Computational Analysis of Bacterial Type III Secreted Signal Sequences and In Silico Identification of New Type III Secreted Proteins

WANG, Yejun

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

in

Biology

The Chinese University of Hong Kong

August 2011

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Thesis/Assessment Committee

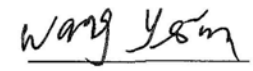
Professor JIANG Liwen (Chairman)

Professor GUO Dianjing (Thesis Supervisor)

Professor NGAI Sai Ming (Committee Member)

Statement:

All the works reported in this thesis were performed by the author, unless stated the otherwise.



WANG, Yejun

Abstract of thesis entitled

Computational Analysis of Bacterial Type III Secreted Signal Sequences and In Silico Identification of New Type III Secreted Proteins

Submitted by Wang Yejun

for the degree of Doctor of Philosophy

at the Chinese University of Hong Kong in August 2011

Type III secretion systems (T3SSs) widely exist in a number of human, plant, and animal bacteria. Their role is to translocate pathogenic proteins into host cells. These pathogenic proteins are called Type III secreted (T3S) effectors, which can cause cellular skeleton changes and can assist bacteria to interact or invade into host cells. Despite the important function of these pathogenic effectors, how they are specifically recognized and secreted by T3SSs is still not clear. Discovery of important features which guide the specific host-pathogen recognition and computational identification of new T3S effectors are therefore of great significance in understanding the mechanisms by which these pathogenic bacteria cause diseases. With these objectives, in this thesis research, I have made following accomplishments. First, I identified a novel amino acid composition (Aac) profile in the N-terminal signal sequences of experimentally validated T3S effectors. These Aac features were adopted in a machine learning model, to effectively predict T3S proteins. Second, I integrated the secondary structure (Sse) and solvent accessibility (Acc) encoded by T3S signal sequences and established a new prediction model with significantly improved prediction performance. Using this

improved model, new T3S proteins were computationally identified from *Salmonella* and selected candidates were further validated experimentally. Finally, I set up a T3SS-related database to integrate all the currently known T3SSs, manually annotated T3SS-related genes, and initiated a web server to accommodate the T3S protein prediction softwares developed in this study. This is so far the most comprehensive database integrated with gene prediction tools for the T3SS research community.

摘要

多種植物、動物以及人體致病菌均表達 III 型分泌系統，并通過後者將致病蛋白轉運至宿主細胞內。經 III 型分泌系統轉運的致病蛋白稱為 III 型效應蛋白，它們能引起宿主細胞骨架改變進而協助細菌入侵或作用於宿主細胞。III 型效應蛋白如何被 III 型分泌系統特異識別、分泌至今尚不清楚。因此，尋找指導 III 型分泌蛋白特異分泌的重要特徵以及計算預測新的 III 型效應蛋白，對於研究細菌致病機理具有重要的意義。圍繞這一目的，在本論文研究中，我進行了以下工作。首先，我收集了一組實驗證實的 III 型效應蛋白，比較這些蛋白序列，發現在其氨基端具有一些共有的氨基酸組成特徵。採用一種機器學習的方法來吸收這些氨基酸組成特徵，開發出高效的 III 型分泌蛋白預測軟件。進一步地，爲了探討 III 型分泌信號序列氨基酸特異組成的可能直接動力，我對 III 型分泌蛋白信號序列肽二級結構、親水性以及三維結構進行了分析並與非 III 型分泌蛋白加以比較，探尋兩者間上述參量的特徵區別。結合二級結構、親水性以及氨基酸組成特徵，對前一部份開發的 III 型分泌蛋白預測模型加以改進，新模型的性能得到顯著改善。利用改進后的軟件，對沙門菌全基因組加以掃描，預測出一組新的 III 型分泌蛋白；部份預測蛋白通過實驗進行了驗證。最後，我建立了一個 III 型分泌系統相關的數據庫。在數據庫中，我收集了當前已知的所有 III 型分泌系統，手工對所有 III 型分泌系統相關基因加以註釋，並為本研究的前兩部份開發的 III 型分泌蛋白預測軟件架設了網絡服務器。

Acknowledgements

First of all, I must give my deepest gratitude to my supervisor, Prof. Guo Dianjing, who led me into the bioinformatics research field, broadened my eyesight and thought, and provided me an ideal, free and active research environment and scientific atmosphere. Without her patient guidance, endless encouragement and constant support, I couldn't have completed my PhD project successfully.

I am grateful to my PhD program committee members Prof. Jiang Liwen and Prof. Ngai Sai Ming, who are sincerely concerned about my research progress, and gave me a lot of valuable suggestions and comments on my seminar, experiments and thesis. Besides, I am also grateful to my external committee member, Prof. Ching Wai Ki, for his constructive comments for my thesis.

I would like to give my special thanks to Dr. Shao Jianlin, who gave me important guidance on statistics and machine learning methods, and to Miss Wang Shanshan and Miss Qi Yan, who helped me a lot in cell culture experiments. I would also like to thank all my labmates in G94, including Zhang Qing, Sun Ming'an, Dr. Wang Wei, Cheng Han, Wu Ting, Dr. Wang Jingxue, Dr. Ma Dongming, Chan Yiuman and Yuan Man, who helped me a lot in either bioinformatics or molecular biology experiment.

Finally, I want to give my thanks to my girl friend and my beloved family for their endless love, understanding and support.

Table of contents

List of Abbreviations

A	Accuracy
ANN	Artificial Neural Network
AUC	the Area Under ROC Curve
BPB	Bi-Profile Bayes
CV	Cross Validation
FN	False Negative
FP	False Positive
MCC	Matthews Correlation Coefficient
NB	Naive Bayes
ROC	Receiver Operating Characteristic
Sn	Sensitivity
Sp	Specificity
SPB	Single-Profile Bayes
SVM	Support Vector Machine
T3S	Type III Secreted
T3SS	Type III Secretion System
TN	True Negative
TP	True Positive

CHAPTER 1

General Introduction

1.1 Type III Secretion Systems

Bacteria can encode at least six types (Type I ~ VI) of protein secretion system, among which Type III and Type IV secretion systems are especially important because they participate in bacterial pathogenesis and symbiosis (Hayes et al., 2010; Galán, 2009; Hueck, 1998; Alvarez-Martinez and Christie, 2009). Many gram negative pathogens encode functional Type III secretion systems (T3SSs), via which a group of pathogenic effectors enter into host cytoplasm, and consequently cause different human, plant or animal diseases, such as plague, typhoid, dysentery, rice blast, bacterial leaf streak and so on (Cornelis, 2000; Schroeder and Hilbi, 2008; Ly and Casanova, 2007; Alfano and Collmer, 2004; Bonas and Van den Ackerveken, 1999).

A typical T3SS contains two sets of proteins, namely, apparatus components and substrate proteins (Fig 1A) (Galán and Wolf-Watz, 2006; Hueck, 1998). T3SS apparatus, which assembles spanning bacterial cell membranes like a syringe, is composed of three contiguous parts: a basal body within bacterial cytoplasm, multi-ring structure spanning inner and outer membranes, and a needle-like hollow filament outside outer membrane. The needle-like filament can contact and insert through host eukaryotic cell membrane (Enninga and Rosenshine, 2009; Izoré, 2011; Schraidt and Marlovits, 2011). With assistance of other accessory proteins, the basal body can specifically recognize substrate proteins, and then translocate them into eukaryotic cytoplasm via T3SS conduit (Lara-Tejero et al.,2011; Ghosp 2004).

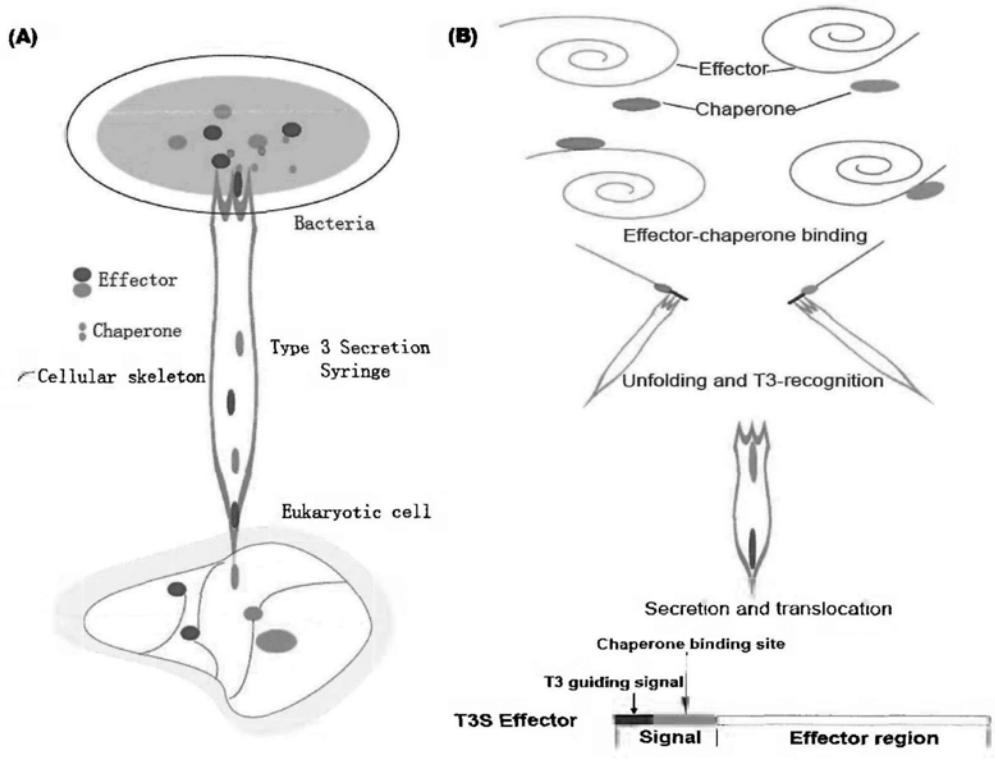


Fig 1. Diagram of Type III Secretion System and Secretion of Type III Secreted Proteins.

(A) Diagram of Type III Secretion System. Type III Secretion System Apparatus is a needle-like structure which spans bacterial inner and outer membrane. It pierces host eukaryotic cell membrane and translocate some specific proteins (Effectors) into host cytoplasm, under the help of other proteins including chaperones. These effectors can induce host cellular skeleton changes, which are beneficial for bacterial invasion.

(B) Mechanism of Type III Secretion. Chaperone can recognize Type III Secreted Effectors and bind to them at specific sites located at the N-terminal end. The binding leads to unfolding of effector and then the effector itself or effector-chaperone complex will be specifically recognized by and secreted through Type III Secretion Conduit. An effector typically contains a signal region and an effector region.

A group of proteins participate in the assembly of T3SSs. These apparatus encoding genes (sometimes together with their regulation genes), the substrate genes, and the chaperones are located in the same chromosomal or plastidial region constituting one or more operons (Galán 1989; Mazurier 2006; Marguerettaz et al, 2011). More substrate genes are scattered rather than clustered together in genomes like apparatus genes (Wood et al., 1996 and 2000; Norris 1998). The T3SS related genes in the same operon or cluster are frequently transmitted into other species horizontally, and they are often transcribed coordinately (He et al., 2004; Heuck 1998; Deane et al., 2010; McDermott et al., 2009; Winnen et al., 2008). In different bacteria species, T3SS apparatus genes are also more conserved than substrates in their DNA or amino acid sequences (He et al., 2004; Heuck 1998).

1.2 Type III Secreted Proteins

Type III secreted (T3S) proteins, frequently called effectors or T3S substrates, are a variety of proteins that can be specifically secreted by T3SS conduits (Fig 1B). However, it is yet not clear about the mechanisms by which T3S proteins are specifically secreted. Although some groups argued with individual molecular evidence that T3S signals were located in mRNA sequences of T3S proteins, most lines of evidence from transgenic and mutational analysis in different bacteria support that the N-terminal amino acids of T3S proteins contain the signals guiding their specific secretion (Anderson and Schneewind, 1997; Ramamurthi and Schneewind 2003; Karavolos et al., 2005; Lloyd et al., 2001 and 2002; Russmann et al., 2002; Schechter et al., 2004; Wang et al., 2008).

Besides their specific recognition by T3SS apparatus, it is also important for T3S proteins to unfold their structure before secretion so that they can be translocated through the rather narrow T3SS conduit (Kubori et al., 1998; Marlovits et al., 2004). It is widely accepted that T3S proteins achieve this goal by binding to specific chanperones (Fig 1B) (Stebbins and Galán, 2003). Most T3S proteins own their one-to-one specific chaperones, while some chaperone may recognize and bind more T3S proteins. Chaperones are proteins located in bacterial cytoplasm and not secreted. After bound by specific chaperone, T3SS protein will unfold its tertiary structure to become linear from aggregating state, and is consequently equipped into the narrow T3SS conduit (Stebbins and Galán, 2001; Akeda and Galán, 2005; Stebbins 2005; Lilic et al., 2006).

1.3 Computational prediction of Type III Secreted Proteins

Bacterial pathogenesis or interaction with host cells is rather a complex process involving a large number of bacterial proteins secreted into host cells. In earlier years after the discovery of T3SSs, only a limited number of T3S effectors were experimentally identified to be significant for bacterial invasion (Heuck 1998; Ghosp 2004). With the increased number of identified T3S proteins, it has been recognized that an unexpected large number of T3S proteins are encoded by bacterial genomes, and the interactions between bacteria and hosts are far more complicated than previously expected.

Due to the low expression abundance and the fine regulation of T3S proteins, as well as the limitations of experimental techniques, it is almost impossible to identify

new T3S proteins solely depending on small-scale experiments. Earlier bioinformatic analysis on the N-terminal amino acid composition (Aac) patterns, G + C contents in their 5' nucleotide encoding region, and the pairwise co-existance of T3S proteins and their chaperones, etc., has led to the identification of a large number of new T3S proteins from a variety of bacteria (Panina et al., 2005; Petnicki-Ocwieja et al., 2002; Tobe et al., 2006). Recently, two groups carefully analyzed different features buried in the N-terminal regions of validated T3S proteins. According to the significant features discovered respectively, T3S protein prediction softwares were developed based on machine learning methods, e.g. Naive Bayes (NB) and Support Vector Machine (SVM) (Samudrala et al., 2009; Arnold et al., 2009). Another group tried to use Artificial Neural Network (ANN) to model position-specific Aac profiles of T3S proteins and obtained better classifying performance (Löwer et al., 2009). Other features including secondary structure (Sse) and solvent accessibility (Acc) were found to contribute to the specific recognition of T3S proteins by T3SS (Yang et al., 2010).

1.4 Project Objectives

The difficulty in new T3S protein discovery based on pure wet-lab experiments requests efficient computational tools for genome-wide *in silico* T3S prediction. Feature identification and extraction are made feasible with the large number of validated T3S proteins. However, a variety of drawbacks exist in the currently available T3S prediction tools, including low accuracy, low specificity, weak inter-species adaptation, etc. Therefore, the major goal of this thesis study is to analyze and identify new distinguishing features in T3S proteins, and to develop T3S

protein prediction software tools with high performance and wide application based on these new features. Specific objectives include:

1. To analyze and extract position-specific features of T3S proteins;
2. To develop feature-based software tools for efficient computational prediction of T3S;
3. To identify new T3S proteins in model bacteria with experimental validation;
4. To develop a database to integrate T3SS related molecular information and to extend the application of our prediction tools for the research community.

CHAPTER 2

High-accuracy Prediction of Bacterial Type III Secreted Effectors Based on Position-specific Amino Acid Composition Profiles

2.1 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 T3SS 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 acids 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 and 2002; Russmann 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-Waltz, 2006; Heuck, 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 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 sequences, general enrichment and depletion of N-terminal amino acid composition, composition frequency of secondary structure elements (coils, 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 softwares developed for T3S protein prediction (Arnold et al., 2009), explores possible sequence-based features exhaustively. In Effective T3, the amino acid composition 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 et al. 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 also be pointed out: (1) 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; (2) the classifying performance was based on train-reclassification results only and no cross-validation was performed; and (3) 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 specificity (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.

I have developed a computational model based on position-specific Aac features for effective T3S protein prediction. I will demonstrate that this model out-performs other current implementations in terms of both sensitivity and selectivity. With this model, a genome-wide prediciton of T3S proteins was also conducted in an important plant pathogen, *Ralstonia solanacearum*.

2.2 Methods and Materials

2.2.1 Data source

A list of experimentally validated type 3 secreted (T3S) proteins from animal pathogens, plant pathogens, and symbiotic bacteria were 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 was 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. Because the number of non-T3S proteins was much larger than that 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 in Wang et al., 2011. The secondary structure (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 solvent accessibility (a combination of 'B' and 'E'). For 5-fold cross-validation, the negative and postive training datasets were

pooled as the fial training datasets and were split into 5 sub-datasets, each containing the same number of positive/negative samples.

2.2.2 Position-specific profiles and feature extraction

The unaligned T3S proteins and non-T3S proteins were used for position-specific feature vextraction. Let vector S = {s1,s2,s3,...,sn} denotes a peptide sequence in which s represents amino acid or other property while 1,2... 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: $p(A_i) = f(A_i)/m_i$, in which $f(A_i)$ and m_i denotes the frequency of amino acid A at position *i* and sequence number at position *i*. For each position, the $p(A_i)$ 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 secondary structure and solvent accessibility 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 more frequently 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 $S = \{s_1,s_2,s_3,...,s_n\}$, where each s_i ($i = 1,2,3,...,n$) 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) and

C_{-1} (non-T3S proteins). The posterior probability of both T3S and non-T3S proteins can be calculated as the occurence 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 postion-specific amino acid compositon, secondary structure and solvent accessibility.

2.2.3 Support vector machine implementation and parameter optimization

R package for SVM, 'e1071', was used to train and build the SVM models (Dimitriadou et al., 2009). Radial basis kernel function was selected for SVM prediction. SVM parameters gamma and cost were optimized using grid search based on 10-fold cross-validation (Scholkopf and Smola, 2002).

2.2.4 Performance assessment

Accuracy (A), Specificity (S_p), Sensitivity (S_n), 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 (T3S) and negative (Non-T3S) proteins correctly classified. S_n (true positive rate) and S_p (true negative rate) represent the percentage of positive proteins (T3S) and that of negative proteins (Non-T3S) correctly classified, respectively. An ROC curve is a plot of S_n versus (1 - S_p), while AUC gives a measure of classifier performance. MCC takes into account true and false positives and negatives, and is generally a balanced measure which can be used

even if the sizes vary significantly between classes.

$A = \frac{TP+TN}{TP+FP+TN+FN}$, $S_p = \frac{TN}{TN+FP}$, $S_n = \frac{TP}{TP+FN}$,

$MCC = \frac{(TP \times TN)-(FN \times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}}$, where *TP*, *TN*, *FP* and *FN* denotes the number of true positives, true negatives, false positives and false negatives, respectively.

2.2.5 Amino acid position shift and frame shift

For position shift test, both insertion and deletion datasets were created. For deletion test, we generated 5 individual datasets with 1, 2, 3, 4 or 5 amino acids deleted respectively at the N-terminal end and excluding starting methionine. For insertion test, one of the 20 amino acids was inserted before the 1st or 2nd 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 amino acid composition bias was found at 1st position for non-T3S proteins and at the 2nd position for T3S proteins. For frame shift experiments, DNA sequences encoding the first 100 amino acids at the N-termini excluding the starting methionine were obtained. For each sequence, two mutations with '-1' and '+1' frame shift 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 re-classified using the optimized BPBAac model.

2.2.6 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 I randomly selected proteins not annotated as T3S from different bacteria species as negative training datasets and the ratio of negative to positive samples was 2:1. I also removed some apparent false positive sequences (e.g., chaperone, 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).

2.2.7 Genome-wide prediction of T3S proteins from *Ralstonia solanacearum*

The recently validated T3S proteins in *Ralstonia solanaceanum* 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 GMI1000 prteome, 3437 chromosome-encoding proteins (Genome ID: NC_003295) and 1676 plasmid-encoding proteins (Genome ID: NC_003296) of *Ralstonia solanacearum* GMI1000 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 tested using the BPBAac model. The cutoff value of BPBAac was set as 0.5. The proteome of Ralstonia was also predicted for T3S protein

candidates using Effective T3 and ANN respectively with originally optimized parameters.

2.3 Results

2.3.1 Distinct position-specific amino acid composition profiles for T3S effectors

The N-terminal amino acids from end were retrieved from T3S and non-T3S proteins respectively and amino acid composition (Aac) was calculated for each position. Significantly distinctive Aac profiles were found between these two types of proteins (Fig 2A and 2B). 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 2A). 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 symbioint). Both groups showed apparent Aac preference profiles different from that of non-T3S proteins (Fig 3A-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 (Fig 3A-B). I also 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 (Fig 3C-F).

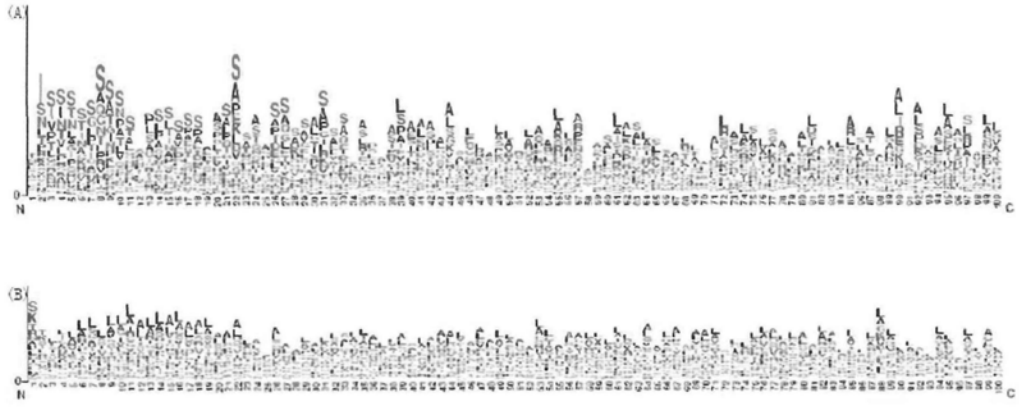


Fig 2. Distinctive N-terminal position-specific Aac feature in T3S proteins.

Amino acid positions are depicted on the horizontal axis. The heights of characters represent the preference or enrichment level. **(A)** Aac preference for T3S proteins. **(B)** Aac preference for non-T3S proteins.

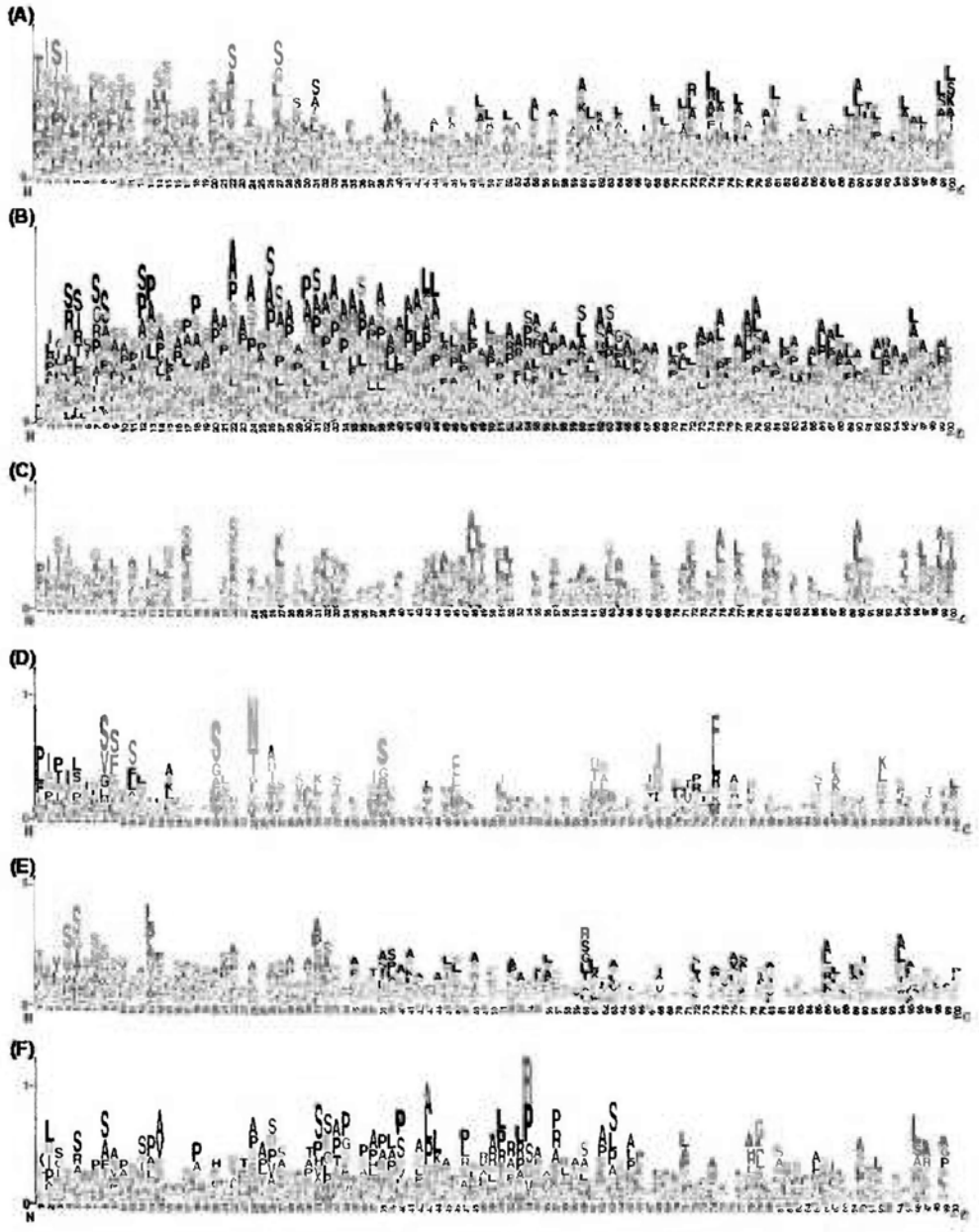


Fig 3. Position-specific amino acid composition (Aac) in different bacteria subgroups.
Horizontal axis: amino acid positions. The heights of characters represent the preference or enrichment level. **(A)-(F)**: T3S protein Aac preference in animal pathogens **(A)**, plant pathogens or symbionts **(B)**, representative individual animal pathogens including *Salmonella* **(C)** and *Citrobacter* **(D)**, and representative plant pathogens including *Pseudomonas* **(E)** and *Xanthomonas* **(F)**.

2.3.2 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, Support Vector Machine (SVM) models were trained for the Aac features (Materials and Methods). Two different SVM models were trained: 1) SPBAac model that only considers the Aac profile of positive T3S training dataset; 2) BPBAac model considers the Aac profiles of both T3S and non-T3S proteins. Table 1 and Fig 4A 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 vs. selectivity) of established feature-based T3S protein prediction methods were reported as 71% vs. 85%, 74% vs. 98% and 65% vs. 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 re-trained 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 4B), although the best performance was achieved when the first 100 amino acid were used (Fig 4B). From this analysis, I 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).

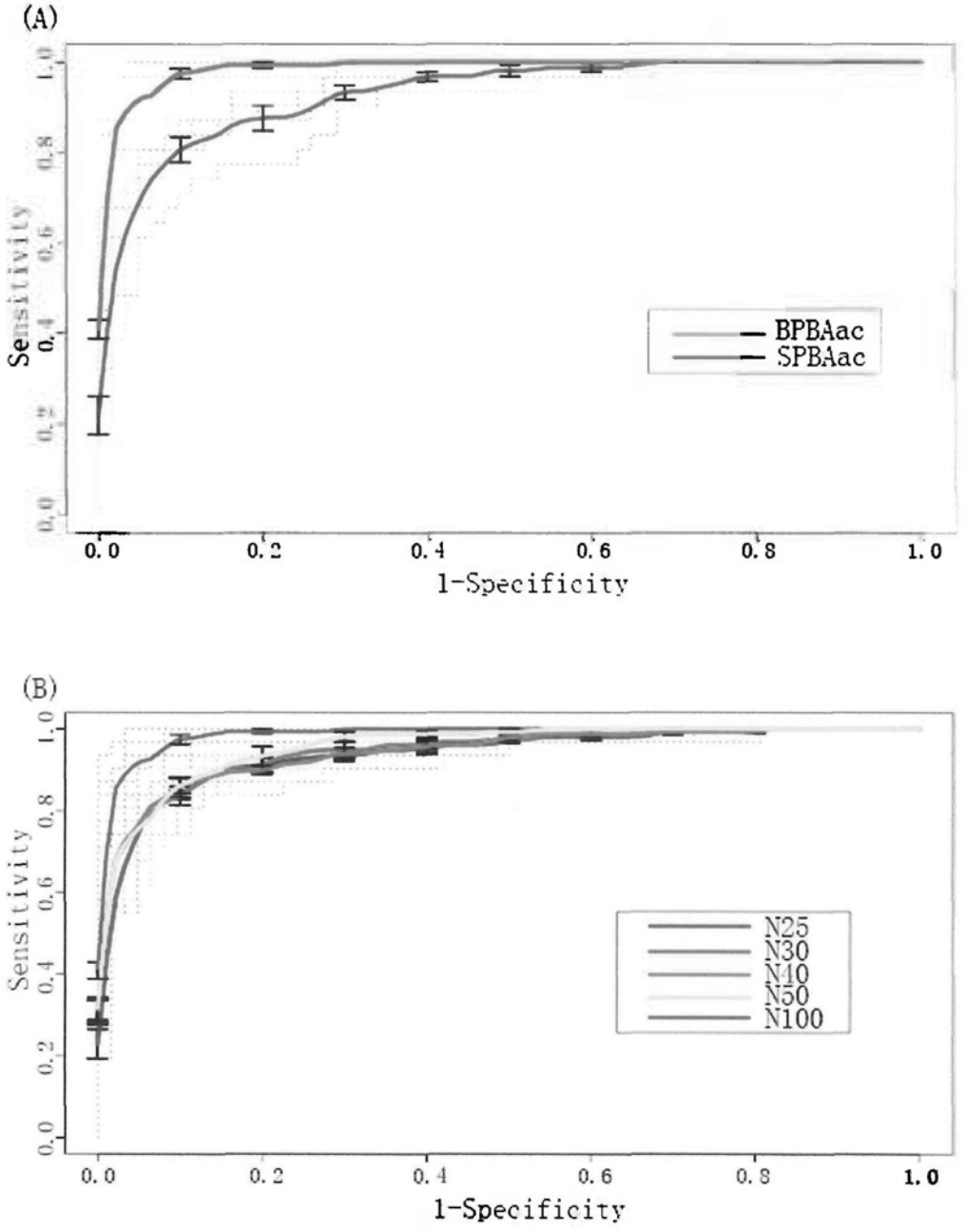


Fig 4. Performance of SVMs based on different feature-extraction models and sequence lengths.

(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.

Table 1. Optimal parameters and corresponding performance of BPB and SPB model based on five-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.
γ: gamma, which was optimized based on 10-fold cross-validation grid search.

2.3.3 The robustness of BPBAac model

The robustness of BPBAac model was examined (1) by randomly selecting sub-datasets with different sample sizes from the training data to re-train the model and to classify the remaining data; and (2) by using the Leave-One-Out strategy. Specifically, T3S and non-T3S proteins from one bacterial genus were eliminated from the test dataset as new testing 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. The 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 5A). For different genera or subgroups, most of the effectors could be recalled and the AUC values did not show significant change (Fig 5B). Taken together, the position-specific Aac profiles were important features for T3S protein identification in different bacteria species.

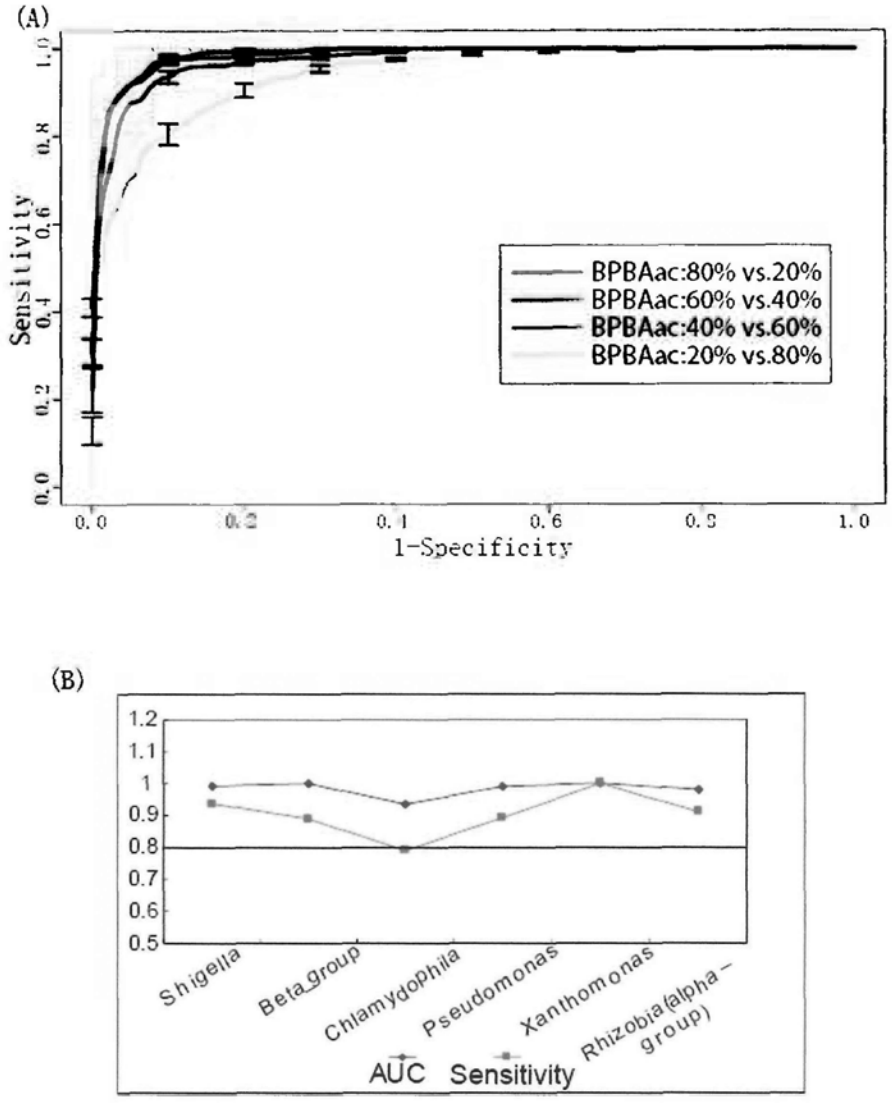


Fig 5. Performance of BPBAac models trained with different datasets.

(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.

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

The secondary structure element (Sse) and solvent accessibility (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 re-trained 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 6). Therefore, although T3S proteins contain Sse and Acc profiles different from those of non-T3S proteins (data shown in Chapter 3), the Aac feature alone is enough to distinguish T3S and non-T3S proteins.

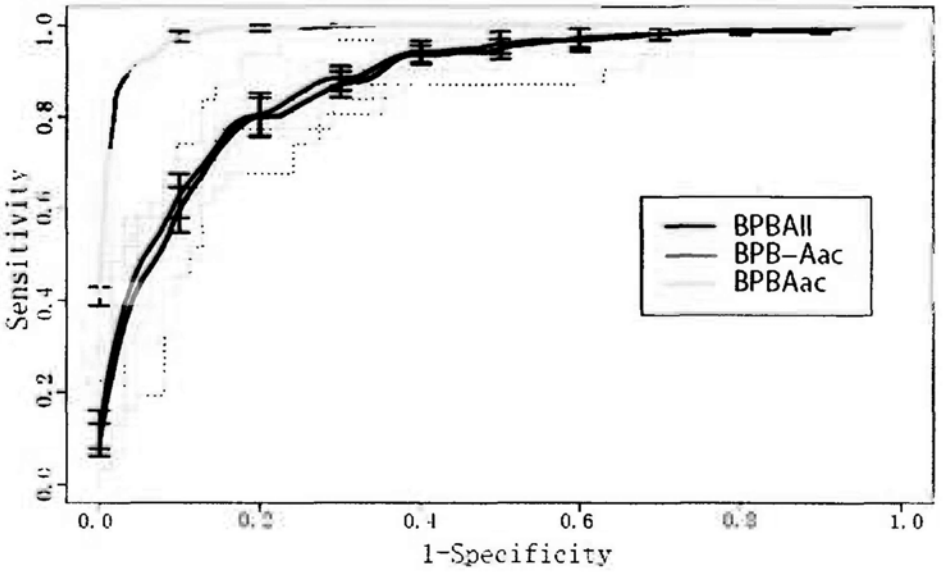


Fig 6. Comparison of ROCs based on different type of features.

All themodels 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.

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

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, form 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 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 1st or 2nd N-terminal amino acid position. For these two types of insertion mutaions, similar proportion of proteins were correctly re-classified (53% and 55% for 1st and 2nd respectively). (Table 2). The re-classifying performance was also influenced by the type of amino acid inserted. Because non-T3S proteins showed significant amino acid preference at the 1st 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 2nd position, and insertion of T3S protein preferred amino acid (e.g., 'I' or 'N') resulted in higher sensitivity.

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

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. N/A: re-classify 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)'.

2.3.6 Performance comparison with current prediction models

The classification performance was compared among BPBAac, Effective T3 and T3SS ANN. First, Effective T3 and ANN were used to re-classify the training datasets used in this research ('BPBAac dataset'), and the performance was compared with cross-validation rates of BPBAac. Table 3 and Fig 7A clearly demonstrated that BPBAac out-performed these two methods in 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: (1) BPBAac was first re-trained using the Effective T3 dataset and ANN dataset respectively before it was used to re-classify BPBAac dataset (Table 3 ad Fig 7B); (2) BPBAac was re-trained using the datasets adopted by Effective T3 and ANN respectively, and the new model was used to re-classify those two datasets (Table 3). As show in Table 3 and Fig 7, BPBAac model consistently performed better than Effective T3 and ANN.

Table 3. Performance comparison among different models

Test Dataset	Method	*Sn* (%)	*Sp* (%)	*A* (%)	AUC(%)	MCC
BPBAac dataset[a]	BPBAac-CV	90.97	97.42	95.27	98.88	0.8929
	BPBAac-ANN	83.77	94.81	91.13	97.52	0.7982
	BPBAac-ET3	93.51	95.45	94.81	98.67	0.8840
	ANN	71.87	93.23	85.64	92.57	0.7182
	Effective T3	82.53	86.63	86.69	89.56	0.6852
ANN dataset[b]	BPBAac	99.62	99.85	-	-	-
	ANN	74	98	-	-	-
Effective T3 dataset[c]	BPBAac	82.73	92.73	-	-	-
	Effective T3	71	85	-	-	-

a. All the methods were used to re-classify BPBAac dataset. BPBAac-CV, BPBAac-ANN and BPBAac-ET3 represent the BPBAac model trained with BPBAac dataset (5-fold cross-validation), ANN dataset, and Effective T3 dataset, respectively. ANN and Effective T3 were trained with their original dataset, respectively.

b. The models were trained with ANN dataset and used to re-classify ANN dataset. The *Sn* and *Sp* values for ANN method were retrieved from reference Löwer,M. and Schneider,G. (2009).

c. The models were trained with Effective T3 dataset and used to classify the same dataset using a 10-fold cross validation. The *Sn* and *Sp* values for Effective T3 method were retrieved from reference Arnold,R. *et al.* (2009).

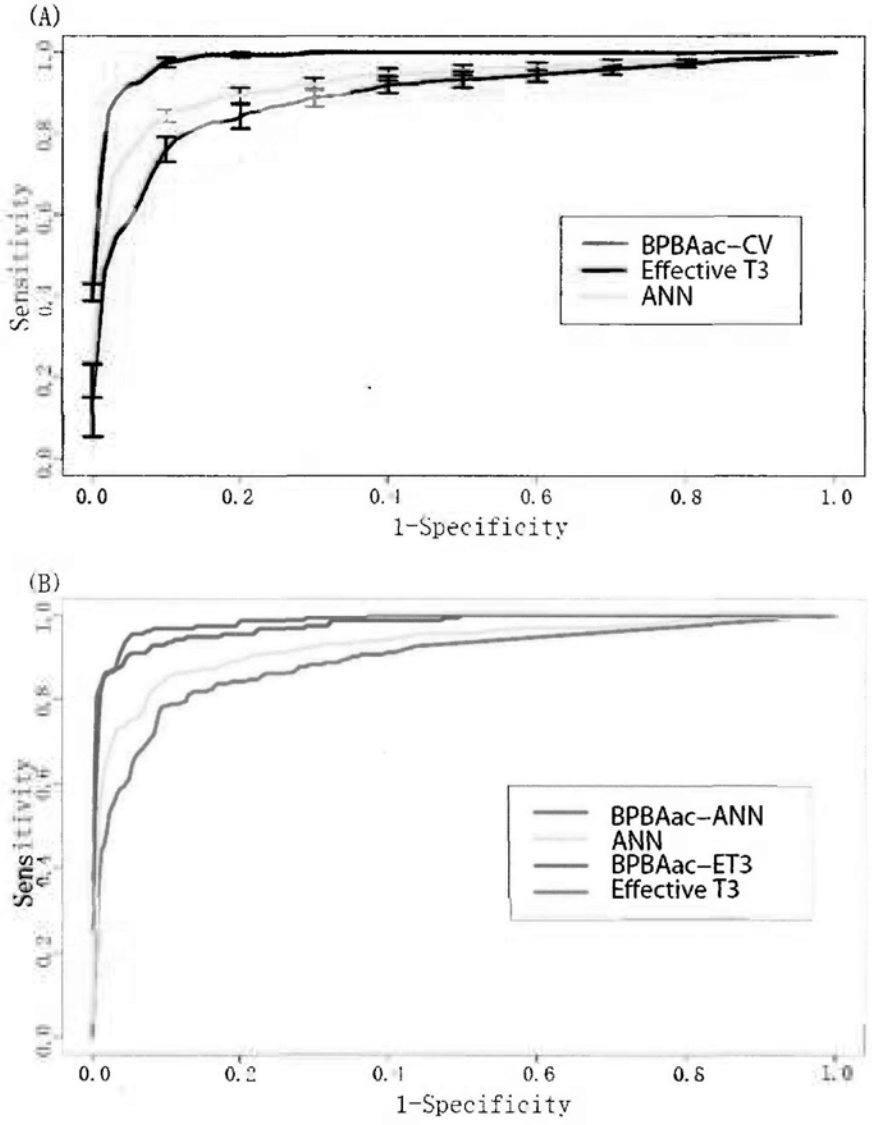


Fig 7. Comparison of performance using different datasets.

(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.

2.3.7 Genome-wide prediction of T3S proteins in Ralstonia 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 Ralstonia solanacearum GMI1000 were used for genome-wide prediction of T3S proteins. Totally, 1.4% (49/3437) chromosome encoding proteins and 2.9% (48/1676). of plasmid encoding proteins were predicted to be T3S proteins (Table S2 in Wang et al., 2011). With a higher recall percentage, a much smaller number of putative T3S candidates was obtained, making validation work more feasible (Table 4). Interestingly, many candidates (38/97, 39.2%) are annotated as 'hypothetical' proteins with unknown function. Some candidates were validated recently (eg., PopF1), closely related with T3SS (e.g, NP_522416.1, Hrp pilus subunit HRPY protein) or originated in bacteriophages (eg., NP_519819.1) (Table S2 in Wang et al., 2011).

Table 4. 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)

2.4 Discussion

2.4.1 Position-specific amino acid composition features in signal regions of T3S proteins

Previous studies have demonstrated that T3S effectors contain conserved N-terminal amino acid composition pattern. For example, Lloyd et al. found that serine and isoleucine were enriched at the N-termini of YopE protein (Lloyd et al., 2002). Petnicki-Ocwieja et al. 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 (Petnicki-Ocwieja et al., 2002). Recently, Samudrala et al. examined sequence-based amino acid composition bias and concluded that the Aac profiles were 'largely uninformative' (Samudrala et al., 2009).

In this study, I 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 (Fig 2A). I 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 (Fig 2A). Using a Bi-profile Bayesian model, I 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 (Fig 4A). Our model achieved great predictive power, and was superior to previous models using sequence-based Aac features (Fig 7A-B). Apart from Effective T3 and ANN, I 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 in Wang et al., 2011). Further exploration of the position-based amino acid composition features may provide important clues about the nature and evolution of the T3SS signals.

2.4.2 Possible features other than Aac in signal regions of T3S proteins

Apart from Acc, other features may also contribute to the mechanisms underlying T3S secretion. Previously, other groups have examined secondary structure (Sse) and solvent accessibility (Acc) (Arnold et. al., 2009; Yang et.al., 2010) but no conclusive remarks can be drawn so far. Arnold et. al. found that neither Sse, nor Acc could improve the classifying performance. Using a combinative feature extraction strategy, Yang et. al. found that the combination of Sse and Acc could improve the model performance. In the present study, I 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 vs. selectivity for Sse and Acc respectively were: 47% vs. 79%, 16% vs. 99%; data not shown). When these two features were combined with Aac, the performance was significantly decreased (Fig 6). Further detailed analysis is being carried out to investigate the possibly contribution of Sse and Acc when they are considered as covariables of Aac. Such analysis may be more realistic because the three types of features are more likely not independent with each other. In this research, however, they were considered independent variables.

2.4.3 Underlying drawbacks and possible limitations of BPBAac

Two methods are frequently adopted for position-specific amino acid composition modeling: Hidden Markov Model (HMM) and sliding window technique. Because T3S proteins contain a long signal bearing sequence, the sliding-window model may become quite complex and sometimes encounter the over-fitting problem (Löwer and Schneider, 2009). I have also attempted a T3S protein HMM, but found its classifying performance inferior. The Bi-profile Bayes model was first proposed by Shao et. al., and it has been successfully used to predict protein methylation sites (Shao et al., 2009). 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, I 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 9 amino acids 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 (Table 2). I 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 frame shifts generated in the signal region (Table 2), I 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.

2.4.4 Application of BPBAac

Finally, I applied the BPBAac model to make a genome-wide prediction of T3S effectors in an important plant pathogen, Ralstonia solanacearum GMI1000. Most of the validated T3S effectors were recalled by BPBAac (Table 4). More importantly, far fewer candidates were predicted by BPBAac than by other models such as ANN and Effective T3 (Table 4). 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.*

The list of newly identified putative T3S candidates may serve as a useful

resource for the research community. Because the Aac features were commonly identified across genus and species (Fig 3A-F), the BPBAac tool could be widely used for efficient T3S effector prediction in various bacteria species.

CHAPTER 3

Identification of New Type III Secreted Proteins Based on Position-specific Sequence-Structure Joint Features

3.1 Introduction

Type III secreted (T3S) effectors represent a group of proteins specifically recognized by and secreted through type III secretion systems (T3SSs) (Galán 2009; Lindeberg and Collmer 2009). The number of T3S effectors varies greatly among different bacterial species with functional T3SSs, while the sequences lack apparent similarity between each other (Hueck 1998; Ghosp 2004). This makes it extremely difficult to identify new T3S effectors by traditional sequence alignment or phylogenetic approaches.

Some important effector coding genes may be clustered with T3SS apparatus elements in the same operon or genomic region (Galán et al., 1989; Jarvis et al., 1995; Huang et al., 1995; Hong and Miller, 1998), which faciliate the earlier identification of effectors in most species (Jarvis et al., 1995; Noel et al., 2002 and 2003). The weak conservation of effectors among closely related species led to slight increase in the number of identified effectors (Kaniga et al., 1995; Hardt and Galán, 1997). Other common features including distinct G + C content and clustering together with chaperones, etc., also help to identify a large number of new effectors genetically scattered in the genomes (Panina et al., 2005; Petnicki-Ocwieja et al., 2002; Tobe et al., 2006).

Two grounding discoveries greatly accelerated the progress of finding new effectors. One is the discovery that the N-terminal peptide sequences of T3S effectors contain both necessary and enough signal information that guide the specific protein secretion (Rüssmann et al., 2002; Lloyd et al., 2001), while the other verified that T3S

effectors can be secreted through T3SS conduits of different bacteria (Rüssmann et al., 2001; Girard et al., 2009). Based on these two discoveries, a lot of new common features were found in the N-terminal signal sequences of T3S proteins, including sequence patterns, amino acid composition frequency, secondary structure composition, etc. (Arnold et al., 2009; Samudrala et al., 2009; Löwer et al., 2009; Yang et al., 2010; Wang et al., 2011). These new features made large-scale computational prediction of new T3S proteins possible (Arnold et al., 2009 and 2010; Samudrala et al., 2009; Löwer et al., 2009; Yang et al., 2010; Wang et al., 2011). Although new features may be useful for the identification of more T3S proteins, it is still difficult to explain the nature or the exact mechanism by which T3S proteins are specifically recognized and secreted. To date, the most important features discovered would be the distinct sequence-based or position-based amino acid composition (Aac) profiles in N-terminal signal regions of T3S proteins (Arnold et al., 2009; Wang et al., 2011). The Aac profiles are not so striking, however, and therefore it becomes infeasible to find out one or more inter-species or even within-species motifs (Wang et al., 2011). The enriched or depleted amino acids in signal sequences do not contain apparently common physical and chemical properties either. To directly interpret the possible connections between relatively unique Aac features and the specificity of protein secretion, several research groups analyzed the second-order structure composition encoded by the primary signal peptide sequences, including the secondary structure (Sse) and water accessibility states (Acc) (Arnold et al., 2009; Samudrala et al., 2009; Wang et al., 2011). Distinctive Sse and Acc features were

noted, and yet these features individually seemed not to contribute to the specific recognition of T3S proteins (Arnold et al., 2009; Samudrala et al., 2009; Wang et al., 2011). One group considered the joint distribution of Sse or Acc and Aac, and provided limited evidence that the Sse and Acc features contribute to the specific secretion of T3S proteins (Yang et al., 2010). Previous tertiary structure studies demonstrated that different T3S effectors (or chaperones) often adopted similar structure though the primary sequences were apparently not conserved (Stebbins and Galán 2001; Singer et al., 2004; Birtalan et al., 2002; Lee et al., 2004; Wulf et al., 2004; Luo et al., 2001). These results provide some clues about the possible connection between the shared 3D structure and the similar secretion mechanism of T3S proteins with weak sequence similarity. However, those studies emphasized on the effector regions rather than the signal regions, and till now most of the 3D structure information of T3S signal regions is not available. Consequently, 3D feature analysis is impossible although the this feature may be important for understanding the basis of T3S protein secretion and the relationship between unique Aac features and the specific recognition of T3S proteins.

According to Aac and other undefined features in the signal regions, T3S proteins could be loosely recognized as a unique protein family. Regarding the origin and the evolution of T3S proteins, Stavrinides et al. proposed a 'terminal re-assortment' hypothesis with genome-wide evidence (Stavrinides et al., 2006). According to this model, the T3S proteins contain signal and effector regions, and these two regions fuse randomly and evolve independently before fusion. Therefore, the formation of

signal sequences may be random and not necessarily present in bacteria, gram-negative bacteria, or bacteria with functional T3SSs. Besides, there might be a group of proteins with similar signals but in fact they are not true T3S effectors although they can be specifically recognized and secreted through functional T3SSs.

Continuing our previous study, in this part, I further analyzed the underlying features that connect distinct Aac profiles and specific recognition of T3S proteins. The new features include Sse, Acc and 3D structure of T3S signal sequences. A new and more effective prediction model based on these features was proposed. Genome-wide T3S prediction was conducted for *Salmonella* and selected predictions were validated experimentally. Finally, based on whole-genome prediction results for various bacteria and yeasts, a hypothesis was proposed for the evolution of T3S signals.

3.2 Materials and methods

The source of T3S and non-T3S protein datasets, the description of prediction methods for secondary structure (Sse) and solvent accessibility (Acc), cross-validation method, Bi-profile Bayes (BPB) method, Support Vector Machine (SVM) implementation, and model performance assessment methods, were all similar to those described in Section 2.2 and Wang et al., 2011. The slight difference is that more non-redundant T3S proteins validated (189) and non-T3S control (385) proteins were included in the final training dataset. Unless specified, the starting methionine is excluded from the N-terminal sequences of T3S or non-T3S proteins.

3.2.1 3D structure prediction and alignment

I-TASSER was used to make de novo prediction of 3D structure for the first 100 amino acids at the N-terminal end of T3S and non-T3S proteins (Wu et al., 2007). The generated model with highest resolution was used for further analysis. To make consequent structure analysis more reliable, two strategies were adopted. Firstly, the resolution cutoff value was set to < 10 angstrom. Secondly, another efficient de novo protein structure prediction tool, MUFOLD was used to predict the structure of filtered protein in parallel (Zhang et al., 2010). The predicted structure was used for further clustering and comparison only if two softwares give significantly similar results (>75% similarity). MultiProt was adopted for structure alignment using a sequence order dependent or independent mode (Shatsky et al., 2004).

3.2.2 Joint feature extraction and model performance comparison

Let vector S = $\{s_1,s_2,s_3,...,s_n\}$ denotes a sequence of peptides, in which s represents amino acid while $_{1,2}$...or $_i$ represents position and $_n$ represents total length of the sequence. For any $_1 \leq _i \leq _n$, s_i has 20 alternatives since it could be any one of the 20 amino acids. Let Sse[s_i] and Acc[s_i] represent the secondary structure element (Sse) and solvent accessibility state (Acc) that s_i takes, respectively. Sse[s_i] belongs to set {C, H, E} and Acc[s_i] belongs to set {B, E}, and consequently for any position $_i$ ($_1 \leq _i \leq _n$), there are 20 x 3 x 2 = 120 types of combination of the three categories of components (amino acid, Sse and Acc). The frequency of each type of combination was calculated for each position of positive training sequences (T3S) and negative training sequences (non-T3S), represented as $P_{+1}(s_i$Sse[s_i]Acc[s_i]) and $P_{-1}(s_i$Sse[s_i]Acc[s_i]), respectively. For each sequence, a feature vector containing $2n$ bi-profile frequencies was obtained for n sequential positions (n was set as 100 in this research):

$\{P_{+1}(s_1$Sse[s_1]Acc[s_1]), $P_{+1}(s_2$Sse[s_2]Acc[s_2]), ..., $P_{+1}(s_n$Sse[s_n]Acc[s_n]),

$P_{-1}(s_1$Sse[s_1]Acc[s_1]), $P_{-1}(s_2$Sse[s_2]Acc[s_2]), ..., $P_{-1}(s_n$Sse[s_n]Acc[s_n])}.

The features were trained with a support vector machine, resulting in a model, namely T3SEpre. SSE-ACC and BPBAac were re-trained with the same dataset with prior parameters suggested by the original paper and the 10-fold cross-validation grid searching results. The performance was compared among T3SEpre, SSE-ACC and BPBAac based on a 5-fold cross-validation evaluation.

3.2.3 Whole-genome T3S protein prediction

Bacteria or yeast whole-genome protein sequences were downloaded from NCBI Genome database (http://www.ncbi.nlm.nih.gov/genome/). The N-teminal up to 100 amino acid position or full-length sequence for peptides with fewer than 100 amino acids were extracted for secondary structure prediction using PSIPRED (McGuffin et al., 2000). The solvent accessibility was predicted using SCRATCH (Cheng et al., 2005). The amino acid sequence, Sse sequence, and Acc sequence were used together for T3SEpre to predict whether the corresponding peptide contains T3S signals (http://biocomputer.bio.cuhk.edu/softwares/T3SEpre). For more specific results, a default cutoff value of 0.5 for T3SEpre was used.

3.2.4 Bacteria, plasmids and cell lines

E.coli DH5alpha and *Salmonella typhimurium* strain SL1344 were used in this research. DH5alpha was commercially available while SL1344 was obtained from Salmonella Genetic Stock Centre (SGSC, http://www.ucalgary.ca/~kesander). The bacteria were cultured on LB plate or in LB broth with or without 100mg/L ampicilin. The plasmids used in this study were summarized in Fig 8 and Table 5. The pMS107 plasmid with Bordetella CyaA gene insertion was gifted by Professor Guy R Cornelis (Focal Area Infection Biology, Biozentrum, University of Basel, Switzerland). A pair of primers (Table 6) were designed to PCR amplify CyaA gene. The pBADB-Myc-His plasmid with an L-arabinose-induced promoter and C-terminal Myc and His double tags, was ordered from Invitrogen (Cat. No. V440-01). CyaA gene fragment was cloned into pBADB-Myc-His plasmid, generating pBADB-CyaA-tag (Fig 8). DNA

sequences encoding N-terminal 100 amino acids of candidate T3S proteins were amplified and cloned into pBADB-CyaA-tag at the 5' end of CyaA sequence, resulting in different constructs (Table 5).

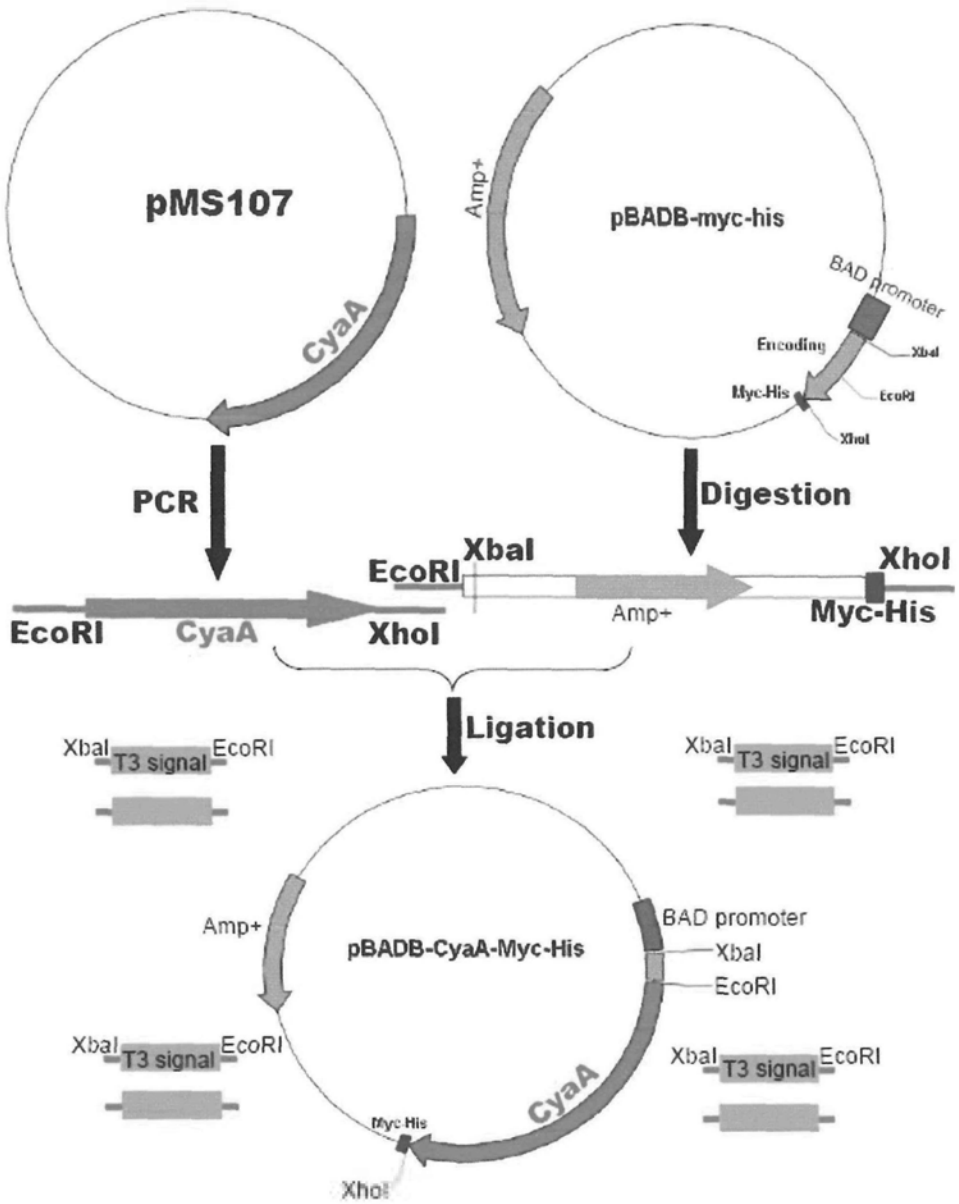


Fig 8. Construction of Cya translocation reporter plasmid.

Plasmid pMS107 containing CyaA fragment was used as template to amplify CyaA gene with EcoRI and XhoI restriction sites. The PCR product was further cloned into plasmid pBADB-Myc-His to get the resulting pBADB-CyaA-tag reporter plasmid. Interesting candidate signal sequences were cloned into pBADB-CyaA-tag plasmid between XbaI and EcoRI sites to obtained different testing plasmids, respectively.

Table 5. Plasmids used in this study

Name	Origin	Property and usage
pMS107	Gifted by Prof. Guy R Cornelis	Containing CyaA gene encoding sequence
pBADB-Myc-His	Ordered from Invitrogen	Amp+; Myc-His double tags; L-arabinose inducing expression
pBADB-CyaA-tag	Constructed in this study	pBADB-Myc-His plasmid inserted with CyaA encoding sequence
pBADB-sipC-CyaA-tag	Constructed in this study	pBADB-CyaA-tag inserted with sequence encoding N-terminal 100 aa of sipC gene at the 5' side of CyaA sequence.
pBADB-slrP-CyaA-tag	Constructed in this study	pBADB-CyaA-tag inserted with sequence encoding N-terminal 100 aa of slrP gene at the 5' side of CyaA sequence.
pBADB-yiiG-CyaA-tag	Constructed in this study	pBADB-CyaA-tag inserted with sequence encoding N-terminal 100 aa of yiiG gene at the 5' side of CyaA sequence.
pBADB-yaaA-CyaA-tag	Constructed in this study	pBADB-CyaA-tag inserted with sequence encoding N-terminal 100 aa of yaaA gene at the 5' side of CyaA sequence.
pBADB-STM1791-CyaA-tag	Constructed in this study	pBADB-CyaA-tag inserted with sequence encoding N-terminal 100 aa of STM1791 gene at the 5' side of CyaA sequence.
pBADB-mdoH-CyaA-tag	Constructed in this study	pBADB-CyaA-tag inserted with sequence encoding N-terminal 100 aa of mdoH gene at the 5' side of CyaA sequence.
pBADB-STM0281-CyaA-tag	Constructed in this study	pBADB-CyaA-tag inserted with sequence encoding N-terminal 100 aa of STM0281 gene at the 5' side of CyaA sequence.

Table 6. Primers used in this study

Name	Sequence	Usage
cyaA-4	5'-GCGAATTCCAGCAATCGCATCAGG-3'	To amplify 5' 4-1215nt of
cyaA-1215	5'-GCTCTAGACTGTCATAGCCGGAAT-3'	cyaA gene
pBADB-5'	5'-ATG CCA TAG CAT TTT TAT CC-3'	To detect pBADB
pBADB-3'	5'-GAT TTA ATC TGT ATC AGG-3'	plasmid
sipC-F	5'-GAGCTCGAGATTAATTAGTAATGTGGGAA-3'	To amplify 5' 4-300nt of
sipC-R	5'-CTGGAATTCAACCTCATTCGCTTTAGT-3'	sipC gene
slrP-F	5'-GAGCTCGAGATTTAATATTACTAATATACAATCTACG-3'	To amplify 5' 4-300nt of
slrP-R	5'-CTGGAATTCACTATTTTCACTCAAAATACAG-3'	slrP gene
yiiG-F	5'-GAGCTCGAGAACCCTGGGCGCAACAGGA-3'	To amplify 5' 4-300nt of
yiiG-R	5'-CTGGAATTCCGCGTAACGCGCCAGACT-3'	yiiG gene
yaaA-F	5'-GAGCTCGAGACTGATTCTGATTTCACCTGC-3'	To amplify 5' 4-300nt of
yaaA-R	5'-CTGGAATTCAAAATCCGCGTCGTTGAACGT-3'	yaaA gene
1791-F	5'-GAGCTCGAGAAGCAACGCCTTTTTTCATCTG-3'	To amplify 5' 4-300nt of
1791-R	5'-CTGGAATTCATGCTGAATGCGTACCGAGA-3'	STM1791 gene
mdoH-F	5'-GAGCTCGAGAAATAAAACAACTGAGTATATTGACG-3'	To amplify 5' 4-300nt of
mdoH-R	5'-CTGGAATTCACGGCCAACCGGGTTGGTTC-3'	mdoH gene
0281-F	5'-GAGCTCGAGAAGCTGGAATGACCGCGTAG-3'	To amplify 5' 4-300nt of
0281-R	5'-CTGGAATTCCGCCAGACAAATCTGCTGG-3'	STM0281 gene

Human liver cancer HepG2 cells were cultured in DMEM supplemented with 10% fetal bovine serum. Cells were grown at 37℃ in a 5% CO2 humidified incubator.

3.2.5 Western blotting

SL1344 strains transfected with different constructs were cultured for 12 h in LB-0.3M NaCl medium containing 100 mg/L ampicilin. The culture was diluted 1:100 fold using fresh LB-0.3M NaCl medium, and grown for another 3 h under slow agitation to obtain an optical density of OD600 0.8~0.9 (Salmonella pathogenecity island 1 (SPI-1) inducing conditions). The fusion proteins with pBAD promoter were induced with 20% L-arabinose during the last 3 hours. Bacterial total proteins were extracted and re-suspened in sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) sample buffer for SDS-PAGE analysis. Protein expression was detected using Western blotting with anti-myc antibody (Invitrogen, Cat. No. R950-25).

3.2.6 Cya Translocation assay

The mechanism of Cya translocation assay was shown in Fig 9 (Sory and Cornelis, 1994). HepG2 cells were plated into 24-well tissue culture plates 1 day before infection. Each well contains 1ml medium, and after 24h culture, the density of adherent cells reached 2 x 10^5 cells per well. HepG2 cells were washed twice, replaced with fresh medium, and used to infect *Salmonella* for 2 hour at a multiplicity of infection (MOI) of 20 (Higashide and Zhou, 2006). After infection, the cells were washed with ice-cold Phosphate-buffered saline (PBS) for three times, and then lysed

in 100 ul of extraction solution (50 mN HCl/0.1% Triton X 100) on ice. The lysate was boiled in a water bath for 5 min, followed by neutralization with 6 ul of 0.5 M NaOH. cAMP was extracted with ethanol. After centrifugation at 11500 x g for 5 min, the supernatant containing cAMP was lyophilized and then quantified using a cAMP ELISA kit (R&D, Cat. No. KGE002B).

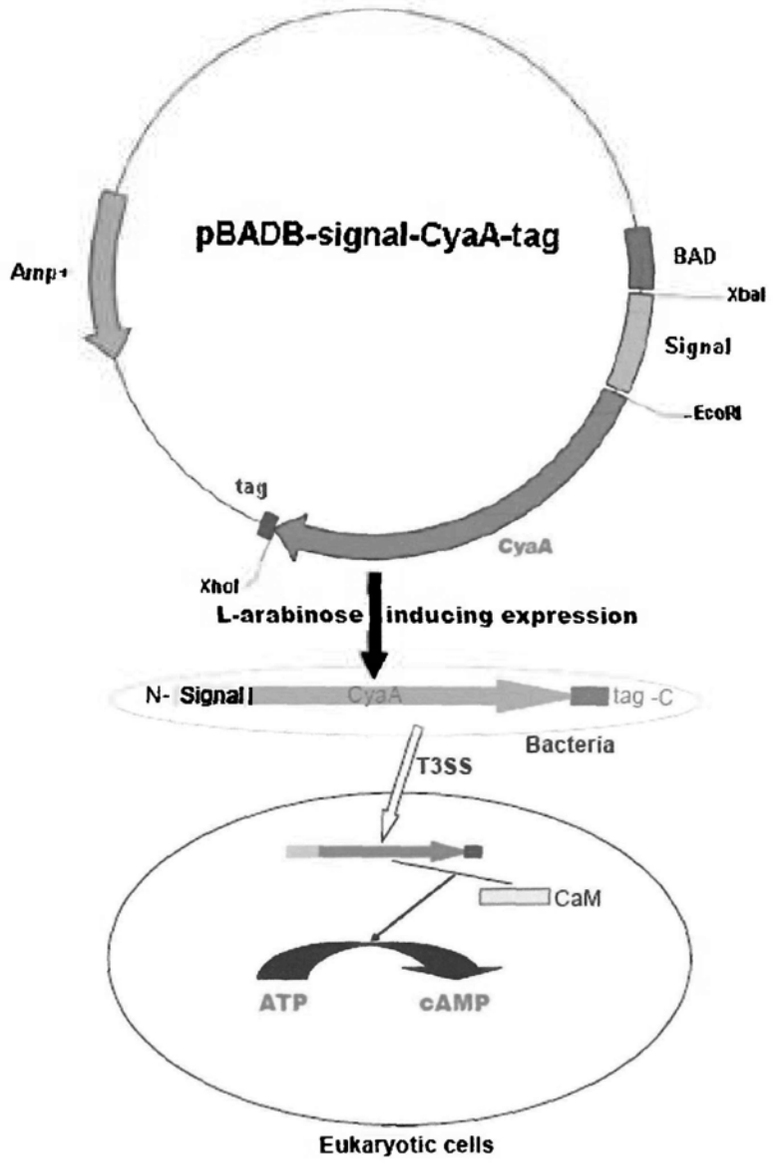


Fig 9. Principles of CyaA translocation assay.

CyaA reporter plasmids inserted with N-terminal candidate signal sequences were transformed into bacteria of functional T3SSs. Under induction of L-arabinose, the mosaic protein fused with N-terminal candidate T3S signals, CyaA polypeptides, and C-terminal Myc-His double tags will be expressed. Under T3SS induction conditions, T3SS apparatus genes will be expressed and assembled. If the signal sequence cloned in reporter plasmid is true T3S signal, it will be specifically recognized by T3SS apparatus, and consequently the fusion protein will be translocated into contacting eukaryotic cells. In cytoplasm of eukaryotic cells, with the assistance of Calmodulin (CaM) protein, CyaA protein will exert its function to catalyze the reaction by which ATP is changed to cAMP. Therefore, the cAMP level will be increased significantly.

3.3 Results

3.3.1 Distinct structural features encoded by N-terminal sequences of T3S proteins

Significantly different Aac profiles were observed between the N-terminal signal sequences of T3S and non-T3S proteins, and apparently enriched serines were found in T3S sequences (Fig 10A and 10B). Further secondary structure comparison suggested significantly enriched coils for most positions in the T3S signal sequences (Fig 10C). This pattern was especially apparent within the first 30 positions (Fig 10C). In contrast, helices were preferred at ~25 positions of non-T3S sequences (Fig 10D). Furthermore, fewer strands were present in the signal regions of T3S proteins compared to non-T3S proteins (Fig 10C and 10D). Solvent accessibility analysis showed that for T3S sequences, most positions were exposed whereas most positions were buried insides for non-T3S sequences (Fig 10E and 10F). Taken together, T3S signal sequences also contain distinctive secondary structure and water accessibility profile apart from Aac. More coils and fewer strands found in the T3S signal regions indicated that the sequences may be more flexible.

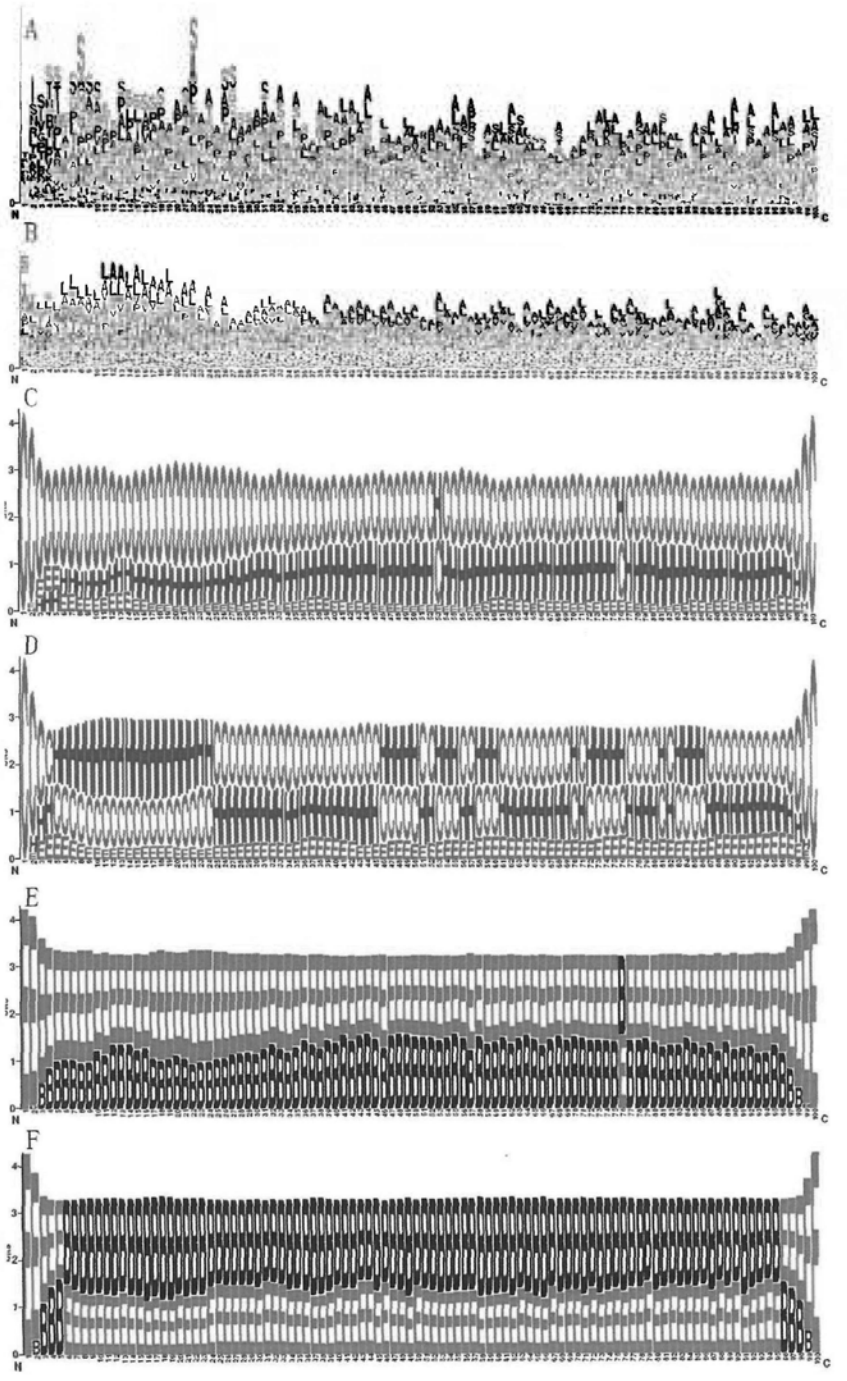


Fig 10. Distinctive N-terminal position-specific Aac, Sse and Acc feature in T3S proteins.
Element positions are depicted on the horizontal axis. The heights of characters represent the preference or enrichment level. **(A)**, **(C)** and **(E)**: Aac, Sse and Acc preference for T3S proteins, respectively. **(B)**, **(D)** and **(F)**: Aac, Sse and Acc preference for non-T3S proteins, respectively.

The primary sequence of a protein determines its secondary structure, solvent accessibility, and other second-order features. These second-order features influence the protein function through higher-level features, e.g. the tertiary structure. To examine whether T3S signal regions may form certain featured conformation, 3D structure was predicted for T3S and non-T3S signal sequences. Among the 189 T3S sequences, 41 had high-resolution (<10 angstrom) and confident (similarity > 75% between I-TASSER and MUFOLD) prediction results. Supervised (protein types being known in the first place) and unsupervised (protein types being unknown in the first place) methods were adopted to cluster the T3S and pairwise non-T3S sequences according to the multiple structure alignment results in sequence ordered or order-independent manner. However, T3S sequences couldn't be distinguished. Therefore, in consequent analysis and model training, 3D structure features were not extracted.

3.3.2 Distinct joint distribution profiles of Sse, Acc and Aac

According to the above results, I hypothesize that Sse and Acc may contribute to the specificity of Type III secretion. Previous studies suggested that individual Sse or Acc features almost made no contribution to the specific recognition of T3S proteins (Arnold et al., 2009; Wang et al., 2011). In these studies, however, the authors supposed the Sse and Acc variables were independent to Aac. Here, Sse, Acc and Aac were considered as covariables dependent on each other, and the joint distribution profiles were observed for each position of signal sequences of T3S and non-T3S proteins.

As shown in Fig 11A, T3S proteins exhibited more apparent joint element preference than non-T3S proteins, since there were significantly fewer elements present in each position of T3S N-terminal sequences. For most positions, the cumulative presence frequency of top 10 and 20 elements were both significantly higher in T3S proteins when compared with non-T3S proteins (Fig 11B). When the ratio of non-T3S and T3S sequence number was decreased to 1:1, the difference was still significant (Fig 11C and 11D), indicating the general joint element preference in T3S proteins was not caused by smaller data size. Non-T3S proteins also showed preference for certain element, especially within the first 25 positions, and yet these element types were apparently different. For example, 'SCe' ('serine-coil-exposed') was most frequently preferred in T3S proteins for most positions, followed by 'TCe' ('threonine-coil-exposed'), 'PCe' ('proline-coil-exposed'), 'NCe' ('asparagine-coil-exposed'), 'GCe' ('glycine-coil-exposed') etc. For non-T3S proteins, 'AHb' ('alanine-helix-buried'), 'LHb' ('leucine-helix-buried'), and 'VHb'

('valine-helix-buried') were more often found (http://biocomputer.cuhk.edu.hk/datasets/Supplemental Table 1).

The position-specific joint element composition features were extracted using a mathematical model adopted previously, namely Bi-profile Bayes (BPB) model, and then Support Vector Machine (SVM) was used to train the features. The parameters were optimized and shown in Table 7. The new classifier, namely T3SEpre, achieved excellent classifying performance (sensitivity of 95.9% and selectivity of 97.7%) according to 5-fold cross-validation results (Table 7). BPBAac is one of the best T3S protein classification softwares, which takes the AAC features into account only. SSE-ACC is another T3S classifier that uses SVM to train sequence-based AAC, SSE and ACC features. A comparison was performed between T3SEpre, BPBAac and SSE-ACC (Table 7 and Fig 12A). As shown, T3SEpre performed significantly better than the other two softwares in terms of sensitivity, specificity, accuracy, MCC and AUC of ROC curve.

The robustness of T3SEpre was further examined using two strategies: (1) Sub-datasets with different size were randomly selected from training data to re-train the model and to classify the remaining data; (2) Leave-One-Out strategy was adopted. Models trained by different sub-datasets performed equally well, and no apparent performance reduction was observed even when only 20% of the original training data were used (Fig 12B). In Leave-One-Out assessment, for different genera or subgroups, almost all the effectors could be recalled and consistently high AUC values were obtained (Fig 12C). The worst performance was achieved using the model trained with

solely plant pathogen proteins to classify animal pathogen proteins, with a recall value of ~81%, which in fact was still quite high (Fig 12C). Therefore, the performance of T3SEpre was stable and robust, and it can be widely applied for T3S protein prediction in different bacteria.

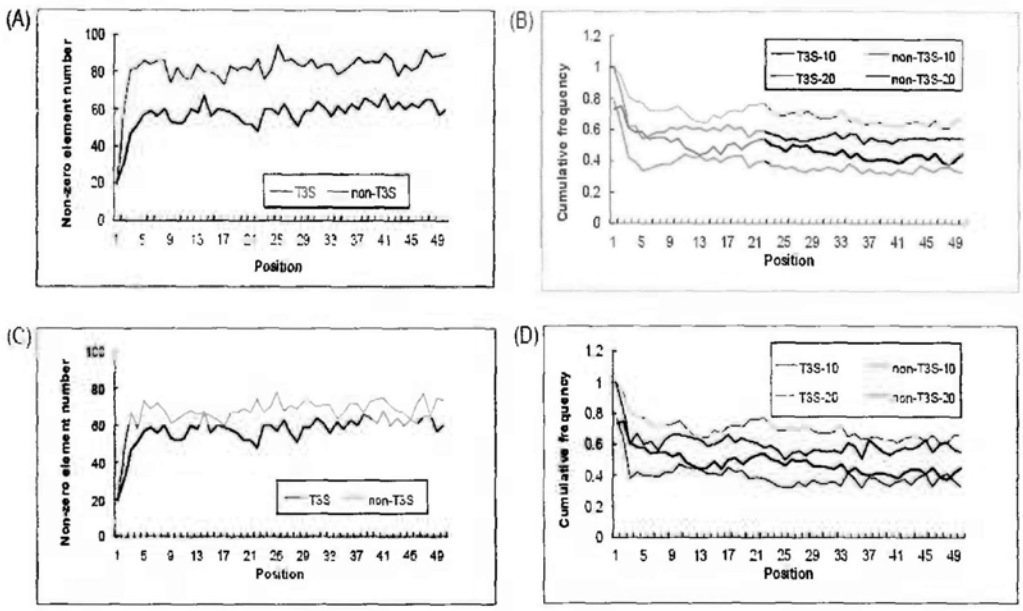


Fig 11. Comparison of preference profile for Aac-Sse-Acc joint features between T3S and non-T3S sequences.

(A) and **(C)**: Total number of non-zero distributed joint features for each position in T3S or non-T3S sequences. Full set of joint features include 120 different elements. The ratio of data size between T3S and non-T3S proteins was about 1:2 in **(A)** and 1:1 in **(C)**. **(B)** and **(D)**: Cumulative frequency of the most enriched 10 (T3S-10 or non-T3S-10) or 20 (T3S-20 or non-T3S-20) joint features in T3S or non-T3S sequences. The ratio of data size between T3S and non-T3S proteins was about 1:2 in **(B)** and 1:1 in **(D)**. Only N-terminal 50 positions of T3S and non-T3S sequences were included for analysis.

Table 7. Optimal parameters and corresponding performance of T3SEpre, BPBAac and SSE-ACC based on five-fold cross-validation

Name	Model[a]	Length[b]	Kernel[c]	C[d]|γ[e]
T3SEpre	BPB	100	RBF	4|0.001
BPBAac	BPB	100	RBF	8|0.001
SSE-ACC	SPB	100	RBF	4|0.008

Name	*Sn* (%) versus *Sp* (%)	*A* (%)	AUC (%)	MCC
T3SEpre	95.9% versus 97.7%	97.09	99.50	0.9347
BPBAac	84.4% versus 94.8%	91.33	96.43	0.8031
SSE-ACC	78.0% versus 95.2%	89.47	94.49	0.7589

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.
γ: gamma, which was optimized based on 10-fold cross-validation grid search.

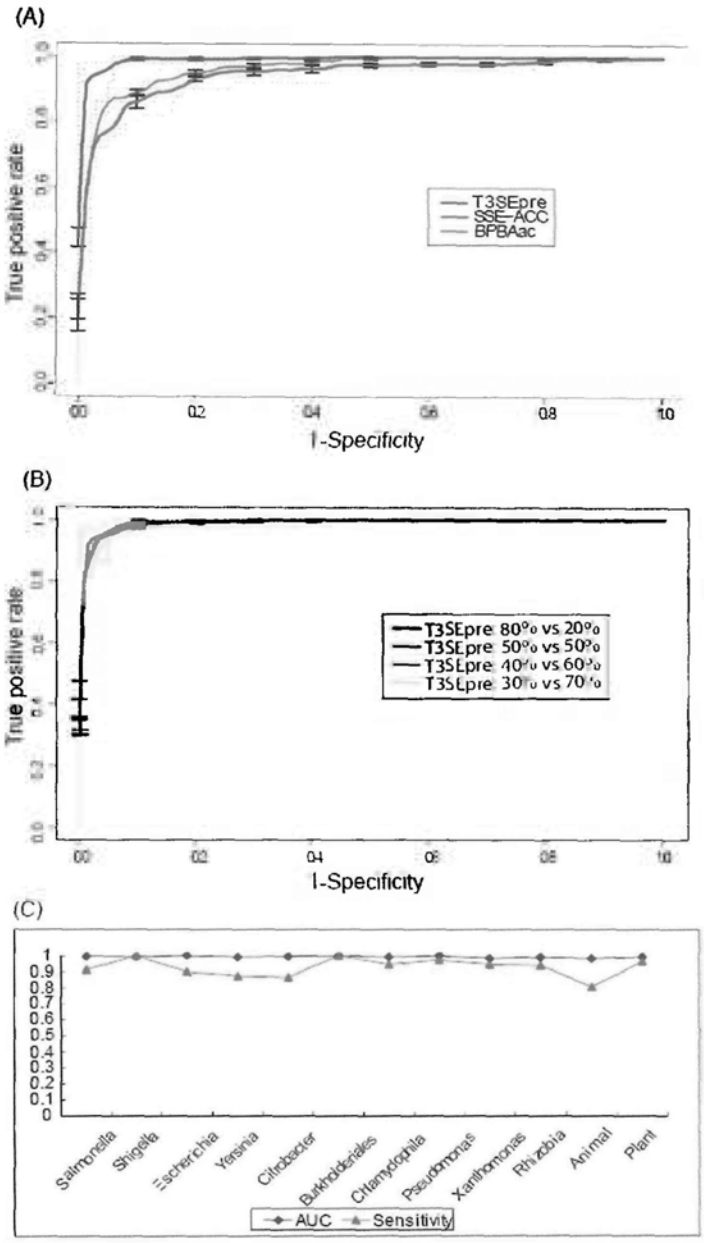


Fig 12. Performance of T3SEpre. (A) ROC curves of different T3S protein prediction softwares based on 5-fold cross validation. All three softwares were retrained with the same datasets used in this research. The parameters were optimized respectively (refer to Table 7). The methods for BPBAac and SSE-ACC were respectively referred to the original descriptions in references Wang et al., 2011 and Yang et al., 2010. **(B)** ROC curves for T3SEpre models with different training-testing data size ratios. 'Xx% vs. Yy%' meant 'the percentage of training data versus. that of testing data'. **(C)** Inter-species/group robustness of T3SEpre to predict T3S proteins. Leave-One-Out strategy was adopted with the exception that, here 'One' meant the data from 'one species/group'. 'Animal' and 'Plant' meant 'animal pathogens/symbioints' and 'animal pathogens/symbioints', respectively.

3.3.3 Identification of new T3S proteins and possible effectors

A list of *Salmonella* T3S proteins were predicted using T3SEpre (Table 8). Most known T3S effectors were correctly predicted, while some new candidates were also identified (Table 8; known SPI-1 effectors were highlighted in red while known SPI-2 effectors were highlighted in blue). A large percentage of the predicted T3S proteins were annotated as 'hypothetical proteins' or with 'unknown function' (Table 8; shown in italic). To further evaluate the prediction power of T3SEpre, some new candidates were selected for experimental validation (Table 8; shown in bold). Western blotting demonstrated that all the fusion proteins were expressed (Fig 13A; protein STM0281-CyaA-tag was also expressed, which was not shown in this figure), and Cya translocation assay confirmed 4/6 newly identified proteins were true T3S proteins (Fig 13B-C). MdoH and YaaA constructs were not found to translocate CyaA protein into host cytoplasm (Fig 13B-C).

Table 8. *Salmonella* T3S proteins predicted with T3SEpre (a strict cutoff, score >= 0.5, was used)

SeqID	Annotation	SVM-Value
Seq2779	sipC	2.06; positive control
Seq1359	sseG	1.83
Seq765	**slrP**	**1.74; validated**
Seq1354	sseC	1.68
Seq1358	sseF	1.37
Seq3891	**yiiG**	**1.18; validated**
Seq2778	sipD	1.13
Seq1989	sopA	1.02
Seq1355	sseD	1.00
Seq1794	sopE2	1.00
Seq1055	sopB	1.00
Seq1352	sseB	1.00
Seq1356	sseE	0.99
Seq1347	ssaB	0.98
Seq1274	katE	0.98
Seq2774	sptP	0.90
Seq3446	ftsY	0.89
Seq4253	*STM4421*	*0.86*
Seq1681	tonB	0.85
Seq272	***STM0281***	***0.82; validated***
Seq4195	hflK	0.75
Seq1973	pduO	0.72
Seq1312	ydiF	0.71
Seq1809	*STM1870*	*0.70*
Seq1730	***STM1791***	***0.70; validated***
Seq2839	sopD	0.69
Seq1931	*STM2005*	*0.68*
Seq2813	ygbI	0.65
Seq1111	**mdoH**	**0.62; not validated**
Seq241	rcsF	0.62
Seq2764	orgC	0.61
Seq5	***yaaA***	***0.59; not validated***
Seq2400	*STM2486*	*0.57*
Seq400	nrdR	0.55

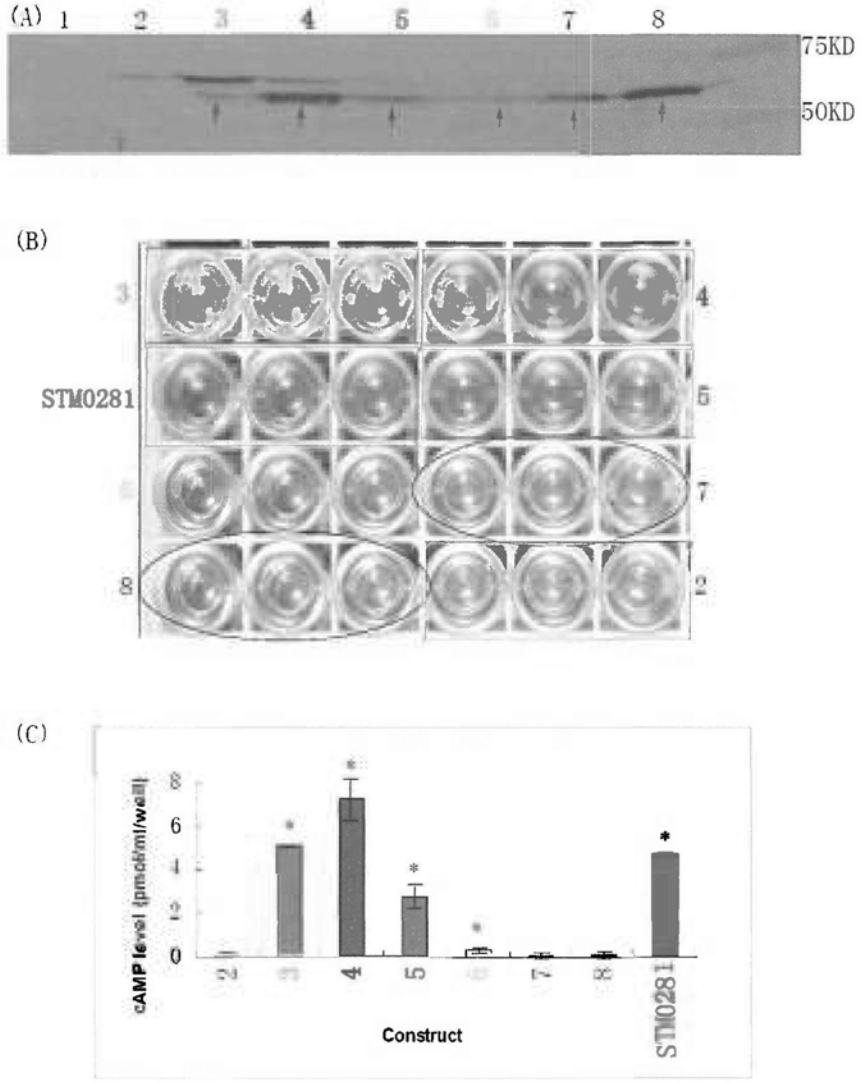


Fig 13. Expression and translocation of predicted T3S proteins.

(A) Western blotting results of candidate fusion proteins in Salmonella. The primary antibody against Myc tag was used. Lane 1 and 2 represent pBADB and pBADB-CyaA plasmid, respectively. Lane 3-8 represent pBADB-CyaA derived constructs inserted with candidate T3S genes sipC, slrP, yiiG, STM1791, yaaA and mdoH, respectively. The expression of STM0281 construct was also detected but in a separated experiment, for which the result was not shown in this picture. Corresponding protein band was indicated by blue (Signal-CyaA-tag) or red arrow (CyaA-tag). **(B)** ELISA coloring results for Cya translocation assays. 2-8 respectively represent different constructs indicated in **(A)**, while STM0281 represent the corresponding construct inserted with STM0281 signal sequence. 3 biological repeats were included for each construct. **(C)** Quantitative and statistic analysis. Stars indicate statistical significance. PBADB-CyaA construct was used as negative control and the construct with sipC was used as positive control.

3.3.4 Wide distribution of T3S proteins in different species

Whole-genome T3S protein prediction was performed to a variety of bacteria and yeast. Some new T3S proteins were also identified from species previously reported to have no T3SSs, such as *Helicobacter* and *Mycobacterium* (Table 9 and website: http://biocomputer.bio.cuhk.edu.hk/T3SEpre_candidates). T3SSs have so far been only found in gram-negative bacteria, and yet a group of T3S candidates were confidently predicted with high scores from gram-positive bacteria and even in yeast (Table 9 and website: http://biocomputer.bio.cuhk.edu.hk/T3SEpre_candidates). Further experimental validation of these predicted proteins are needed before any conclusive remarks can be made.

Table 9. Potential T3S proteins in representative species predicted with T3SEpre (5 with highest prediction scores were given for each species; for total sets of prediction and detailed sequence annotation, please refer to http://biocomputer.bio.cuhk.edu.hk/T3SEpre_candidates.)

Species	SeqID	SVM-Value
Agrobacterium (NC_003062)	Seq225	1 8262186013
(Gram -, no reported T3SS)	Seq153	1 8182696817
	Seq1738	1 3438005288
	Seq556	1 3053632004
	Seq1265	1 1505788212
Helicobacter (NC_000915)	Seq1414	1 2728691623
(Gram -, no T3SS)	Seq1463	0 81876863031
	Seq595	0 79814466522
	Seq559	0 76127749423
	Seq1438	0 74734863579
Mycobacterium (NC_002755)	Seq292	2 5282200382
(Gram -, no T3SS)	Seq175	2 1134425698
	Seq4137	2 091326393
	Seq2598	1 8061094469
	Seq3425	1 7706735128
Staphylococcus (NC_013450)	Seq357	1 1964289751
(Gram +, no T3SS)	Seq1352	1 1653620221
	Seq946	0 99871685539
	Seq920	0 82115720581
	Seq1676	0 70904028709
Streptococcus (NC_011900)	Seq1084	0 86992793954
(Gram +, no T3SS)	Seq166	0 86910627471
	Seq662	0 73099842015
	Seq1625	0 63822324187
	Seq908	0 60030423151
Yeast (S288c, no T3SS)	Seq1417	2 9246207038
	Seq5249	2 9232109626
	Seq840	2 7428577459
	Seq2394	2 6276802073
	Seq3386	2 3665151381

3.4 Discussion

3.4.1 Structural features for T3S protein recognition

Several lines of evidence suggested that the N-terminal sequences contain signals guiding the specific recognition and secretion of T3S proteins (Karavolos et al., 2005; Lloyd et al., 2001 and 2002; Russmann et al., 2002; Schechter et al., 2004; Wang et al., 2008). The molecular basis of this specificity, however, remains to be determined. Several groups attempted to find sequence-based specific T3S signal features. However, it is difficult to find common domains or motifs within a certain bacterial genus or closely related genera due to the high diversity of sequences (Buchko et al., 2010; Petnicki-Ocwieja et al., 2006). Recently, both sequence-based and position-based amino acid enrichment and depletion were noticed in the N-terminal region of T3S proteins (Arnold et al., 2009; Wang et al., 2011). Computational models based on these primary sequences derived features can well classify the T3S and non-T3S proteins, demonstrating amino acid sequences of T3S proteins at least encode part of the T3S specificity. Furthermore, some second-order elements including Sse and Acc were analyzed to search for more direct and specific features (Arnold et al., 2009; Wang et al., 2011). Although differences were found between the T3S and non-T3S proteins, these features were not considered responsible for the specificity because they independently did not improve the performance of T3S protein classifier. Most recently, one group attempted to analyze the combinatorial features of Aac and Sse/Acc (Yang et al., 2010), but the model did not outperform other models which only based on Aac features, therefore not leading to any valuable conclusion. In this

research, the Aac, Sse and Acc were considered as inter-dependent co-variables, and I analyzed the position-specific joint distribution profiles of three features, and found that integrating these new combinatorial features could significantly improve the prediction performance. I therefore reasoned that Sse and Acc represent some T3S-specific features which are partly encoded by Aac.

For tertiary structure, although the structure homology was proposed for T3S effectors lacking sequence similarity, it was merely suitable for the effector region but not for the N-terminal signal region. Typical T3S effector structure studies frequently neglected the N-terminal region becuase of its flexibility and it is difficult to resolve the structure (Galán and Wolf-Watz, 2006; Stebbins and Galán 2001; Singer et al., 2004; Birtalan et al., 2002; Lee et al., 2004; Wulf et al., 2004). The 3D structure information of the T3S signal region is therefore not available. In this research, I computationally modeled and compared the 3D structure of T3S signals. However, the overall prediction resolution for T3S sequences were much lower than those for the non-T3S sequences, which indirectly indicate the specificity of T3S signals and the lack of resolved structure for similar peptide fragments. Furthermore, difficulty in clustering the T3S signal sequences based on 3D structure could partly reflect the flexibility of T3S signals (Galán and Wolf-Watz, 2006), which is consistent with the observation that coils were apparently enriched in T3S signal sequences (Fig 10). Taken together, it seems difficult to explain the specificity of Type 3 secretion based on the 3D structure of T3S signals. The 3D structure features were therefore not included for the classification. The model, T3SEpre, based on the combinatorial

features of Aac, Sse and Acc, serve as an excellent classifier in distinguishing T3S and non-T3S proteins. Future advancement in the 3D structure analysis may finally help explain the molecular mechanism of T3S signal recognition and specific secretion.

3.4.2 The formation and evolution of T3S signal sequences

Apart from the specific features embedded in the T3S signal sequences, how these sequences are formed and evolved also remains an enigma. In many bacteria speccies, some T3S effectors were obtained together with T3SS apparatus through a horizontal transfer event (Heuck 1998). For these effectors, the signal sequences seemed to co-evolve with T3SS apparatus genes. However, more effectors were found to be scattered in the bacterial genomes. In model species such as *Salmonella*, different effectors function coordinately in the host-bacteria interactions (Kubori and Galán 2003). It is interesting to investigate how these scattered effectors can be fine-tuned for gene expression, exert necessary function and meanwhile contain the exactly T3SS-recognized signals.

Inspired by the 'terminal re-assortment' hypothesis proposed by Stavrinides et al., I partitioned a full-length T3S protein into 2 parts: the N-terminal signal part and the C-terminal function part. Gene expression data under different conditions were analyzed to observe whether T3S proteins were co-expressed with known T3SS apparatus or other related genes. I found that T3S signals are widely located within different bacterial genes (e.g. yiiG, STM1791 and STM0281 genes validated in this research, which were not co-regulated with SPI-1 or SPI-2 apparatus genes). Some

gram-positive bacteria and even yeasts also encode T3S signal-containing genes (Table 9). Therefore, T3S signal could be formed randomly and evolve independently with T3SS apparatus. A protein with putative T3S signals is not necessarily an effector, because T3S effector must contain a functional domain and must be co-regulated with T3SS apparatus as well as other relevant genes for expression. For this reason, the candidate T3S proteins predicted from gram-positive bacteria or yeasts should be called T3S proteins (or substrates) rather than effectors.

3.4.3 Application of T3SEpre

Finally, T3SEpre, a T3S protein prediction software was developed. The source codes and R package can be found from the following website: http://biocomputer.bio.cuhk.edu.hk/softwares/T3SEpre. Although this model outperformed other softwares as evidenced by both in silico analysis and wet-lab experiments, different models may have their own advantages and disadvantages. Users are advised to choose the most suitable software for individual needs.

CHAPTER 4

T3DB: an Integrated Database for Bacterial Type III Secretion System

4.1 Introduction

Although T3SSs have been widely identified in many bacteria species, quite limited number of T3SSs have been extensively investigated. Till now, T3SSs from only 5 animal pathogens (*Salmonella*, *Shigella*, *Yersinia*, pathogenic *E. coli* and *Citrobacter*) and 2 plant pathogens (*Pseudomonas* and *Xanthomonas*) have been well studied (Lara-Tejero et al., 2011; Tree et al., 2009; Ogawa et al., 2008; Shao et al., 2008; Mundy et al., 2004; Hauseret et al., 2009; Kay et al., 2009). The relatively low sequence similarity, and frequent horizontal transfer among bacteria makes it difficult to identify T3SS orthologs (Pallen et al., 2005; Desvaux et al., 2006). Most importantly, different nomenclature and categorization method/terms have created confusion and difficulty in literature search as well as in data interpretation using relevant information explored in the parallel genera (Pallen et al., 2005; Desvaux et al., 2006). A unified platform integrating various source of information is needed to facilitate the T3SS related research.

Such an integrated platform has been initiated previously by Pallen MJ et.al., and a database (http://3base.bham.ac.uk) has been created, aiming to formulate a taxonomy for type-III secretion, and to help identify new T3SSs in newly-sequenced bacterial genomes (Pallen et al., 2005). However, till recently, the database only contains annotated T3SSs from 4 species. In another similar database DTTSS (http://sdbi.sdut.edu.cn/ttss), only very limited number of T3SSs have been annotated, and a great number of non-T3SS genes have been mis-annotated as T3SS genes. Other databases, such as T3SEdb (http://effectors.bic.nus.edu.sg/T3SEdb) and Effective

(http://effectors.org) mainly store subsets of validated and predicted T3SS genes (mainly effectors) respectively (Tay et al., 2010; Jehl et al., 2011). A systematic function annotation of individual T3SS effector coding genes, the relationship between these effectors and their chaperones, and the structural components and regulators of corresponding T3SSs may provide useful multiple-aspects datasets for further studies on gene evolution, chaperone-effector interaction, T3SS-related regulatory network, and bacteria-host interaction, etc. Regretfully, no such information can be found in any of the current databases. In addition, none of the current platform implements various computational software tools for Type 3 effector prediction.

In this part, I developed a Type 3 secretion system-related Database (T3DB), aiming to annotate all the T3SS related genes and to set up an integrated platform for the T3SS community. In this database, T3SS genes, which include apparatus (including accessories), effectors (proteins secreted through T3SS conduit), chaperones, and transcription regulators, are collected and classified into one of the five phylogenetic types (He et al., 2004). For each gene, its genome coordinate, gene accession, gene alias, nucleic acid sequence, protein sequence, and detailed function were annotated. The T3 orthologs were also annotated. All these data were manually annotated, and experimental evidence was provided for each effector, chaperone or transcription regulator. The database can be freely browsed, searched or downloaded. Besides, web servers were included for BPBAac and T3SEpre, the two T3SS effector prediction tools described in chapter 2 and 3 of this thesis. The links to other T3SS effector prediction servers were also provided.

4.2 Database construction and implementation

The framework for T3DB construction involves 4 steps: (1) identification of T3SS containing bacterial genera and species; (2) T3SS gene identification and categorization; (3) T3SS gene annotation; (4) ortholog annotation.

First, a text based literature search strategy was adopted to obtain a comprehensive list of T3SS related publications. 'Type III Secretion System', 'Type 3 Secretion System', 'TTSS' and 'T3SS' were respectively used as key word to search the Pubmed database. This search resulted in more than 3000 non-redundant publications. The abstract of each publication was scanned manually, and the bacteria genera and species were recorded and examined. Some bacteria may contain not yet reported candidate T3SSs. Instead of using comprehensive sequence alignments to find these candidates, we only included potential T3SS candidates based on literature reviews in which the authors presented sequence alignments, genome localization, and phylogenetic evidence. This procedure generated a list of 26 bacteria genera from different classes, even from different phyla (Fig 14). The phylogenetic relationship between these bacterial genera was annotated from Bergey's Manual (Garrity et al., 2005) (Fig 14). For each genus, the model species and strains with the most adequate experimental data and molecular information were further selected. The genomes (chromosomes and plasmids) of most of the selected model strains have been sequenced, and the current release contains 75 model species. The host type (animal or plant) and interaction type (pathogenesis or symbiosis) were annotated for each species according to Bergey's Manual (Garrity et al., 2005).

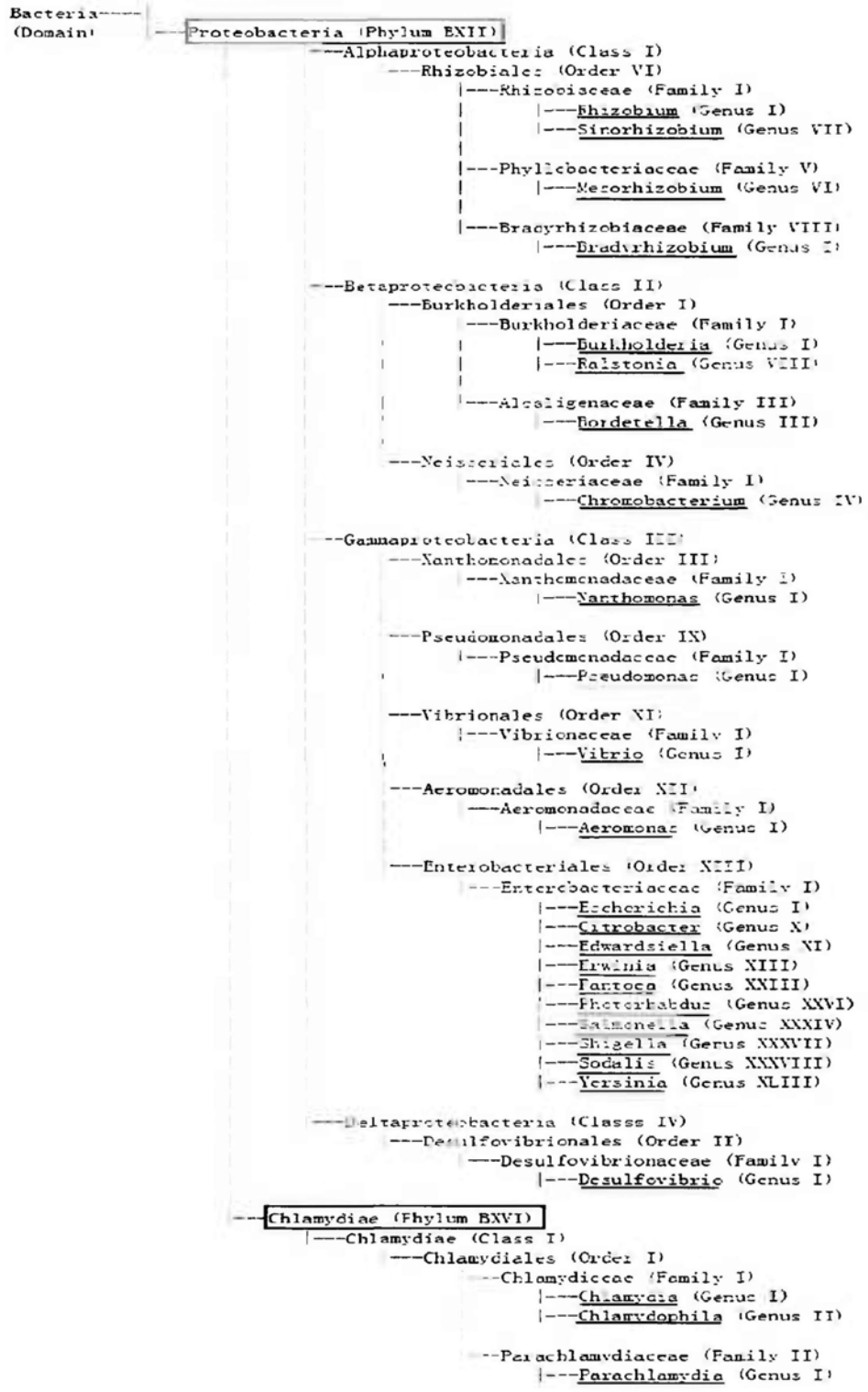


Fig 14.List and phylogenetic relationship of bacterial genera containing functional T3SSs.
The phylogenetic relationship was referred to Bergey's Manual (Garrity et al., 2005). Each genus name was underscored while its phylum name was indicated in rectagle.

I then collected all the T3SS-related genes for each selected model strain. Because it is much easier to identify orthologs in bacteria that belong to the same genus, each genus was therefore used as key words in combination with either 'Type III Secretion System', 'Type 3 Secretion System', 'TTSS' , or 'T3SS' respectively to search relevant literature in the Pubmed database. Each literature hit was manually curated and genes related to T3SS were collected, together with their bacterial host strain, alias, gene accession, and detailed function. Furthermore, the candidate gene sequences and their genomic coordinates were tracked and compared, and T3SS orthologs in different species or strains were identified. Because a strain may contain more than one T3SS, and genes with similar sequence, structural, function and genomic clustering features among different T3SSs in the same strain can not be accurately defined as paralogs or orthologs, we used a new term 'T3 ortholog' to specify this case. Specifically, any genes with the above-stated similar features among different T3SSs in the same or different strains were collectively termed 'T3 orthologs'. A non-redundant T3SS gene set was obtained for each genus after filtering out the redundant orthologs. Each gene in the non-redundant T3SS gene set was assigned a unique name, in the form of 'XXX-YYY', where 'YYY' is the traditional gene name for that gene in most studied strains and 'XXX' describes the genus. The genus name was included in the gene name so that users can easily distinguish the genus from which the candidate gene originates. Even in the same genus, the orthologs in different strains may have different names. After a unique nomenclature was set for each ortholog cluster in a genus, the other names representing the same gene were considered as alias. For strains with more than

one T3SS, the nomenclature for genes was in the form of 'XXX-ZZZ-YYY', where 'XXX' and 'YYY' represent genus and gene name respectively, and 'ZZZ' describes the T3SS name. Each T3SS were classified into one of the five putative categories (He et al., 2004) according to phylogenetic analysis of the conserved T3SS proteins among all bacteria that contain T3SSs.

The T3SS genes annotated in T3DB are divided into 4 major categories: Apparatus (Category I), Chaperone (Category II), Effector (Category III), and Transcription Regulator (Category IV). Apparatus genes encode those that assemble the needle-like structure as well as accessory genes. Genes in this category are further sub-classified into different function clusters (Fig 15). Chaperone genes encode proteins serve as chaperones in assisting effector proteins to secrete through T3SS conduit. Effectors genes encode proteins specifically secreted through T3SS conduit. Some effectors themselves also function as structure proteins, such as those translocon proteins (e.g., Sal-SPI1-SipB and SipC). In such cases, they were classified into 'Translocon' in Category I. T3SS transcription regulators were collected as an independent category. For categories, chaperone, effector and transcription regulator, at least one reference with experimental evidence was required to support the function annotation. For category apparatus, sequence similarity and genomic organization were used as evidence, for which two conditions must be both met. For bacteria that contain multiple-T3SSs, some effectors cannot be precisely classified to a specified T3SS; in such case, the name of the orthologous gene cluster adopts 'XXX-YYY' instead of 'XXX-ZZZ-YYY' system.

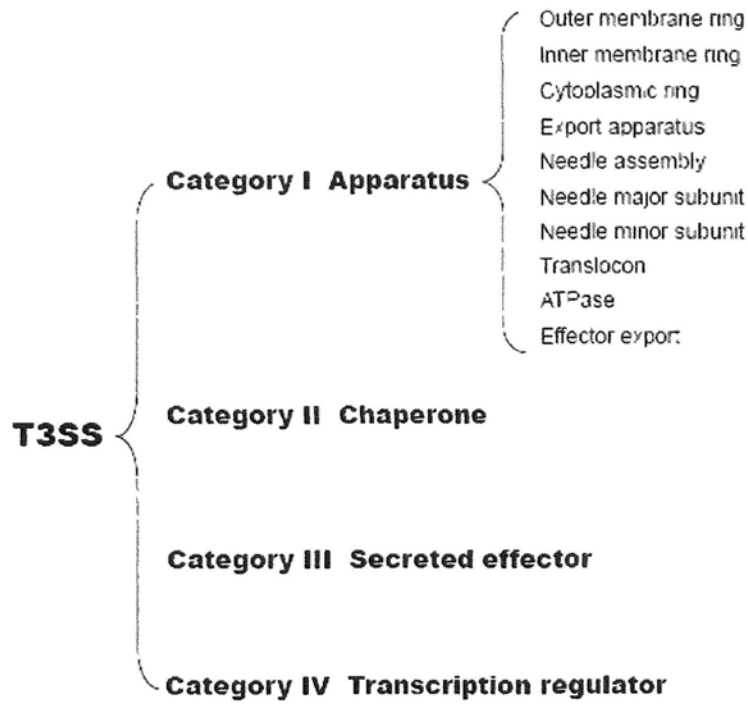


Fig 15. Categories and sub-categories (or function clusters) of T3SS genes.

For gene annotation, orthologous genes in different strains within the same genus adopt the same gene names. To distinguish these genes from different strains, a unique ID was assigned to each gene. The ID is represented by T3X, where 'X' is one of the four characters ('A': Apparatus; 'C': Chaperone; 'E': Effector; 'R': Regulator), followed by 11 numerical numbers representing different phyla (1), classes (1), orders (1), families (1), genera (2), strains (2), T3SSs in the same strain (1), and the individual genes (2), respectively. It should be noted that when more than T3SSs are presented in a single strain and it is not able to determine which T3SS the gene belong to, the corresponding number is replaced by a character 'x'. For each gene, the genome type (chromosome or plasmid), genome ID and gene coordinates in genome (if available), strand direction, nucleic acid and protein sequences, major function category, detailed function annotation, structure information, within-genera/strain and inter-genera 'T3 ortholog' relationship and reference PMIDs were all annotated.

In the last step, a 'T3 ortholog' cluster (see previous description) was created for each gene. For T3SS proteins, the sequence similarity among orthologs in different genera, especially distantly-related genera, was very low. Specific alignment strategies with less stringent thresholds were adopted. Besides, the within-genome synteny information was considered. In the current release of T3DB, the T3 orthologs were annotated based on literature using different comparison methods. Some orthologs with extremely low sequence similarity were annotated as 'T3 orthologs' if they share high similarity in structure (Structure Orthologs) or in function (Function Orthologs) based on publication.

The database was created and maintained using MySQL. The interactive interfaces were written in PHP scripts. Two integrated web servers respectively for BPBAac (Wang et al., 2011) and T3SEpre were implemented using PHP and Javascript. The local packages for BPBAac and T3SEpre were written and implemented with R and Perl.

4.3 Database Usage

The T3DB user interfaces include: (1) browse interface, (2) search interface, (3) download interface and (4) software interface.

In browse interface, bacterial genera with T3SSs and their phylogenetic relationship between genera was shown in a tree (Fig 16A). Users can select and click any interested genus to browse the genus page (Fig 16B). In each genus page, one can skim the basic information about the bacteria genus including host type, interaction type, T3SS number, T3SS names, locations, classes, and the representative strains and their host species in 'Basic information' field. Besides, two alternative browse routes were provided for users to further browse T3SS genes in the genus: by strain and by genus-conserved ortholog cluster. In the "browse by strain" mode, clicking the interested bacteria strain will bring out a new interface showing the full list of T3SS genes in this specified strain (Fig 16C). Users can learn about the gene distribution and corresponding molecular information in the selected strain. In the "browse by ortholog cluster" mode, a general introduction of the number and the function of non-redundant T3SS genes in the genus are provided. Each gene name icon represents a unique genus-conserved ortholog cluster. After selecting an interested gene cluster, a list of T3DB records of all the annotated orthologs in different strains within this genus will appear (Fig 16D). The alternative browsing modes are designed to cater for different research interest: either on bacteria or on specific genes. To learn more details about the individual gene in a particular strain, one should click the interested accession record, which leads to the final annotation page for this gene (Fig 16E). General

information includes gene accession, alias, genomic coordinators, sequence and major function category. The 3D-structure accession (if any) was also linked. The gene was annotated with detailed function, and the references (PMIDs) were included for tracking purpose. Finally, the known intra- or inter-genus T3 orthologs were annotated and within-database links were provided for efficient accession to these orthologs (Fig 16E).

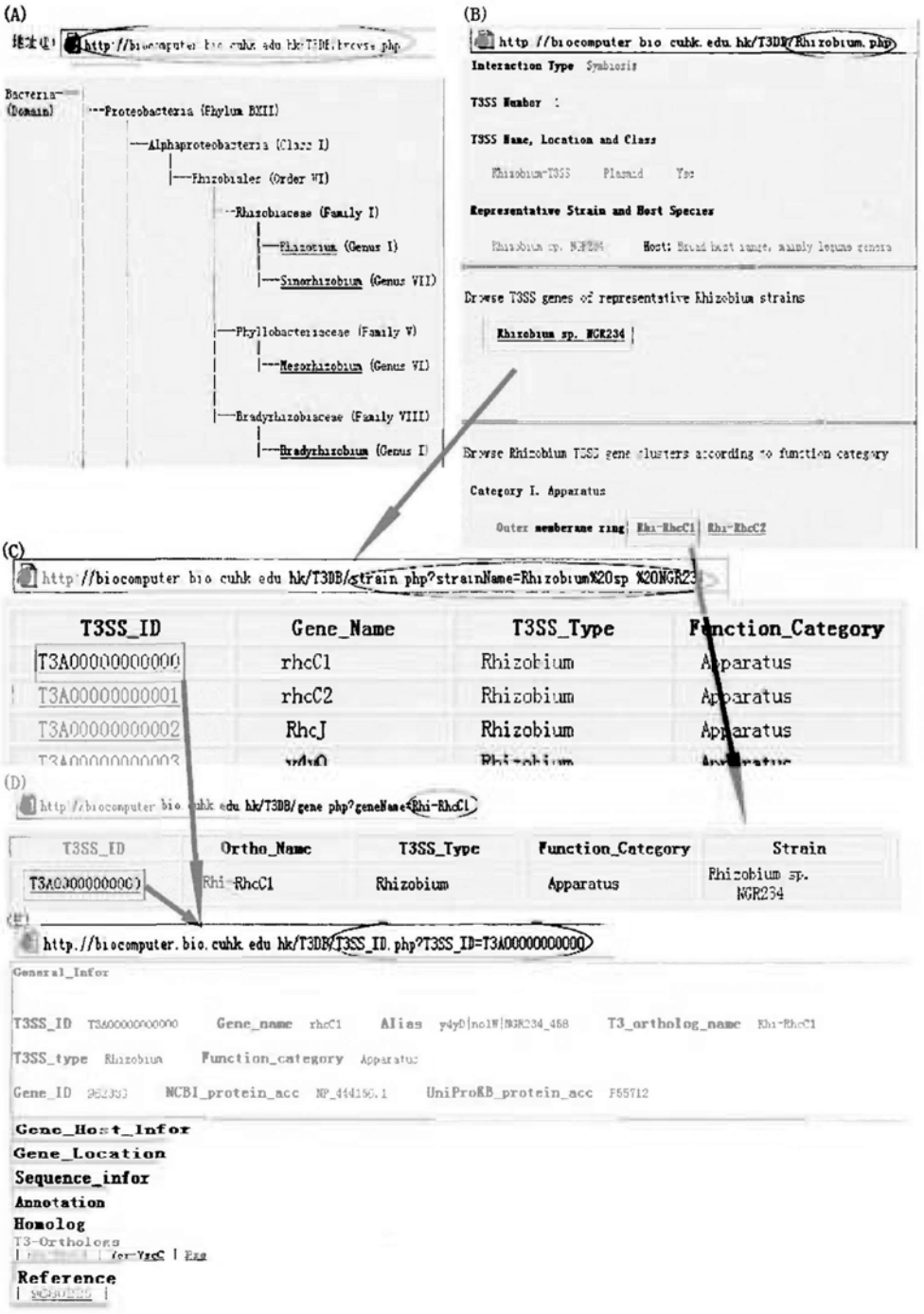


Fig 16. Browse interfaces of T3DB. **(A)** Overall browse page. **(B)** Browse page for individual bacterial genus (eg. Rhizobium). **(C)** or **(D)** page will be shown up when the corrsponding icon indicated within red rectangle in **(B)** is clicked, respectively. Arrows indicate the relationship between pomped out pages and corresponding icons. **(E)** shows the major items annotated in final annotation page for each individual gene. Please Refer to http://biocomputer.bio.cuhk.edu.hk/T3DB/browse.php for details.

As shown in Fig 17A-B, T3DB provides multiple search modes. Users can search an interested T3SS gene through its gene accession, T3ID or gene name. A Blast program was integrated for the users to input interested nucleic acid or protein sequence for similarity search. Users may also download gene list and sequences by bacterial genus (or strain), by T3 ortholog cluster, or by function category (Fig 18A-D).

Because T3SS effectors play important roles in host-bacteria interactions, it is of great significance to identify new genes encoding T3SS effectors. Efficient in silico prediction tools with high sensitivity and high specificity have been developed recently. Two T3SS prediction softwares, BPBAac and T3SEpre, were integrated into the database (Fig 19A-B). For BPBAac, users may input a sequence or upload a FASTA file to make prediction. For T3SEpre, users need to input a sequence, its secondary structure, and its solvent accessibility. For both softwares, the specificity and sensitivity of the prediction can be freely defined by users. Besides, links to other T3SS effector prediction softwares are also provided.

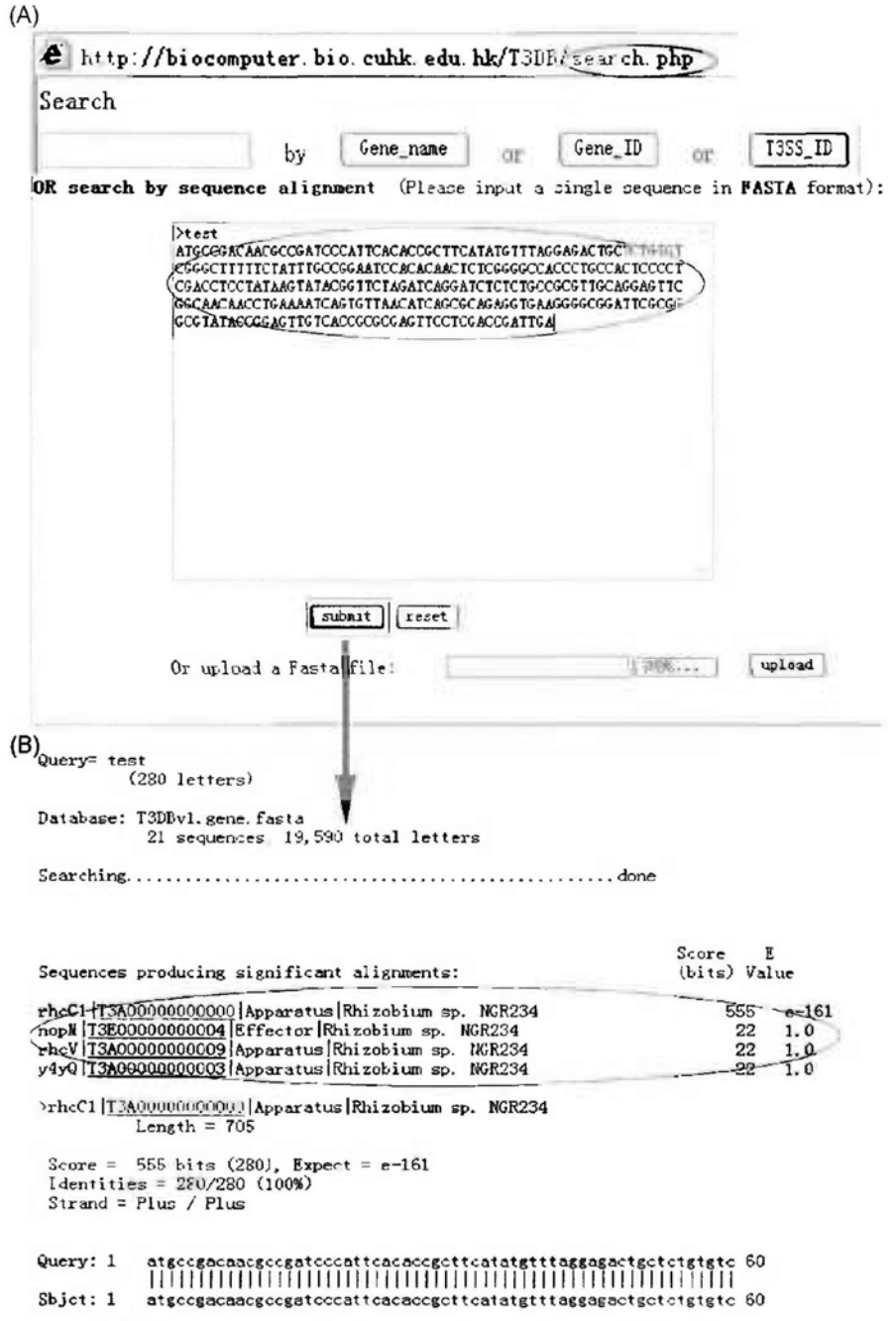


Fig 17. Search interfaces of T3DB.

(A) Search page. Gene_name (including alias), Gene_ID (including protein access) as well as T3SS_ID could be used for searched. A blast program was also integrated for aligning a unknown sequence against stored T3SS genes. Circled sequence in **(A)** is an example; when clicking icon 'submit', a new page with alignment result will be pompted out, e.g., **(B)**. Please Refer to http://biocomputer.bio.cuhk.edu.hk/T3DB/search.php for details.

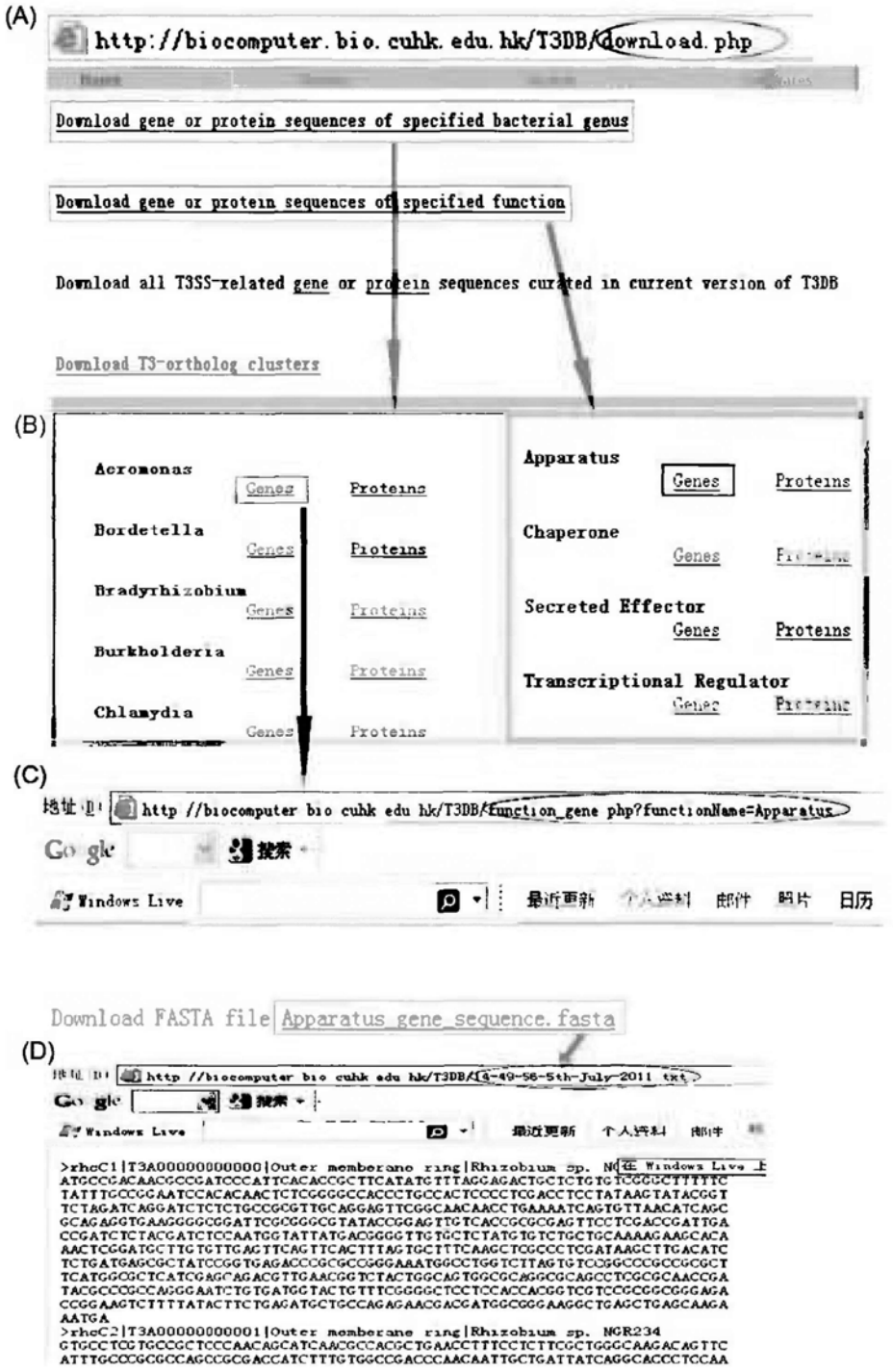


Fig 18. Download interfaces of T3DB.

(A) and **(B):** Alternative download modes. **(C)** and **(D)**: example of download manipulatation. Please Refer to http://biocomputer.bio.cuhk.edu.hk/T3DB/download.php for details.

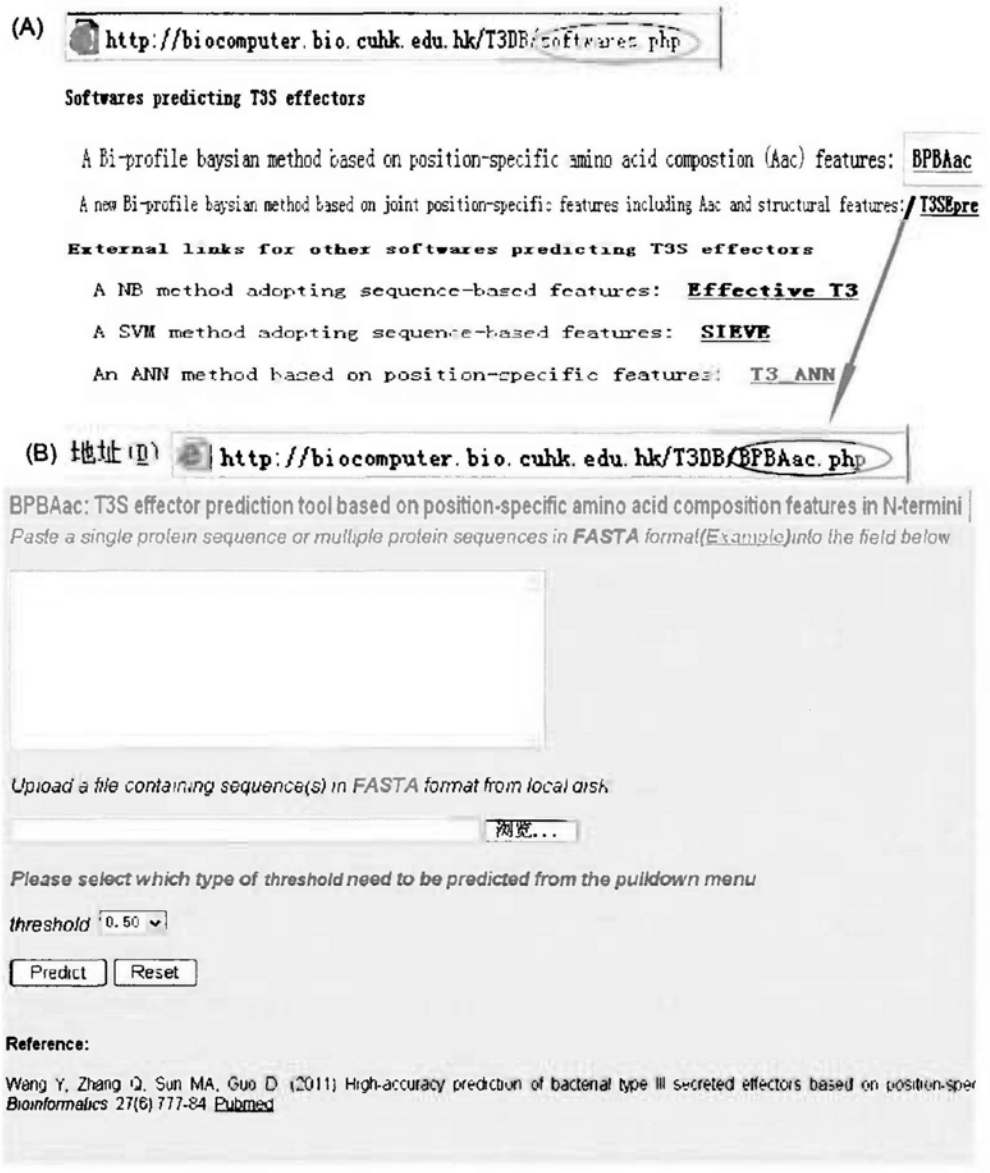


Fig 19. Software interfaces of T3DB.

(A) Major interface of softwares (BPBAac, T3SEpre and external softwares can be linked). **(B)** Interface of BPBAac software. Please Refer to http://biocomputer.bio.cuhk.edu.hk/T3DB/softwares.php for details.

4.4 Discussion

T3SS has received continuous high research attention due to the important roles it plays in bacterial pathogenesis and symbiosis. An integrated platform for data storage, data analysis, and knowledge inter-change may greatly facilitate the T3SS related study in the research community. In the current release of T3DB, one can find that the function of Rhizobium NopA, NopB and NopX proteins have been well studied and annotated (as translocon). These Nop proteins, however, have not yet been identified and studied in other genera, e.g. Sinorhizobium, Mesorhizobium and Bradyrhizobium. Through searching T3DB, Nop orthologs most likely encoding T3SS translocons, were identified in these 3 genera. Furthermore, when comparing the T3SS apparatus genes in the two Bradyrhizobium model strains, no NopX ortholog was found in strain Bradyrhizobium elkanii USDA61. According to published result(Okazaki et al., 2009), the T3SS in the USDA61 strain is supposed to be functional. This raises an interesting question as to whether NopX is functionally necessary for T3SSs in all rhizobia. Based on the ortholog clusters, one may also study the phylogenetic relationship among different T3SSs, or the co-evolutionary relationship between different functional categories. The manually-curated high quality effector and chaperone data are useful for feature study and evolution study of these special protein groups.

In the future, we plan to extend the T3DB in the following directions: First, to analyze more model strains using bioinformatics and comparative genomics strategies, and to include the T3SS data from non-model strains as well. Second, a transcription regulatory network for different T3SSs will be constructed. Third, two types of

protein-protein interaction networks will be integrated. One is to describe the interactions between T3SS proteins and other bacterial proteins, and the other is to describe the interactions between T3SS effectors and host proteins. We hope that T3DB can make important contribution to T3SS related research in the future.

CHAPTER 5

Conclusions and Perspectives

5.1 Contributions and conclusions from this thesis research

In summary, I have made the following contributions and conclusions through my thesis research:

1) I developed two high-performance software tools for effective prediction of bacterial Type III Secreted proteins. BPBAac only considered position-specific amino acid composition features, while T3SEpre also integrated second-order structure-based features besides primary sequences. Both softwares outperformed other implementations with similar applications. I also constructed a relational database, T3DB, to integrate, annotate, and analyze the molecular information of T3SS. Besides, two web servers were implemented in for BPBAac and T3SEpre, respectively.
2) With the assistance of T3SEpre, I identified a list of new T3S proteins in *Salmonella*, and selected candidates were validated experimentally.
3) The fact that computational models based on composition bias features could effectively recognize T3S proteins indicated that the position-based amino acid preference at least partly contributes to the specificity of T3S signals.
4) I demonstrated that second-order structure based features also contribute to the specificity of T3S signals.
5) Nearly half of the validated T3S proteins could still be recognized independent of position shift, suggesting that the T3S signals can tolerate position shift to certain level.

5.2 Future perspectives

Future study will be directed towards understanding the mechanism of T3S signal recognition and T3S signal formation. Protein-protein interaction experiments combined with molecular modeling will provide new clues about how T3S proteins are specifically recognized. As for the formation and evolution of T3S signals, the most intriguing question is how the effector genes, which are scattered along the genome and far away from T3SS apparatus gene cluster, are co-regulated and specifically recognized by T3SS apparatus. Comparative genomics and gene regulatory network analysis in model bacteria species are likely helpful in addressing this challenging question.

Reference

Akeda Y, Galán JE. (2005). Chaperone release and unfolding of substrates in type III secretion. Nature, 437, 911-915.

Alfano JR, Collmer A. (2004). Type III secretion system effector proteins: double agents in bacterial disease and plant defense. Annu. Rev. Phytopathol., 42, 385-414.

Alvarez-Martinez CE, Christie PJ. (2009). Biological diversity of prokaryotic type IV secretion systems. Microbiol. Mol. Biol. Rev., 73, 775-808.

Anderson DM., Schneewind O. (1997). A mRNA signal for the type III secretion of Yop proteins by Yersinia enterocolitica. Science, 278, 1140-1143.

Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M, Rattei T. (2009). Sequence-based prediction of type III secreted proteins. PLoS pathogens, 5, e1000376.

Arnold R, Jehl A, Rattei T. (2010). Targeting effectors: the molecular recognition of Type III secreted proteins. Microb. Infect., 12, 346-358.

Bingle LE, Bailey CM, Pallen MJ. (2008). Type VI secretion: a beginner's guide. Curr. Opin. Microbiol., 11, 3–8.

van Eerde A, Hamiaux C, Pérez J, Parsot C, Dijkstra BW. (2002). Three-dimensional secretion signals in chaperone-effector complexes of bacterial pathogenesis. Mol. Cell, 9, 971-980.

Bonas U, van den Ackerveken G. (1999). Gene-for-gene interactions: bacterial avirulence proteins specify plant disease resistance. Curr. Opin. Microbiol., 2, 94-8.

Buchko GW, Niemann G, Baker ES, Belov ME, Smith RD, Heffron F, Adkins JN, McDermott JE. (2010). A multi-pronged search for a common structural motif in the secretion signal of Salmonella enterica serovar Typhimurium type III effector proteins. Mol. Biosyst., 6, 2448-2458.

Cheng J, Randall AZ, Sweredoski MJ, Baldi P. (2005). SCRATCH: a protein structure and structural feature prediction server. Nuceic Acids Res., 33, W72–W76.

Cornelis GR. (2000). Molecular and cell biology aspects of plague. Proc. Natl. Acad. Sci. USA., 97, 8778-83.

Crooks GE, Hon G, Chandonia JM, Brenner SE. (2004). WebLogo: a sequence logo generator. Genome Res., 14, 1188–1190.

Deane JE, Abrusci P, Johnson S, Lea SM. (2010). Timing is everything: the regulation of type III secretion. Cell. Mol. Life Sci., 67, 1065-75.

Desvaux M, Hébraud M, Henderson IR, Pallen MJ. (2006). Type III secretion: what's in a name? Trends Microbiol., 14, 157-160.

Dimitriadou E. (2009). e1071: Misc Functions of the Department of Statistics (e1071). R package version 1.5–1.9. TU Wien.

Eddy SR. (1998). Profile hidden Markov models. Bioinformatics, 14, 755–763.

Ehrbar K, Friebel A, Miller SI, Hardt WD. (2003). Role of the Salmonella pathogenicity Island 1 (SPI-1). protein InvB in type III secretion of SopE and SopE2, two Salmonella effector proteins encoded outside of SPI-1. 185, 6950-6967.

Enninga J, Rosenshine I. (2009). Imaging the assembly, structure and activity of type III secretion systems. Cell. Microbiol., 11, 1462-70.

Fath MJ, Kolter R. (1993). ABC transporters: bacterial exporters. Microbiol. Rev., 57, 995–1017.

Fischer W, Haas R, Odenbreit S. (2002). Type IV secretion systems in pathogenic bacteria. Int. J. Med. Microbiol., 292, 159–168.

Galán JE. (2009). Common themes in the design and function of bacterial effectors. Cell. Host Microbe., 5, 571-9.

Galán JE, Curtiss R 3rd. (1989). Cloning and molecular characterization of genes whose products allow Salmonella typhimurium to penetrate tissue culture cells. Proc. Natl. Acad. Sci. USA., 86, 6383-7.

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

Garrity GM. (2005). Bergey's Manual of Systematic Bacteriology, 2nd Edition, Published by Springer, New York.

Ghosp P. (2004). Process of protein transport by the type III secretion system. Microbiol. Mol. Biol. Rev., 68, 771-795.

Girard F, Crepin VF, Frankel G. (2009) Modelling of infection by enteropathogenic Escherichia coli strains in lineages 2 and 4 ex vivo and in vivo by using Citrobacter rodentium expressing TccP. Infect. Immun., 77, 1304-1314.

Hardt WD, Galán JE. (1997) A secreted Salmonella protein with homology to an avirulence determinant of plant pathogenic bacteria. Proc. Natl. Acad. Sci. USA., 94, 9887-9892.

Hauser AR. (2009). The type III secretion system of Pseudomonas aeruginosa: infection by injection. Nat. Rev. Microbiol., 7, 654-65.

Hayes CS, Aoki SK, Low DA. (2010). Bacterial contact-dependent delivery systems. Annu. Rev. Genet., 44, 71-90.

He SY, Nomura K, Whittam TS. (2004). Type III protein secretion mechanism in mammalian and plant pathogens. Biochim Biophys Acta., 1694, 181-206.

Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC, Ala'Aldeen D. (2004). Type V protein secretion pathway: the autotransporter story. Mol. Biol. Rev., 68, 692–744.

Higashide W, Zhou D. (2006). The first 45 amino acids of SopA are necessary for InvB binding and SPI-1 secretion. J. Bacteriol., 188, 2411-2420.

Hong KH, Miller VL. (1998). Identification of novel Salmonella invasion locus homologous to Shigella ipgDE. J. Bacteriol., 180, 1793-1802.

Huang HC, Lin RH, Chang CJ, Collmer A, Deng WL. (1995). The complete hrp gene cluster of Pseudomonas syringae pv. syringae 61 includes two blocks of genes required for harpinPss secretion taht are arranged colinearly with Yersinia ysc homologs. MPMI, 8, 733-746.

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

Izoré T, Job V, Dessen A. (2011). Biogenesis, Regulation, and Targeting of the Type III Secretion System. Structure, 19, 603-12.

Jarvis KG, Girón JA, Jerse AE, McDaniel TK, Donnenberg MS, Kaper JB. (1995) Enteropathgenic Escherichia coli contains a putative type III secretion system necessary for the export of proteins involved in attaching and effacing lesion

formation. Proc. Natl. Acad. Sci. USA., 92, 7996-8000.

Jehl MA, Arnold R, Rattei T. (2011). Effective--a database of predicted secreted bacterial proteins. Nucleic Acids Res., 39, D591-595.

Kaniga K, Trollinger D, Galán JE. (1995). Identification of two targets of the type III protein secretion system system encoded by the inv and spa loci of Salmonella typhymurium that have homology to the Shigella IpaD and IpaA proteins. J.Bacteriol. 177, 7078-7085.

Karavolos MH, Roe AJ, Wilson M, Henderson J, Lee JJ, Gally DL, Khan CM. (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.

Kay S. (2009). How Xanthomonas type III effectors manipulate the host plant. Curr. Opin. Microbiol., 12, 37-43.

Kim JH, Lee J, Oh B, Kimm K, Koh I. (2004) Prediction of phosphorylation sites using SVMs. Bioinformatics, 20, 3179–3184.

Kubori T, Matsushima Y, Nakamura D, Uralil J, Lara-Tejero M, Sukhan A, Galán JE, Aizawa SI. (1998). Supramolecular structure of the Salmonella typhymurium type III protein secretio system. Science, 280, 602-605.

Kubori T, Galán JE. (2003). Temporal regulation of salmonella virulence effector function by proteasome-dependent protein degration. Cell, 115, 333-342.

Kuhle V, Hensel M. (2004). Cellular microbiology of intracellular Salmonella enterica: functions of the type III secretion system encoded by Salmonella pathogenicity island 2. Cell. Mol. Life Sci., 61, 2812-2826.

Lara-Tejero M, Kato J, Wagner S, Liu X, Galán JE. (2011). A sorting platform determines the order of protein secretion in bacterial type III systems. Science, 331, 1188-91.

Lee CC, Wood MD, Ng K, Andersen CB, Liu Y, Luginbühl P, Spraggon G, Katagiri F. (2004). Crystal structure of the type III effector AvrB from Pseudomonas syringae. Structure, 12, 487-494.

Lilic M, Vujanac M, Stebbins CE. (2006). A common structural motif in the binding of virulence factors to bacterial secretion chaperones. Mol. Cell, 21, 653-664.

Lindeberg M, Collmer A. (2009) Gene ontology for type III effectors: capturing

processes at the host-pathogen interface. Trend. Microbiol., 17, 304-311.

Lloyd SA, Norman M, Rosqvist R, Wolf-Watz H. (2001). Yersinia YopE is targeted for type III secretion by N-terminal, not mRNA, signals. Mol. Microbiol., 39, 520–531.

Lloyd SA, Sjöström M, Andersson S, Wolf-Watz H. (2002). Molecular characterization of type III secretion signals via analysis of synthetic N-terminal amino acid sequences. Mol. Microbiol., 43, 51–59.

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

Luo Y, Bertero MG, Frey EA, Pfuetzner RA, Wenk MR, Creagh L, Marcus SL, Lim D, Sicheri F, Kay C, Haynes C, Finlay BB, Strynadka NC. (2001). Strcutural and biochemical characterization of the type III secretion chaperones CesT and SigE. Nat. Struct. Biol., 8, 1031-1036.

Ly KT, Casanova JE. (2007). Mechanisms of Salmonella entry into host cells. Cell. Microbiol., 9, 2103-11.

Marguerettaz M, Pieretti I, Gayral P, Puig J, Brin C, Cociancich S, Poussier S, Rott P, Royer M. (2011). Genomic and evolutionary features of the SPI-1 type III secretion system that is present in Xanthomonas albilineans but is not essential for xylem colonization and symptom development of sugarcane leaf scald. Mol. Plant Microbe Interact., 24, 246-59.

Marie C, Broughton WJ, Deakin WJ. (2001). Rhizobium type III secretion systems: legume charmers or alarmers? Curr. Opin. Plant Biol., 4, 336-342.

Marlovits TC, Kubori T, Sukhan A, Thomas DR, Galán JE, Unger VM. (2004). Structural insights into the assembly of the type III secretion needle complex. Science, 306, 1040-1042.

Mazurier S, Lemunier M, Hartmann A, Siblot S, Lemanceau P. (2006). Conservation of type III secretion system genes in Bradyrhizobium isolated from soybean. FEMS Microbiol. Lett., 259, 317-25.

McDermott JE, Taylor RC, Yoon H, Heffron F. (2009). Bottlenecks and hubs in inferred networks are important for virulence in Salmonella typhimurium. J. Comput. Biol. 16, 169-80.

McGuffin LJ, Bryson K, Jones DT. (2000). The PSIPRED protein structure prediction server. Bioinformatics, 16, 404–405.

Meyer D, Cunnac S, Guéneron M, Declercq C, Van Gijsegem F, Lauber E, Boucher C, Arlat M. (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 MB, Chesnokova O, Dahlbeck D, Clark ET, Rossier O, Bonas U, Staskawicz BJ. (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, Tamura N, Iwabuchi M. (2010). Genome-wide identification of a large repertoire of Ralstonia solanacearum type III effector proteins by a new functional screen. MPMI, 23, 251–262.

Mundy R, Petrovska L, Smollett K, Simpson N, Wilson RK, Yu J, Tu X, Rosenshine I, Clare S, Dougan G, Frankel G. (2004). Identification of a novel Citrobacter rodentium type III secreted protein, EspI, and roles of this and other secreted proteins in infection. Infect. Immun., 72, 2288-302.

Noël L, Thieme F, Nennstiel D, Bonas U. (2002) Two novel type III-secreted proteins of Xanthomonas campestris pv. vescatoria are encoded within the hrp pathogenicity island. J. Bacteriol., 184, 1340-1348.

Noël L, Thieme F, Gäbler J, Büttner D, Bonas U. (2003) XopC and XopJ, two novel type III effector proteins from Xanthomonas campestris pv. vescatoria. J. Bacteriol., 185, 7092-7102.

Norris FA, Wilson MP, Wallis TS, Galyov EE, Majerus PW. (1998). SopB, a protein required for virulence of Salmonella dublin, is an inositol phosphate phosphatase. Proc. Natl. Acad. Sci. USA., 95, 14057-14059.

Ogawa M, Handa Y, Ashida H, Suzuki M, Sasakawa C. (2008). The versatility of Shigella effectors. Nat. Rev. Microbiol., 6, 11-16.

Okazaki S, Zehner S, Hempel J, Lang K, Göttfert M. (2009). Genetic organization and functional analysis of the type III secretion system of Bradyrhizobium elkanii. FEMS. Microbiol. Lett., 295, 88-95.

Pallen MJ, Beatson SA, Bailey CM. (2005). Bioinformatics, genomics, and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. FEMS Microbiol. Rev., 29, 201-229.

Panina EM, Mattoo S, Griffith N, Kozak NA, Yuk MH, Miller JF. (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, Schneider DJ, Tam VC, Chancey ST, Shan L, Jamir Y, Schechter LM, Janes MD, Buell CR, Tang X, Collmer A, Alfano JR. (2002). Genomewide identification of proteins secreted by the Hrp type III protein secretion system of Pseudomonas syringae pv. DC3000. Proc. Natl. Acad. Sci. USA, 99, 7652-7657.

Ramamurthi KS, 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, Igwe EI, Sauer J, Hardt WD, Bubert A, Geginat G. (2001). Protection against murine listeriosis by oral vaccination with recombinant Salmonella expressing hybrid Yersinia type III proteins. J. Immunol., 167, 357-365.

Rüssmann H, Kubori T, Sauer J, Galán JE. (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, Heffron F, McDermott JE. (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, Smola AJ. (2002). Learning with Kernels. Cambridge: MIT Press.B., Smola, A.J. (2002). Learning with Kernels. Cambridge: MIT Press.

Schechter LM, Roberts KA, Jamir Y, Alfano JR, Collmer A. (2004). Pseudomonas syringae type III secretion system targeting signals and novel effectors studied with a Cya translocation reporter. J Bacteriol., 186, 543-555.

Schraidt O, Marlovits TC. (2011). Three-dimensional model of Salmonella's needle complex at subnanometer resolution. Science, 331, 1192-5.

Schroeder GN, Hilbi H. (2008). Molecular pathogenesis of Shigella spp.: controlling host cell signaling, invasion, and death by type III secretion. Clin. Microbiol. Rev., 21, 134-56.

Shao F. (2008). Biochemical functions of Yersinia type III effectors. Curr. Opin. Microbiol., 11, 21-29.

Shao J, Xu D, Tsai SN, Wang Y, Ngai SM. (2009). Computational identification of protein methylation sites through bi-profile Bayes feature extraction. PloS one, 4, e4920.

Shatsky M, Nussinov R, Wolfson HJ. (2004). A method for simultaneous alignment of multiple protein structures. Proteins, 56, 143-156.

Singer AU, Desveaux D, Betts L, Chang JH, Nimchuk Z, Grant SR, Dangl JL, Sondek J. (2004). Crystal structures of the type III effector protein AvrPphF and its chaperone reveal residues required for plant pathogenesis. Structure, 12, 1669-1681.

Sory MP, Cornelis GR. (1994). Translocation of a hybrid YopE-adenylate cyclase from Yersinia enterocolitica into Hela cells. Mol. Microbiol., 14, 583-394.

Stavrinides J, Ma W, Guttman DS. (2006). Terminal reassortment drives the quantum evolution of type III effectors in bacterial pathogens. PLoS. Pathog., 2, e104.

Stebbins CE. (2005). Structural microbiology at the pathogen-host interface. Cell. Microbiol., 7, 1227-1236.

Stebbins CE, Galán JE. (2001). Maintenance of an unfolded polypeptide by a cognate chaperone in bacterial type III secretion. Nature, 414, 77-81.

Stebbins CE, Galán JE. (2003). Priming virulence factors for delivery into the host. Nat. Rev. Mol. Cell. Biol., 4, 738-743.

Sun GW, Gan YH. (2010). Unraveling type III secretion systems in the highly versatile Burkholderia pseudomallei. Trends Microbiol., 18, 561-568.

Tay DM, Govindarajan KR, Khan AM, Ong TY, Samad HM, Soh WW, Tong M, Zhang F, Tan TW. (2010). T3SEdb: data warehousing of virulence effectors secreted by the bacterial Type III Secretion System. BMC Bioinformatics., 11, Suppl 7:S4.

Tobe T, Beatson SA, Taniguchi H, Abe H, Bailey CM, Fivian A, Younis R, Matthews S, Marches O, Frankel G, Hayashi T, Pallen MJ. (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.

Tree JJ, Wolfson EB, Wang D, Roe AJ, Gally DL. (2009). Controlling injection:

regulation of type III secretion in enterohaemorrhagic Escherichia coli. Trends Microbiol., 17, 361-70.

Viprey V, Del Greco A, Golinowski W, Broughton WJ, Perret X. (1998). Symbiotic implications of type III protein secretion machinery in Rhizobium. Mol. Microbiol., 28, 1381-1389.

Wang Y, Hou Y, Huang H, Liu GR, White AP, Liu SL. (2008). Two oral HBx vaccines delivered by live attenuated Salmonella: both eliciting effective anti-tumor immunity. Cancer Lett, 263, 67-76.

Wang Y, Zhang Q, Sun MA, Guo D. (2011). High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. Bioinformatics, 27,777-784.

Winnen B, Schlumberger MC, Sturm A, Schüpbach K, Siebenmann S, Jenny P, Hardt WD. (2008). Hierarchical effector protein transport by the Salmonella typhimurium SPI-1 type III secretion system. PLoS One, 3, e2178.

Wood MW, Rosqvist R, Mullan PB, Edwards MH, Galyov EE. (1996). SopE, a secreted protein of Salmonella dublin, is translocated into the target eukaryotic cell via a sip-dependent mechanism and promotes bacterial entry. Mol. Microbiol., 22, 327-338.

Wood MW, Jones MA, Watson PR, Siber AM, McCormick BA, Hedges S, Rosqvist R, Wallis TS, Galyov EE. (2000). The secreted effector protein of Salmonella dublin, SopA, is translocated into eukaryotic cells and influences the induction of enteritis. Cell. Microbiol., 2, 293-303.

Wulf J, Pascuzzi PE, Fahmy A, Martin GB, Nicholson LK. (2004). The solution structure of type III effector protein AvrPto reveals conformational and dynamic features important for plant pathogenesis. Structure, 12, 1257-1268.

Wu S, Skolnick J, Zhang Y. (2007). Ab initio modelling of small proteins by iterative TASSER simulations. BMC Biol., 5, 17.

Yang Y, Zhao J, Morgan RL, Ma W, Jiang T. (2010). Computational prediction of type III secreted proteins from gram-negative bacteria. BMC bioinformatics, 11 Suppl 1, S47.

Zhang J, Wang Q, Barz B, He Z, Kosztin I, Shang Y, Xu D. (2010). MUFOLD: A new solution for protein 3D structure prediction. Proteins, 78, 1137-1152.