing, Sparse representation, Markov transfer matrix, Signal peptide, Feature extraction, Secretory proteins.

INTRODUCTION

Many proteins of model organisms can be found released into the culture medium supernatant by a number of different secretory pathways. In this process, the signal peptide is a functionally special segment that guides newly synthesized proteins through the secretory pathway. A growing application of signal peptides in the biotechnological industries concerns the expression and secretion of proteins [1].

Feature analysis and identification of signal peptides has the potential to increase our understanding of protein transport mechanisms, and could contribute to the industrial-scale production of important natural proteins. However, the amino acid composition and the length of the signal peptide vary significantly according to the specific species. In addition, many signal peptides from the same species are also different, which has caused considerable difficulties and challenges in attempts to identify the signal peptide [2].

The original signal peptide sequence is represented by amino acid symbols, which cannot be directly calculated by the analysis algorithms. To facilitate data representation and processing, the sequence of symbols must be converted to numerical data expressed as a feature vector prior to calculation. The main purpose to feature extraction methodology is to obtain a set of numerical features that are most effective for identification, and that reflect the intrinsic properties of the investigated object from several aspects or in the context of a particular aspect. So, feature extraction plays a key role in an intelligent algorithm for signal peptide identification.

Compressive sensing/sampling theory (CS) [3-5], as an abstract mathematical idea concerning sparse representation, was proposed by Donoho and Candès et al. in 2006. CS has breach traditional thinking in linear sampling methods of information theory, and established a new theoretical framework for signal sampling and processing based on sparse

representation and optimization theory. Past studies of CS theory demonstrated that as long as the high-dimensional signal is compressible on a transform basis, the signal could be projected onto a low-dimensional space by a measurement matrix, which is irrelevant to the transform basis. Further, if the signal expression is sparse enough, the sparsest representation has an acceptable discriminating effect. CS theory performs very well in the context of acquiring high-density information, and has now been developed and applied in various areas where researchers have exploited it for feature extraction and recognition, including sparse representation-based classification algorithm for medical images [6], mathematical framework of cost-effective genotyping protocols to detect severe genetic diseases [7], and techniques applied to spectrum hole identification for wideband cognitive radios [8], etc.

In this paper, we have developed a new technique of feature extraction for signal peptides using the effective representation of sparse signals. The objective is to quantify and transform some important attributes of signal peptide sequence, such as amino acid composition, sequence order, etc., into sparse representations, and to use CS technology to extract high-density discriminatory information as the feature vector.

MATERIALS AND METHODS

Since the symbol sequence of a signal peptide cannot be directly used as computational data, generally, the first and necessary step of an analysis is to pre-process the original sequence to form a numerical feature vector. The task of feature extraction requires us to formulate an effective mathematical expression of sequences that can truly reflect the intrinsic attribute of the signal peptide. Fortunately, CS is present as a promising theory to satisfy our demands.

The projection process of CS is the dimensional conversion, which can preserve such important information as composition or structure of the signal and in the meantime has the potential of reducing data redundancy. So CS is considered a potential effective method of feature extraction.

2.1 Compressive Sensing (CS) Theory

Compressive sensing uses transformed space to describe the signal sampling, in which the mass sampling of sparse signals is compressed into projected information with less data (referred to as observation values). A commonly used approach for obtaining effective observation values is random linear projection. Conversely, the original signal can be faithfully reconstructed with the observations by solving a series of optimization problems or approximation problems [9, 10].

Supposing a discrete-time signal $x \in R^N$ of length N can be expanded by a set of orthogonal basis vectors $\Psi = [\Psi_1, \Psi_2, ..., \Psi_N]$, that is:

$$\mathrm{x} = \sum_{i=1}^{N} \psi_i \theta_i = \psi\theta \tag{1}$$

Where Ψ is an orthogonal matrix with each column $\Psi_i (i = 1, 2, ..., N)$ corresponding to a basis function of sparse transform, which can be Wavelet transform, Discrete-Cosine transform, or Fourier transform, depending on the application. Then, $\boldsymbol{\theta}=[\theta_1, \ldots, \theta_N]^T$ is the transformed vector consisting of the N coefficient expression defined as $\theta_i = \langle \boldsymbol{x}, \boldsymbol{\psi}_i \rangle$ (i=1,2 ,…, N). Assuming that the sampling signal $\boldsymbol{x}$ only contains $K (K << N)$ non-zero coefficients on the orthogonal basis Ψ, signal $\boldsymbol{x}$ is then generally considered to be sparse or compressible.

Then signal $\boldsymbol{x}$ can be compressed by an $M \times N$ ($M<N$) measurement matrix Φ, which obtains a low-dimensional vector, and is expressed as the following:

$$\boldsymbol{s} = \Phi \boldsymbol{x} \tag{2}$$

Where $\boldsymbol{s} \in R^M$ indicates the projected vector of M linear measurement components. Substitute equation (1) into equation (2), so that we have

$$\mathrm{s} = \Phi \psi\theta = \Theta\theta \tag{3}$$

Now, the original signal $\boldsymbol{x}$ has been reduced with the ratio M/N. It is worthwhile realizing that the measured signal $\boldsymbol{s}$ is not the exact value of signal $\boldsymbol{x}$, but is essentially the projected value from high-dimension to low-dimension. Analysis from the perspective of traditional mathematical theory, informs us that the measurement vector projected by CS is the optimal combination value of the original signal. In other words, the measurement value is a less volume of

high-density information including all of the signals in the original sampling.

Here the measurement matrix needs to satisfy the condition called Restricted Isometry Property (RIP) [4]: That is $1-\varepsilon \le \|\Phi v\|_2 / \|v\|_2 \le 1+\varepsilon$. Thus, the linear measurement should have stable energy properties. The restricted isometry of the observation matrix furnished a theoretical guarantee for faithfully reconstructing the compressible signal from observation values. An equivalent condition of RIP is Incoherence, which means that the measure matrix Φ and the sparse matrix Ψ cannot be representation each other. According to previous studies, the measurement matrix selected as a Gaussian random matrix will has a high probability of incoherence with an arbitrary sparse base, and also allows the conditions of restricted isometry to be satisfied [11, 12].

2.2 Markov Transition Matrix of the Signal Peptide

Markov chain is a widely applied mathematic model that reveals the collection of state distribution on a sequence [13]. Typically, the signal peptides are described using limited symbols to denote 20 kinds of natural amino acids. Should these residues on the chain be regarded as state parameters, it follows that the sequences of the amino acids will express a series of transition of state. In this way, a finite stationary Markov model can be constructed based on symbol distribution to reflect the intrinsic relationship and further permits detection of the comprehensive information of signal peptide sequences.

In order to quantitatively describe the state transition behavior on a given signal peptide, we defined a 20×20 Markov matrix whose rows and columns were denoted by amino acids to represent the frequency of occurrence of each dipeptide. Assume that $\boldsymbol{U}(i, j)=\{(R_i, R_j), z\}$, that is to say in the frequency matrix $\boldsymbol{U}$, the element value located in the *i-th* row (denoted by amino acid R_i), *j-th* column (denoted by amino acid R_j) is numerical *z*. Wherein R_i is the previous residue of a dipeptide and R_j is the latter, *z* is the transition frequency from R_i to R_j through the full sequence. Thus, we find that the pair-wise residues (R_i,R_j) in matrix $\boldsymbol{U}$ correspond to their respective denotation, and give the assignment $\boldsymbol{U}(i, j)=z$. Thus, the Markov matrix that reflects the composition of the dipeptide and the series of state relations on a sequence can be constructed.

For example, the signal peptide of the secreted protein AIDA_ECOLI from Gram-negative bacteria is as follows: "MNKAYSIIWSHSRQAWIVASELARGHGFVLAKNTLLVLAVV". After we successively process the pair-wise residues from the beginning terminal to the end, this symbol sequence converted into a unified Markov matrix has the following design:

Table 1. Markov transfer frequency matrix of signal peptide AIDA_ECOLI

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C																				
S									1		1	1			1					
T																1				
P																				
A		1										1	1				1		1	1
G											1							1		
N			1										1							
D																				
E																1				
Q					1															
H		1				1														
R						1				1										
K					1		1													
M							1													
I															1		1			1
L					3											1	1			
V					1											2				
F																	1			
Y		1																		
W		1													1					

Obviously, the frequency of occurrence of all dipeptides are included in the matrix, wherein the numerical cell implies that the dipeptide was present in the sequence, whereas the blank cell means that the dipeptide never occurred and had

a default value of zero. In the rest part of this paper, the Markov Matrix in shortened form will refer to the Markov transfer frequency matrix.

2.3 Feature Extraction by Compressive Sensing

According to the construction principle of the Markov Matrix, it is reasonable to understand that the number of digits in matrix $\boldsymbol{U}$ is not more than L-1 (where L is the length of a given signal peptide), because the maximum number of dipeptides in the full sequence is L-1. In view of the facts that the signal peptide is generally composed of 15 to 60 amino acids [2], L is dramatically smaller than the size of the Markov matrix (i.e. $L << 400$). One could perceive that an important characteristic of the matrix $\boldsymbol{U}$ is that the non-zero elements are very sparse and the remaining elements are zero.

For subsequent analysis and recognition, only small portions of these data in matrix U are needed. However, most of the useless redundancy will be discarded when permitted. Such property is consistent with that of sparse signals (wherein only very few coefficients are non-zero relative to the signal length). Therefore it is most applicable for the spare matrix $\boldsymbol{U}$ to exploit the randomly projected method by CS technique.

We can obtain a 400-dimensional digital signal by row expansion of the Markov Matrix. In virtue of the sparse property of the expansion signal, we can directly use the unit orthogonal basis as the sparse base. Besides, in order to satisfy the RIP condition mentioned in Section 2 above, (as a high probability is possible), the independent and identically distributed Gaussian random matrix is selected as the measurement matrix.

Under these conditions, then we can perform the operation of random projection according to equation (3). The result of the inner product expressed by a low dimensional observed signal is essentially the extracted feature vector of the signal peptide. Figure 1 shows the key steps of feature extraction.

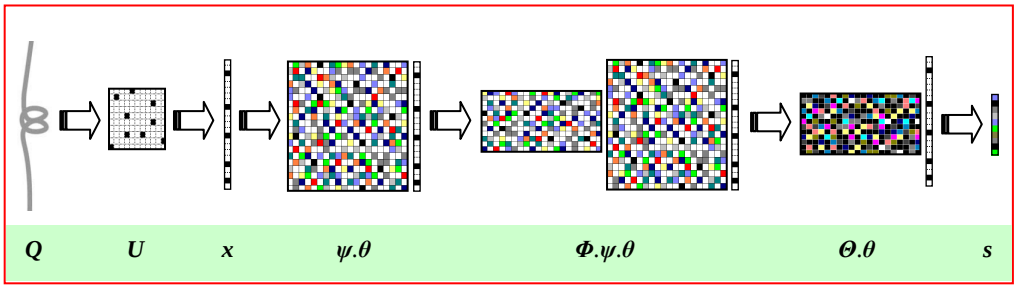


Figure 1. Key steps involved in the feature extraction by compressive sensing

The inputs are symbol sequences of signal peptides and the final outputs are the feature vectors. Expatiations of the denotations in fig. 1 are as follows:

$\boldsymbol{Q}$: denotes the original signal peptide sequence expressed by amino acid symbols;

$\boldsymbol{U}$: is a 20×20 Markov transition frequency matrix;

$\boldsymbol{x}$: is a 400-dimensional expansion signal of matrix $\boldsymbol{U}$;

$\boldsymbol{\psi}$: is a 400×400 sparse basis, unit orthogonal basis $\boldsymbol{E}$ is selected in this paper;

$\boldsymbol{\theta}$: denotes the conversion signal of $\boldsymbol{x}$ on the sparse basis $\boldsymbol{\psi}$, if $\boldsymbol{\psi} = \boldsymbol{E}$, then $\boldsymbol{\theta} = \boldsymbol{x}$;

$\boldsymbol{\Phi}$: is a M×400 measurement matrix, here Gaussian random matrix obeys the uniform distribution on [5, 0.1] is selected;

$\boldsymbol{\Theta}$: is a M×400 filter matrix, $\boldsymbol{\Theta} = \boldsymbol{\Phi} \times \boldsymbol{\psi}$.

$\boldsymbol{s}$: denotes the m-demensional measurement signal after compression, which is the expected feature vector of $\boldsymbol{Q}$;

Remarkable advantages of the processing sensor theory including an ability to maintain sufficiently important information when sparse signals are compressed. Consequently, the fundamental nature of our method, and its main properties are also vested in maintaining the dipeptide composition and transition behavior of the Markov matrix $\boldsymbol{U}$, which are finally combined in the optimal low-dimensional observed signal (vector $\boldsymbol{s}$). This vector is the desired effective information that reflects the intrinsic properties of the signal peptide, and the subsequent identification of the secretory protein can be successfully performed using this high-density information.

The Markov matrix in our method contains the occurrence frequency of amino acid residues, and at the same time reflects the sequence order of amino acids and the inherent structural information of the dipeptide. Essentially, it can be regarded as a comprehensive method of multiple feature extraction, including amino acid composition [14],

sequence order method [15] and sequence wavelet decomposition [16, 17].

NUMERICAL EXPERIMENTS AND RESULTS DISCUSSION

Benchmark datasets used in our experiments were obtained from the following website: http://www.cbs.dtu.dk/ftp/signalp [18], which was released by Nielsen et al. We chose three different species: (1) Eukaryotes; (2) Gram +ve bacteria; and (3) Gram -ve bacteria. For the secretory proteins, the extended sequences of the signal peptides were given in the dataset, which included the sequence of the signal peptide plus the nearest 30 amino acids of the mature protein.

Such simplified fragments are reasonable because they maintain the characteristics of signal peptides without neglecting the effect of neighboring sequences. But for the non-secretory proteins of negative sample, since no signal peptide exists, the data sets gives the first 70 amino acids of each sequence. Detailed information of the datasets is (Table 2) below:

Table 2. Numbers of protein sequences in three datasets

Species	Secreted proteins	Non-secreted proteins	Total
Eukaryotes	1009	269	1278
Gram- bacteria	265	186	451
Gram+ bacteria	140	64	204
Total	1414	519	1933

Since a few entries in the original dataset were found to contain uncertain residues denoted as “X” that were included in the sequence, X might be glutamic acid, proline, or glycine. In order to improve the quality of the data and to obtain a result without bias, such sequences were removed manually. Therefore protein sequences listed in Table 2 excluded four entries: KV4A_MOUSE and CAS1_SHEEP (Eukaryotes), GUN5_THEFU (Gram +ve bacteria), OMPH_YEREN (Gram -ve bacteria).

3.1 Effectiveness of the CS Feature Vector

We extracted 20D (M = 20) numerical feature vectors of the protein sequences (Table 2) using new methods that we proposed in Section 2. Further, in order to properly and objectively examine the effectiveness of the CS features (for comparisons with other features, we named ours as the CS feature), we evaluated the new features according to the following indices: that is the trace of intra-class scatter matrix $tr(S_w)$ and the trace of the inter-cluster scatter matrix $tr(S_b)$.

A small intra-cluster distance and a large inter-cluster distance will induce an acceptable recognition result. Therefore the smaller the ratio of $tr(S_w)/tr(S_b)$, the better the recognition performance. The indices are defined as:

$$S_w = \sum_{k=1}^{C}\sum_{i=1}^{N_k}(\boldsymbol{x}_i^{(k)} - \boldsymbol{m}_k)(\boldsymbol{x}_i^{(k)} - \boldsymbol{m}_k)^T \tag{4}$$

$$S_b = \sum_{k=1}^{C} N_k(\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^T \tag{5}$$

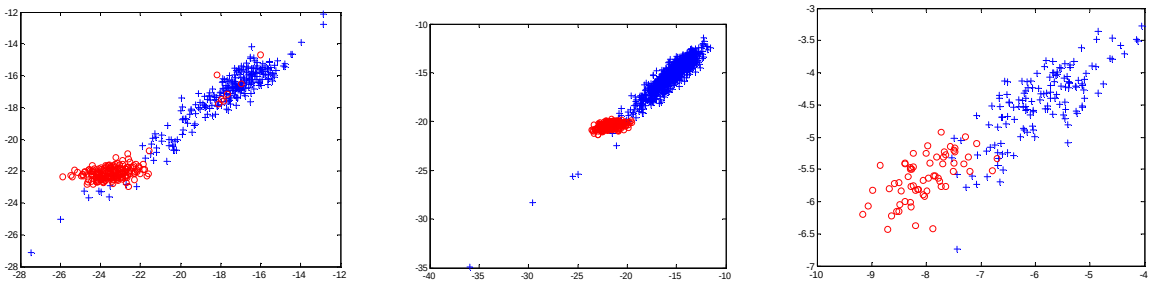
Where denotes the feature vectors of proteins, C is the number of clusters, Nk is the number of samples belonging to class k, $\mathbf{m}_k = [m_{k1}, m_{k2}, \cdots, m_{kL}]^T$ is the mean vector of class k, and m is the mean vector of all samples in one dataset. Table 3 shows the values of this performance index. As compared with the traditional features of amino acid composition and scaled wavelet energy, the compressive sensing features have the best anticipated separability of the three types of datasets.

Table 3. Indicators of different feature vectors

Species	$tr(S_w)/tr(S_b)$		
	Amino acid composition features	Scale wavelet energy features	Compressive sensing features
Eukaryotes	20.7469	7.6608	0.5180
Gram -ve bacteria	12.7370	4.9372	0.5758
Gram +ve bacteria	11.8218	6.6286	1.9722

In order to visualize the distribution of CS feature vectors, and at the same time, keep their distance relationship among original samples as far as possible, we project the CS features onto 2D space using a linear mapping method [19], from which we can get 2D data from the 20D input; the projected results are shown in Figure 2. We can see that

the secretory and the non- secretory proteins can be substantially distinguished. The separability of Eukaryotes is the best example of the three species. However, Gram -ve and Gram +ve datasets are somewhat overlapping, and such visual phenomena are consistent with the separability indicators in Table 3.



(a) **(b)** **(c)**

Figure 2. Two-dimensional mapped distribution of the CS features, (a) Eukaryotes, (b) Gram -ve bacteria, (c) Gram +ve bacteria

3.2 Recognition Effect of CS Feature Vector

We employ the popular fuzzy clustering algorithms (FCM) which has been increasingly and widely used by investigators in theoretical and practical applications to accomplish the task of recognition. Here we used two indices; (1) Identification accuracy and (2) Partition entropy to evaluate the clustering performance based on different features.

Accuracy gives the percentage of all true identifications. Partition entropy criterion is based on entropy function, which is defined in Equation (6):

$$J(\cdot) = -\frac{1}{n}\sum_{k=1}^{C}\sum_{i=1}^{n} u_{ik}\log_2(u_{ik}) \tag{6}$$

Where n is the number of protein samples in a given dataset, C is the number of clusters, and u_{ik} denotes the membership degree of the *i-th* samples belonging to the *k-th* cluster. The clustering performance is improved if the value of the function $J(\cdot)$ is smaller.

The discrimination of CS features is examined and compared with previous features including amino acid compositions and scale wavelet energy. Proteins should be partitioned into two categories of secreted and non-secreted types, which are regularly separated among the three species. The identification results on three datasets are shown in Table 4.

Table 4. Clustering performance based on different features

Feature Vectors	Eukaryotes		Gram -ve bacteria		Gram +ve bacteria	
	accuracy	J(·)	accuracy	J(·)	Accuracy	J(·)
Amino acid composition features	66%	1.0000	75%	1.0204	64%	0.9999
Scale wavelet energy features	76%	0.6210	70%	0.6299	62%	0.6810
CS features	94%	0.2197	90%	0.1488	84%	0.2286

Most of the current published prediction algorithms need specific training samples, and relatively, the recognition accuracy for unsupervised learning is generally low. Owing much to the advantage of CS theory in capturing important information, the proposed new CS features can achieve a high degress of recognition, and in the case of unsupervision the performance of CS is much better than the other two features in Table 4.

CONCLUSION

Appropriate feature expression is an important factor that influences the effectiveness of recognition. In this paper, we have described a novel technique to extract the feature vector of signal peptides by introducing the powerful theory of CS. Previous studies have shown that CS has excellent performance features in the context of acquiring high-density and important information when compressing sparse signals. Therefore, CS feature vectors as optimal high-density information, includes all of the original signals. In the numerical experiments, high recognition accuracy has been achieved based on CS feature vectors even in the case of unsupervised leaning without training samples. In addition, the Markov transfer frequency matrix used in our method is not the only way to quantifiably represent signal peptide information. Alternatives with meaningful formulations are desirable and should be exploited in the future.

Acknowledgements
This research was partially supported by Program for National Natural Science Foundation of China (Grant No.: 11271163, 31371721), Specialized Research Fund for the Doctoral Program of Higher Education (Grant No.: 20120093120016), the Natural Science Foundation of Jiangsu Province of China (Grant No.: BK20131102, BK20130117), Fundamental Research Funds for the Central Universities (Grant No.: JUSRP211A23).

REFERENCES

[1] Schallmey, M.; Singh, A.; Ward, O. P. *Can. J. Microbiol.*, **2004**, 50(1), 1-17.
[2] Nielsen, H.; Engelbrecht, J.; Brunak, S.; Heijne, G. V. *Int. J. Neural Syst.*, **1997**, 8, 581-599.
[3] Donoho, D. *IEEE Trans. Inform. Theory.*, **2006**, 52(4), 1289-1306.
[4] Candès, E.; Romberg, J.; Tao, T. *IEEE Trans. Inform. Theory.*, **2006**, 52(2), 489-509.
[5] Candès, E. J.; Wakin, M. B. *IEEE Signal Processing Magazine*, **2008**, 25(2), 21-30.
[6] Cao, H. B.; Deng, H. W.; Li, M.; Wang, Y. P. *IEEE Trans. on Nanobioscience*, **2012**, 11(2), 111-118.
[7] Erlich, Y.; Gordon, A.; Brand, M.; Hannon, G. J.; Mitra, P. P. *IEEE Trans. Inform. Theory*, **2010**, 56(2), 706-723.
[8] Tian, Z.; Giannakis, G. *IEEE International Conference on Acoustics, Speech and Signal*, **2007**, 4, IV-1357-1360.
[9] Candès, E.; Tao, T. *IEEE Trans. Inform. Theory*, **2006**, 52(12), 5406-5425.
[10] Tropp, J. A.; Gilbert, A. C. *IEEE Trans. on Inform. Theory*, **2007**, 53(12), 4655-4666.
[11] Candes, E. J.; Tao, T. *IEEE Trans. Inform. Theory*, **2005**, 51(12), 4203-425.
[12] Candes, E.; Romberg, J.; Tao, T. *Comm. Pure Appl. Math.*, **2006**, 59(8), 1207-1223.
[13] Elfeki, A.; Dekking, M. A Markov *Journal Mathematical Geology*, **2001**, 33(5), 569-589.
[14] Shen, H. B.; Chou, K. C. Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, **2006**, 22(14), 1717-1722.
[15] Chou, K. C. *Proteins: Struct., Funct., Genet.*, **2001**; 43(3), 246-255.
[16] Liò, P. *Bioinformatics*, **2003**, 19(1), 2-9.
[17] Gao, C. F.; Qiu, Z. X.; Wu, X. J.; Tian, F. W.; Zhang, H.; Chen, W. *Protein Pept. Lett.*, **2011**, 18(8), 831-838.
[18] Nielsen, H.; Engelbrecht, J.; Brunak, S.; Heijne, G. V. *Protein Eng.*, **1997**, 10(1), 1-6.
[19] Bian, Z. Q.; Zhang, X. G. *Pattern Recognition*. 2nd ed. Beijing: Tsinghua University Press, **2000**, 185-198.