BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 6 2011, pages 777–784
doi:10.1093/bioinformatics/btr021

Sequence analysis

Advance Access publication January 13, 2011

High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles

Yejun Wang, Qing Zhang, Ming-an Sun and Dianjing Guo*

School of Life Sciences and the State Key Lab for Agrobiotechnology, The Chinese University of Hong Kong (CUHK), Shatin, New Territories, Hong Kong

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Bacterial type III secreted (T3S) effectors are delivered into host cells specifically via type III secretion systems (T3SSs), which play important roles in the interaction between bacteria and their hosts. Previous computational methods for T3S protein prediction have only achieved limited accuracy, and distinct features for effective T3S protein prediction remain to be identified.

Results: In this work, a distinctive N-terminal position-specific amino acid composition (Aac) feature was identified for T3S proteins. A large portion (∼50%) of T3S proteins exhibit distinct position-specific Aac features that can tolerate position shift. A classifier, BPBAac, was developed and trained using Support Vector Machine (SVM) based on the Aac feature extracted using a Bi-profile Bayes model. We demonstrated that the BPBAac model outperformed other implementations in classification of T3S and non-T3S proteins, giving an average sensitivity of ∼90.97% and an average selectivity of ∼97.42% in a 5-fold cross-validation evaluation. The model was also robust when a small-size training dataset was used. The fact that the position-specific Aac feature is commonly found in T3S proteins across different bacterial species gives this model wide application. To demonstrate the model's application, a genome-wide prediction of T3S effector proteins was performed for *Ralstonia solanacearum*, an important plant pathogenic bacterium, and a number of putative candidates were identified using this model.

Availability: An R package of BPBAac tool is freely downloadable from: http://biocomputer.bio.cuhk.edu.hk/softwares/BPBAac.

Contact: djguo@cuhk.edu.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 27, 2010; revised on November 18, 2010; accepted on January 9, 2011

1 INTRODUCTION

Six types of secretion systems have been identified in Gram-negative bacteria, two of which (type I and type II) have been studied extensively (Bingle *et al.*, 2008; Fath and Kolter, 1993; Fischer *et al.*, 2002; Henderson *et al.*, 2004; Hueck, 1998). The type III secretion system (T3SS) has been widely adopted by different bacteria, such as animal pathogens *Salmonella*, *Shigella* and *Vibrio*, plant pathogens *Pseudomonas*, *Xanthomonas*, and *Ralstonia*, and some symbiotic bacteria such as *Rhizobia* (Hueck, 1998). T3SSs play important roles in host–pathogen interactions that are often mediated by T3S effectors specifically secreted into host cells through the type III secretion conduits (Galán and Wolf-Watz, 2006).

Previous studies have shown that the first 100 amino acid at the N-terminal region may contain the signal peptides and chaperone-binding sequences needed to guide the secretion of T3S proteins (Karavolos *et al.*, 2005; Lloyd *et al.*, 2001, 2002; Rüssmann *et al.*, 2002; Schechter *et al.*, 2004; Wang *et al.*, 2008). Most known T3S proteins have at least one chaperone, which mediates its secretion through the extremely narrow T3S conduit (Stebbins and Galán, 2001). Unlike most other signal peptides, T3S signals are not cleaved after secretion. Due to low sequence similarity and lack of common features among different T3S signal sequences, the established prediction methods used for identifying signal peptides do not apply to T3S signals (Galán and Wolf-Watz, 2006; Hueck, 1998). Computational prediction of T3S protein has long been considered to be a particularly difficult challenge.

Computational approaches have been attempted to predict T3S proteins based on sequence similarity, consensus patterns, gene-adjacent sequence features, etc. (Panina *et al.*, 2005; Petnicki-Ocwieja *et al.*, 2002; Tobe *et al.*, 2006). Different machine learning algorithms, e.g. Naive Bayes (NB), Artificial Neural Network (ANN) and Support Vector Machine (SVM) (Arnold *et al.*, 2009; Löwer and Schneider, 2009; Samudrala *et al.*, 2009; Yang *et al.*, 2010), have also been adopted to identify the general signal features. Some important features, including G+C content of the primary DNA sequence, general enrichment and depletion of N-terminal amino acid composition (Aac), composition frequency of secondary structure elements (coil, helices or strands) and water accessibility states (exposed or buried) have been identified and used for *in silico* prediction (Arnold *et al.*, 2009; Löwer and Schneider, 2009; Samudrala *et al.*, 2009; Yang *et al.*, 2010). Effective T3, one of the earliest software developed for T3S protein prediction (Arnold *et al.*, 2009), explores possible sequence-based features exhaustively. In Effective T3, the Aac and property preference within the signal region (not position specific) was represented in two reduced alphabets (Arnold *et al.*, 2009), which may lead to loss of signal information buried in individual amino acid. In addition, no position-specific features were analyzed in Effective T3. An ANN model proposed by Löwer and Schneider (2009) adopts a sliding window technique (with a window width of 25) and an optimal model is obtained based on the signal sequence located within the first 30 amino acids at the N-terminal end (Löwer and Schneider, 2009). Although this model achieved high selectivity (98%), its sensitivity was rather low (74%). Some drawbacks of the ANN model should

*To whom correspondence should be addressed.

also be pointed out: (i) the training dataset was not validated and it contains wrongly annotated non-T3S proteins, including chaperones located in cytoplasm and a number of validated flagella proteins not secreted through T3SSs. In addition, some proteins with high homology were not excluded; (ii) the classifying performance was based on train-reclassification results only and no cross-validation was performed; and (iii) its complexity makes it difficult to interpret the biological implications. Most recently, a SVM model, SSE-ACC, was proposed to learn features using Aac-Sse and Aac-Acc (Aac, Sse and Acc represents amino acid composition, secondary structure and solvent accessibility, respectively) combination frequencies using SVM (Yang *et al.*, 2010). These features, however, was trained from only one plant pathogen genus and then used to predict the T3S effectors in *Rhizobium*. The authors reported a significantly increased selectivity (91%) with a trade-off of apparently lowered sensitivity (65%). Therefore, new features need to be identified and used for more effective T3S protein identification.

We here propose a computational model based on position-specific Aac profile for effective T3S protein prediction. We demonstrate that our model outperformed other current implementations and achieved both high sensitivity (97.42%) and high selectivity (90.97%) in a 5-fold cross-validation experiment. A genome-wide prediction of T3S effectors in an important plant pathogen, namely *Ralstonia solanacearum*, was also conducted.

2 METHODS

2.1 Data source

A list of experimentally validated T3S proteins from animal pathogens, plant pathogens and symbiotic bacteria was manually annotated from a literature search. A list of non-T3S proteins were randomly selected from different bacteria, followed by removal of the known effectors and their homologs. For T3S and non-T3S proteins, only one representative was selected as the training sequence for each orthologous or paralogous cluster. JAligner, an alignment tool implementing Smith–Waterman algorithm was used to make a pairwise alignment for any two T3S or non-T3S proteins (http://jaligner.sourceforge.net/). The ratio between the pairwise score and self-score is calculated. A sensitive cutoff, 0.15, was set for identifying paralogs or orthologs (Arnold *et al.*, 2009). In total, 154 non-redundant T3S peptides obtained were subsequently used as positive dataset. As the number of non-T3S proteins was much larger than the number of positive proteins, 308 peptides were randomly selected from the negative peptide pool to form final negative training set, to overcome the imbalance between positive and negative datasets (Arnold *et al.*, 2009; Kim *et al.*, 2004). Details of these two datasets and the reference for each T3S protein can be found in the Supplementary Materials (Text S1). The Sse (represented as a combination sequence of 'C', 'H' or 'E' of each sequence) was predicted using PSIPRED (McGuffin *et al.*, 2000), and SCRATCH (Cheng *et al.*, 2005) was used to predict the Acc (a combination of 'B' or 'E'). For 5-fold cross-validation, the negative and positive training datasets were pooled as the final training datasets and were evenly split into five sub-datasets, each containing the same number of positive/negative samples.

2.2 Position-specific profiles and feature extraction

The unaligned T3S proteins and non-T3S proteins were used for position-specific feature extraction. Let vector $S=s_1,s_2,s_3,\ldots,s_n$ denotes a peptide sequence in which s represents amino acid or other properties while $1,2,\ldots$ or i represents position and n represents the total sequence length. For m sequences, the position-specific occurrence of a certain amino acid A is described as: $p(\mathrm{A}_i)=f(\mathrm{A}_i)/m_i$, in which $f(\mathrm{A}_i)$ denotes the frequency of amino acid A at position i. For each position, the $p(\mathrm{A}_i)$ of different amino acids form a position set (or profile), and for a sequence S with a length of n, n values (extracted from each position set) comprise a composition vector. Similar profiles and feature vectors were extracted for corresponding Sse and Acc in T3S or non-T3S peptides. WebLogo was adopted to exhibit the position-specific preference profiles (Crooks *et al.*, 2004).

For feature extraction, both the Bi-profile Bayes (BPB) method (Shao *et al.*, 2009) and the more frequently applied Single-profile Bayes method (SPB) were adopted as appropriate. These two methods are similar except that BPB takes into consideration the features of negative training dataset. Simply, given a protein sequence $S=s_1,s_2,s_3,\ldots,s_n$, where each S_i $(i=1,\ldots,n)$ denotes an amino acid at position i, and n denotes the sequence length, S can be classified as one of the two classes: C_1 (T3S proteins) or C_{-1} (non-T3S proteins). The posterior probability of both T3S and non-T3S proteins can be calculated as the occurrence of each amino acid at each position in the training dataset. More details about the BPB method can be found in Shao *et al.* (2009). The BPB and SPB signatures were extracted for position-specific amino acid composition, Sse and Acc.

2.3 SVMs implementation and parameter optimization

R package for SVM, 'e1071', was used to train and build the SVM models (http://cran.r-project.org/). Radial basis kernel function $K(s_i,s_j)=\exp(-\gamma||s_i-s_j||^2)$ was selected for SVM prediction. SVM parameter γ and penalty parameter C were optimized using grid search based on 10-fold cross-validation (Scholkopf and Smola, 2002).

2.4 Performance assessment

Accuracy (A), Specificity (Sp), Sensitivity (Sn), Receiver Operating Characteristic (ROC) curve, the area under *ROC* curve (AUC) and Matthews Correlation Coefficient (MCC) were utilized to assess the predictive performance. In the following formula, A denotes the percentage of both positive instances (T3S) and negative instances (non-T3S) correctly predicted. Sn (true positive rate) and Sp (true negative rate), respectively, represent the percentage of positive instances (T3S) and the percentage of negative instances (non-T3S) correctly predicted. An ROC curve is a plot of Sn versus (1−sp), and is generated by shifting the decision threshold. AUC gives a measure of classifier performance. MCC takes into account true and false positives and false negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.

$$A=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}},Sp=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},Sn=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\mathrm{MCC}=\frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FN}\times\mathrm{FP})}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}}$$

where TP, TN, FP and FN denotes the number of true positives, true negatives, false positives and false negatives, respectively.

2.5 Amino acid position shift and frameshift

For position shift test, both insertion and deletion datasets were created. For deletion test, we generated five individual datasets with 1, 2, 3, 4 or 5 amino acids deleted respectively at the N-terminal end excluding starting methionine. For insertion test, one of the 20 amino acids was inserted before the first or second amino acid position respectively for each mutated sequence, and in total 40 mutated sequences were generated for each T3S or non-T3S protein. These two positions were selected because apparent Aac bias was found at the first position for non-T3S proteins and at the second position for T3S proteins. For frameshift experiments, DNA sequences encoding the first 100 amino acids at the N-terminal excluding the starting methionine were obtained. For each sequence, two mutations with '−1' and '+1' frameshift were created, respectively. The mutated sequences were translated into peptides, with all the encountered stop codons replaced with methionine (Arnold *et al.*, 2009). The resulting sequences were reclassified using the optimized BPBAac model.

2.6 Comparison with available methods

The original datasets used for Effective T3 (Arnold *et al.*, 2009) and ANN (Löwer and Schneider, 2009) were collected from the relevant reports. For Effective T3, no detailed gene accessions or sequences of negative dataset were available (Arnold *et al.*, 2009), so we randomly selected proteins not annotated as T3S from different bacteria species as negative training datasets and ratio of negative to positive samples was 2:1. We also removed some apparent false positive sequences (e.g. chaperones, flagella proteins, etc.) from the ANN training dataset. Effective T3 and ANN were implemented with the optimized parameters suggested by their respective authors (Arnold *et al*., 2009; Löwer and Schneider, 2009).

2.7 Genome-wide prediction of T3S proteins from *R.solanacearum*

The recently validated T3S proteins in *R.solanacearum* were annotated from Mukaihara, *et al*. (2010). In total, 47 validated T3S proteins were retrieved from GMI1000 and these validated T3S proteins were also included for the final BPBAac training. For prediction of T3S proteins from GMI proteomes, 3437 chromosome-encoding proteins (Genome ID: NC_003295) and 1676 plasmid-encoding proteins (Genome ID: NC_003296) of *R.solanacearum* GMI1000 (Salanoubat *et al*., 2002) were downloaded from the NCBI genome database (http://www.ncbi.nlm.nih.gov/sites/genome). The first 100 amino acids (excluding methionine) from the N-terminal end were retrieved from each protein. The feature vectors were constructed and then tested using the BPBAac model. The cutoff value of BPBAac was set as 0.5. The proteomes of *Ralstonia* were also predicted for T3S protein candidates using Effective T3 and ANN, respectively with originially optimized parameters.

3 RESULTS

3.1 Distinct position-specific Aac profiles for T3S effectors

The N-terminal amino acids were retrieved from T3S and non-T3S proteins, respectively and Aac was calculated for each position. Significantly distinctive Aac profiles were found between these two types of proteins (Fig. 1A and B). For T3S proteins, the 20 types of amino acids were not evenly distributed at each amino acid position, especially for the first 50 positions (Fig. 1A). Consistent with previous observation using sequence-based method (Arnold *et al.*, 2009), serine was enriched in most of the first 50 positions. Contrarily, leucine was found to be selectively enriched in certain positions, e.g. position 13, 14, 15 etc., but not completely 'depleted' as described in an earlier report (Arnold *et al.*, 2009).

The T3S effectors were further split into an animal-pathogen group and a plant-pathogen group (including *Rhizobium*, a plant symbiont). Both groups showed apparent Aac preference profiles different from that of non-T3S proteins (Supplementary Figure S1A and B). For each position, most of the enriched/depleted amino acids were similar between two groups, although isoleucine, asparagine and threonine were more often preferred by animal pathogens whereas alanine, proline and arginine were more enriched in plant pathogens (Supplementary Figure S1A andB). We manually checked the validated T3S effectors for individual genera or species and found that their overall Aac profiles were similar and apparently different from those in non-T3S proteins (Supplementary Figure S1C–F).

Table 1. Optimal parameters and corresponding performance of BPB and SPB model based on 5-fold cross-validation

Name	Model[a]	Length[b]	Kernel[c]	C[d]|γ[e]
BPBAac	BPB	100	RBF	8|0.002
SPBAac	SPB	100	RBF	2|0.001
Name	*Sn* (%) versus *Sp* (%)	*A* (%)	AUC (%)	MCC
BPBAac	90.97 versus 97.42	95.27	98.88	0.8929
SPBAac	71.61 versus 94.51	86.88	93.02	0.6979

[a]Mathematic model used for feature extraction. [b]N-terminal sequence length used for feature extraction. [c]SVM kernel function. RBF: radial basis function. [d]C: cost, which was optimized based on 10-fold cross-validation grid search. [e]γ: gamma, which was optimized based on 10-fold cross-validation grid search.

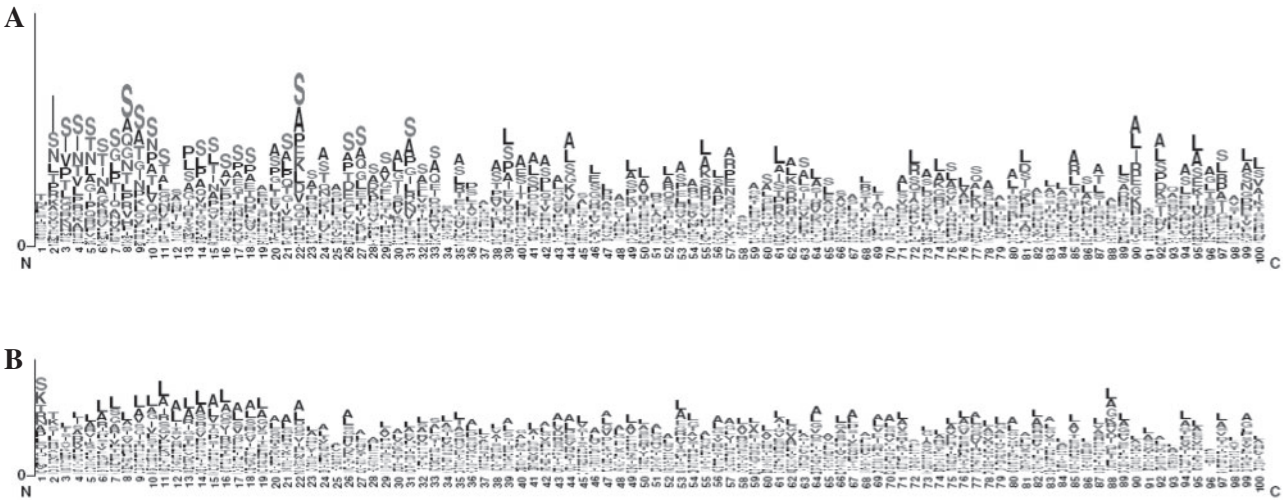

Fig. 1. Distinctive N-terminal position-specific Aac feature in T3S proteins. Amino acid positions are depicted on the horizontal axis. The heights of characters represent the preference or enrichment level. (**A**) Aac preference for T3S protein. (**B**) Aac preference for non-T3S proteins.

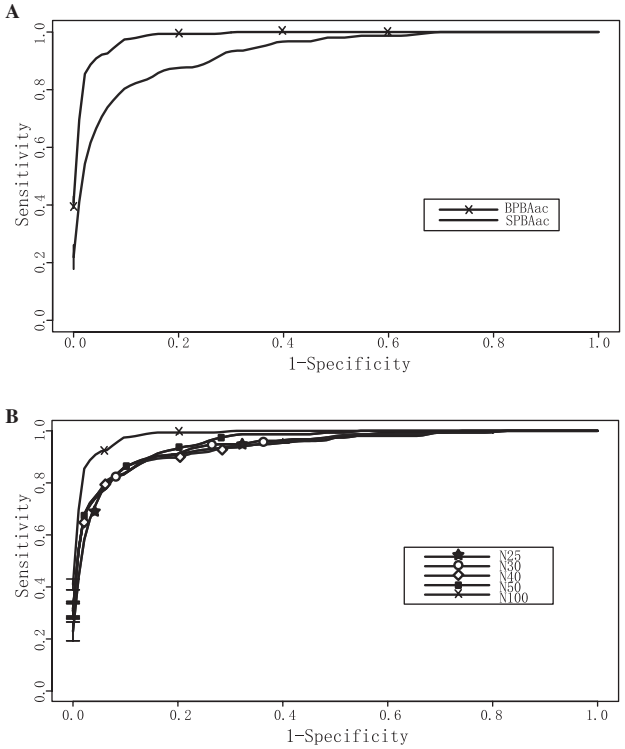

Fig. 2. Performance of SVMs based on different feature-extraction models and sequence lengths. (**A**) ROC curves of SVM classifiers based on BPB and SPB models, respectively. The length was 100 amino acid. (**B**) ROC curves of BPB SVMs based on different lengths of N-terminal sequences. The curves were based on a 5-fold cross-validation results.

3.2 N-terminal position-specific Aac features can be used to classify T3S and non-T3S proteins

In order to further investigate whether this position-specific Aac preference is a general feature for T3S effectors, SVM models were trained for the Aac features (Section 2). Two different SVM models were trained: (i) SPBAac model that only considers the Aac profile of positive T3S training dataset; (ii) BPBAac model considers the Aac profiles of both T3S and non-T3S proteins. Table 1 and Figure 2A showed that BPB model outperformed SPB model significantly. SPB model achieved high selectivity (94.51%) and an acceptable sensitivity (71.61%), while BPB model achieved both high selectivity (97.42%) and high sensitivity (90.97%) in a 5-fold cross-validation. The accuracy, AUC of ROC curve and MCC value of BPB were all larger than those of SPB (Table 1). The best predictive power (sensitivity versus selectivity) of established feature-based T3S protein prediction methods were reported as 71% versus 85%, 74% versus 98% and 65% versus 91% for Effective T3, ANN and SSE-ACC, respectively (Arnold *et al*., 2009; Löwer and Schneider, 2009; Yang *et al*., 2010). Therefore, the position-specific amino acid profiles can serve as independent and effective features for T3S and non-T3S protein classification. The fact that BPB model outperformed SPB model indicates the important contribution of the negative training data.

Previous computational modeling studies showed that T3S signals were mainly located within the first 30 or 25 amino acid positions (Arnold *et al*., 2009; Löwer and Schneider, 2009). In order to optimize the length of signal sequence, BPB models were retrained and compared using N-terminal sequences containing the first 25, 30, 40, 50 and 100 amino acid positions, respectively (named BPBAac-N25, N30, N50 and N100 model, respectively). As shown by the ROC curves, model using N-terminal 25 or 30 positions achieved good performance (Fig. 2B), although the best performance was achieved when the first 100 amino acid were used (Fig. 2B). From this analysis, we conclude that sequences beyond the first 30 amino acids also contain important signals to guide protein secretion. Other optimized parameters, such as kernel function, gamma and cost values, were also tested (Table 1).

3.3 The robustness of BPBAac model

The robustness of BPBAac model was examined by (i) randomly selecting sub-datasets with different sample sizes from the training data to retrain the model and to classify the remaining data; and (ii) using the Leave-One-Out strategy. Specifically, T3S effectors and non-T3S proteins from one bacterial genus were eliminated from the test dataset, and then the remaining data were used to train the model and to classify the testing dataset. This process was repeated using different bacteria genus. Our results showed that models trained using different sub-datasets also performed equally well, and no apparent reduction in performance was observed even when only 40% of the original training data were used (Fig. 3A). For different genera or subgroups, most of the effectors could be recalled and the AUC values did not show significant change (Fig. 3B). Taken together, the position-specific Aac profiles were important features for T3S protein identification in different bacteria species.

3.4 Aac feature alone is enough to distinguish T3S and non-T3S proteins

The Sse and Acc of N-terminal amino acids have also been reported as useful features to distinguish T3S proteins from non-T3S ones (Yang *et al*., 2010). In order to examine whether these two features can improve the classifying performance of BPBAac, Sse and Acc BPB features were extracted and trained individually. In addition, models were retrained using combinations of any two types or all three types of features, respectively. The results showed that neither Sse nor Acc was able to improve the performance (Fig. 4). Therefore, although T3S proteins contain Sse and Acc profiles different from those of non-T3S proteins (data available on request), the Aac feature alone is enough to distinguish T3S and non-T3S proteins.

3.5 Some T3S effectors contain N-terminal position-specific Aac features that can tolerate position shifts and frameshifts

Some T3S effectors may contain amino acid insertions/deletions, which lead to position shift in the signal regions. For example, various forms of amino acid deletions or insertions were found in XopO, a T3S protein, from different *Xanthomonas* strains. When the BPBAac model was applied to two XopO homologs which bear amino acid position shift, both were correctly classified. To examine whether the Aac feature is sensitive to amino acid position shifts, deletions and insertions were introduced to T3S or non-T3S proteins, respectively. As shown in Table 2, after introduction of position shifts (amino acid deletion), ∼50% of the T3S proteins retained their Aac

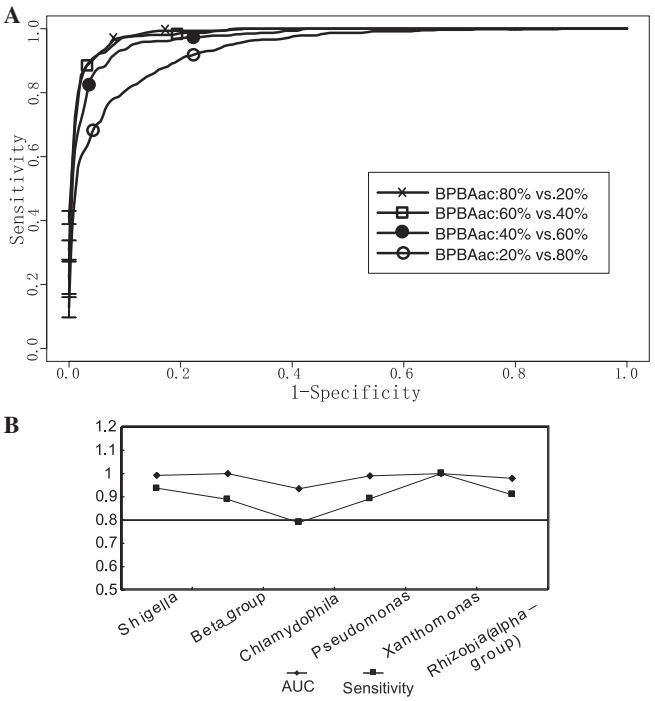

Fig. 3. Performance of BPBAac models trained with different datasets. (**A**) ROC curves of SVM classifiers trained with different sub-training dataset. 'Training versus Test' denotes the 'percentage of training data' versus 'percentage of test data'. The curves and performance are based on average 5-fold cross-validation results. (**B**) The Leave-One-Out test results. The positive and negative datasets from representative species or groups were extracted. The remaining training datasets were used for model retraining, and the retrained model was used to classify the extracted datasets. *Pseudomonas* and *Xanthomonas* were adopted as representatives of plant pathogens; *Shigella*, Beta_group and *Chlamydophila* were adopted as representatives of animal pathogens. AUC and sensitivity (recall) values were represented by solid diamond and rectangle, respectively.

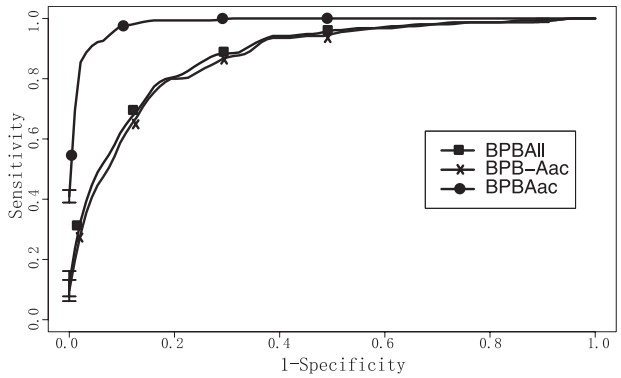

Fig. 4. Comparison of ROCs based on different type of features. All the models used parameters optimized for BPBAac. 'BPBAll' denotes a model based on the combination of all three types of features; 'BPB-Aac' denotes a model based on the other features except 'Aac'; 'BPBAac' denotes a model based on 'Aac' feature only. All the curves were obtained based on a 5-fold cross-validation results.

Table 2. Effects of position shift and frameshift on reclassification performance

Mutation	Method	Sn (%)	Sp (%)
No mutation	NA[a]	100.00	100.00
Deletions	First deletion	53.90	91.88
	First to second deletions	48.05	96.10
	First to third deletions	44.16	93.18
	First to fourth deletions	44.81	97.72
	First to fifth deletions	38.96	97.40
Insertions	First insertion[b]	52.82 ± 3.56	93.69 ± 1.36
	Second insertion[b]	54.87 ± 3.13	92.99 ± 1.01
Frameshifts	+1[c]	12.67 (19/150)	92.67 (278/300)
	−1[c]	14.00 (21/150)	94.33 (283/300)

[a]NA: reclassify all the original training data using BPBAac model. [b]The sensitivity and selectivity were both represented as mean ± SD for insertions with 20 types of amino acids. [c]The sensitivity and selectivity were both represented as 'percentages (number of correctly predicted proteins/total number of proteins)'.

feature, and more T3S proteins lost their Aac feature profiles with increased number of position shifts (Table 2).

Amino acid insertions were also examined by inserting different amino acids before the first or second N-terminal amino acid position. For these two types of insertion mutations, similar proportion of proteins were correctly reclassified (53 and 55% for first and second, respectively) (Table 2). The reclassifying performance was also influenced by the type of amino acid inserted. As non-T3S proteins show significant amino acid preference at the first position, insertion of non-T3S proteins preferred amino acid (e.g. 'S' or 'K') resulted in higher selectivity. On the other hand, T3S proteins showed significant amino acid preference at the second position, and insertion of T3S protein preferred amino acid (e.g. 'I' or 'N') resulted in higher sensitivity (data not shown).

As some T3S effectors are insensitive to frameshifts, some researchers have argued that the signals may locate within the mRNA sequences rather than the amino acid sequences (Mudgett *et al*., 2000; Ramamurthi and Schneewind, 2003; Rüssmann *et al*., 2002). It was recently suggested that few effectors (10%) maintain sequence-based Aac profiles when frameshifts occur (Arnold *et al.*, 2009). To test whether the position-specific Aac feature is sensitive to frameshift, we created frameshifts (both '−1' and '+1'shifts) for both T3S and non-T3S proteins. It was found that ∼13% and ∼93% of frameshifted T3S and non-T3S were correctly classified by BPBAac, respectively (Table 2). Non-T3S proteins were more insensitive to frameshift is likely due to the fact that such proteins contain much fewer amino acid preference features (Fig. 1B). Some T3S effectors such as AvrBs2 of *Xanthomonas*, that are known to be tolerant to frameshifts in this research have also been confirmed by wet-lab experiments (Mudgett *et al*., 2000).

3.6 Performance comparison with current prediction models

The classification performance was compared among BPBAac, EffectiveT3 and T3SS ANN. First, EffectiveT3 and ANN were used to reclassify the training datasets used in this research ('BPBAac dataset'), and the performance was compared with cross-validation rates of BPBAac. Supplementary Table S1 and Figure 5A clearly demonstrated that BPBAac outperformed these two methods in

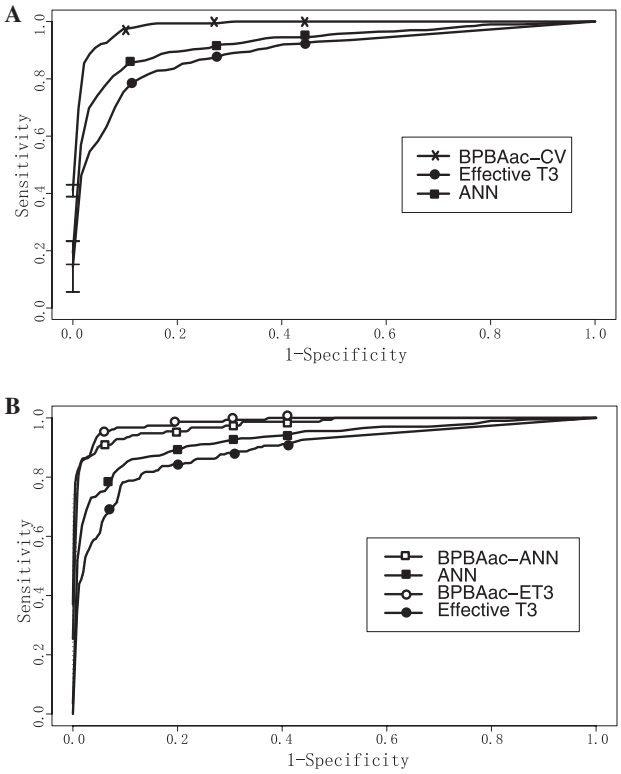

Fig. 5. Comparison of performance using different datasets. (**A**) ROC curves using original training dataset. BPBAac-CV: BPBAac model based on an average 5-fold cross-validation training and testing result. (**B**) BPBAac-ANN and BPBAac-ET3 were BPBAac models trained with ANN and Effective T3 datasets, respectively. ANN and Effective T3 were trained with their original datasets, respectively. All the four models were used to reclassify BPBAac dataset.

terms of sensitivity, specificity, accuracy, MCC value or AUC value of ROC curve.

To make a fair comparison among different models, other two strategies were adopted: (i) BPBAac was first retrained using the EffectiveT3 dataset and ANN dataset respectively before it was used to reclassify BPBAac dataset (Supplementary Table S1 and Fig. 5B); (ii) BPBAac was retrained using the datasets adopted by EffectiveT3 and ANN respectively, and the new model was used to reclassify those two datasets (Supplementary Table S1). As shown in Table S1 and Figure 5, BPBAac model consistently performed better than Effective T3 and ANN.

3.7 Genome-wide prediction of T3S effectors in *R.solanacearum*

Ralstonia solanacearum is a very important pathogenic bacterium that causes severe bacterial wilt to a wide range of potential host plants, including crop and fruit plants. Currently, the information about the T3S effectors in this genus is limited. As an application of the BPBAac model, proteins encoded by chromosome and plasmid of *R.solanacearum* GMI1000 were used for genome-wide prediction of T3S proteins. As shown in Supplementary Table S2, 1.4% (49/3437) chromosome encoding proteins and 2.9% (48/1676) of plasmid encoding proteins were predicted to be T3S effectors. With a higher recall percentage, a much smaller number of putative T3S candidates was obtained, making validation work more feasible (Table 3). Interestingly, many candidates (38/97, 39.2%) are annotated as 'hypothetical' proteins with unknown function. Some candidates were validated recently (e.g. PopF1), closely related with T3SS (e.g. NP_522416.1, Hrp pilus subunit HRPY protein) or originated in bacteriophages (e.g. NP_519819.1) (Supplementary Table S2).

Table 3. Genome-wide prediction of T3S proteins in *R.solanacearum* using different models

Model	Recall % (*n*/*N*)	Chromosomal gene % (*n*/*N*)	Plasmidial gene % (*n*/*N*)
BPBAac	93.6 (44/47)	1.4 (49/3437)	2.9 (48/1676)
Effective T3	57.4 (27/47)	9.6 (331/3437)	11.4 (191/1676)
ANN	68.1 (32/47)	10.7 (368/3437)	12.7 (213/1676)

4 DISCUSSION

Previous studies have demonstrated that T3S effectors contain conserved N-terminal Aac pattern. For example, Lloyd *et al*. (2002) found that serine and isoleucine were enriched at the N-termini of YopE protein. Petnicki-Ocwieja *et al*. (2002) reported two consensus patterns (i.e. enrichment and deletion) at the N-terminal end of *Pseudomonas* Hrp-secreted proteins and a group of new T3S effectors were identified based on these patterns. Recently, Samudrala *et al*. (2009) examined sequence-based Aac bias and concluded that the Aac profiles were 'largely uninformative'. In this study, we carefully examined the position-specific Aac profiles within the N-terminal sequences of T3S and non-T3S proteins and identified distinctive amino acid enrichment/depletion profiles for T3S proteins. We found that although the first 30 N-terminal amino acid positions are most informative, important signal information were also embedded within the sequences beyond the first 30 positions. Using a Bi-profile Bayesian model, we extracted the position-specific Aac feature within the first 100 amino acid position and used it as an efficient classifier to distinguish T3S from non-T3S proteins. Our model achieved great predictive power and was superior to previous models using sequence-based Aac features. Apart from Effective T3 and ANN, we also compared the BPBAac model with SIEVE, one of the earliest prediction methods adopting sequence-based features (Samudrala *et al.*, 2009) and the BPBAac model also performed better (Supplementary Table S3). Further exploration of the position-based Aac features may provide important clues about the nature and evolution of the T3SS signals.

Apart from Aac, other features may also contribute to the mechanisms underlying T3S secretion. Previously, other groups have examined Sse and Acc (Arnold *et al*., 2009; Yang *et.al.*, 2010) but no conclusive remarks can be drawn so far. Arnold *et al.* (2009) found that neither Sse, nor Acc could improve the classifying performance. Using a combinative feature extraction strategy, Yang *et al.* (2010) found that the combination of Sse and Acc could improve the model performance. In the present study, we did find distinctive Sse and Acc profiles between positive and negative dataset. However, when individually trained using Sse and Acc, the model failed to show good performance (sensitivity versus

selectivity for Sse and Acc, respectively were: 47% versus 79%, 16% versus 99%; data not shown). When these two features were combined with Aac, the performance was significantly decreased (Fig. 4). Further, detailed analysis is being carried out to investigate the most relevant positions more exposed to the solvent and their Sse organization, and to predict the disorder of these structural regions. Such analysis may provide insights into whether and how Sse and Acc may contribute to the specific secretion of T3S proteins.

Two methods are frequently adopted for position-specific Aac modeling: Hidden Markov Model (HMM) and sliding window technique. As T3S proteins contain a long signal bearing sequence, the sliding window model may become quite complex and sometimes encounter the overfitting problem (Löwer and Schneider, 2009). We have also attempted a T3S protein HMM (Eddy, 1998), but found its classifying performance inferior (data not shown). The BPB model was first proposed by Shao *et al.* (2009), and it has been successfully used to predict protein phosphorylation sites. One main advantage of the BPB model is that it considers the features of both positive and negative training samples. Compared to sliding windows and HMMs concerning amino acid insertion, deletion or match, an underlying drawback of our position-based Aac model (BPBAac) is that amino acid insertions or deletions were neglected. Unexpectedly, our model performed very effectively despite this potential drawback. After a careful examination of homologous T3S effectors from closely related bacterial strains, we found that insertions or deletions within the first 100 N-terminal positions were very rare. Only one protein, XopO from *Xanthomonas*, was found to have two homologs (GenBank accession: AAV74207.1 and CAJ22686.1) with nine amino acid deletions/insertions, and both homologs were correctly classified. Interestingly, a large portion (∼50%) of T3S proteins exhibit distinct position-specific Aac features that can tolerate position shifts. We therefore hypothesize that position shifts seldom happen within the signal sequences of T3S proteins; and in the case that they happen, many T3S proteins manage to maintain the original Aac features. This may partially explain the high performance of the BPBAac model where position shift was not taken into consideration. Together with the observation in this research that some T3S proteins (∼13%) could resist frameshifts generated in the signal region, we presume that bacteria may adopt strategies at both protein and mRNA levels to resist the negative mutations that may destroy T3S signals during the course of evolution.

Finally, we applied the BPBAac model to make a genome-wide prediction of T3S effectors in an important plant pathogen, *R.solanacearum* GMI1000. Most of the validated T3S effectors were recalled by BPBAac (Table 3 and Supplementary Table S1). More importantly, far fewer candidates were predicted by BPBAac than by other models such as ANN and Effective T3 (Table 3). A significant proportion of the predictions overlapped with those of ANN and Effective T3, indicating they are most likely true T3S proteins, while a large number were 'hypothetical' proteins with unknown function. These candidates are especially interesting because they have so far received little attention. One candidate, PopF1, had been validated as T3S effector protein but not included in our training dataset (Meyer *et al.*, 2006). Another candidate, NP_522416.1, was a Hrp pilus subunit HRPY protein, which could be possibly secreted via T3SS conduit. In addition, one of these candidates originated in bacteriophages. It is known that some T3S proteins originate from phages, such as SopE in *Salmonella*. Further experiments are being carried out to validate some of the newly predicted T3S proteins. We hope that the list of newly identified putative T3S candidates may serve as a useful resource for the research community. As the Aac features were commonly identified across genus and species, we believe that the BPBAac tool can be widely used for efficient T3S effector prediction in various bacteria species.

ACKNOWLEDGEMENTS

We are grateful for the constructive suggestions made by Dr Shao Jianlin at School of Life Sciences, CUHK. We would also like to thank Dr David Wilmshurst, CUHK's Academic Editor, for improving the article's English.

Funding: Hong Kong's University Grants Committee (AoE Plant & Agricultural Biotechnology Project AoE-B-07/09); CUHK's Institute of Plant Molecular Biology and Agrobiotechnology.

Conflict of interest: none declared.

REFERENCES

Arnold,R. *et al*. (2009) Sequence-based prediction of type III secreted proteins. *PLoS pathogens*, **5**, e1000376.

Bingle,L.E. *et al*. (2008) Type VI secretion: a beginner's guide. *Curr. Opin. Microbiol*., **11**, 3–8.

Cheng,J. *et al*. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nuceic Acids Res.*, **33**, W72–W76.

Crooks,G.E. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Fath,M.J. and Kolter,R. (1993) ABC transporters: bacterial exporters. *Microbiol. Rev.*, **57**, 995–1017.

Fischer,W. *et al*. (2002) Type IV secretion systems in pathogenic bacteria. *Int. J. Med. Microbiol.*, **292**, 159–168.

Galán,J.E. and Wolf-Watz,H. (2006) Protein delivery into eukaryotic cells by type III secretion machines. *Nature*, **444**, 567–73.

Henderson,I.R. *et al*. (2004) Type V protein secretion pathway: the autotransporter story. *Mol. Biol. Rev.*, **68**, 692–744.

Hueck,C.J. (1998) Type III protein secretion systems in bacterial pathogens of animals and plants. *Mol. Biol. Rev.*, **62**, 379–433.

Karavolos,M.H. *et al*. (2005) Type III secretion of the Salmonella effector protein SopE is mediated via an N-terminal amino acid signal and not an mRNA sequence. *J. Bacteriol.*, **187**, 1559–1567.

Kim,J.H. *et al*. (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.

Lloyd,S.A. *et al*. (2001) Yersinia YopE is targeted for type III secretion by N-terminal, not mRNA, signals. *Mol. Microbiol.*, **39**, 520–531.

Lloyd,S.A. *et al*. (2002) Molecular characterization of type III secretion signals via analysis of synthetic N-terminal amino acid sequences. *Mol. Microbiol.*, **43**, 51–59.

Löwer,M. and Schneider,G. (2009) Prediction of type III secretion signals in genomes of gram-negative bacteria. *PLoS ONE*, **4**, e5917.

McGuffin,L.J. *et al*. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.

Meyer,D. *et al*. (2006) PopF1 and PopF2, two proteins secreted by the Type III protein secretion system of *Ralstonia solanacearum*, are translocators belonging to the HrpF/NopX family. *J. Bacteriol.*, **188**, 4903–4917.

Mudgett,M.B. *et al*. (2000) Molecular signals required for type III secretion and translocation of the Xanthomonas campestris AvrBs2 protein to pepper plants. *Proc. Natl Acad. Sci. USA*, **97**, 13324–13329.

Mukaihara,T. *et al*. (2010) Genome-wide identification of a large repertoire of *Ralstonia solanacearum* type III effector proteins by a new functional screen. *Mol. Plant Microbe Interact.*, **23**, 251–262.

Panina,E.M. *et al*. (2005) A genome-wide screen identifies a Bordetella type III secretion effector and candidate effectors in other species. *Mol. Microbiol.*, **58**, 267–279.

Petnicki-Ocwieja,T. *et al*. (2002) Genomewide identification of proteins secreted by the Hrp type III protein secretion system of Pseudomonas syringae pv. tomato DC3000. *Proc. Natl Acad. Sci. USA*, **99**, 7652–7657.

Ramamurthi,K.S. and Schneewind,O. (2003) Yersinia yopQ mRNA encodes a bipartite type III secretion signal in the first 15 codons. *Mol. Microbiol.*, **50**, 1189–1198.

Rüssmann,H. *et al*. (2002) Molecular and functional analysis of the type III secretion signal of the Salmonella enterica InvJ protein. *Mol. Microbiol.*, **46**, 769–779.

Salanoubat,M. *et al*. (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature*, **415**, 497–502.

Samudrala,R. *et al*. (2009) Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. *PLoS pathogens*, **5**, e1000375.

Scholkopf,B. and Smola,A.J. (2002) *Learning with Kernels*. MIT Press, Cambridge.

Schechter,L.M. *et al*. (2004) Pseudomonas syringae type III secretion system targeting signals and novel effectors studied with a Cya translocation reporter. *J. Bacteriol.*, **186**, 543–555.

Shao,J. *et al*. (2009) Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS ONE*, **4**, e4920.

Stebbins,C.E. and Galán,J.E. (2001) Maintenance of an unfolded polypeptide by a cognate chaperone in bacterial type III secretion. *Nature*, **414**, 77–81.

Tobe,T. *et al*. (2006) An extensive repertoire of type III secretion effectors in Escherichia coli O157 and the role of lambdoid phages in their dissemination. *Proc. Natl Acad. Sci. USA*, **103**, 14941–14946.

Wang,Y. *et al*. (2008) Two oral HBx vaccines delivered by live attenuated Salmonella: both eliciting effective anti-tumor immunity. *Cancer Lett.*, **263**, 67–76.

Yang,Y. *et al*. (2010) Computational prediction of type III secreted proteins from gram-negative bacteria. *BMC bioinformatics*, **11** (Suppl. 1), S47.