Bioinformatics, 35(12), 2019, 2051–2057
doi: 10.1093/bioinformatics/bty931
Advance Access Publication Date: 8 November 2018
Original Paper

Sequence analysis

DeepT3: deep convolutional neural networks accurately identify Gram-negative bacterial type III secreted effectors using the N-terminal sequence

Li Xue[1,†], Bin Tang[2,†], Wei Chen[3,*] and Jiesi Luo[4,*]

[1]School of Public Health, Southwest Medical University, Luzhou, Sichuan 646000, PR, China, [2]Basic Medical College of Southwest Medical University, Luzhou, Sichuan 646000, PR, China, [3]Integrative Genomics Core, City of Hope National Medical Center, Duarte, CA 91010, USA and [4]Key Laboratory for Aging and Regenerative Medicine, Department of Pharmacology, School of Pharmacy, Southwest Medical University, Luzhou, Sichuan, 646000, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on June 26, 2018; revised on October 22, 2018; editorial decision on November 2, 2018; accepted on November 7, 2018

Abstract

Motivation: Various bacterial pathogens can deliver their secreted substrates also called effectors through Type III secretion systems (T3SSs) into host cells and cause diseases. Since T3SS secreted effectors (T3SEs) play important roles in pathogen–host interactions, identifying them is crucial to our understanding of the pathogenic mechanisms of T3SSs. However, the effectors display high level of sequence diversity, therefore making the identification a difficult process. There is a need to develop a novel and effective method to screen and select putative novel effectors from bacterial genomes that can be validated by a smaller number of key experiments.

Results: We develop a deep convolution neural network to directly classify any protein sequence into T3SEs or non-T3SEs, which is useful for both effector prediction and the study of sequence-function relationship. Different from traditional machine learning-based methods, our method automatically extracts T3SE-related features from a protein N-terminal sequence of 100 residues and maps it to the T3SEs space. We train and test our method on the datasets curated from 16 species, yielding an average classification accuracy of 83.7% in the 5-fold cross-validation and an accuracy of 92.6% for the test set. Moreover, when comparing with known state-of-the-art prediction methods, the accuracy of our method is 6.31–20.73% higher than previous methods on a common independent dataset. Besides, we visualize the convolutional kernels and successfully identify the key features of T3SEs, which contain important signal information for secretion. Finally, some effectors reported in the literature are used to further demonstrate the application of DeepT3.

Availability and implementation: DeepT3 is freely available at: https://github.com/lje00006/DeepT3.

Contact: weichen2@coh.org or ljs@swmu.edu.cn

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

1 Introduction

Bacteria can manipulate their surrounding environment through the secretion of proteins into other living organisms and into the extracellular milieu. In Gram-negative bacteria this process is mediated by different types of secretion systems from Types I to VIII secretion system (T1SS–T8SS) (Desvaux *et al.*, 2009). Among these, the Type III secretion system (T3SS) is a specialized protein delivery system that has been widely encoded in many Gram-negative bacteria, such as *Escherichia, Shigella, Yersinia, Salmonella and Pseudomonas* etc. (Cornelis, 2006; He *et al.*, 2004). The T3SS creates a contiguous channel through the bacterial and host membranes, allowing injection of specialized bacterial effector proteins directly to the host cell (Burkinshaw and Strynadka, 2014). These T3SS secreted effector (T3SE) proteins act as virulence factors within the host and are able to alter and manipulate vital host cell functions, such as signal transduction (Schroeder and Hilbi, 2008) and innate immune response (Engel and Balachandran, 2009). Due to the key role of T3SEs in the establishment and maintenance of host–pathogen interactions, there is considerable research interest in the identification of T3SS effectors. To date, quite a number of T3SEs have been identified by both *in vitro* and *in silico* methods. Although the experimental approaches are the best way to identify novel T3SEs, implementation of these approaches is often expensive and time-consuming. Therefore, in recent years, many computational methods have been proposed to predict T3SEs.

We emphasize the categorizing of available computational approaches into two types: (i) alignment-based, where known T3SEs are aligned from given bacterial genome to search homology proteins; and (ii) learning-based, where T3SEs are scored and predicted from a training model by considering different sequence-derived features. The methods based on sequence alignment were originally proposed in (Panina *et al.*, 2005; Petnicki-Ocwieja *et al.*, 2002), but T3SEs evolve very quickly and their amino acid sequences are highly variable to adapt in different hosts and defend against immune system attacks (Ma and Guttman, 2008). Thus, many T3SEs have no homology with other effectors in public databases. The low sequence similarity between different T3SEs leads to the poor performance of these methods. The learning-based methods are expected to perform better than the alignment-based tool because different sequence and structure features are considered, together with the contributions of machine learning models. Based on the general features of T3SEs, different machine learning methods have been adopted, e.g. Naïve Bayes (NB) (Arnold *et al.*, 2009), artificial neural network (ANN) (Lower and Schneider, 2009), support vector machine (SVM) (Dong *et al.*, 2013; Samudrala *et al.*, 2009; Wang *et al.*, 2011; Yang *et al.*, 2010) and random forest (RF) (Yang *et al.*, 2013).

Arnold *et al.* (2009) developed the first universal *in silico* program, EffectiveT3, by incorporating frequencies of amino acids, short peptides and residues with certain physic-chemical properties of N-terminal sequences. Löwer and Schneider (Lower and Schneider, 2009) used the sliding-window technique to capture signal features among the first 30 amino acids of the effector sequences. Sato *et al.* (2011) introduced N-terminal instability, codon adaptation index and ProtParam indices to refine the discriminatory power of the classifier. Yang *et al.* (2010) extracted amino acid composition, secondary structure and solvent accessibility information of N-terminal sequences to construct the learning model. Wang *et al.* (2011) developed a BPBAac model based on position-specific amino acid composition. Dong *et al.* (2013) used the profile-based k-spaced amino acid pair composition to represent the N-terminal sequences and they called this new method, BEAN. Most recently, we constructed the conservation profiles of the N-terminal sequences using the position-specific scoring matrix (Yang *et al.*, 2013). With the combination of other features, including amino acid composition, solvent accessibility information, secondary structure and six physic-chemical properties, our method gives very high accuracy in classifying T3SEs and non-T3SES.

One drawback in standard machine learning approaches is that the features have to be predefined, the appropriate choice of features affects the prediction accuracy and there is limited flexibility for model changes or updates. These drawbacks are overcome by deep learning techniques that allow for the reduction of feature engineering: the model learns to extract features as a natural consequence of the process of fitting the model's parameters to the available data. With the common availability of data and an ever-increasing computing power, deep learning approaches proved to be very efficient and outperformed traditional machine learning approaches. Over the past few years, deep learning has been relatively widely used by the bioinformatics, computational biology and medical informatics community (Angermueller *et al.*, 2016; Min *et al.*, 2017; Miotto *et al.*, 2017). While the deep networks, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and long short-term memory cells have been exploited for gene expression regulation (Singh *et al.*, 2016), regulatory genomics (Alipanahi *et al.*, 2015; Kelley *et al.*, 2016; Liu *et al.*, 2018), protein structure prediction (Heffernan *et al.*, 2017), protein sequence classification (Szalkai and Grolmusz, 2018), protein subcellular localization (Almagro Armenteros *et al.*, 2017) and peptide design (Veltri *et al.*, 2018).

In this article, we present a method based on deep CNNs to predict the T3SEs using the N-terminal sequences. The deep learning model depends solely on protein primary sequences, requiring no prior knowledge to define rules for the feature engineering. We demonstrate that the deep CNN models accurately predict the T3SEs for different bacterial species and accurately capture the important signal information for T3SE secretion. Based on these deep CNN models, we develop the tool DeepT3. Performance evaluation and comparison show that DeepT3 outperforms the state-of-the-art tools in two different test sets. When applied to some candidate effectors reported in the literature, DeepT3 identifies the T3SEs with remarkable accuracy. DeepT3 facilitates the understanding of T3SS pathogenic mechanisms and the development of potential therapeutics.

2 Materials and methods

2.1 Dataset

In our previous work (Yang *et al.*, 2013), we carried out an extensive literature search to obtain information of all secretory proteins from Types I to VIII and downloaded the corresponding sequences from Swiss-Prot and TrEMBL (UniProt, 2008). In addition, we collected other experimentally verified T3SEs from the works of Wang *et al.* (2011) and Tay *et al.* (2010) to increase the comprehensiveness of the T3SE data. In total, the original dataset consists of 662 T3SEs and 787 non-T3SEs. CD-HIT (Huang *et al.*, 2010) with the sequence identity cutoff of 30% was used to remove similar sequences. Then proteins with <100 amino acids were further skipped. Finally, we obtained a balanced dataset that included 283 T3 and 313 non-T3 proteins. We developed the deep-learning-based prediction tool using this dataset for training.

We created the negative sample (non-T3SEs) by choosing secreted proteins from T1SS to T8SS in Gram-negative bacteria after

removing the known T3SEs and their homologs. Compared with the other non-T3SEs, the proteins from other secretion systems are more difficult to distinguish from T3SEs and make the prediction model more reasonable and reliable (Luo *et al.*, 2015). However, we also collected other non-T3 proteins to investigate the influence of negative samples. To obtain a 1:3 ratio of positive to negative samples, 849 negative samples were randomly selected from the previous work of Dong *et al.* (2013), which was compiled from eight well-studied Gram-negative bacterial proteomes by several criteria. The pairwise sequence identity among negative samples was also controlled as ≤30% using CD-HIT.

Reports have shown that N-terminal residues appear to provide the targeting information for protein translocation (Bendtsen *et al.*, 2004; Casper-Lindley *et al.*, 2002; Yang *et al.*, 2010). Some effectors, e.g. AvrBs2 in *Xanthomonas* (Casper-Lindley, *et al.*, 2002), Tir in *Enteropathogenic Escherichia coli* (EPEC; Crawford and Kaper, 2002) and PopD in *Pseudomonas aeruginosa* (Tomalka *et al.*, 2012), only depends on the first ∼50 residues to be secreted or trans-located. Costa *et al.* (2012) further found that the conserved chaperone-binding domain covers the first 25–45 residues. But there are still many other effectors that require the first ∼100 residues for secretion. For example, the translocation signals in some Yops are located in the first 50–100 residues (Schesser *et al.*, 1996; Sory *et al.*, 1995). It also has been proven that the region between residues 15 and 78 of SopE is responsible for binding the chaperone InvB (Lee and Galan, 2003). Taken together, the maximal secretion or translocation may require the first 100 amino acids, in which the signal peptides of T3SEs may be contained. Thus, only the first 100 residues were extracted from the dataset in all following calculations.

2.2 Overview of the DeepT3 deep learning model

The deep CNN in DeepT3 consisted of a hierarchical architecture that used raw N-terminal sequence as input and predicted the probability of T3SEs (Fig. 1). The deep CNN model in DeepT3 consisted of convolution layers, rectification layers, pooling layers and fully connected layers. Each input sequence was converted to a one-hot matrix with 20 rows and 100 columns. The twenty rows corresponded to the twenty amino acids G, A, V, L, I, P, F, Y, W, S, T, C, M, N, Q, D, E, K, R and H. Fifty filters were used in each convolution layer. Each filter was a 20×12 matrix. These filters automatically extracted predictive features from input sequences during model training. After convolution, the rectified linear units were used to output the filter scanning results that were above the thresholds, which were learned during model training. Max pooling was applied in the pooling layer to reduce variance and increase translational invariance by computing the maximum value of a feature over a region. All the pooling results were combined in one vector, resulting in a vector with a size of 2200. The vector was batch-normalized before inputting it into the fully connected layer. A fully connected layer was employed in our model with 650 nodes. A dropout layer was added following the fully connected layer to avoid over-fitting. The output layer applied the softmax function to predict the probability of T3SE and non-T3SE. The deep CNN was implemented with MXNet libraries (https://mxnet.incubator.apache.org/).

2.3 Training of deep CNN

The proposed models were optimized for the cross entropy loss function using mini-batch stochastic gradient descent with Adam updates and regularized by dropout with a 0.5 dropout rate.

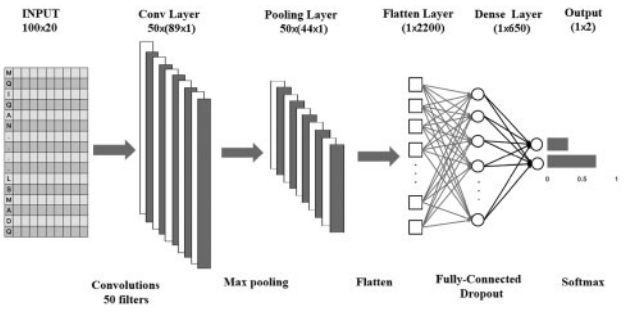

Fig. 1. The architecture of deep CNN for T3SE prediction. The network accepts features of proteins of first 100 amino acids as input. The first convolutional layer performs 50 convolutions with 20 × 12 filter on the one- hot matrix, producing 50 feature maps of size 1 × 89. The second pooling layer performs 1 × 2 spatial pooling of each feature map using the max value, producing 50 feature maps of size 1 × 44. All the pooling results are joined together into one vector by flattening. The hidden features in this vector are fully connected to a hidden layer of 650 nodes, which are fully connected to 2 output nodes to predict the probability of T3SE and non-T3SE. The output node uses softmax function as activation function, whereas all the nodes in the other layers use rectified linear function as activation function

Random search was used to optimize a set of important hyper-parameters, such as:

i. Learning rate: The step size that the optimizer should take in the parameter space while updating the model parameters.
ii. Batch Size: The number of training example to consider before updating the parameters.
iii. Maximum Epochs: The total number of iterations over the model training.
iv. Early Stopping Patience: The number of epochs to stop the model training, when the validation loss does not improve.

We used learning rate to 0.005, momentum to 0.9, maximum number of training epochs to 100, early stopping patience to 10 and the optimal value of the batch size tuned on the cross-validation set was found to 80.

2.4 Generating motifs learned by deep CNN

A filter can be visualized as a sequence motif detector. This helps to understand which amino acids the filter prefers at each sequence position.

Each time the filter is moved, it produces an output by taking the sum of the element-wise product of the input and the filter position-specific weights. The filter unfavorable amino acids will suppress the output while preferred amino acids will increase the output. When a subsequence matches a filter preference, this filter will be activated and will produce a positive output for the subsequence position (Jurtz, *et al.*, 2017). For each test N-terminal sequence, we sought the activated position that had the maximum convolution value among all the filters. The 12-bp (filter width) subsequence starting at this position of the test sequence was extracted. All of the 12-bp subsequences with the maximum convolution value for each sequence in the test set were pooled together and aligned. The frequencies of twenty amino acids at each position were then calculated, and the position weight matrix representing the motif was derived (Wang *et al.*, 2018).

2.5 Comparison with conventional machine models and published methods

To evaluate DeepT3, we compared its prediction performance with that of the following learning models that previously showed

state-of-the-art T3SE prediction, i.e. SVM, RF, ANN, k-nearest neighbor (KNN) and NB from the Scikit-learn library (Pedregosa *et al.*, 2011). For each model, we set parameters to provide the highest accuracy on the training datasets. For SVM, we considered the radial basis function (RBF) as the kernel function, and two parameters, the regularization parameter C and the kernel width parameter γ were optimized by using a grid search approach. It could identify good parameters based on exponentially growing sequences of (C, γ) ($C = 2^{-2}, 2^{-1}, \ldots, 2^{9}$ and $\gamma = 2^{-6}, 2^{-5}, \ldots, 2^{5}$). For RF, the two parameters, *ntree* (the number of trees to grow) and *mtry* (the number of variables randomly selected as candidates at each node), were optimized using a grid search approach; the value of *ntree* was from 500 to 3000 with a step length of 500, and the value of *mtry* was from 2 to 40 with a step length of 2. For NN, we tested the feed-forward back-propagation network with two and three hidden layers, each with an optimal number of neurons, and using sigmoid functions for hidden layers. The learning rate was set to 0.0001 and the weight decay to −0.001. For KNN, we chose Euclidean distance as distance function and set the number of neighbors (K) in the set {3, 5, 7, 9, 11, 13, 15, 17, 19, 21 and 23}. The K value with the highest prediction performance was kept.

We also compared three published algorithms that were previously used for T3SEs prediction, i.e. EffectiveT3 (Arnold *et al.*, 2009), BPBAac (Wang *et al.*, 2011) and BEAN2.0 (Dong *et al.*, 2015). We retrieved predictive results for tested T3SEs from their web sites and tool. Default parameters were used for all methods.

2.6 Performance evaluation

Cross-validation is a common method for estimating the performance of a classification model. In this study, this process partitions the labeled data into five non-overlapping equally sized sets, and trains the predictor on the union of four of these before testing on the remaining set. This is repeated five times such that each of the five sets is used as the test set exactly once, and the average performance parameters are recorded.

For two-class classification problems, six parameters, namely, Sensitivity (SN), Specificity (SP), Precision (PRE), Accuracy (ACC), *F*-value and Matthew's correlation coefficient (MCC), are used to evaluate the overall predictive performance of classification models. These parameters are defined in the Supplementary Material. Additionally, the receiver operating characteristic (ROC) curve, which is a plot of the true-positive rate versus the false-positive rate, is depicted to visually measure the comprehensive performance of different methods. The area under the curve (AUC) is also provided in each of the ROC plots. The maximum value of the AUC is 1.0, which denotes a perfect prediction. A random guess gives an AUC value of 0.5.

3 Results

3.1 The features learned by the deep learning model captured the leader sequence of T3SEs

We analyzed the amino acid occurrences (including those over-represented and underrepresented) on each position of T3SS effectors. We examined the first 100 N-terminal residues of T3SEs and non-T3SES with the Two Sample Logo program (Vacic *et al.*, 2006), and studied the differences among the two groups of proteins with respect to their amino acid preferences (Supplementary Fig. S1). For the N-terminus, remarkable consensus was found in T3SE sequences, while amino acid residues tended to be more disordered in non-T3SE sequences. Specifically, the N-terminal sequences of T3SEs (from Positions 6 to 17) showed a significant overrepresentation of serine and proline residues, with lysine, leucine and alanine largely absent. There was no significant motif pattern in the remaining N-terminal sequences, except for some positions, which showed an enrichment of arginine or leucine residue.

One of the distinctive advantages of the deep learning model is the ability to automatically extract predictive features from inputs during the model training. We explored features that were learned in our deep CNN by investigating the test sequences that activated the filters in the convolution layer. When the filter is slid over N-terminal sequence, it functions as motif detector and becomes activated when certain position matches its preference. We first visualized the positions within the N-terminal sequence that have high importance for the prediction of T3SEs (Fig. 2a). The position importance is determined by the fraction of filters activated at the position. We observed that the majority of filters were activated when they convolved a continuous region including the positions from 6 to 17. This result supports above sequence analysis. Furthermore, these activation sequences were aligned together to obtain the learned motif represented by the position weight matrix for T3SEs. The result showed that the discovered amino acid motif revealed by the deep CNN was almost identical to known motif for T3SEs (Fig. 2b). The similarities between motifs were compared by Tomtom (Gupta *et al.*, 2007) using the Pearson correlation coefficient. The Tomtom *P*-value was 9.91e-15.

3.2 T3SEs prediction by the deep CNN models

The deep CNN models consisted of convolution layers, rectification layers, pooling layers and fully connected layers. Theses interrelated architectural factors may determine the performance of the convolutional network models. Consequently, we carefully designed and sized the model architecture to make it appropriate for our purpose (Supplementary Figs S2 and S3). In addition, the length of input sequence was optimized to achieve the best characterization of the N-terminal sequence. Eleven models were constructed with eleven different sequence lengths (length = 50, 60, 70, 80, 90, 100, 110, 120, 130, 140 and 150). The prediction results for the eleven models are shown in Supplementary Figure S4. As seen from the curves, the best performance is obtained by using the length of 100 amino acids, suggesting that the first 100 residues of T3SE are adequate as input sequences for deep CNN models.

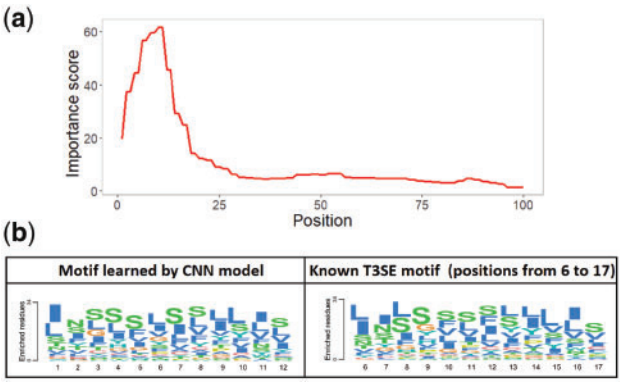

Fig. 2. Visualizing and understanding DeepT3. **(a)** Relative importance of N-terminal 100 residues revealed by deep CNN model. The position importance is determined by the fraction of filters activated at the position. **(b)** Example of motif learned by the deep CNN model and compared with the known motif of T3SEs. The sequence signatures automatically learned by the deep CNN model of DeepT3 captured the known T3SEs motif

The deep CNN model was first trained using a balanced dataset of 283 T3SEs and 313 non-T3SEs (Training Dataset 1), and negative samples were obtained by selecting secreted proteins from other secretion systems. In order to measure the capability of the CNN model to predict the T3SEs, we used the testing strategy based on 5-fold cross-validation. In the cross-validation approach, the average SN, SP, PRE, *F*-score, ACC and MCC were 0.837, 0.836, 0.823, 0.830, 0.837 and 0.674, respectively. Another way of an unbiased estimation of classifier's accuracy is to use the ROC plot that is a popular way of analyzing the overall performance of a classifier. It describes the true-positive rate as a function of false-positive rate for different trade-offs between sensitivity and specificity. The AUC is commonly used as a summary measure of diagnostic accuracy. The ROC plot for a combined 5-fold cross-validation curve (Fig. 3a) and the corresponding average AUC value (Table 1) also support the conclusion that the deep CNN method is robust and is particularly suitable for classifying T3SEs and non-T3SEs.

To investigate whether the negative samples affect the performance of CNN model, we trained another predictor using the non-secreted proteins as negative samples. In more detail, from 7143 non-Type III proteins in eight well-studied Gram-negative bacterial proteomes (*E.coli* O157: H7, *Salmonella enterica serovar*, *Typhimurium*, *Pseudomonas syringae* DC3000, *Yesinia pestis* bv. Antiqua, *Chlamydia trachomatics*, *Shigella flexneri*, *Yesinia enteroclitica* and *Burkholderia pesudomallei*) (Dong *et al.*, 2013), 849 protein sequences were randomly selected to generate an unbalanced training data set with a 1:3 ratio of positive to negative samples (Training Dataset 2). To deal with the imbalanced positive/negative ratio, we trained the deep CNN by maximizing AUC, which is an unbiased measure for class-imbalanced data. To maximize AUC of our deep CNN model, we formulated the AUC function in a ranking framework, approximate it by a polynomial approximation (Wang *et al.*, 2016). Through the 5-fold cross-validation test on the new training dataset, the deep CNN model also achieved a good performance. The average SN, SP, PRE, *F*-score, ACC, MCC and AUC were 0.786, 0.945, 0.875, 0.821, 0.889, 0.746 and 0.933, respectively (Table 1 and Fig. 3b).

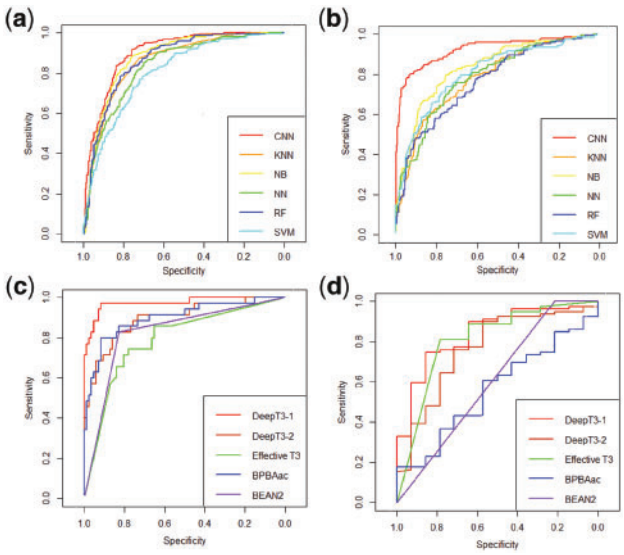

Fig. 3. Comparison of ROC curves for T3SEs prediction using different machine learning models and algorithms. (a) Five-fold cross-validation tests using different machine learning models on training dataset 1. (b) Five-fold cross-validation tests on training dataset 2. **(c)** Performance evaluation of two DeepT3 models and several algorithms on an independent test set. **(d)** Performance evaluation on the *P.syringae* test set

Finally, we compared the prediction results of the CNN models with SVM, RF, ANN, KNN and NB models. We used the same one-hot encoded vector as input to train other models. Using two training datasets based on the 5-fold cross-validation test, the performance of different methods is shown in Table 1 and Figure 3. The CNN model outperformed other models in terms of accuracy, specificity and precision (Table 1). That is, CNN model yielded an average accuracy, specificity and precision of 0.837, 0.836 and 0.820, respectively, which is better than the next best results of 0.813, 0.775 and 0.780, respectively, obtained by the NB model, on the Training Dataset 1. The CNN model further achieved higher accuracy than other methods on the Training dataset 2. Although NB models also yielded very high sensitivity of 0.855 and 0.914, respectively, the specificity was less than 0.800. The ROC analysis further shows that the models based on CNN outperforms other models on both training datasets (Fig. 3a and b).

3.3 Performance comparison with existing tools

We compared the performance of the deep CNN models with three published methods, EffectiveT3 (Arnold *et al.*, 2009), BPBAac (Wang *et al.*, 2011) and BEAN2 (Dong *et al.*, 2015) on an independent dataset. For this purpose, we prepared this dataset by retrieving proteins that have a pairwise identify of <60% by CD-HIT and 25–60% by BLAST with those in the whole training dataset. The testing samples consisted of 35 T3SEs and 86 non-T3SEs. The results showed that the DeepT3-1 model based on training Dataset 1 achieved a higher sensitivity and accuracy compared with the other three types of methods, while having next best specificity (Table 2 and Fig. 3c). DeepT3-1 achieved the best accuracy value of 0.926, which is 20.7, 6.3 and 7.4% higher than EffectiveT3, BPBAac and BEAN2. The good performance of DeepT3 was also confirmed in another independent dataset from a plant pathogen *P.syringae*. The 85 Type III effectors and 14 non-Type III effectors that are not included in all models were collected from Baltrus *et al.*'s (2011) work. It indicated that the DeepT3-1 model still performed much better than EffectiveT3, BPBAac and BEAN2. DeepT3-1 gave the highest accuracy of 0.884 (Table 2) and highest AUC value of 0.838 (Fig. 3d). Overall, these results suggest that deep-learning-based method outperforms the state-of-the-art methods for the prediction of T3SEs.

3.4 Practical performance of DeepT3 on candidate effectors reported in the literature

Finally, we used some candidate effectors to assess the practical feasibility of DeepT3 models. Yang *et al.* (2010) obtained 17 effectors from rhizobial bacterial, which have been verified as T3SEs by wet-bench experiments. Deng *et al.* (2012) have experimentally confirmed 22 new T3SEs from EPEC, including C_0814/NleJ and LifA. Dong *et al.* (2013) have collected 24 newly identified Type III effectors from literature.

Through a comprehensive survey of ~20 000 bacterial genomes, Hu *et al.* (2017) have identified 174 non-redundant T3SSs from 109 genera and 5 phyla. They subsequently have extended their search to identify Type III effectors, resulting in 519 experimentally validated non-redundant effectors. By excluding the proteins occurring in our dataset and with <100 amino acids, the remaining T3SEs were predicted by DeepT3 (Table 3). The DeepT3-1 model correctly predicted 11 from 13, 17 from 19, 18 from 23 and 213 from 259 candidate T3SEs, respectively. The DeepT3-2 model also showed

Table 1. The performance of various classifiers based on the 5-fold cross-validation tests

Method	Training	PRE	SN	SP	*F*-score	ACC	MCC	AUC
CNN	1	**0.823 ± 0.018**	0.837 ± 0.035	**0.836 ± 0.018**	**0.830 ± 0.023**	**0.837 ± 0.019**	**0.674 ± 0.040**	**0.901 ± 0.014**
	2	**0.875 ± 0.021**	0.786 ± 0.063	**0.945 ± 0.014**	**0.821 ± 0.030**	**0.889 ± 0.014**	**0.746 ± 0.035**	**0.933 ± 0.030**
NB	1	0.780 ± 0.051	**0.855 ± 0.045**	0.775 ± 0.080	0.814 ± 0.016	0.813 ± 0.025	0.634 ± 0.041	0.881 ± 0.017
	2	0.468 ± 0.016	**0.914 ± 0.054**	0.487 ± 0.025	0.619 ± 0.024	0.628 ± 0.020	0.396 ± 0.055	0.833 ± 0.035
RF	1	0.801 ± 0.067	0.746 ± 0.095	0.823 ± 0.081	0.767 ± 0.036	0.786 ± 0.028	0.579 ± 0.063	0.869 ± 0.018
	2	0.706 ± 0.038	0.660 ± 0.034	0.893 ± 0.033	0.682 ± 0.036	0.782 ± 0.029	0.491 ± 0.036	0.766 ± 0.014
SVM	1	0.712 ± 0.052	0.756 ± 0.060	0.718 ± 0.077	0.731 ± 0.038	0.736 ± 0.039	0.476 ± 0.076	0.806 ± 0.033
	2	0.700 ± 0.079	0.559 ± 0.072	0.877 ± 0.049	0.617 ± 0.051	0.772 ± 0.031	0.467 ± 0.070	0.807 ± 0.028
NN	1	0.750 ± 0.068	0.742 ± 0.104	0.762 ± 0.113	0.739 ± 0.041	0.753 ± 0.036	0.513 ± 0.067	0.844 ± 0.014
	2	0.656 ± 0.082	0.593 ± 0.089	0.841 ± 0.059	0.618 ± 0.057	0.759 ± 0.038	0.448 ± 0.085	0.787 ± 0.022
KNN	1	0.799 ± 0.020	0.728 ± 0.050	0.833 ± 0.028	0.761 ± 0.026	0.783 ± 0.018	0.566 ± 0.035	0.860 ± 0.027
	2	0.683 ± 0.064	0.507 ± 0.060	0.883 ± 0.029	0.581 ± 0.055	0.759 ± 0.031	0.427 ± 0.076	0.782 ± 0.026

The bold values indicate the best prediction results.

Table 2. Comparison results of our DeepT3, EffectiveT3, BPBAac and BEAN2

Method	PRE	SN	SP	*F*-score	ACC	MCC	AUC
Independent dataset							
DeepT3-1	0.825	**0.943**	0.919	**0.880**	**0.926**	**0.830**	**0.974**
DeepT3-2	0.643	0.771	0.825	0.701	0.810	0.569	0.896
Effective T3	0.542	0.839	0.741	0.658	0.767	0.521	0.803
BPBAac	**0.944**	0.548	**0.988**	0.694	0.871	0.656	0.902
BEAN2	0.674	0.935	0.835	0.784	0.862	0.706	0.865
P.syringae dataset							
DeepT3-1	0.905	0.962	0.429	0.932	**0.884**	**0.472**	**0.838**
DeepT3-2	**0.913**	0.924	0.500	0.918	0.860	0.437	0.763
Effective T3	0.906	0.906	0.428	0.906	0.838	0.334	0.810
BPBAac	0.875	0.494	**0.571**	0.631	0.505	0.046	0.562
BEAN2	0.883	**0.988**	0.083	**0.938**	0.884	0.271	0.607

The bold values indicate the best prediction results.

Table 3. The prediction results of different methods on the newly effectors

Method	Yang *et al.*	Deng *et al.*	Dong *et al.*	Hu *et al.*
DeepT3-1	11/13	17/19	18/23	213/259
DeepT3-2	10/13	15/19	17/23	194/259
Effective T3	11/13	16/19	13/23	209/259
BPBAac	11/13	9/19	3/23	150/259
BEAN2	11/13	16/19	16/23	203/259

The bold values indicate the best prediction results.

high accuracy. Meanwhile, these newly effectors were also predicted by EffectiveT3, BPBAac and BEAN2. The comparison results of DeepT3 with them are listed in Table 3. It further proves the strong power of DeepT3 in identifying T3SS effectors.

4 Conclusion

We presented a deep convolution neural network to directly classify a protein sequence into Type III effectors or non-Type III effectors. To our knowledge, this is the first deep learning method that can directly identify T3SEs from the primary sequences rather accurately without using feature input. DeepT3 skips steps of feature extraction from the sequence objects, and feature selection for determining 'effective features', leveraged by the use of a CNN. In addition, the deep CNN model in DeepT3 accurately learned known motifs of the leader sequences *de novo* during model training, suggesting that the high performance of DeepT3 resulted from its ability to capture *de facto* sequence features affecting effectors secretion.

The performance on the benchmark datasets and the comparative study between the CNN and other machine learning methods proved the deep learning as a better predictor. And on the independent test datasets, the CNN model is more accurate in identifying T3SE of target proteins and clearly demonstrates the advantages of the deep CNN method for T3SEs prediction. We anticipate that DeepT3 will be used as a powerful tool for hypothesis-driven experimental studies on novel T3SS effectors and their biological functions.

Acknowledgements

We would like to acknowledge the members of Center for Bioinformatics and Systems Biology at Wake Forest School of Medicine.

Funding

This work has been supported by the National Natural Science Foundation of China [No. 21803045].

Conflict of Interest: none declared.

References

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.

Almagro Armenteros,J.J. *et al.* (2017) DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, **33**, 3387–3395.

Angermueller,C. *et al.* (2016) Deep learning for computational biology. *Mol. Syst. Biol.*, **12**, 878.

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

Baltrus,D.A. *et al.* (2011) Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates. *PLoS Pathog.*, **7**, e1002132.

Bendtsen,J.D. *et al.* (2004) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng., Des. Sel.*, **17**, 349–356.

Burkinshaw,B.J. and Strynadka,N.C. (2014) Assembly and structure of the T3SS. *Biochim. Biophys. Acta*, **1843**, 1649–1663.

Casper-Lindley,C. *et al.* (2002) Direct biochemical evidence for type III secretion-dependent translocation of the AvrBs2 effector protein into plant cells. *Proc. Natl. Acad. Sci. USA*, **99**, 8336–8341.

Cornelis,G.R. (2006) The type III secretion injectisome. *Nat. Rev. Microbiol.*, **4**, 811–825.

Costa,S.C. *et al*. (2012) A new means to identify type 3 secreted effectors: functionally interchangeable class IB chaperones recognize a conserved sequence. *mBio*, **3**, e00243-11.

Crawford,J.A. and Kaper,J.B. (2002) The N-terminus of enteropathogenic Escherichia coli (EPEC) Tir mediates transport across bacterial and eukaryotic cell membranes. *Mol. Microbiol.*, **46**, 855–868.

Deng,W. *et al*. (2012) Quantitative proteomic analysis of type III secretome of enteropathogenic Escherichia coli reveals an expanded effector repertoire for attaching/effacing bacterial pathogens. *Mol. Cell. Proteomics*, **11**, 692–709.

Desvaux,M. *et al*. (2009) Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol.*, **17**, 139–145.

Dong,X. *et al*. (2015) BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. *Database*, **2015**, bav064.

Dong,X.B. *et al*. (2013) Using weakly conserved motifs hidden in secretion signals to identify type-III effectors from bacterial pathogen genomes. *PLoS One*, **8**, e56632.

Engel,J. and Balachandran,P. (2009) Role of Pseudomonas aeruginosa type III effectors in disease. *Curr. Opin. Microbiol.*, **12**, 61–66.

Gupta,S. *et al*. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.

He,S.Y. *et al*. (2004) Type III protein secretion mechanism in mammalian and plant pathogens. *Biochim. Biophys. Acta*, **1694**, 181–206.

Heffernan,R. *et al*. (2017) Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, **33**, 2842–2849.

Hu,Y.M. *et al*. (2017) A global survey of bacterial type III secretion systems and their effectors. *Environ. Microbiol.*, **19**, 3879–3895.

Huang,Y. *et al*. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Jurtz,V.I. *et al*. (2017) An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics*, **33**, 3685–3690.

Kelley,D.R. *et al*. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.

Lee,S.H. and Galan,J.E. (2003) InvB is a type III secretion-associated chaperone for the Salmonella enterica effector protein SopE. *J. Bacteriol.*, **185**, 7279–7284.

Liu,Q. *et al*. (2018) Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics*, **34**, 732–738.

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

Luo,J. *et al*. (2015) A sequence-based two-level method for the prediction of type I secreted RTX proteins. *Analyst*, **140**, 3048–3056.

Ma,W. and Guttman,D.S. (2008) Evolution of prokaryotic and eukaryotic virulence effectors. *Curr. Opin. Plant Biol.*, **11**, 412–419.

Min,S. *et al*. (2017) Deep learning in bioinformatics. *Brief. Bioinform.*, **18**, 851–869.

Miotto,R. *et al*. (2017) Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform* , doi: 10.1093/bib/bbx044.

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

Pedregosa,F., *et al*. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

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

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

Sato,Y. *et al*. (2011) Meta-analytic approach to the accurate prediction of secreted virulence effectors in gram-negative bacteria. *BMC Bioinformatics*, **12**, 442.

Schesser,K. *et al*. (1996) Delineation and mutational analysis of the Yersinia pseudotuberculosis YopE domains which mediate translocation across bacterial and eukaryotic cellular membranes. *J. Bacteriol.*, **178**, 7227–7233.

Schroeder,G.N. and Hilbi,H. (2008) Molecular pathogenesis of Shigella spp.: controlling host cell signaling, invasion, and death by type III secretion. *Clin. Microbiol. Rev.*, **21**, 134–156.

Singh,R. *et al*. (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, **32**, i639–i648.

Sory,M.P. *et al*. (1995) Identification of the YopE and YopH domains required for secretion and internalization into the cytosol of macrophages, using the cyaA gene fusion approach. *Proc. Natl. Acad. Sci. USA*, **92**, 11998–12002.

Szalkai,B. and Grolmusz,V. (2018) SECLAF: a webserver and deep neural network design tool for hierarchical biological sequence classification. *Bioinformatics*, 2487–2489.

Tay,D.M. *et al*. (2010) T3SEdb: data warehousing of virulence effectors secreted by the bacterial Type III Secretion System. *BMC Bioinformatics*, **11** (**Suppl 7**), S4.

Tomalka,A.G. *et al*. (2012) A translocator-specific export signal establishes the translocator-effector secretion hierarchy that is important for type III secretion system function. *Mol. Microbiol.*, **86**, 1464–1481.

UniProt,C. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.

Vacic,V. *et al*. (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, **22**, 1536–1537.

Veltri,D. *et al*. (2018) Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 2740–2747.

Wang,M. *et al*. (2018) DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res.*, **46**, e69.

Wang,S. *et al*. (2016) AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics*, **32**, i672–i679.

Wang,Y. *et al*. (2011) High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics*, **27**, 777–784.

Yang, X.J. *et al*. (2013) Effective identification of gram-negative bacterial type III secreted effectors using position-specific residue conservation profiles, *PLoS One*, **8**, e84439.

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