A new feature selection method for computational prediction of type III secreted effectors

Yang Yang* and Sihui Qi

Department of Computer Science and Engineering,
Shanghai Maritime University,
1550 Haigang Ave., Shanghai 201306, China
Email: yangyang@shmtu.edu.cn
Email: yangy09@gmail.com
Email: qsh_0123@126.com
*Corresponding author

Abstract: The type III secretion system (T3SS) is a specialised protein delivery system that plays an important role in pathogenic bacteria. However, the secretion mechanism has not been fully understood yet. Especially, the identification of type III secreted effectors is a notoriously challenging problem which has attracted a lot of research interests in recent years. In this paper, we introduce a machine learning method using amino acid sequence features for predicting T3SEs. We use a topic model called HMM-LDA to select useful features, and conduct experiments on *Pseudomonas syringae* as well as some other bacterial genomes. The cross-validation results on *P. syringae* data set show an improved prediction accuracy with the reduced feature set. The experimental results on the test sets also demonstrate that the accuracy of the proposed method is comparable to or better than the accuracies achieved by other available T3SE prediction tools.

Keywords: type III secretion system; effector; topic model; computational prediction; feature selection.

Reference to this paper should be made as follows: Yang, Y. and Qi, S. (2014) 'A new feature selection method for computational prediction of type III secreted effectors', *Int. J. Data Mining and Bioinformatics*, Vol. 10, No. 4, pp.440–454.

Biographical notes: Yang Yang received her PhD in Computer Science in 2009 from Shanghai Jiao Tong University, China. She is an Associate Professor in the Department of Computer Science, Shanghai Maritime University, China. Her research interests focus on machine learning and bioinformatics.

Sihui Qi is a graduate student in the Department of Computer Science in Shanghai Maritime University. She obtained her BS from Liaocheng University, Shandong, in 2010. Her main research interests are data mining and bioinformatics.

1 Introduction

The type III secretion system (T3SS) is among the most complex secretion systems in Gram-negative bacteria (Galán and Collmer, 1999), which is indispensable for the pathogenesis of a large variety of plant and animal pathogens, such as *Pseudomonas*, *Erwinia*, *Xanthomonas*, *Ralstonia*, *Salmonella*, *Yersinia*, *Shigella* and *Escherichia*, etc. (He et al., 2004; Cornelis, 2006). Using T3SS, these pathogens inject virulence proteins, so-called type III secreted effectors (T3SEs), directly into the host cells. These T3SEs are transported across the two membranes of the Gram-negative bacteria and the membrane of the host cell through a needle-like apparatus. Then they target their specific host substrates and promote disease development in the host cell. Recently, T3SS and T3SEs are also found in non-pathogenic bacteria, such as the microsymbiont rhizobia (Viprey et al., 1998).

Researchers have been exploring the working principle and mechanism of T3SS for over a decade. Although the structure of T3SS apparatus has been uncovered, the precise mechanism underlying the secretion process has not been fully understood yet. In recent years, more and more emphases have been put on the studies of T3SEs. It is not only because of their important functions for the virulence of pathogens, but also because they could provide hints in discovering the working principle and mechanism of TTSS. These effector proteins must have some unique characteristics that could be recognised by the T3SS and trigger the secretion process. However, there is no defined secretion signal that has been found in the known effectors. Moreover, the number of T3SEs that have been confirmed is only several hundreds. The plant pathogen *Pseudomonas syringae* has been a model organism for the study of type III effectors. Thus far, over two hundreds of T3SEs have been identified and confirmed in *P. syringae* strains, more than the total number of effectors identified from all other bacterial species. Therefore, we conjecture that a large portion of T3SEs in other bacteria remain unknown.

A lot of efforts have been dedicated to identifying novel T3SEs and searching the secretion signal both by wet-bench and computational methods. The wet-bench methods, e.g., functional screen and protein secretion assay (Guttman et al., 2002), are accurate but labor-intensive, which cannot deal with high-throughput screening. As the sequencing techniques have gained breakthrough for the past decade, and a large number of sequenced genomes for plant and animal pathogens became available, bioinformatics approaches are in great demand for accelerating the verification of T3SEs. Some computational tools for the prediction of T3SEs have been developed. (Vinatzer et al., 2005; Arnold et al., 2009; Löwer and Schneider, 2009; Yang et al., 2010; Wang et al., 2011; Sato et al., 2011) The ultimate goal of a prediction system is to produce an accurate effector candidate list that could help increase the efficiency of wet-bench experimental verification and discover the signals that direct the secretion.

The computational methods typically follow two trends: domain knowledge-based and sequence-based. The domain knowledge-based methods utilise biological features (Yip et al., 2009; Jin et al., 2008; Huang and Pavlovic, 2008), including searching conserved regulatory motifs in the promoters (Ferreira et al., 2006) (e.g., *P. syringae* has a motif called the hrp box), identifying genes in vicinity to chaperone homologues (Panina et al., 2005), predicting unstability of N-terminus and non-optimal codon usage (Sato et al., 2011), etc. These methods have some limitations: a) Not all genes preceded

by the TTSS-related regulatory motif are T3SEs, and some effectors do not have the promoters at all; b) Some effectors function without a chaperone, or are encoded at separate loci; c) The domain knowledge is usually not available but needs to be calculated by computational tools, which can not provide accurate information. In this paper, we focus on the sequence-based methods.

The sequence-based methods mainly study the amino acid or nucleotide sequences (Dong et al., 2010; Haddow et al., 2011). Comparisons of the amino acid sequences of known T3SEs show great sequence diversity. This is because they evolve fast in order to adapt to different hosts and respond to the resistance from the host immune systems (Ma and Guttman, 2008). Researchers have detected amino acid composition biases in T3SEs, especially in the N-terminus, such as an overall amphipathic amino acid composition, an over-representation of serine and glutamine, and the absence of acidic residues (Petnicki-Ocwieja et al., 2002). A typical type III effector usually contains a secretion/translocation signal in the N-terminus and a functional domain in the C-terminus. The secretion signal is believed to be contained in the first 50 or 100 amino acids (Schechter et al., 2004; Lloyd et al., 2002; Wang et al., 2011). Actually, the first 15 amino acids are most essential. However, these features are not accurate enough to identify new effectors because some effectors do not possess these features at all (Schechter et al., 2006). Besides, many known effectors are identified by homology search, which fails to find novel type of effectors. Recently, some machine learning methods have been proposed for the prediction of T3SEs (Arnold et al., 2009; Löwer and Schneider, 2009; Yang et al., 2010; Wang et al., 2011). They attempt to extract features from protein sequences and perform prediction based on these features. Arnold et al. (2009) used the frequencies of amino acids as well as the frequencies from two reduced alphabets, i.e., they mapped amino acids to groups according to the amino acid properties. They also computed the frequencies of di- and tri-peptides from each of the alphabets. Löwer and Schneider (2009) used sliding-window technique to extract features. Yang et al. (2010), Yang (2011) used amino acid composition, *k*-mer composition, as well as SSE-ACC method (amino acid composition in terms of their different secondary structures and solvent accessibility states). Wang et al. (2011) proposed a position-specific feature extraction. The position-specific occurrence time of each amino acid is recorded, and then the profile is analysed to compose features. All of these methods aim to represent the amino acid composition, order and position information as feature vectors.

In this paper, we regard the protein sequences as text written in a certain kind of biological language. The residues and peptides, i.e., *k*-mers (*k*-tuple amino acid sequences) are the words composing the text. Since the number of *k*-mers would be very large when k increases, we conduct feature reduction instead of using all the k-mers as features. In order to eliminate the noisy words effectively, we adopt a topic model, HMM-LDA, whose advantage over other LDA models is that it introduces both syntax states and topics. We have carried out in-depth exploration on selecting informative words, and conducted a series of experiments to examine their discriminative ability. The experiments on *P. syringae* data set show an improved prediction accuracy with the reduced feature set. Furthermore, we have applied the method to a variety of bacterial genomes. The results demonstrate that the new method has comparable or better performance than the existing T3SE prediction tools.

2 Methods

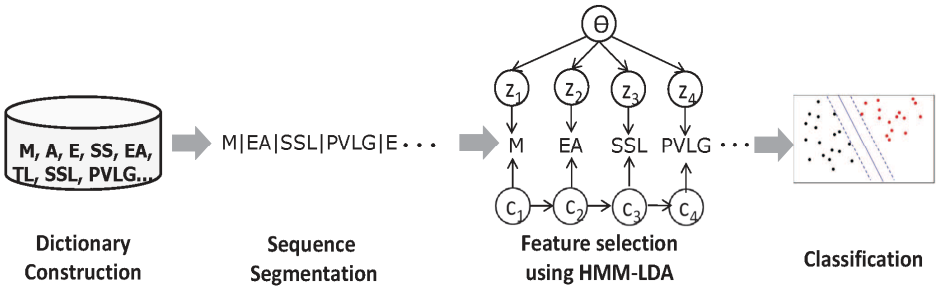
In this paper, we model the protein sequences as a kind of biological language, which is composed of words without any space or punctuation in between the words. Considering each single amino acid is the smallest unit in protein sequence, we assume each amino acid as a single-character word, and *k*-mers can be regarded as multi-word lexemes in natural language. Similar to Chinese text, in which segmentation is an important and basic step for text processing, we first perform a segmentation process on the amino acid sequences in order to separate the character strings into meaningful words or phrases.

Obviously, there are many differences between natural languages and protein sequences. Protein sequences have a smaller alphabet, but are much longer than text sentences. In text, words are minimal independent and meaningful language units, and natural languages usually have predefined dictionaries. However, protein sequences are written in an unknown language to us at the present state, whose words are not delineated. Any combination of letters with arbitrary length may be a word. So we first need to build a dictionary, which is the basis of segmentation. Therefore, our method consists of four steps: 1) Construct a dictionary, i.e., word set; 2) Segment the protein sequences by matching the words in the dictionary; 3) Run HMM-LDA model on the segmented sequences and select informative words from the dictionary; 4) Create feature vectors and conduct the classification. Figure 1 shows the pipeline of our method. In the following, we describe the four steps respectively in details.

2.1 Dictionary construction

In natural languages, words are generally the combinations of characters that frequently appear in the text. Thus in our study, the occurrence time of each *k*-mer in the data set is recorded and the frequent ones are put into the dictionary. To avoid encountering unknown words, all 20 amino acids should be included in the dictionary. For the *k*-mers ($k > 1$), we only preserve a certain portion for each value of *k* according to their occurrence times.

Figure 1 Flowchart of the new method (see online version for colours)



2.2 Segmentation

Segmentation is the process of matching sequences with words in the dictionary. In this step, we use the segmentation method proposed in (Yang et al., 2008). We first consider the segmentations which generate the least number of segments, i.e., long words are

preferred to be matched. This is based on the consideration that longer strings contain more sequence information. The idea is similar to the maximum match (MM) algorithm (Wong and Chan, 1996) widely used in Chinese word segmentation. It is not enough to consider solely the number of segments, because there are still multiple ways of segmentation with the same number of segments. Then we assign a weight for each word in the dictionary and add a maximum weight product criterion to ensure the unique best segmentation. For any given sequence, the segmentation which has the biggest weight product is selected.

2.3 HMM-LDA and feature reduction

At this stage, we already have built a dictionary, but not all of the words in the dictionary are necessary to be the features, like the auxiliary words in language text, e.g., "in", "some", "however". Therefore, we need to further condense the feature set. In this step, by using the topic model, we introduce a latent topic layer into the original segmented sequences.

The topic model is a kind of statistical model in the realm of machine learning and natural language processing. It is able to discover the implicit topic information in the document. Over the last decade, topic models have been researched extensively. Besides in text automatic classification, information retrieval and other related applications of natural language processing, they have also been successfully applied in image segmentation and classification, social network analysis, etc. In the realm of bioinformatics, some researchers have used topic models and obtained good results. For example, in the study of protein remote homology detection, Liu et al. (2008) used latent semantic indexing model, and Yeh and Chen (2010) used latent topic vector model. Their topic models have higher accuracies than word-based models.

The latent Dirichlet allocation (LDA) (Blei et al., 2003), perhaps the most common topic model currently in use, describes each document in a corpus as generated from a mixture of topics, and each topic is characterised by a word distribution. The HMM-LDA model further extends this topic mixture model by separating syntactic words from content words whose distributions depend primarily on local context and document topic, respectively. The major difference between LDA and HMM-LDA is that each word is generated independently in LDA model, while there is local dependencies between nearby words in HMM-LDA model. We have experimented both original LDA and HMM-LDA models, and the latter one performs better (See Section 3.5). That may be because the HMM-LDA model discovers both syntactic classes and semantic topics in the document, and it is more helpful to eliminate the noisy words, thus we used HMM-LDA in our study.

After building the HMM-LDA model over the segmented protein sequences, we obtain latent topic information of the words, which can be utilised for feature selection, i.e., we would like to select the informative words as features.

Intuitively, two types of words can be removed from the feature set, unusual words and widespread words. We have proposed an algorithm (Qi et al., 2011) to eliminate these two kinds of words by setting two key parameters. One is a lower bound of word frequency used to eliminate the unusual words, and the other is a lower bound for frequency difference, which is used to remove the words that appear nearly equally on multiple topics. Specifically, this algorithm aims to search the words specific to some topics, i.e., to keep the words which are assigned to some topics with high probability but

do not pervasively assigned to many topics. This algorithm obtains better accuracy compared with the method using the feature set without this feature reduction step, but the criterion of selecting word is a little rigorous.

Here we introduce a simpler but more effective algorithm, shown in Algorithm 1, which mainly considers the rare topics and unusual words. First, the appearance time of the most popular word of each topic (n_t = max $n_{w,t}$, where $n_{w,t}$ is the number of times that word w has been assigned to topic t) is recorded. Then all the topics are sorted according to n_t. Some rare topics, i.e., the topics with low ranks, can be discarded. The parameter m is used to determine the number of topics to be considered. The third step, also the major step, is selecting words whose occurrence times exceed a certain value, the parameter x. Since the HMM-LDA model has already distinguished syntax words and topic words, and most widespread used or equally distributed words are syntax words, we use Algorithm 1 in this paper and the experimental results also demonstrate its efficacy.

Algorithm 1

Input: Word set $\mathcal{W}$

Output: Reduced word set $\mathcal{W}$'

Set $\mathcal{W}' = \phi$

Sort $t \in \mathcal{T}$ in descending order according to n_t, where $n_t = \max_{w \in \mathcal{W}} n_{w,t}$.

Let L be the list of sorted topics.

for each topic t that ranks top-m of L **do**
  for each word w **do**
    if $n_{w,t} > x$, and w is not in $\mathcal{W}$ **then**
      Add w to $\mathcal{W}'$.
    end if
  end for
end for

2.4 Classification

After the feature reduction procedure, we calculate the appearance time of each word in the feature set (based on the segmentation result and HMM-LDA model), thus construct the feature vectors. We use the support vector machines (SVMs) as the classifiers, which are widely used in bioinformatics because of their excellent and stable performance in the classification tasks. Here, we used the RBF kernel function. A grid search in logarithmic space was performed to find optimal values for the complexity parameters C and γ.

3 Results and discussions

3.1 Data set

Pseudomonas syringae is a model organism in plant pathology and has by far the largest number of putative and confirmed effectors. We have collected a total of 283 confirmed

Pseudomonas effectors from databases and literatures, belonging to three strains, *P. syringae* pv. tomato strain DC3000, *P. syringae* pv. syringae strain B728a and *P. syringae* pv. *phaseolicola strain* 1448A. Since homology search has been a major means to discover putative effectors, the sequence similarity of this data set is very high. Considering that the redundancy of the data set would result in overestimation on the accuracy of the classifier, we clustered these effectors with sequence similarity over 60% and kept only the representative sequence of each cluster in our data set, leaving 108 positive samples.

The negative data set was extracted from the genome of *P. syringae* pv. tomato strain DC3000 because it has been intensively investigated for the research of TTSS. We excluded all the proteins related to T3SS, as well as the hypothetical proteins. (Note that this set may still contain some unknown effectors.) And then we selected randomly from the remaining samples to constitute the negative set, which has 760 samples, thus there is a total of 868 samples of this data set.

In order to examine the generalisation ability of the prediction system, we prepared test sets of other bacterial genomes. The first test set is composed of type III effectors from rhizobia. The type III secretion system has been shown to play an important role during the nodulation process of several rhizobial species. As multiple rhizobial species have the T3SS apparatus, the function and mechanism of T3SS in nodulation have received a lot of attention in the research field of plant-microbe interactions (Marie et al., 2003). However, only a few rhizobial T3SEs have been confirmed. Therefore, computational tools are in great demand to detect novel secreted proteins in rhizobia. Although the biological effects of T3SS are different in rhizobia and *P. syringae*, we have discovered that they have similar secretion mechanisms (Yang et al., 2010). 12 confirmed T3SEs were collected from four rhizobial strains for test.

The second test set consists of confirmed effectors from multiple species including both plant pathogens and animal pathogens, such as *Salmonella enterica*, *Yersinieae*, *Shigella* and *Escherichia coli*. Although the TTSS mechanism has great diversity among different species, we want to check the universality of our prediction system. This test set is mainly consisted of a T3SE database we have maintained before as well as the data used in Wang et al. (2011) and Sato et al. (2011), including a total of 194 samples.

As mentioned before, the secretion signal is believed to be contained in the first 50 or 100 amino acids. The N-terminal 100 amino acids have been demonstrated useful for the identification of effectors (Yang et al., 2010; Wang et al., 2011). Thus in this study, the first 100 amino acids are used for feature creation.

3.2 Experimental settings and evaluation criteria

In the experiments, we used HMM-LDA model from the Matlab Topic Modeling Toolbox 1.4 (Steyvers and Griffiths, 2011). As in LDA, the number of topics has great impact on the performance of HMM-LDA. The optimum number of topics was searched in the range from 5 to 95. We found that the highest precision (72.2%) was obtained when the number of topics is 55 on the validation dataset. The other parameters used in the HMM-LDA model is set to be default, i.e., number of syntactic states is 12, $\alpha = 50/T$ (T is the number of topics), and $\beta = 0.01$.The parameter m is set to be 50. x is set to be $n_t/10$, which is a relatively small threshold to alow keeping more words. (See Algorithm 1 for the calculation of n_t.)

Our implementation of the support vector machines adopted LibSVM version 2.8 (Chang and Lin, 2001). The kernel parameter γ and C are set as 2^{-5} and 2^4, respectively.

In order to provide reliable predictions for future wet bench analysis, we used three metrics to evaluate the performance of the proposed method, including precision (*P*), recall (*R*) and total accuracy (*TA*). The precision is the ratio of the samples correctly classified into the positive class compared to the total number of samples classified into the positive class. And the recall measures ratio of samples classified as positive among all positive examples. These two metrics are used to measure the prediction quality of effectors, and *TA* is used to measure the overall prediction quality, i.e., the ratio of the test samples the system classifies correctly. They can be defined in terms of the number of true positives (*TP*), the number of false positives (*FP*), and the number of false negatives (*FN*) as follows.

$$P=\frac{TP}{TP+FP} \tag{1}$$

$$R=\frac{TP}{TP+FN} \tag{2}$$

$$TA=\frac{TP+TN}{TP+FP+TN+FN}. \tag{3}$$

3.3 Cross-validation results on Pseudomonas syringae

The first 100 N-terminal amino acid residues were retrieved from effectors and non-effectors, and converted to feature vectors. We conducted five-fold cross-validation on the 868 samples from *P. syringae*.

In our method, dictionary construction, the first step, directly influences the segmentation result in the second step. Besides, the size of the dictionary also influences the feature reduction procedure in the third step and the size of the final feature set. Therefore, in the experiments, we need to estimate the proper size of the dictionary. Firstly, a maximum word length *MaxLen* should be set, which specifies the set of *k*-mers from which words are selected. In our previous studies on the prediction of T3SEs, we found that due to the diversity of the effector sequences, long *k*-mers could not improve the prediction performance. In fact, *k*-mers ($k > 4$) has little help on improving accuracy but largely increases the computational complexity. Therefore, we only include single amino acids, di-mers and tri-mers in the dictionary. All 20 amino acids, the basic unit of the biological language, are included in the dictionary to avoid encountering unknown words. Di-mers and tri-mers are selected according to their appearance times. We have conducted a series of experiments to search the proper number of di-mers and tri-mers. The numbers of di-mers and tri-mers are searched from 50 to 350, respectively. Figure 2 shows the precisions of different combinations of numbers of di-mers and tri-mers using Algorithm 1.

From Figure 2, we can find that di-mers play a more important role in the prediction than tri-mers. Adding a few number of tri-mers (about 50) can improve the prediction accuracy. However, as the number of tri-mers increases, the accuracy decreases. Moreover, not all the di-mers are useful. Obviously, when the number of di-mers exceeds 250, we can not get a satisfying precision.

Figure 2 Cross validation results with different number of words

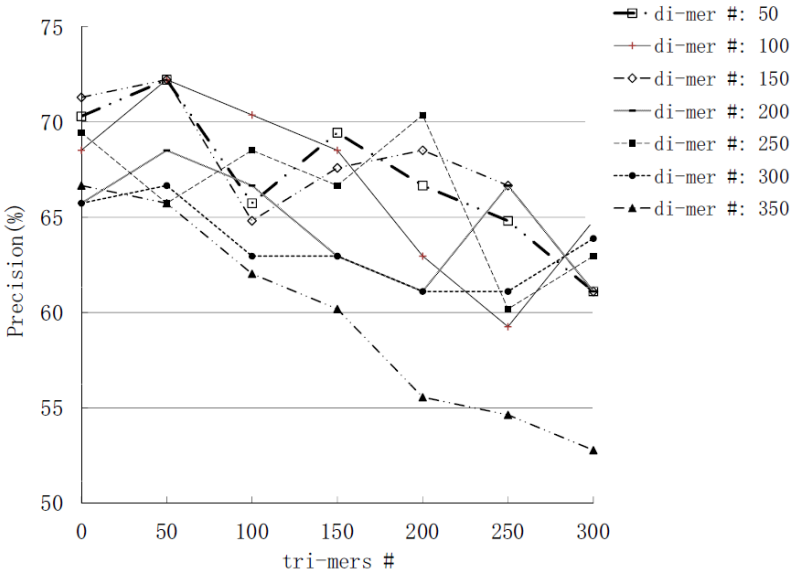


Table 1 Cross validation results obtained using different number of words

# Words in Dictionary		*P (%)*	*R (%)*	*TA (%)*	*# words selected*	
70	(20+50+0)	70.3	77.5	94.0	70	(20+50+0)
120	(20+50+50)	72.2	76.5	93.8	70	(20+45+5)
170	(20+100+50)	**72.2**	70.3	92.7	125	(20+90+15)
220	(20+100+100)	71.3	70.6	92.7	135	(20+93+22)
170	(20+150+0)	71.2	73.5	92.8	161	(20+141+0)
220	(20+150+50)	**72.2**	**83.9**	**94.8**	176	(20+142+14)
470	(20+250+200)	70.4	72.4	93.0	308	(17+238+53)

The first column and last column show the dimensions of feature vectors before and after feature reduction, and the three numbers in the parentheses are the numbers of single amino acids, di-mers and tri-mers, respectively.

There are 7 combinations which obtain precisions higher than 70%. In Table 1, we list the detailed precisions, recalls, total accuracies, and the numbers of dimensions before and after feature selectiong (using HMM-LDA) of the 7 combinations for a comprehensive comparison. It can be observed that the best accuracy is obtained when the numbers of single amino acids, di-mers and tri-mers are 20, 150 and 50, respectively. For all these 7 groups, the numbers of tri-mers have been greatly reduced, while there is only a little decrease in the numbers of di-mers. This result further illustrates that long *k*-mers don't have the expected effect in the classification. We have performed a further analysis on the tri-mers. The occurrence times of all the tri-mers from the dictionary are recorded in the N-terminal 100 amino acids of the positive and negative sets respectively. The ratio of these two values ranges from 0 to 0.47. We sort the ratios in ascending order, and find that the top 10 and last 8 values are all from the removed tri-mers, while the selected tri-mers' ratios are distributed in the interval from 0.03 to 0.14. More specifically, both the tri-mers that appear zero or only one time in the effector sets are

discarded, and the tri-mers that appear frequently both in the positive and negative sets are also discarded. How the selected *k*-mers relate with the secretion signals is under further investigation.

3.4 Performance on other bacterial genomes

In order to evaluate the generalisation ability and universality of the prediction system, we conducted prediction on other bacterial genomes that encode T3SSs. The training data consist of 868 proteins from *Pseudomonas syringae* used in the cross validation. The two test sets are described in Section 3.1.

In the first test set, all the effectors are from rhizobia. Our method correctly identifies 10 of the 12 effectors, generating an accuracy of 83.3%. The second test set is composed of multiple bacterial species. The type III secreted effectors exhibit great diversity and have a wide variety of functions across strains and species. Thus the cross-species prediction is a difficult job. The prediction system obtains a recall of 43.8% by recognising 85 of the 194 effectors. This result suggests that although the secreted proteins are diverse across species, there are common patterns on N-terminal amino acid sequences.

Table 2 Result Comparison on Pseudomonas syringae using five-fold cross-validation

Method	*Dimension*	*TA (%)*	*P (%)*	*R (%)*
di-mer	400	93.2	65.7	76.3
tri-mer	8000	91.6	32.4	**100**
frequency	220	93.7	**72.2**	75.7
LDA	197	94.2	71.3	80.2
HMM-LDA	176	**94.8**	**72.2**	83.9

3.5 Performance comparison with other methods

We examined the performance of the new method by comparing the prediction accuracies of multiple methods. Table 2 lists the number of dimensions, total accuracy (TA), precision and recall of five methods, respectively. The first and second methods use all the di/tri-mers without feature reduction. The feature vectors are created by counting the occurrence times of all the di/tri-mers overlapingly. The third method records the occurrence times of all the dictionary words as features, i.e., which has the initial step of feature reduction by *k*-mer frequency. And the last two methods utilise LDA and HMM-LDA for feature selection respectively on the basis of the dictionary, i.e., feature set condensed by the frequency criterion.

Table 2 clearly shows that the reduced feature set achieves good performance. The tri-mer method has the biggest number of dimensions, but its performance is the worst. The reason why the tri-mer method has such high recall and low precision is that the number of false negative is zero while the number of false positive is very big. We also conducted experiments using *k*-mers with $k > 3$, and the accuracy is even worse. On the contrary, the HMM-LDA method has only 176 dimensions, but it has the best classification performance.

In addition, we have compared our method with two publicly available computational tools for T3SE prediction, EffectiveT3 and BPBAac. EffectiveT3 (Arnold et al., 2009), the first universal in silico prediction program for the identification of novel TTSS

effectors, which takes into account frequencies of amino acids and short peptides (di-mers and tri-mers) as features and uses Naive Bayesian Classifier. Another SVM-based computational T3SE prediction model called BPBAac (Wang et al., 2011), which has been newly developed and has high prediction accuracy. The result comparison is shown in Table 3, in which the thresholds are set to be default, i.e., 0.99 for NB classifier in EffectiveT3 and 0.5 for SVMs in BPBAac and our methods.

Our method has the highest accuracy in predicting rhizobial effectors, while BPBAac performs the best in prediction of the hybrid data set. Overall, our method has a comparable performance with other available computational tools for T3SE prediction.

Table 3 Result Comparison on the two test sets

Method	*# recognised effectors*		*Accuracy (%)*	
	Test set 1	*Test set 2*	*Test set 1*	*Test set 2*
EffectiveT3	8	84	66.7	43.3
BPBAac	8	102	66.7	**52.6**
Our method	10	85	**81.3**	43.8

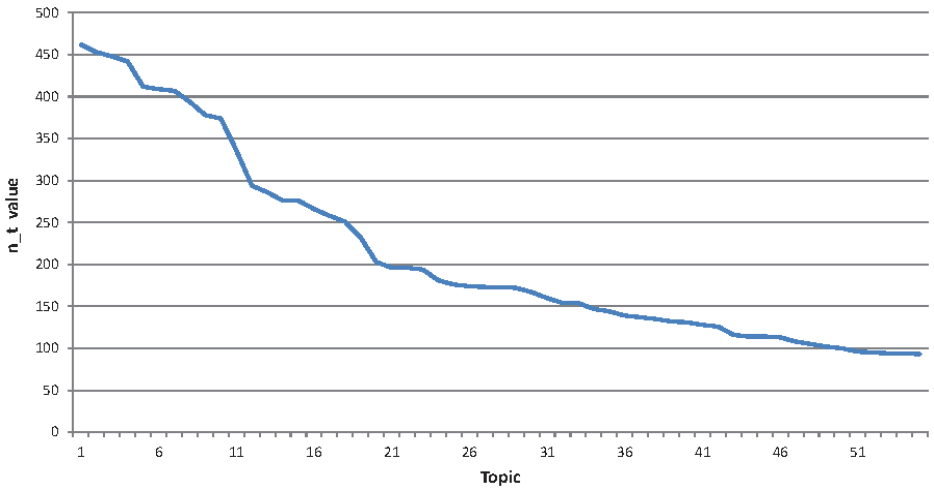
3.6 Discussion

This paper proposes a new feature selection method, which consists of two key points. The first point is to eliminate noisy topics, and the second point is to remove rare words.

For natural language, a topic is what a text, a paragraph or a sentence is about. It is a particular subject that the text write about. Each topic has its symbolic words. In Algorithm 1, we sort the topics according to n_t in descending order. A topic with a low rank means that none of the words has been assigned to it with many times, i.e., the topic may have no symbolic word at all. Therefore, we regard such topics as noisy topics.

As a further illustration, we plot the n_t values for the sorted list of all the 55 topics used in our experiments in Figure 3.

Figure 3 n_t values of topics in descending order (see online version for colours)



We can observe from this figure that n_t decreases more slowly at the lower-ranking topics. n_t values of the last five topics are all less than 100 and nearly equal. Therefore, it is a proper choice to consider the first 50 topics for feature selection in the experiments.

In the word selection step, we removed the words which are not assigned to any topic with more than x times. We did not set a constant value for this threshold x, but used $n_t/10$ instead. After running the HMM-LDA model, we examined the matrix containing the number of times word i has been assigned to topic j. The matrix is very sparse. We found that the numbers of times that the words are assigned to high-ranking topics have great differences. For example, for the first topic, the most popular word 'I' has been assigned to the topic over 450 times, while the least frequent word 'S' has been assigned just once. As for the low-ranking topics, the most popular words have been assigned for a little more than 100 times, and the least frequent words have been assigned once or several times. Generally, each topic has 5~15 words. The threshold $n_t/10$ can screen the rare words and keep the symbolic words for each topic.

In the present study, the experimental results have demonstrated the effectiveness of the selected words. However, given the word set and topic information, it is still difficult to determine the specific signal or characteristics that directly relate to the secretion process. For protein sequences, the latent layer revealed by the topic models could be secondary or spatial structure, function domain or other biochemical properties. The association among the symbolic words of the topics and the latent biological characteristics are considered for the future study in deciphering the precise secretion mechanism.

4 Conclusion

In this paper, we use machine learning approaches to predict proteins secreted via the type III secretion system. We extract features from the N-terminal amino acid sequences by regarding the sequences as text documents and k-mers as words. Firstly, a dictionary is constructed according to k-mer frequency. Secondly, the protein sequences are segmented into non-overlapping k-mers. Then we model a latent topic layer between the document and words. Each protein sequence is a mixture of a number of topics, and each word is assigned to a certain topic. The topic model called HMM-LDA is adopted for feature selection. At last, by using the state-of-art classifier, support vector machines, we constructed the system to distinguish T3SEs and non-T3SEs.

A five-fold cross-validation was conducted to examine the prediction accuracy on *Pseudomonas syringae* data set. The best accuracy is achieved when there is a total of 220 words in the dictionary, including 20 amino acids, 150 di-mers and 50 tri-mers. After feature selection using HMM-LDA, the number of features reduces to 176. The prediction accuracy is much better than using di-mers and tri-mers without sequence segmentation, and even better than using the dictionary words (the feature set before the feature reduction by HMM-LDA). Besides *Pseudomonas syringae*, we also conducted predictions on two test sets including type III effectors from multiple bacterial genomes. Rhizobium is an important kind of bacteria for symbiotic study. Our method successfully identifies 10 of 12 confirmed rhizobial effectors, better than two other computational tools, EffectiveT3 and BPBAac. As for a hybrid data set including 194 effectors from a variety of strains, our method also has a comparable performance with other prediction systems.

Thus far, a large portion of T3SEs still remain unknown. Bioinformatics tools are of great importance for high-throughput recognition of T3SEs and exploration of their characteristics. We believe that this new computational method can be widely used for efficient prediction of T3SEs in various bacteria species. By modeling the protein sequences into a kind of biological language, discovering secretion signals according to the selected informative words and latent topic information is a subject for further investigation.

Acknowledgements

We would like to thank Professor Wenbo Ma at the Department of Plant Pathology and Microbiology, UC, Riverside, for providing the effector data sets. 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).

References

Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, H., Horn, M. and Rattei, T. (2009) 'Sequence-based prediction of type III secreted proteins', *PLoS Pathogens*.

Blei, D., Ng, A. and Jordan, M. (2003) 'Latent dirichlet allocation', *The Journal of Machine Learning Research*, Vol. 3, pp.993–1022.

Chang, C.C. and Lin, C.J. (2001) *LIBSVM: A Library for Support Vector Machines*, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cornelis, G.R. (2006) 'The type III secretion injectisome', *Nature Reviews Microbiology*, Vol. 4, No. 11, pp.811–825.

Dong, Q., Zhou, S. and Liu, X. (2010) 'Prediction of protein-protein interactions from primary sequences', *International Journal of Data Mining and Bioinformatics*, Vol. 4, No. 2, pp.211–227.

Ferreira, A.O., Myers, C.R., Gordon, J.S., Martin, G.B., Vencato, M., Collmer, A., Wehling, M.D., Alfano, J.R., Moreno-Hagelsieb, G., Lamboy, W.F. et al. (2006) 'Whole-genome expression profiling defines the HrpL regulon of Pseudomonas syringae pv. tomato DC3000, allows de novo reconstruction of the Hrp cis element, and identifies novel coregulated genes', *Molecular Plant-Microbe Interactions*, Vol. 19, No. 11, pp.1167–1179.

Galan, J. and Collmer, A. (1999) 'Type iii secretion machines: bacterial devices for protein delivery into host cells', *Science*, Vol. 284, No. 5418, p.1322.

Guttman, D.S., Vinatzer, B.A., Sarkar, S.F., Ranall, M.V., Kettler, G. and Greenberg, J.T. (2002) 'A functional screen for the type III (Hrp) secretome of the plant pathogen *Pseudomonas syringae*', *Science*, Vol. 295, No. 5560, pp.1722–1726.

Haddow, C., Perry, J., Durrant, M. and Faith, J. (2011) 'Predicting functional residues of protein sequence alignments as a feature selection task', *International Journal of Data Mining and Bioinformatics*, Vol. 5, No. 6, pp.691–705.

He, S.Y., Nomura, K. and Whittam, T.S. (2004) 'Type III protein secretion mechanism in mammalian and plant pathogens', *BBA-Molecular Cell Research*, Vol. 1694, Nos. 1–3, pp.181–206.

Huang, P. and Pavlovic, V. (2008) 'Protein homology detection with biologically inspired features and interpretable statistical models', *International Journal of Data Mining and Bioinformatics*, Vol. 2, No. 2, pp.157–175.

Jin, R., Si, L. and Chan, C. (2008) 'A bayesian framework for knowledge driven regression model in micro-array data analysis', *International Journal of Data Mining and Bioinformatics*, Vol. 2, No. 3, pp.250–267.

Liu, B., Wang, X., Lin, L., Dong, Q. and Wang, X. (2008) 'A discriminative method for protein remote homology detection and fold recognition combining top-n-grams and latent semantic analysis', *BMC Bioinformatics*, Vol. 9, No. 1, p.510.

Lloyd, S., Sjostrom, M., Andersson, S. and Wolf-Watz, H. (2002) 'Molecular characterization of type III secretion signals via analysis of synthetic N-terminal amino acid sequences', *Molecular Microbiology*, Vol. 43, No. 1, pp.51–59.

Löwer, M. and Schneider, G. (2009) 'Prediction of Type III Secretion Signals in Genomes of Gram-Negative Bacteria', *PloS one*.

Ma, W. and Guttman, D.S. (2008) 'Evolution of prokaryotic and eukaryotic virulence effectors', *Current Opinion in Plant Biology*, Vol. 11, No. 4, pp.412–419.

Marie, C., Deakin, W.J., Viprey, V., Kopcinska, J., Golinowski, W., Krishnan, H.B., Perret, X. and Broughton, W.J. (2003) 'Characterization of Nops, nodulation outer proteins, secreted via the type III secretion system of NGR234', *Molecular Plant-Microbe Interactions*, Vol. 16, No. 9, pp.743–751.

Panina, E., Mattoo, S., Griffith, N., Kozak, N., Yuk, M. and Miller, J. (2005) '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.

Petnicki-Ocwieja, T., Schneider, D.J., Tam, V.C., Chancey, S.T., Shan, L., Jamir, Y., Schechter, L. M., Janes, M.D., Buell, C.R., Tang, X. et al. (2002) 'Genomewide identification of proteins secreted by the Hrp type III protein secretion system of Pseudomonas syringae pv. tomato DC3000', *Proceedings of the National Academy of Sciences*, Vol. 99, No. 11, p.7652.

Qi, S., Yang, Y. and Song, A. (2011) 'Feature reduction using a topic model for the prediction of type iii secreted effectors', *Neural Information Processing*, pp.155–163.

Sato, Y., Takaya, A. and Yamamoto, T. (2011) 'Meta-analytic approach to the accurate prediction of secreted virulence effectors in gram-negative bacteria', *BMC Bioinformatics*, Vol. 12, No. 1, p.442.

Schechter, L.M., Roberts, K.A., Jamir, Y., Alfano, J.R. and Collmer, A. (2004) '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.

Schechter, L.M., Vencato, M., Jordan, K.L., Schneider, S.E., Schneider, D.J. and Collmer, A. (2006) 'Multiple approaches to a complete inventory of *Pseudomonas syringae* pv. tomato DC3000 type III secretion system effector proteins', *Molecular Plant-Microbe Interactions*, Vol. 19, No. 11, pp.1180–1192.

Steyvers, M. and Griffiths, T. (2011) *Matlab Topic Modeling Toolbox 1.4.* Software available at http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

Vinatzer, B.A., Jelenska, J. and Greenberg, J.T. (2005) 'Bioinformatics correctly identifies many type III secretion substrates in the plant pathogen *Pseudomonas syringae* and the biocontrol isolate P. fluorescens SBW25', *Molecular Plant-Microbe Inteructions*, Vol. 18, No. 8, pp.877–888.

Viprey, V., Del Greco, A., Golinowski, W., Broughton, W.J. and Perret, X. (1998) 'Symbiotic implications of type III protein secretion machinery in Rhizobium', *Molecular Microbiology*, Vol. 28, No. 6, pp.1381–1389.

Wang, Y., Zhang, Q., Sun, M. and Guo, D. (2011) 'High-accuracy prediction of bacterial type III secreted (T3S) effectors based on position-specific amino acid composition profiles', *Bioinformatics*.

Wong, P. and Chan, C. (1996) 'Chinese word segmentation based on maximum matching and word binding force', *Proceedings of the 16th Conference on Computational Linguistics*, Vol. 1, pp.200–203.

Yang, Y. (2011) 'A comparative study on sequence feature extraction for type iii secreted effector prediction', *Proceeding of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Vol. 3, pp.1560–1564.

Yang, Y., Lu, B. and Yang, W. (2008) 'Classification of protein sequences based on word segmentation methods', *Proceedings of the 6th Asia-Pacific Bioinformatics Conference*, Kyoto, Japan, 14–17 January.

Yang, Y., Zhao, J., Morgan, R., Ma, W. and Jiang, T. (2010) 'Computational prediction of type III secreted proteins from gram-negative bacteria', *BMC Bioinformatics*, Vol. 11 (Suppl 1), S47.

Yeh, J. and Chen, C. (2010) 'Protein remote homology detection based on latent topic vector model', *Proceedings of 2010 International Conference on Networking and Information Technology (ICNIT)*, pp.456–460.

Yip, K., Cheung, L., Cheung, D. and Jing, L. (2009) 'A semi-supervised approach to projected clustering with applications to microarray data', *International Journal of Data Mining and Bioinformatics*, Vol. 3, No. 3, pp.229–259.