Sequence-Based Feature Extraction for Type III Effector Prediction

Tingting Sui, Yang Yang, and Xiaofeng Wang

***Abstract*—The type III secretion system (T3SS) is a complex structure which allows gram-negative pathogens to destroy eukaryotic cell biology by injecting virulence factors directly into the host cell cytoplasm. Composed of around 30 proteins, T3SS is among the most complex secretion systems identified in Gram-negative bacteria. Since type III secreted effectors (T3SEs) are essential for the pathogenicity, identification of T3SEs is one of the core problems in computational biology. This paper puts forward a new method for the prediction of T3SEs. The method is a sequence-based approach which can extract useful features from amino acid sequences. By calculating the frequency of the features from different segments of protein sequences, the data set is represented by the feature vectors and classified by Support Vector Machine (SVM). The experimental results show superiority over other available approaches on classification accuracy.**

***Index Terms*—Type III secreted effector prediction, sequence-based approach, feature extraction, type III secretion, word segmentation, hybrid feature system.**

I. Introduction

The type III secretion system (T3SS) is a complex mechanism which directs the delivery of virulence proteins (effectors) into the host cells [1]. Upon translocation, type III secreted effectors (T3SEs) modulate diverse host cell processes unanimously to ensure the host-microbe interaction [2]. Researchers have found that T3SS serves as an essential component for the pathogenesis of a large variety of plant and animal bacterial pathogens [3]. Therefore, great research interests have been attracted to the study of T3SS.

The T3SS proteins include regulatory proteins, structure proteins, effectors and chaperones [3], [4]. The former two types of proteins function for controlling the expression of T3SS and building the system [3]. The structure of T3SS usually contains a needle-like apparatus and bases embedded in the inner and outer bacterial membranes [5]. Although the structure of T3SS has become unambiguous, we have not understood the precise secretion mechanism completely [6]. On one hand, T3SEs mimic eukaryotic virulence proteins in both function and structure [7]. On the other hand, the variation and biological evolution enable T3SEs' sequences to be more diversified [8]. These characteristics make the recognition of T3SEs become more difficult. Therefore, the study of T3SS should put the crucial step in the identification of T3SEs.

Up to now, researchers have applied multiple computational approaches to identify T3SEs, e.g., gene-adjacent features-based method, sequence similarity-based method, etc. [9]-[11]. Especially, various machine learning algorithms have been applied in this area, such as Artificial Neural Network [12], Naive Bayes Algorithm [13], Hidden Markov Model (HMM) [14] and Support Vector Machine (SVM) [15]. However, classifiers have limited capabilities in improving the prediction accuracy. Therefore, how to represent proteins appropriately becomes more important. Researchers have developed a lot of feature extraction methods, such as the features extracted from amino acid sequences or annotation data, e.g., secondary structure and solvent accessibility [15], [16]. At first, amino acid composition (AAC) was used to predict T3SEs because researchers have detected amino acid composition biases in T3SEs [17]. Later, the approaches based on amino acid pair composition (AAPC) or motif emerged [18], [19]. However, the classification results are not satisfying because no defined consensus motifs or features have been discovered for T3SEs. Most recently, some effective approaches have been proposed. R. Arnold, S. Brandmaier, et al. proposed Effective-T3 using amino acid composition as features [13]. Löwer and Schneider proposed a method based on sliding-window model combined with Artificial Neural Networks [12]. Y. Wang, Q. Zhang, M. Sun, and D. Guo developed the BPBAac, which also adopts the sliding-window technique and creates position-specific Aac profiles for classification [20]. These methods mainly utilize the frequency and position information of single amino acids, but they neglect peptides or longer amino acid subsequences, which may contain secretion signal. S. Qi, Y. Yang, and A. Song regarded the protein sequences as text and introduced topic models to extract informative words (k-tuples) [21], while the words with most discriminative ability are not necessarily selected. Therefore, new features need to be identified, especially the signal subsequence for T3SEs, which could help to build more effective T3SE predictor.

In this paper, we propose a new hybrid method to identify potential T3SEs using both the subsequence frequency and position information of amino acid sequences with support vector machine (SVM) classifier. Moreover, we carry out a comprehensive and profound exploration on the classification performance of several popular methods recently developed. The experimental results suggest that our

Manuscript received December 27, 2012; revised February 28, 2013. This work was supported by the National Natural Science Foundation of China (Grant No. 61003093), the Science Foundation for The Excellent Youth Scholars of Shanghai Municipality and the Science & Technology Program of Shanghai Maritime University (Grant No. 20110009).

Tingting Sui is with the Institute of Computer Application and Technology, Shanghai Maritime University, CO 201306 CHN (e-mail: suisui61@163.com).

Yang Yang and Xiaofeng Wang are with the Department of Computer Science and Engineering of Shanghai Maritime University, CO 201306 CHN (e-mail: yangy09@gmail.com, xfwang@shmtu.edu.cn).

DOI: 10.7763/IJBBB.2013.V3.206

method outperforms most of the present methods in the prediction of T3SEs.

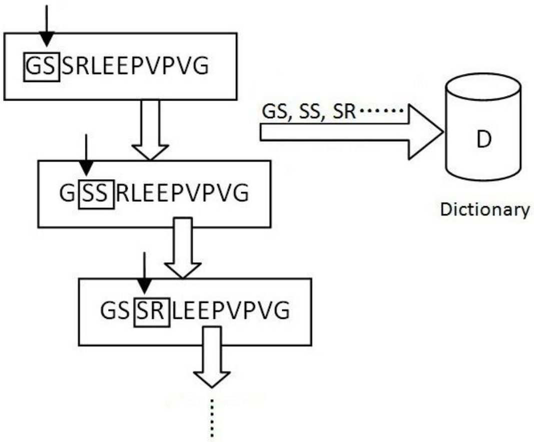


Fig. 1. The initial scanning process.

II. Method

Protein sequences consist of consecutive amino acids. We assume them as a certain sort of biological language which has a corresponding dictionary D [22]. In order to constitute D appropriately, attention should be paid to how to extract representative features from protein sequences. Since researchers have shown that the first 100 amino acids which contain translocation information are enough for secretion [23]-[25], we only focus on N-terminals instead of full-length protein sequences for calculation.

The rest of this section falls into three parts. Section II-A introduces our feature extraction process. At first, 20 amino acids [26] are included in $\mathcal{D}$ and adopted as features. Then, we make an extension of the length of the features existed in D from one to k, where k is an integer and definitely larger than one. We refer to it as k-tuple. In order to abate worthless k-tuples and reduce the dimension, we set a proper threshold and select the representative features. The dimension reduction part is described in Section II-B. Finally, in Section II-C, we build a hybrid prediction system with all the features existed in D and the features are extracted from three segments of the N-terminal, respectively.

A. Feature Extraction

Although amino acid composition method is simple and convenient, it discards the order information of neighboring residues and only calculates the occurrence time of each single amino acid. Therefore, we use the k-tuples which can represent some order information in the classification of protein sequences. For example, the sequences 'AIC' and 'CIA' are separated into three single amino acids, 'A', 'I' and 'C', as well as the sequence 'ACI' while using the amino acid composition method. However, if we use 2-tuple characteristics, 'AIC' is represented by 'AI' and 'IC', and 'CIA' is represented by 'CI' and 'IA'.

Fig. 1 shows the initial scanning process, which scans the k-tuple in an overlapping manner, and adds all the present k-tuples into the dictionary D. Thus, any k-tuple, even occurring very few times, is recorded in D. Some k-tuples may be useless or even noisy items, which lead to more computational time and the decline of prediction accuracy. Therefore, the development of an efficient dimension reduction method is crucial.

B. Dimension Reduction

Since the dimensionality of this method expands rapidly as k increases, k should be assigned to a relatively small value to avoid the intractable computation. Typically, for amino acids, four is the maximum distance between local interactions [16].Therefore, the maximal length of k is defined as four and every k-tuple with k no bigger than the maximal length will be checked based on certain criterion.

An intuition is that the most frequently presented strings in one class which seldom appear in another class are useful words with discriminative ability, thus k-tuples' occurrence times are recorded. K-tuples, whose frequencies are different in two classes, can be put into the dictionary. Considering the dimension disaster, the selected words' appearance times should be recounted while cutting down the size of $\mathcal{D}$. The two steps of our dimension reduction method are introduced in the following.

1) Word variance: Calculating the words' variances is to cut off the apparently worthless words. Firstly, the k-tuples' occurrence times should be recorded for each class in the training set. The average frequency of each word can be obtained from (1),

$$x_j = \frac{\sum_{i=1}^{N} count_i(j)}{N}, \quad for \quad i = 1,...,N \quad and \quad j = 1,...,M \tag{1}$$

where x_j denotes the average frequency of the jth word, $count_i(j)$ represents the times that word j occurs in protein sequence i, N denotes the amount of protein sequences, M represents the total number of the words.

In our study, we apply (1) to the training set and get the features' average frequencies for both positive and negative training samples. In order to further reveal the discriminative ability of the features, we use word variance as the criterion. The bigger the value of the variance, the more useful the word would be. The variance is defined in (2),

$$dis_i = (x_i^+ - x_i^-)^2 \times 100\%, \tag{2}$$

where dis_i exhibits the degree of deviation of word i in two classes, x_i^+ denotes the average frequency of word i in T3SEs and x_i^- represents the average frequency of word i in non-T3SEs. Using (2), we can choose discriminative words with apparent ease.

In addition, the variances multiplied by k are recorded as w_i which would be used as the ranking weight for the further dimension reduction. The weight is given by the following equation,

$$w_i = (x_i^+ - x_i^-)^2 \times k \times 100\%, \tag{3}$$

where w_i means the weight of word i, and k is the length of the features.

2) Word frequency: After the first step of dimension

reduction, the size of D shrinks a lot. However, cutting the number of features in D is still a major task while building the dictionary. We record the frequency for each k-tuple appearing in the training set of T3SEs again. However, the manner of feature extraction we used before can only roughly choose the words of the sequences, which is not biologically meaningful. Because proteins are linear polymer chains of amino acids, chain-based automatic word segmentation method is much more suitable for model protein sequences [22]. Therefore, this time we count the occurrence time in a different manner which is to first segment the sequences into words, and then calculate the average frequency of the word extracted. Through this process of word extraction, we can get independent and meaningful biological language units, shown in Fig. 2.

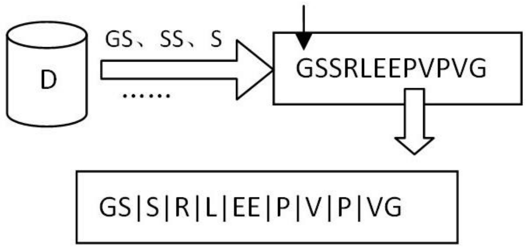


Fig. 2. New method for calculating frequency.

TABLE I: THE EXPERIMENTAL DATA SET

Data set	No. of positive samples	No. of negative samples	Total No.
I	108	760	868
II	210	1000	1210

Compared with Fig. 1, the new way of calculating frequency appears to be more meaningful. The sequences are segmented into features like English texts which are composed of words [22], [27]-[29]. The segmentation is determined by the match degree and the weights that have been calculated in (3). From Fig. 2, we can see that we segmented the subsequence ‘GSS’ into ‘GS’ and ‘S’, because the weight of ‘GS’ is higher than both ‘G’ and ‘GSS’ [27]. Then, using (1), we may further choose the features which are informative for classification and have great influence on global performance.

C. Hybrid Features

In our study, we use N-terminal 100 amino acids instead of full-length sequences, while some researchers have pointed out that rich secretion or translocation may only require the first 15 or 50 amino acids [15], [23], [24]. Therefore, the maximum length is fixed to 100 and each sequence is divided into three parts: the first part is from one to fifteen amino acids, the second part is from 16 to 50 and the third part is from 51 to 100. Then, we can further utilize the position information of amino acid sequences.

After feature extraction and dimension reduction, we create the hybrid feature vectors. In the hybrid model, the data set T is represented as

$$\mathcal{T}=\begin{bmatrix} a_{11}^1 & a_{12}^1 & \cdots & \cdots & a_{1N}^1 \\ a_{21}^1 & a_{22}^1 & \cdots & \cdots & a_{2N}^1 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ a_{i1}^1 & a_{i2}^1 & \cdots & \cdots & a_{iN}^1 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ a_{M1}^1 & a_{M2}^1 & \cdots & \cdots & a_{MN}^1 \\ a_{11}^2 & a_{12}^2 & \cdots & \cdots & a_{1N}^2 \\ a_{21}^2 & a_{22}^2 & \cdots & \cdots & a_{2N}^2 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ a_{i1}^2 & a_{i2}^2 & \cdots & \cdots & a_{iN}^2 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ a_{M1}^2 & a_{M2}^2 & \cdots & \cdots & a_{MN}^2 \\ a_{11}^3 & a_{12}^3 & \cdots & \cdots & a_{1N}^3 \\ a_{21}^3 & a_{22}^3 & \cdots & \cdots & a_{2N}^3 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ a_{i1}^3 & a_{i2}^3 & \cdots & \cdots & a_{iN}^3 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ a_{M1}^3 & a_{M2}^3 & \cdots & \cdots & a_{MN}^3 \end{bmatrix} \tag{4}$$

where the $a_{i1}^1\ a_{i2}^1\ ......a_{iN}^1$ denote the features’ frequencies of the first part of the sequence i, $a_{i1}^2\ a_{i2}^2\ ......a_{iN}^2$ denote the features’ frequencies of the second part of the sequence i, $a_{i1}^3\ a_{i2}^3\ ......a_{iN}^3$ denote the features’ frequencies of the third part of the sequence i, N is the size of dictionary $\mathcal{D}$, and M represents the number of protein sequences. Each row stands for a protein sequence while each column represents the performance of only one k-tuple. The performance is discussed in Sec. III

III. EXPERIMENTAL RESULTS

A. Data Set

In order to test and verify the accuracy of our hybrid method, two data sets are collected. One is the Pseudomonas syringae which acts as the model organism of T3SS with the most verified T3SEs. Through deep study [16] of the three strains of Pseudomonas syringae, including P. syringae pv. phaseolicola strain 1448A, P. syringae pv. syringae strain B728a and P. syringae pv. tomato strain DC3000, 283 effectors have been confirmed. However, the sequence similarity of those effectors is very high and over 61%. This is due to the fact that the majority of the verified effectors are homologs. Therefore, the redundant samples were eliminated and 108 positive samples left. Similarly, we extracted the negative data set from the genome of P. syringae pv. Tomato strain DC3000 and removed all the protein sequences related to T3SS, as well as the hypothetical proteins (Note that some unidentified effectors still remain in this set.). Considering the imbalance of the two sets, we selected the sequences from the remaining samples to constitute the negative set at random and kept the ratio of the size of the negative set to the positive set’s quantity as 7:1. The numbers of the data sets are listed in Table I. The data set is used for cross-validation to estimate the performance of our prediction system and select the best parameters as well. Another data set was extracted in the same way as those in the former data set. The positive samples were collected from various pathogenic bacteria, e.g., Rhizobium, E. coli, Yersinia, Salmonella, etc.. The total number of positive samples is 210 while the number of negative samples is 1000. Although the secretion mechanism of T3SS has great diversity across species, there should be some characteristics in common. Therefore, this set of data is used for examining the generalization ability of the prediction system.

B. Experimental Settings and Evaluation Criteria

Support Vector Machine (SVM) is a state-of-the-art classifier that employs an optimizer to identify an optimal separating hyperplane to discriminate two classes of interest. Due to its good performance, SVM is the favorite supervised learning method of the bioinformatics researchers [30]. In our experiment, we used the SVM as our learning machine and adopted the implementation of LibSVM version 3.31 [31] with RBF kernel.

Since we adopt 7-fold cross-validation to assess the method, the data set is divided into seven groups of approximately equal size. After seven rounds of training and test, in which we choose six groups as training set and the other one as test set, we have got seven prediction results. Then, the performance value is obtained by calculating the average of those results. For measuring the effectiveness of our proposed approach, we use sensitivity (*Sens*), specificity (*Spec*) and total accuracy (*TA*) as evaluation criterion.

The *Sens*, *Spec* and *TA* are obtained by solving the following equations:

$$Sens = \frac{TP}{TP + FN} \tag{5}$$

$$Spec = \frac{TN}{TN + FP} \tag{6}$$

$$TA = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

where *Sens* is the ratio of true positives (*TP*) to the sum of true positives and false negatives (*FN*), *Spec* is the number of true negatives (*TN*) divided by the sum of true negatives and false positives (*FP*).

In order to measure the overall prediction quality, we use the ratio of correct predictions compared to the total size of the data set to define *TA* as (7).

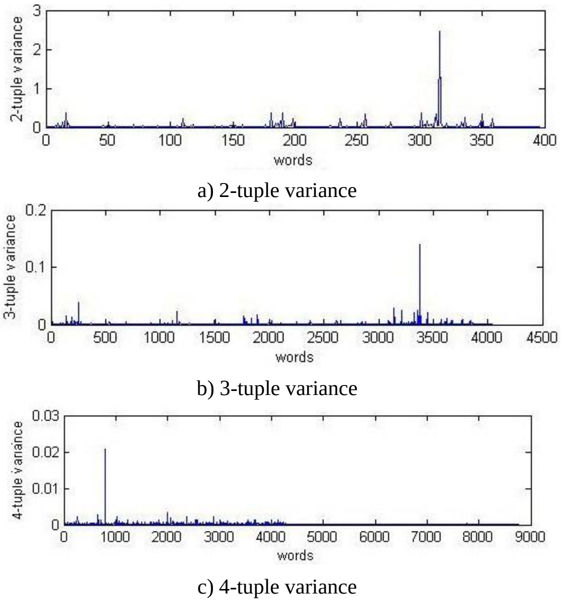
C. Results

In this experiment, we firstly found all the k-tuples occurred in the sequences of data set I, where k was set from 1 to 4. Then, we obtained the variance by solving (2). Fig. 3 depicts the variance distribution of k-tuples. Obviously, it can be observed that only a small portion of them have high variances which are discriminative for the prediction.

In order to reduce the dimension, we sorted all the words according to their variances in descending order. Then, we set a threshold on the basis of the gradient of the variance curve. The words whose variances are higher than the corresponding threshold can be kept. After a series of experiments, the threshold values were set at 0.12, 0.27, 0.07 (note that the value of k is 2, 3 and 4, respectively) according to the biggest decreasing gradient and these thresholds can indeed achieve the best accuracy. The distribution of the words and the prediction performance is displayed in Table II.

TABLE II: PERFORMANCE OF THE DIMENSION REDUCTION

Method	Word length	Threshold	Dimension	Total	TA (%)	Sens (%)	Spec (%)
Dimension Reduction with Word Variance	1	0	20	473	87.5	82.6	94.1
	2	0.12	270				
	3	0.27	135				
	4	0.07	48				
Dimension Reduction with Word Frequency	1	0	20	302	90.0	86.4	94.4
	2	0.18	167				
	3	0.07	85				
	4	0.03	30				



a) 2-tuple variance

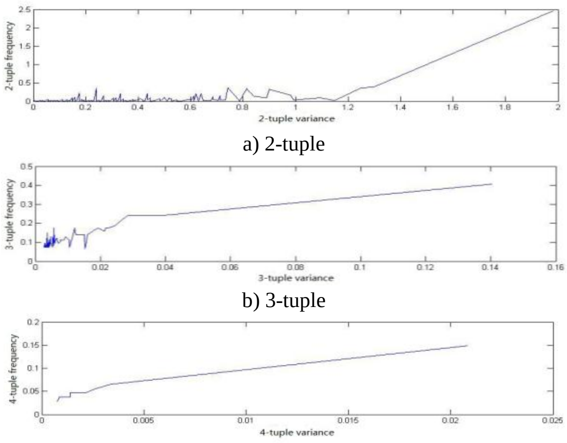
b) 3-tuple variance

c) 4-tuple variance

Fig. 3. Word variance distribution.

In the second step of dimension reduction, we recalculate the average frequencies of the features in ascending order of the k-tuples' variances. In Fig. 4 the horizontal axis denotes word variance, and the vertical axis denotes the average frequency of the corresponding word. It can be observed that some words' average frequencies are very low though their variances are relatively high. Considering that rare words may be useless for the prediction, the features were sorted according to the average frequency in descending order and chosen according to the threshold. The threshold values were set to be 0.18, 0.07, 0.03 (note that the value of k is 2, 3 and 4, respectively) according to the biggest decreasing gradient. As a result, the size of $\mathcal{D}$ shrinks to 302. The prediction results are shown in Table II. Apparently, the performance is improved, with increases of 2.5% on the total accuracy and 3.8% on sensitivity. The results suggest that the feature reduction works well and the words existing in $\mathcal{D}$ are discriminative.

Moreover, according to (4), we take the position information into consideration. The results suggest that the prediction accuracy can be apparently improved through this system. The accuracy is about 7.5% improved compared with the result produced using the former method, the sensitivity is 8.8% higher than that obtained after the former two processes and the specificity is up to 100%, as shown in Table III. It demonstrates that the hybrid features extracted separately from different parts of the N-terminal can contribute to the prediction performance.



a) 2-tuple

b) 3-tuple

c) 4-tuple

Fig. 4. Association chart of word average frequency and word variance.

The above results imply that the method with dimension reduction and hybrid features performs much better than using the original feature set. Since we have got the best parameters and discriminative features through the prediction system, we use the data set II to examine the generalization ability of our method. *TA* is used to evaluate the overall prediction performance of the method, which we have mentioned before. It can be observed in the second line of Table III that the accuracy (91.2%) was obtained when the size of $\mathcal{D}$ is only 302, which demonstrates the generalization ability of our method.

D. Performance Comparison with Current Prediction Models

In order to demonstrate the effectiveness of our method, we have compared a total of 8 methods as listed in Table IV. The method abbreviations and their corresponding description are in the following:

1) AAC: Amino acid compositions;
2) AAPC: Amino acid pair composition;
3) MT: Using motifs identified by MEME as features;
4) AAC+AAPC: The combination of Methods 1 and 2;
5) BPBAac: The method proposed by Y. Wang, Q. Zhang, M. Sun, and D. Guo. adopts the sliding-window technique and creates position-specific Aac profiles for classification [20];
6) Effective-T3: The method proposed by R. Arnold, S. Brandmaier, *et al.* [13];
7) SWANN: The method proposed by M. Löwer and G. Schneider. [12];
8) OM: The method proposed in this paper.

To make a fair comparison among these models, we used the same data set (data set II) to predict T3SEs. The results are summarized in Table IV. Obviously, our method has advantages over most of the other methods in classification accuracy, and amino acid pair composition (AAPC) performs the best among the former three sequence-based methods (method 1, 2, 4). Method 4 (AAC+AAPC) and method 3 (MT) have similar prediction accuracy. This observation suggested that the k-tuples can contain more information related to T3SS signals. Overall, the BPBAac method performs the best on identifying T3SEs with the highest sensitivity (91.8%), while our method has the highest specificity (95.6%) and total accuracy (91.2%), which indicates that BPBAac has a higher false positive rat than that of our method. Since our goal is to find novel effectors, our method with a low false positive rate could help reduce the cost of our future wet-bench experiments for validating the predicted effectors. In all, our method can be a useful computational tool for identifying T3SEs.

TABLE III: PREDICTION PERFORMANCE WITH THE HYBRID FEATURE METHOD

Data set	TA (%)	Sens (%)	Spec (%)
I	97.5	95.2	100
II	91.2	72.4	95.6

TABLE IV: METHOD LIST AND RESULT COMPARISON

No.	Method Abbr.	Sens (%)	Spec (%)	TA (%)
1	AAC	50.7	91.4	84.0
2	AAPC	54.5	92.0	85.3
3	MT	52.1	91.5	84.5
4	AAC+AAPC	52.3	91.6	84.6
5	BPBAac	**91.8**	90.9	91.1
6	Effective-T3	68.9	88.9	86.7
7	SWANN	63.0	94.9	88.1
8	OM	72.4	**95.6**	**91.2**

IV. CONCLUSION

This paper proposes an efficient method for extracting discriminative features and predicting type III secreted effectors using machine learning approaches. We extract features from the protein sequences and use support vector machines to classify T3SEs. We firstly record all the k-tuples present in the T3SS sequences in an overlapping manner. Then, to avoid the dimension disaster and select useful features, dimension reduction is conducted in two steps. The first step is to choose the features according to the word variance. The bigger the value of the variance, the more useful the word would be. Considering the similarity between protein sequences and natural language text, we segment the protein sequences into words according to the feature weight. By counting frequencies of the words, the final feature set is established and constitutes dictionary D. However, compared with the prediction method based only on extracted features, it is more effective to add the position information into consideration. Therefore, the protein sequences are converted into hybrid feature vectors by counting frequencies of the features separately from different parts of the N-terminal.

To demonstrate our method, we have conducted a series of experiments on two data sets, and the results show high accuracy and universal property especially by using hybrid features. We also made a comparison with other popular approaches. The comparison results show that our method has superiority over most of the existing methods. As a future work, we will keep exploring more specific signals and structural information so as to advance the understanding of type III secretion system.

REFERENCES

[1] J. Galán and A. Collmer, "Type iii secretion machines: bacterial devices for protein delivery into host cells," *Science*, vol. 284, no. 5418, pp. 1322–1328, 1999.

[2] J. Alfano and A. Collmer, "Type iii secretion system effector proteins: double agents in bacterial disease and plant defense," *Annu. Rev. Phytopathol.*, vol. 42, pp. 385–414, 2004.

[3] D. Büttner and S. He, "Type iii protein secretion in plant pathogenic bacteria," *Plant physiology*, vol. 150, no. 4, pp. 1656–1664, 2009.

[4] A. Hauser, "The type iii secretion system of pseudomonas aeruginosa: infection by injection," *Nature Reviews Microbiology*, vol. 7, no. 9, pp. 654–665, 2009.

[5] L. Worrall, E. Lameignere, and N. Strynadka, "Structural overview of the bacterial injectisome," *Current opinion in microbiology*, vol. 14, no. 1, pp. 3–8, 2011.

[6] A. Loquet *et al.*, "Atomic model of the type iii secretion system needle," *Nature*, 2012.

[7] P. Dean, "Functional domains and motifs of bacterial type iii effector proteins and their roles in infection," *FEMS microbiology reviews*, vol. 35, no. 6, pp. 1100–1125, 2011.

[8] J. Lewis, D. Guttman, and D. Desveaux, "The targeting of plant cellular systems by injected type iii effector proteins," *Seminars in cell & developmental biology*, vol. 20, no. 9. Elsevier, pp. 1055–1063, 2009.

[9] T. Petnicki-Ocwieja, D. Schneider, V. Tam, S. Chancey, L. Shan *et al.*, "Genomewide identification of proteins secreted by the hrp type iii protein secretion system of pseudomonas syringae pv. tomato dc3000," in *Proc. of the National Academy of Sciences*, vol. 99, no. 11, 2002, pp. 7652–7657.

[10] E. Panina, S. Mattoo, N. Griffith, N. Kozak, M. Yuk, and J. Miller, "A genome-wide screen identifies a bordetella type iii secretion effector and candidate effectors in other species," *Molecular microbiology*, vol. 58, no. 1, pp. 267–279, 2005.

[11] T. Tobe *et al.*, "An extensive repertoire of type iii secretion effectors in escherichia coli o157 and the role of lambdoid phages in their dissemination," in *Proc. of the National Academy of Sciences*, vol. 103, no. 40, 2006, pp. 14941–14946.

[12] M. Löwer and G. Schneider, "Prediction of type iii secretion signals in genomes of gram-negative bacteria," *PloS one*, vol. 4, no. 6, pp. e5917, 2009.

[13] R. Arnold, S. Brandmaier, F. Kleine, P. Tischler, E. Heinz, S. Behrens, A. Niinikoski, H. Mewes, M. Horn, and T. Rattei, "Sequence-based prediction of type iii secreted proteins," *PLoS pathogens*, vol. 5, no. 4, pp. e1000376, 2009.

[14] M. Vencato *et al.*, "Bioinformatics-enabled identification of the hrpl regulon and type iii secretion system effector proteins of pseudomonas syringae pv. phaseolicola 1448a," *Molecular plant-microbe interactions*, vol. 19, no. 11, pp. 1193–1206, 2006.

[15] Y. Yang, J. Zhao, R. Morgan, W. Ma, and T. Jiang, "Computational prediction of type iii secreted proteins from gram-negative bacteria," *MC bioinformatics*, vol. 11, no. Suppl 1, pp. S47, 2010.

[16] Y. Yang, "A comparative study on sequence feature extraction for type iii secreted effector prediction,", in *Proc. Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE,* vol. 3. 2011, pp. 1560–1564.

[17] M. Bhasin and G. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23 262–23 266, 2004.

[18] E. Zaslavsky *et al.*, "A combinatorial optimization approach for diverse motif finding applications," *Algorithms for Molecular Biology*, vol. 1, no. 1, pp. 13, 2006.

[19] M. Shamim, M. Anwaruddin, and H. Nagarajaram, "Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs," *Bioinformatics*, vol. 23, no. 24, pp. 3320–3327, 2007.

[20] Y. Wang, Q. Zhang, M. Sun, and D. Guo, "High-accuracy prediction of bacterial type iii secreted effectors based on position-specific amino acid composition profiles," *Bioinformatics*, vol. 27, no. 6, pp. 777–784, 2011.

[21] S. Qi, Y. Yang, and A. Song, "Feature reduction using a topic model for the prediction of type iii secreted effectors," in *Proc. Neural Information Processing*, Springer, 2011, pp. 155–163.

[22] Y. Yang and B. Lu, "Extracting features from protein sequences using chinese segmentation techniques for subcellular localization," in *Proc. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, IEEE*, 2005, pp. 1–8.

[23] C. Casper-Lindley, D. Dahlbeck, E. Clark, and B. Staskawicz, "Direct biochemical evidence for type iii secretion-dependent translocation of the avrbs2 effector protein into plant cells," in *Proc. of the National Academy of Sciences*, vol. 99, no. 12, 2002, pp. 8336–8341.

[24] M. Mudgett, O. Chesnokova, D. Dahlbeck, E. Clark, O. Rossier, U. Bonas, and B. Staskawicz, "Molecular signals required for type iii secretion and translocation of the xanthomonas campestris avrbs2 protein to pepper plants," in *Proc. of the National Academy of Sciences*, vol. 97, no. 24, 2000, pp. 13324–13329.

[25] L. Schechter, K. Roberts, Y. Jamir, J. Alfano, and A. Collmer, "Pseudomonas syringae type iii secretion system targeting signals and novel effectors studied with a cya translocation reporter," *Journal of bacteriology*, vol. 186, no. 2, pp. 543–555, 2004.

[26] K. Nishikawa, Y. Kubota, and O. Tatsuo, "Classification of proteins into groups based on amino acid composition and other characters. i. angular distribution," *Journal of biochemistry*, vol. 94, no. 3, pp. 981–995, 1983.

[27] W. Jin, "Chinese segmentation disambiguation," in *Proc. of the 15th conference on Computational linguistics-Volume 2. Association for Computational Linguistics*, 1994, pp. 1245–1249.

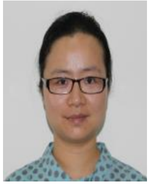
[28] M. Ahmed and L. Khan, "Sisc: A text classification approach using semi supervised subspace clustering," in *Proc. IEEE International Conference on Data Mining Workshops*, 2009, pp. 1–6.

[29] M. Ahmed, L. Khan, N. Oza, and M. Rajeswari, "Multi-label asrs dataset classification using semi-supervised subspace clustering," in *Proc. Conference on Intelligent Data Understanding*, 2010.

[30] K. Krishna, P. Ganesan, H. Mehrnaz, K. Kai-Uwe, and M. Thomas, "Blprot: prediction of bioluminescent proteins based on support vector machine and relieff feature selection," *BMC Bioinformatics*, vol. 12.

[31] C. Chang and C. Lin, *LIBSVM: a library for support vector machines*, 2011.

Tingting Sui was born in 1988. She received the B.S. in computer science and technology from Shanghai Maritime University, China, in 2011. She is currently doing her master degree in Shanghai Maritime University, where she is researching aspects of bioinformatics, machine learning, and data mining.



Yang Yang is an associate professor in the Department of Computer Science and Engineering of Shanghai Maritime University. She received her B.S. and Ph.D. in computer science from Shanghai Jiao Tong University in 2003 and 2009, respectively.

She was a visiting researcher in the Department of Computer Science in the University of California, Riverside during 2007-2009 and 2012-2013. Her research interests include machine learning, bioinformatics and data mining.

Xiaofeng Wang is a professor in the Department of Computer Science, Shanghai Maritime University. He is currently the dean in the Department of Computer Science of Shanghai Maritime University, His research interests include data mining and knowledge discovery, grid computing and application, bioinformatics and intelligent information processing.