Briefings in Bioinformatics, 23(1), 2022, 1–11

https://doi.org/10.1093/bib/bbab420
Problem Solving Protocol

T4SEfinder: a bioinformatics tool for genome-scale prediction of bacterial type IV secreted effectors using pre-trained protein language model

Yumeng Zhang†, Yangming Zhang†, Yi Xiong, Hui Wang, Zixin Deng, Jiangning Song and Hong-Yu Ou

Corresponding authors: Hong-Yu Ou, State Key Laboratory for Microbial Metabolism, Joint International Laboratory on Metabolic & Developmental Sciences and School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai, 200030, China. E-mail: hyou@sjtu.edu.cn; Jiangning Song, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Victoria 3800, Australia. E-mail: Jiangning.Song@monash.edu
†Yumeng Zhang and Yangming Zhang contributed equally to this work.

Abstract

Bacterial type IV secretion systems (T4SSs) are versatile and membrane-spanning apparatuses, which mediate both genetic exchange and delivery of effector proteins to target eukaryotic cells. The secreted effectors (T4SEs) can affect gene expression and signal transduction of the host cells. As such, they often function as virulence factors and play an important role in bacterial pathogenesis. Nowadays, T4SE prediction tools have utilized various machine learning algorithms, but the accuracy and speed of these tools remain to be improved. In this study, we apply a sequence embedding strategy from a pre-trained language model of protein sequences (TAPE) to the classification task of T4SEs. The training dataset is mainly derived from our updated type IV secretion system database SecReT4 with newly experimentally verified T4SEs. An online web server termed T4SEfinder is developed using TAPE and a multi-layer perceptron (MLP) for T4SE prediction after a comprehensive performance comparison with several candidate models, which achieves a slightly higher level of accuracy than the existing prediction tools. It only takes about 3 minutes to make a classification for 5000 protein sequences by T4SEfinder so that the computational speed is qualified for whole genome-scale T4SEs detection in pathogenic bacteria. T4SEfinder might contribute to meet the increasing demands of re-annotating secretion systems and effector proteins in sequenced bacterial genomes. T4SEfinder is freely accessible at https://tool2-mml.sjtu.edu.cn/T4SEfinder_TAPE/.

Key words: Type IV secreted effectors; pre-trained language model; sequence analysis; deep learning

Introduction

Type IV secretion systems (T4SSs) are multiprotein nanomachines widely distributed in both Gram-negative and Gram-positive bacteria [1]. According to the transported substrates and their functions, T4SSs can be divided into the following three categories: conjugation systems, DNA-uptake and -release systems and effector translocator systems [2]. Unlike conjugative [3] and transformation systems mainly related to horizontal

Yumeng Zhang is a master student in the School of Life Sciences & Biotechnology, Shanghai Jiao Tong University.
Yangming Zhang is a master student in the School of Life Sciences & Biotechnology, Shanghai Jiao Tong University.
Yi Xiong is an associate professor in the School of Life Sciences & Biotechnology, Shanghai Jiao Tong University.
Hui Wang is a professor in the State Key Laboratory of Pathogens and Biosecurity, Beijing Institute of Microbiology and Epidemiology.
Zixin Deng is a professor in the School of Life Sciences & Biotechnology, Shanghai Jiao Tong University.
Jiangning Song, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Victoria 3800, Australia.
Hong-Yu Ou is a professor in the School of Life Sciences & Biotechnology, Shanghai Jiao Tong University.
Submitted: 14 July 2021; **Received (in revised form):** 31 August 2021

gene transfer, another type of T4SSs that delivers effector proteins to target cells plays a crucial role in the virulence of pathogens [4]. Pathogenic bacteria employ T4SSs to translocate type IV secreted effectors (T4SEs) or protein-DNA complexes into eukaryotic cytoplasm that disrupt the signal transduction of the host cells and cause various diseases [1, 3, 5]. Previous studies have investigated specific T4SSs in *Legionella pneumophila*, *Coxiella burnetii*, *Brucella melitensis* and other pathogenic bacteria that transfer T4SEs into human cells [6]. The effector proteins secreted by Dot/Icm T4SS in *L. pneumophila* are reported to target cellular pathways controlling the intermembrane transport [7] and help protect the pathogens.

Identification of T4SEs that can be translocated into the host cells, with their pathogenic mechanism, has been attracting an increasing interest due to the clinical significance. Many experimental approaches [8] have been designed to discover the existence of effector proteins in the eukaryotic cytoplasm, including enzyme fusion, immunofluorescence detection, and other proteomics methods. Although experiments can provide reliable inference on effector proteins, it takes considerable time and cost for the verification of all T4SE candidates.

After functional T4SEs were experimentally discovered progressively, the biological features of verified effector proteins allowed the development of computational approaches for the prediction of T4SEs. Apart from traditional sequence similarity-based methods, various machine learning and deep learning algorithms [9–17] were utilized to train classification models to distinguish potential secreted effectors from numerous non-effector proteins. Amino acid composition, position-specific scoring matrix (PSSM) [18] and structural information are commonly extracted to characterize the representation for protein sequences, while support vector machine (SVM) [19], random forest (RF) [20] and convolutional neural network (CNN) [21] might be the most popular machine learning classifiers for T4SE prediction. PSSM is generally acknowledged to capture the conservation patterns in biological sequences, and accordingly prediction tools that use PSSM as the features usually outperform other strategies. However, the generation of PSSM requires extensive search of similar sequences in large protein sequence databases like Uniref50 [22] by PSI-BLAST [18]. The currently available bioinformatics tools are incapable of making both rapid and accurate predictions for T4SEs in pathogens. An online annotation tool can facilitate the screening of putative T4SEs and the characterization of their pathogenic mechanism; however, most of the recent studies, especially those that have applied deep learning techniques, did not provide such a large-scale web service.

Meanwhile, the accelerated development of deep learning methods has introduced a novel way to explore and interpret protein sequences [23] through the analogy between natural language and biological language. Statistical language models [24] can estimate the distribution of each amino acid over the protein sequence by learning from the contextual information, which is consistent with the intuition of amino acid interaction. Transformer [25], an innovative model architecture based on the attention mechanism, has established a state-of-the-art approach in language modeling. Bidirectional Encoder Representations from Transformers (BERT) [26] has provided a standard pipeline in natural language processing containing unsupervised pre-training and fine-tuning on downstream tasks. In addition to the success in machine translation [27] and question answering systems [28], pre-trained language models have achieved remarkable success to explore and model the biological sequence data [29–34], especially in protein structure prediction.

In this study, we formulated the prediction of type IV secreted effectors (T4SEs) as a particular downstream task based on the pre-trained language model of protein sequences and accordingly developed T4SEfinder, a novel genome-scale tool for identifying T4SEs. It aims to take the advantage of long-time pre-training to capture the biological representation. In particular, its support of high-throughput computational classification might allow us to generate new insights about the taxonomy distribution of T4SEs and their functions.

Materials and methods

Data integration of experimentally verified T4SEs

We have recently updated the type IV secretion system database SecReT4 v2.0 [35] (Supplementary Table S1) and then used the experimentally verified T4SEs as the positive samples in the training dataset. We have also integrated the positive training samples used in three recent studies on T4SE prediction [12, 15, 16]. It is noteworthy that most of the effector proteins in the SecReT4 database could be found in the common training set; however, 121 newly added T4SEs are not similar to the T4SEs in the previous dataset (BLASTp identities $<$60%). Therefore, we obtained the final 518 T4SEs in our training dataset after using CD-HIT [36] to remove the homologous sequences (BLASTp identities $\geq$60%). The taxonomy composition of the 518 T4SEs in the training dataset can be found in Supplementary Table S2. For the negative samples in our training dataset, the entire set of non-effectors in two studies [9, 37] were gathered together, resulting in 1584 non-effector proteins following the same procedures to eliminate the sequence redundancy. The complete flowchart to construct the dataset is shown in Supplementary Figure S1.

For benchmark testing, the independent test dataset in this study consists of 20 T4SEs and 150 negative non-effectors after removal of similar sequences against the training dataset according to the sequence identity threshold (60%). The positive T4SEs were obtained from the UniProt [38] database and S4TE [39], while the negative samples were derived from *Vibrio parahaemolyticus* serotype O3: K6 strain RIMD 2210633 [40], similar to Bastion4 [15], DeepT4 [12] and CNNT4SE [16]. The performance of the developed T4SEfinder method was benchmarked against several existing tools on the same test dataset, such as Bastion4 [15] and CNN-T4SE [16].

The deep learning architecture of T4SEfinder

T4SEfinder implements an end-to-end prediction process starting from protein sequences in the FASTA format to the predicted probabilities of T4SEs. Figure 1 illustrates the major functional modules inside the deep learning-based prediction model. As can be seen, the feature extraction module employs a protein-encoding method based on a pre-trained BERT model from Tasks Assessing Protein Embeddings (TAPE) [29], termed TAPEBert in this study. Pfam [41] was selected as the pre-training corpus for TAPE, and the objective of self-supervision training is to predict the masked amino acid residues more accurately. An intuitive option for further classification is to use a multi-layer perceptron (MLP) [42] that outputs prediction results of the target sequences. The parameters of the MLP model obtained from 5-fold cross-validation jointly determine the outcome of T4SEfinder. We also attempted to select support vector machine (SVM) as the downstream classifier following the encoder so that we have two pre-trained-based models, TAPEBert_MLP and TAPEBert_SVM.

The composition of the position-specific scoring matrix (PSSM) and convolutional neural network (CNN) turns out to be

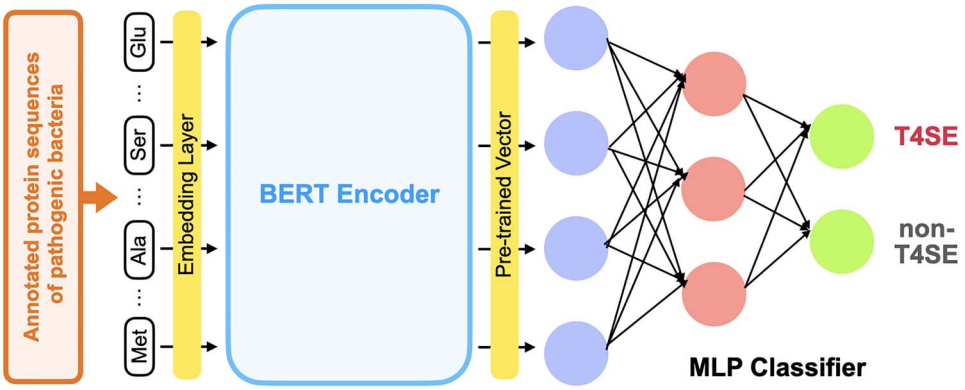

Figure 1. The model architecture of T4SEfinder (TAPEBert_MLP) to predict T4SEs. The TAPEBert_MLP model in T4SEfinder combines the TAPEBert pre-trained protein language model and a multi-layer perceptron (MLP) as the major architecture. The input of the model is protein sequences of arbitrary length that are composed of 20 common amino acids. The pre-trained BERT encoder is used to extract biological features, while MLP generates the final classification result.

a sensible choice for T4SE prediction [16]. Accordingly, we also constructed the PSSM_CNN model. The C-terminus of T4SEs has shown significant preference in amino acids (Supplementary Figure S2), thus potentially affecting their biological functions. A Bi-directional Long Short Term Memory Network (BiLSTM) [43] is adopted to incorporate the global and local representation of protein sequence from the pre-trained language model and the PSSM profile of the last 30 amino acid residues at the C-terminus. The model is referred to as HybridBiLSTM due to the fusion of distinct features. The prediction frameworks of PSSM_CNN and HybridBiLSTM are displayed in Supplementary Figures S3 and S4, respectively.

Sequence encoding strategies in this study

We can decompose the pipeline of T4SE prediction into the feature extraction module and the classification module. The prediction for T4SEs can be symbolized in the following formula:

$$\mathrm{P}\left(\mathrm{T4SE} \mid \mathrm{x}\right) = g\left(\mathrm{f_{enc}}\left(\mathrm{x}\right)\right) \tag{1}$$

where x represents the input protein sequence, $\mathrm{f_{enc}}$, g stand for the models used in the sequence encoding and classification, respectively.

At the stage of feature extraction, the PSSM profiles generated by PSI-BLAST and the pre-trained protein language model lead to two distinct encoding strategies for protein sequence, which are briefly introduced below.

Position-specific scoring matrix

In this study, we perform PSI-BLAST search against the UniprotKB/Swiss-Prot database with the default parameters (*i.e.*, e-value = 10 and num_iterations = 3) to find the distant evolutionary relationships of the protein sequences. We assume that PSSM(x) is a L × 20 matrix characterizing the output of PSI-BLAST. f_{enc}^{PSSM} transforms the original matrix into a 20 × 20 matrix and can be expressed as follows:

$$f_{enc}^{\mathrm{PSSM}} = \sum_{i=1}^{L} \mathrm{PSSM_i(x)} \times I\left(\mathrm{x_i} == a\right) \ \text{for each a} \in \mathrm{A} \tag{2}$$

where L is the length of a protein sequence; I(·) is the indicator function; A represents the set of 20 common amino acid residues; i denotes the row number of the PSSM profile and the position in the protein sequence.

Protein pre-trained language model

We apply a pre-trained language model for protein sequence, namely TAPE [29], as a feature extraction strategy. This protein language model encodes the protein sequences and outputs a 768-dimensional embedding vector:

$$f_{\mathrm{enc}}^{\mathrm{LM}} = \mathrm{TAPEBertEncoder}\left(\mathrm{x}\right) \tag{3}$$

where the TAPEBert encoder comprises input embedding, positional encoding and stacked transformers with the multi-head attention, layer norm layers and residual connections [25].

In addition, the loss function of this masked language model is optimized at the pre-training stage of protein language modeling.

$$\mathcal{L}_{MLM} = -\sum_{\hat{x} \in m(x)} \log p\left(\hat{x} | x_{\backslash m(x)}\right) \tag{4}$$

where m(x) and $x_{\backslash m(x)}$ denote the masked amino acid residues from the entire protein sequence and the rest sequence, respectively.

Model training

We introduce the architectures of four mentioned models (i.e., TAPEBert_MLP, TAPEBert_SVM, PSSM_CNN and HybridBiLSTM), and discuss the training process in this section.

$$\mathrm{P_{TAPEBert\_MLP}}\left(\mathrm{T4SE}|\mathrm{x}\right) = \mathrm{softmax}\left(\mathrm{MLP}\left(f_{enc}^{LM}(\mathrm{x})\right)\right),$$

$$\mathrm{MLP}\left(\cdot\right) = \mathrm{fc_2}\left(\mathrm{Dropout}\left(\mathrm{ReLU}\left(\mathrm{fc_1}\left(\cdot\right)\right)\right)\right). \tag{5}$$

where softmax and ReLU denote the activation functions, Dropout represents the dropout layer to avoid overfitting, $\mathrm{fc_1}$ and $\mathrm{fc_2}$ are two distinct fully connected layers.

$$\mathrm{P_{TAPEBert\_SVM}}\left(\mathrm{T4SE}|\mathrm{x}\right) = \mathrm{SVM}\left(f_{enc}^{LM}(\mathrm{x})\right),$$

$$\mathrm{SVM}\left(\cdot\right) = \mathrm{sigmoid}\left(w^T\phi\left(\cdot\right) + b\right). \tag{6}$$

where sigmoid denotes the sigmoid function for binary classification, w and b are weights and bias, respectively, $\phi(\cdot)$ represents the Gaussian kernel function.

$$P_{\text{PSSM\_CNN}}(\text{T4SE}|x) = \text{CNN}\left(f_{\text{enc}}^{\text{PSSM}}(x)\right),$$

$$\text{CNN}(\cdot) = \text{softmax}\left(\text{MLP}\left(\text{ConvBlock}_2\left(\text{ConvBlock}_1(\cdot)\right)\right)\right),$$

$$\text{ConvBlock}(\cdot) = \text{MaxPool}\left(\text{ReLU}\left(\text{BN}\left(\text{Conv}(\cdot)\right)\right)\right). \quad (7)$$

where MaxPool, BN, Conv denote the max pooling layer, batch normalization [44] layer and convolution layer in the convolutional neural network, respectively.

$$P_{\text{HybridBiLSTM}}(\text{T4SE}|x) = f_{\text{cls}}\left(\text{Attention}(Q, K, V)\right),$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

$$K = V = \text{BiLSTM}\left(f_{\text{enc}}^{*}(x_{C30}), [h_0, h_0]\right),$$

$$Q = \text{Dropout}(K),$$

$$f_{\text{enc}}^{*}(\cdot) = \text{Conv}\left(\text{Normalize}\left(\text{PSSM}(\cdot)\right)\right)$$

$$h_0 = \text{Dropout}\left(\text{fc}\left(f_{\text{enc}}^{\text{LM}}(x)\right)\right). \quad (8)$$

where $f_{\text{cls}}, f_{\text{enc}}^{*}, f_{\text{enc}}^{\text{LM}}$ represent the classifier, the encoding function for the last 30 amino acid residues at C-terminus, and the pre-trained language model, respectively. d_k equals the dimension of the hidden state, while $[h_0, h_0]$ denotes the concatenation of the initial hidden state.

All of our models were implemented with the PyTorch deep learning framework, and the cross-entropy loss function was adopted to train the classifiers. The Adam optimizer with a cosine annealing schedule helped to improve the prediction performance. Dropout, weight decay and an early stopping strategy for monitoring the validation F1-score with the patience of 15 epochs were employed to prevent overfitting. The hyperparameters of deep learning models used in this study are listed in Supplementary Table S3.

Performance assessment

Six common measures for classification are used to evaluate the performance of T4SEfinder and other prediction tools for T4SEs. These include accuracy (ACC), sensitivity (SN), specificity (SP), precision (PR), F1-score and Matthew's correlation coefficient (MCC) and are formulated below:

$$\text{ACC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}},$$

$$\text{SN} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$

$$\text{SP} = \frac{\text{TN}}{\text{TN}+\text{FP}},$$

$$\text{PR} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$

$$\text{F1}-\text{score} = \frac{2}{1/\text{SN}+1/\text{PR}},$$

$$\text{MCC} = \frac{(\text{TP}\times\text{TN})-(\text{FN}\times\text{FP})}{\sqrt{(\text{TP}+\text{FN})\times(\text{TP}+\text{FP})\times(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})}}. \quad (9)$$

where TP, TN, FP and FN denote the numbers of true positives, true negatives, false positives and false negatives, respectively.

The receiver-operating characteristic (ROC) curve and the precision-recall curve are also effective evaluation methods. ROC curve visualizes the changes in the true-positive rate and false-positive rate in response to the varying discrimination threshold, while the precision-recall curve depicts the tradeoff between the precision and recall for different thresholds. The areas under the curves are termed AUC and AUPRC (equivalent to average precision), respectively. For imbalanced classes, the precision-recall curve is a sensible and more suitable choice for class-imbalance data than the ROC curve.

Results

Application of pre-trained model for protein sequence embedding

Features related to the position-specific scoring matrix (PSSM) are mentioned as typical choices for protein sequence embedding. In this regard, the protein language model has brought new insight into biological sequence classification. We visualized the predictive capabilities of both PSSM and TAPEBert embeddings to distinguish T4SEs from non-effector proteins. We transformed the original PSSM profile by summing up the rows of the same amino acid, and normalizing the data with the sequence length, thereby generating 400-dimension embedding vectors. We also took the average of the last hidden layer in the pre-trained TAPEBert encoder as the sequence embedding. Figure 2 shows a two-dimensional projection of the embedding space using t-SNE [45]. We can see that most of the effector proteins were clustered in a group with either the PSSM or TAPEBert embedding. The clustering result indicates that the application of the pre-trained model in sequence embedding may achieve a comparable result as the state-of-the-art methods based on PSSM profiles.

Performance evaluation on repeated 5-fold cross-validation

We further employed four different machine learning and deep learning models for the identification of T4SEs. TAPEBert_MLP and PSSM_CNN model were designed to compare the performance of models trained using two alternatives of protein sequence embedding. Prediction results with different classifiers following TAPEBert embedding were measured as well. We also inspected the performance of combining the features from the pre-trained model and the PSSM profiles in HybridBiLSTM. For each of the predictors, a procedure of repeated 5-fold cross-validation (10 times) was employed to assess their generalization performances on the validation datasets. The learning curves for TAPEBert_MLP in Supplementary Figure S5 monitored the training loss, validation loss and F1-score during the training progress. The performance evaluation results are documented in Table 1, and the corresponding ROC curves and precision-recall curves are illustrated in Figure 3.

After a comprehensive analysis of the observed performance of different classifiers on repeated 5-fold cross-validation, several important observations could be made to evaluate the predictors. Compared to the CNN model based on the PSSM features, TAPEBert_MLP achieved a better accuracy (90.4 versus 89.9%) and comparable F1-score (0.797 versus 0.794). In addition, another advantage of pre-trained-based models is the reduction in false discovery rate (the precision was about 3% higher). When we assessed the capabilities of the classification head after the TAPEBert embedding, evident improvement (about 1% of accuracy and 0.03 of F1-score) could be achieved through use of the

Figure 2. The two-dimensional projection of the embedding vectors generated by TAPEBert encoder and PSSM profiles using t-SNE. The green and blue points displayed in the figure represent the T4SEs encoded by TAPEBert and PSSM profiles, respectively. The grey points denote the non-effectors in our dataset. The majority of the T4SEs are observed to cluster in a group by both encoding strategies.

Table 1. Performance of various classifiers in this study evaluated by the ten-time repeated 5-fold cross-validation. ACC: Accuracy; SN: sensitivity; SP: specificity; PR: precision; F1: F1-score; MCC: Matthews correlation coefficient

Method	ACC	SN	SP	PR	F1	MCC
TAPEBert_MLP	90.4 ± 1.4%	76.8 ± 4.1%	94.8 ± 1.7%	83.2 ± 4.3%	0.797 ± 0.028	0.736 ± 0.037
TAPEBert_SVM	89.3 ± 1.7%	71.2 ± 4.5%	95.2 ± 1.3%	83.0 ± 4.2%	0.766 ± 0.039	0.701 ± 0.049
PSSM_CNN	89.9 ± 1.8%	79.2 ± 4.3%	93.4 ± 2.5%	80.0 ± 5.6%	0.794 ± 0.033	0.729 ± 0.045
HybridBiLSTM	91.3 ± 1.0%	80.1 ± 4.4%	95.0 ± 1.7%	84.3 ± 3.9%	0.820 ± 0.022	0.764 ± 0.028

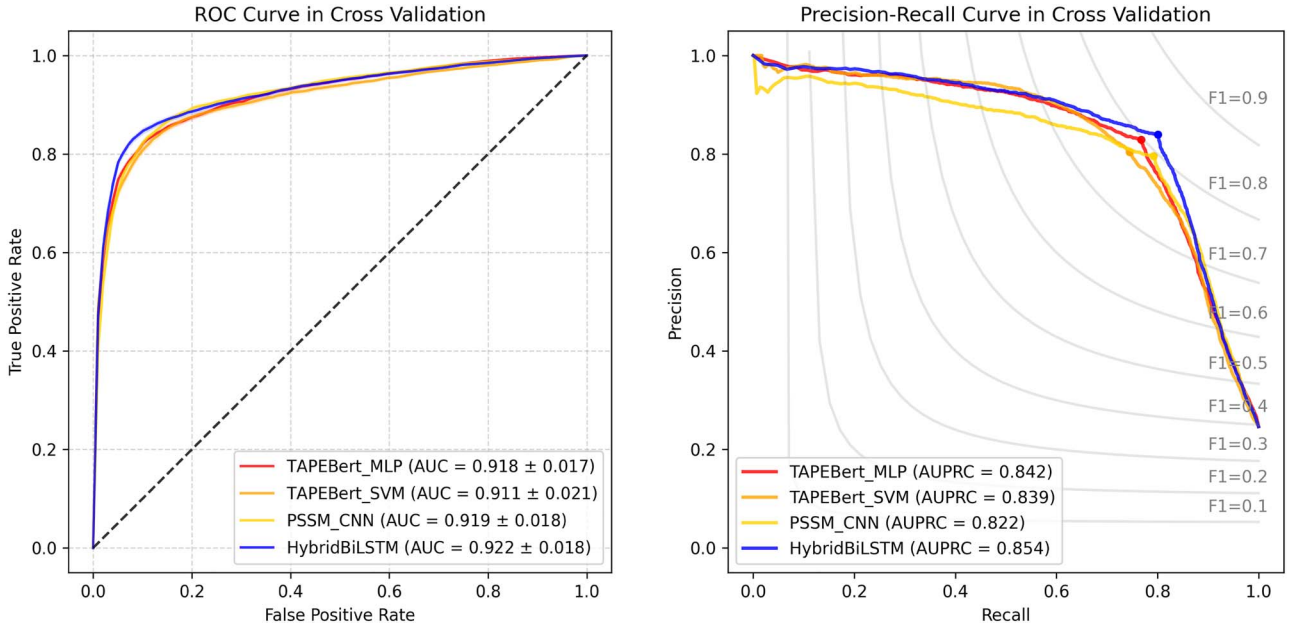

Figure 3. ROC curves and precision-recall curves of the four different models used by T4SEfinder (TAPEBert_MLP, TAPEBert_SVM, PSSM_CNN and HybridBiLSTM) in the ten-time repeated 5-fold cross-validation. In the left panel, the black dashed line denotes the ROC curve with the predicted labels selected randomly. The means and the standard deviations of AUC are listed in the legend. In the right panel, the gray lines represent the contour lines of the F1-score, where the points achieving the best F1-scores are indicated.

Table 2. Performance comparison of the models in T4SEfinder and other existing tools on the independent test dataset. ACC: Accuracy; SN: sensitivity; SP: specificity; PR: precision; F1: F1-score; MCC: Matthews' correlation coefficient

Method	ACC	SN	SP	PR	F1	MCC
T4SEpre_psAac	91.8%	65.0%	95.3%	65.0%	0.650	0.603
T4SEpre_bpbAac	90.0%	70.0%	92.7%	56.0%	0.622	0.570
DeepT4	86.5%	75.0%	88.0%	45.5%	0.566	0.513
BastionX	92.9%	100.0%	92.0%	62.5%	0.769	0.758
CNNT4SE_Vote	98.2%	85.0%	100.0%	100.0%	0.919	0.913
TAPEBert_MLP	96.5%	90.0%	97.3%	81.8%	0.857	0.838
TAPEBert_SVM	95.9%	80.0%	98.0%	84.2%	0.821	0.798
PSSM_CNN	91.8%	90.0%	92.0%	60.0%	0.720	0.693
HybridBiLSTM	95.3%	90.0%	96.0%	75.0%	0.818	0.796

MLP classifier instead of the SVM classifier. By incorporating the TAPEBert embedding and local PSSM features, the HybridBiLSTM model achieved the highest accuracy (91.3%) and F1-score (0.820).

In addition, the selection of protein sequence identity cut-off can change the size of the training dataset, thereby having an influence on the result of cross-validation. To examine the impact of such influence on the model performance, we removed the sequence redundancy in the positive and negative samples according to varying BLASTp identities (25, 30, 40, 50 and 60%), and train the TAPEBert_MLP models through the same repeated 5-fold cross validation process on the resulting datasets. The validation accuracy, F1-score and MCC under different sequence identity thresholds are shown as boxplots in Supplementary Figure S6.

Performance comparison on the independent test set

Several bioinformatics tools are currently available for T4SE prediction. We assessed the performance and generalization ability of our proposed method and the existing classifiers on the independent test dataset. Table 2 presents the performance comparison of our models and another five published predictors.

As can be seen from Table 2, TAPEBert_MLP attained an accuracy of 90.0%, F1-score of 0.957 and MCC of 0.838, respectively. The TAPEBert_MLP model also achieved a better trade-off between the sensitivity and precision based on the independent dataset. The comprehensive superiority of TAPEBert_MLP reflects the state-of-the-art in prediction for T4SEs as a single model. The ROC curves and precision-recall curves in Figure 4 provide an alternative to examine the robustness of our models and other existing tools for T4SE identification.

Whole-genome detection for T4SEs in pathogenic bacteria

Previous experimental studies have characterized the Dot/Icm secretion systems that are associated with effector translocation in *L. pneumophila* (Supplementary Figure S7) and *C. burnetii* [46, 47]. These two species of pathogenic bacteria have become the major source of the T4SEs archived by SecReT4 v2.0 (Supplementary Figure S8). In this study, we have scanned all the annotated proteins in 5 *L. pneumophila* and 5 *C. burnetii* chromosomes using the prediction models of T4SEfinder. The label of each protein sequence is confirmed according to T4SEs collected in SecReT4 [35] database. We compared the prediction performance of the models in Figure 5 and Supplementary Table S4.

Table 3. F1-score and elapsed time of different methods in this study for predicting T4SEs encoded by all the annotated genes of *Coxiella burnetii* RSA 493[a]

Method	Elapsed time[b]	F1-score
TAPEBert_MLP	0:01′04”	0.455
TAPEBert_SVM	0:01′06”	0.468
PSSM_CNN	1:18′33”	0.494
HybridBiLSTM	1:18′43”	0.516
BastionX	5:00′09”	0.412

[a]The *C. burnetii* RSA 493 genome contained 1657 annotated protein-coding genes.
[b]All experiments were performed on a Linux server with two Intel Xeon Gold 5117 CPUs of 14 cores and one GeForce RTX 2080 SUPER (8G) GPU except BastionX, which was used its web server available at https://bastionx.erc.monash.edu/.

HybridBiLSTM achieved the highest AUC and F1-score thanks to the feature fusion. Meanwhile, TAPEBert_MLP attained equivalent accuracy in overall prediction results to PSSM_CNN and approached the leading performance in *L. pneumophila*, which corroborated the effectiveness of the pre-trained model TAPEBert. To evaluate the computational efficiency, we recorded the elapsed time for predicting 1657 annotated proteins in *C. burnetii* RSA 493 in Table 3. As can be seen, TAPEBert embedding helped reduce the prediction time by ~100 times in contrast with the PSSM-based methods. All prediction results for the whole-genome detection of putative T4SEs are available at https://tool2-mml.sjtu.edu.cn/T4SEfinder_TAPE/download.html.

The genome-scale detection of pathogenic bacteria might depict a novel understanding of the relationship between effector proteins and genomic properties. The VirB secretion system in *Brucella* also plays a crucial role in protecting the pathogen [48], but the number of experimentally verified T4SEs is limited. To screen the potential effector proteins out, we visualized the genome maps [49] of *B. melitensis* ATCC 23457 chromosome II (NCBI accession number: NC_012442) with the probability distribution for each annotated gene to encode a T4SE, and labeled the genomic location of putative T4SE-encoding genes in Figure 6.

Implementation of an online tool to facilitate T4SE identification

As an implementation of the developed T4SEfinder method, we have developed an online web server to facilitate the community-wide efforts for T4SE prediction. Figure 7 illustrates how to use the online version of T4SEfinder. Specifically, users can submit one or multiple protein sequences in the FASTA format and choose one of the available prediction models among TAPEBert_MLP, PSSM_CNN and HybridBiLTSM. The

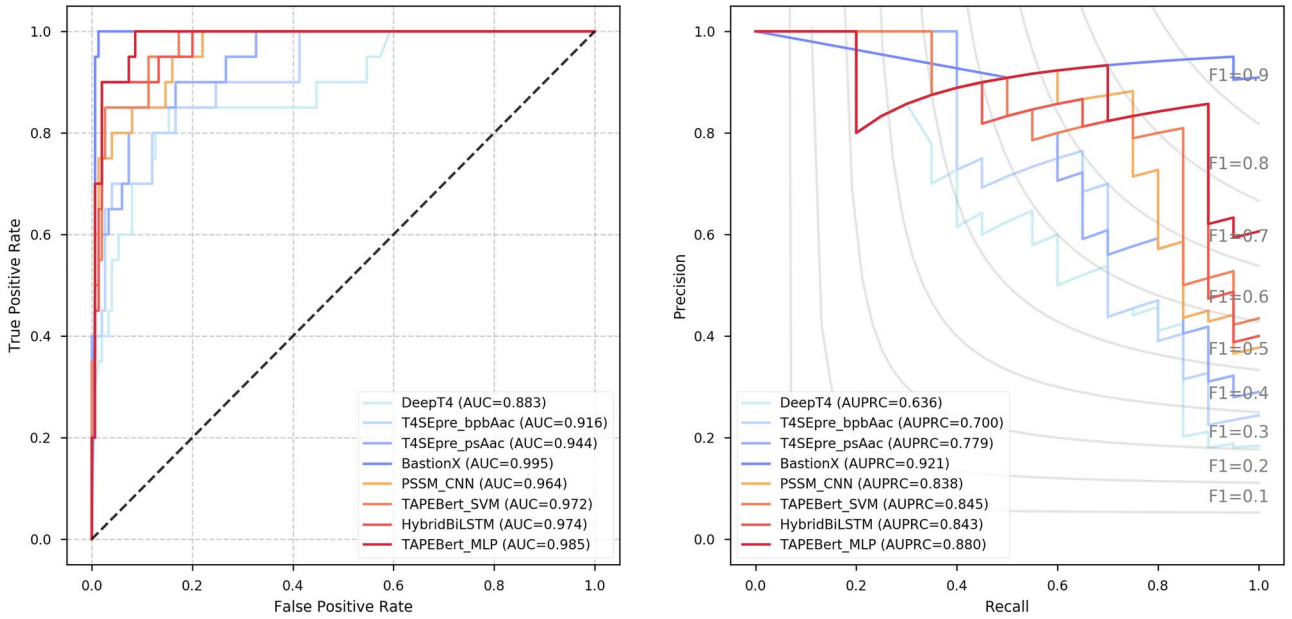

Figure 4. ROC curves and precision-recall curves for the performance evaluation of the models in T4SEfinder and other T4SE prediction tools on the independent test dataset. TAPEBert_MLP outperformed all of the simple classifiers that used amino acid composition as the feature, and approached t prediction results as BastionX, a relatively accurate PSSM-based predictor in AUC (0.985 versus 0.995) and AUPRC (0.880 versus 0.921).

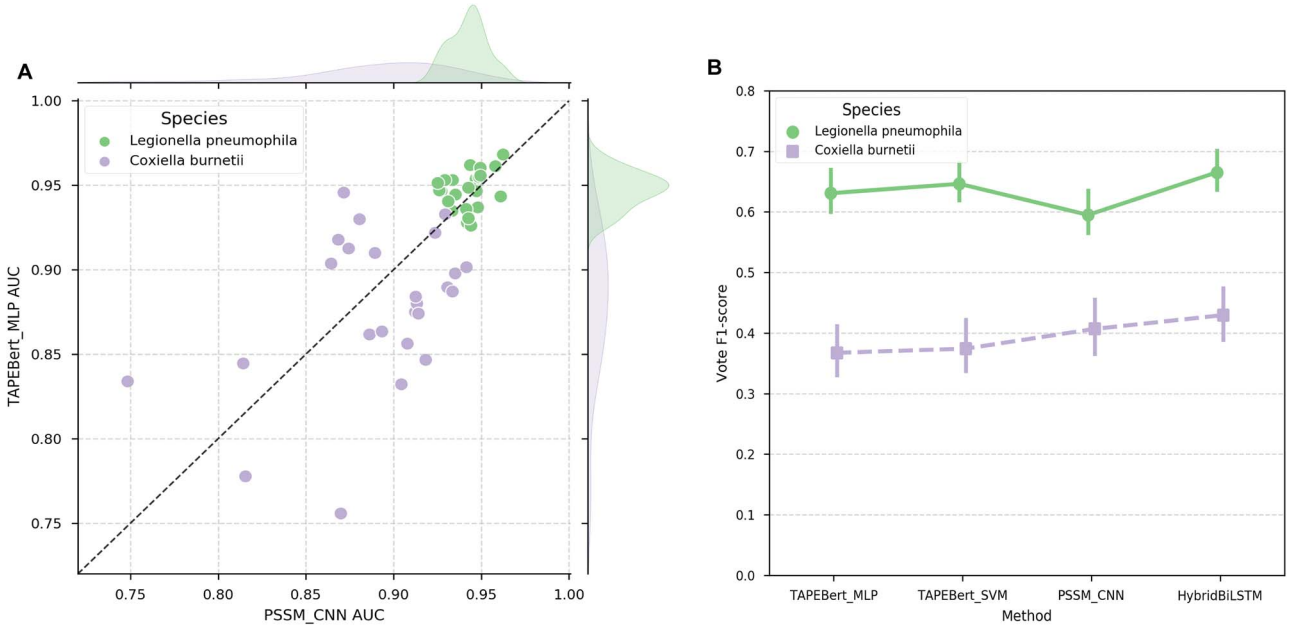

Figure 5. Whole-genome annotation results of T4SEs in *Legionella pneumophila* and *Coxiella burnetii*. A. The AUC distributions resulting from the T4SE prediction in *L. pneumophila* and *C. burnetii* by TAPEBert_MLP and PSSM_CNN are compared (Supplementary Table S4). Each point represents the AUC of a certain strain predicted by TAPEBert_MLP and PSSM_CNN with the same training set. The density plots along the X-axis and Y-axis represent the density distribution of the AUC. TAPEBert_MLP appeared to provide equivalent annotation results as PSSM_CNN. B. The F1-scores obtained by the models in T4SEfinder and the voting strategy are compared. The error bars indicate the standard error in predicting different strains of the pathogen.

prediction result will be generated in the form of a DataTable. For the predicted T4SEs, users can subsequently search the similar effectors or target proteins in the SecReT4 database (Supplementary Figure S9).

Discussion

In this study, we have developed an efficient web server T4SEfinder for genome-scale T4SE prediction, an alternative to the currently available PSSM-based methods. The employed TAPEBert model derived from the long-time pre-training assists to capture the biological representation of protein sequences. The procedure of feature extraction from the TAPEBert model is accelerated by the graphics processing unit (GPU) to realize the rapid prediction. To the best of our knowledge, this study represents the first use case for the pre-trained sequence embedding in the field of bacterial secretion systems and effector proteins.

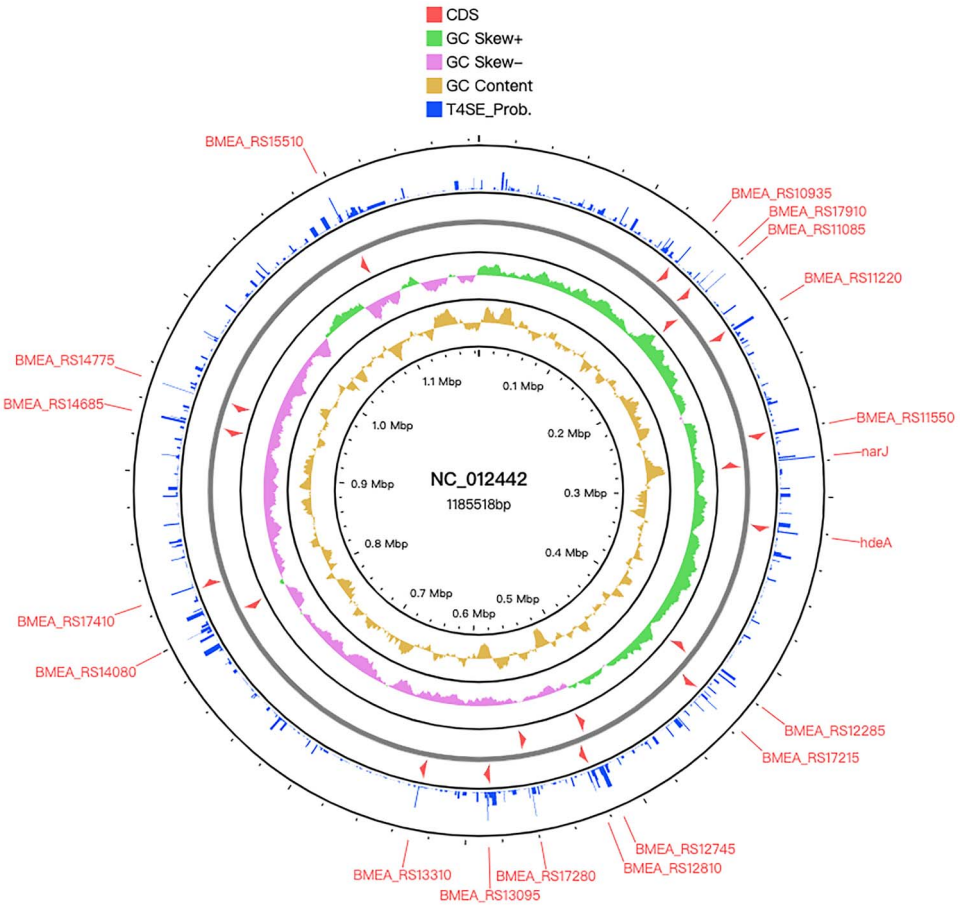

Figure 6. Visualization of *Brucella melitensis* ATCC 23457 chromosome II with the detected T4SEs. The locations of the CDS regions encoding T4SEs predicted by T4SEfinder are visualized by CGView Server. The blue bars indicate the probability distribution to encode T4SEs along the chromosome. The putative T4SE-encoding genes (with the predicted probability over 0.5) are marked out with red arrows as well as the gene names.

To confirm the effectiveness of the pre-trained BERT model, four candidate models for predicting T4SEs were compared through repeated 5-fold cross-validation. We terminated the training process when the validation F1-score had no improvement for 15 epochs in all experiments to avoid overfitting. In addition, it might be difficult to eliminate the difference between sensitivity and specificity due to the different numbers of the positive samples and the negative samples; however, we tried to balance the trade-off between them. TAPEBert_MLP model achieved a slightly higher accuracy, F1-score and MCC than PSSM_CNN, and was considered as a more robust predictor by measuring the standard error of each evaluation metric. Performance improvement could be achieved by HybridLSTM through integrating the PSSM-based features with the pre-trained sequence embedding. The results suggested that leveraging the pre-trained language model of protein sequences contributed to the more precise classification of T4SEs. In the phase of the independent test, a more straightforward method (TAPEBert_MLP) outperformed the best model in cross-validation (HybridLSTM) because of insufficient training data to support a more complex model. TAPEBert_MLP also demonstrated an attractive advantage in computational efficiency for the whole-genome T4SE detection. Therefore, we have selected TAPEBert_MLP as the default prediction algorithm for the T4SEfinder web server.

We have also compared the capability for T4SE prediction of T4SEfinder and other five existing tools on the independent test set. Among the compared tools, T4SEpre_psAac and T4SEpre_bpbAac [10] used the feature of amino acid composition; DeepT4 [12] only encoded the protein sequence at N-terminus and C-terminus by one-hot vectors. Note that these methods did not use the PSSM-based features, thus offering the advantage in computing speed; however, they were not proper genome-scale detection tools for T4SEs due to the poor predictive performance. Bastion4 [15] provided an ensemble model that took into consideration all of the sequence encoding, PSSM profiles and structure description. CNN-T4SE [16] also integrated various types of features to generate a more comprehensive prediction. Although the addition of PSSM encoding has helped to improve the accuracy significantly for Bastion4 and CNN-T4SE, the running time of PSI-BLAST is a tricky problem for large-scale prediction. In comparison with the other prediction tools, T4SEfinder (TAPEBert_MLP) achieves a better balance between the prediction accuracy and the computing speed so that we can regard it as the state-of-the-art approach in genome-scale detection for T4SEs.

In the near future, T4SEfinder will be further developed and upgraded to maintain the prediction efficiency, and pursue higher accuracy as well. Ensemble learning-based tools such as Bastion4 has promoted the model performance

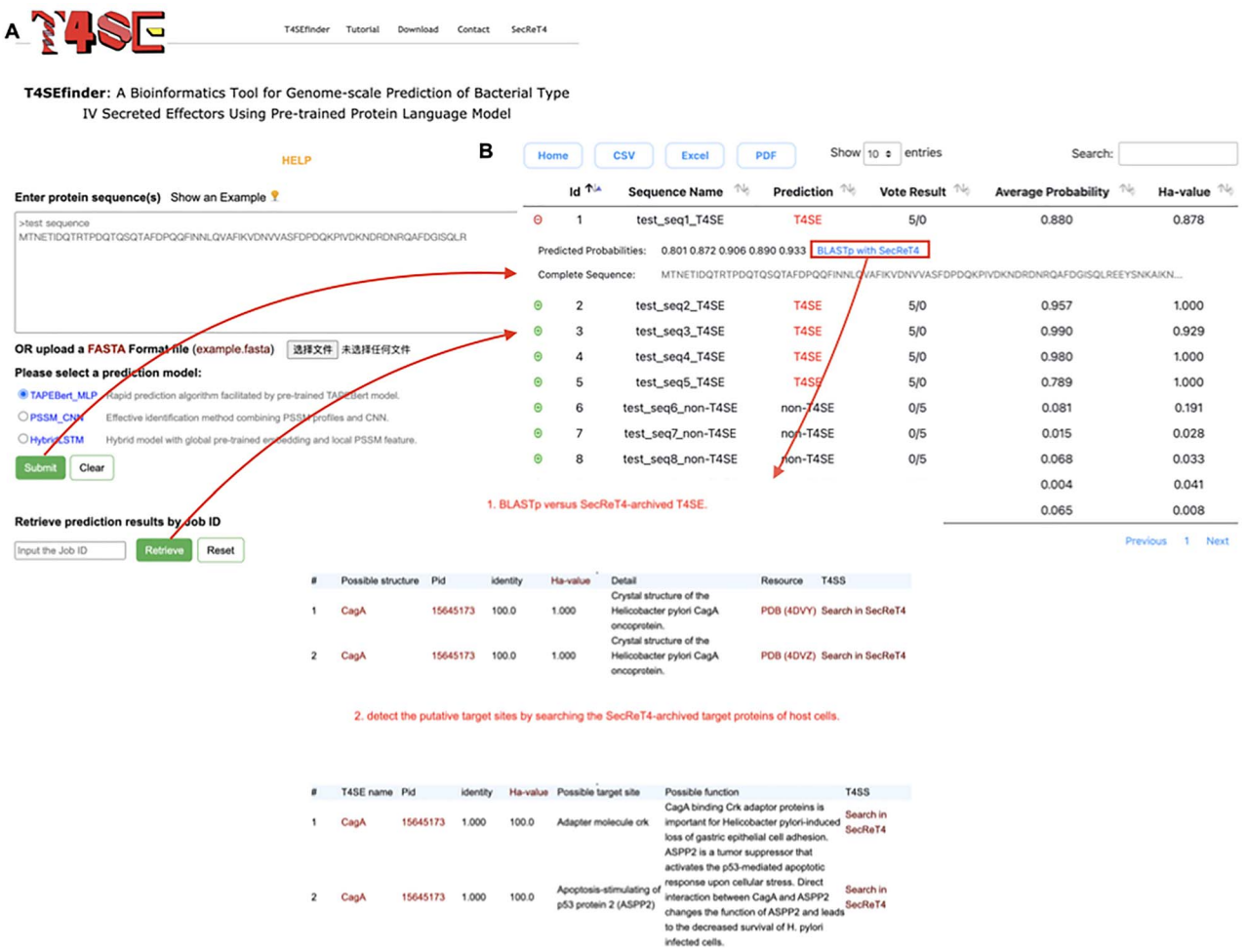

Figure 7. User interface of the web service of T4SEfinder. **A**. Users can upload a FASTA formatted protein sequence(s) and select the corresponding model to predict T4SEs. They can also retrieve the completed task according to the Job ID. **B.** The web page for prediction results: Show both the predicted probabilities and the similarity (Ha-value) compared by known T4SEs. Users can download the results in the csv or excel format, and conduct subsequent analysis for the putative T4SEs. **C.** The webpage for subsequent analysis: Search the similar effectors and target proteins in the SecReT4 database.

successfully; T4SEfinder will also combine other relevant biological features and classifiers. On the other hand, updating the pre-trained model in the feature extraction module to obtain more effective embedding of biological information may also prove useful for improving the general performance. In addition, the attention mechanism in the Transformer-based protein language model [50] may provide an advanced method for biological interpretation of deep learning techniques.

Conclusion

We have developed a publicly available web server T4SEfinder to facilitate community-wide efforts for T4SE prediction. By using the pre-trained language model of protein sequences, T4SEfinder has the capability of detecting the T4SEs in all annotated proteins encoded by a bacterial whole genome in minutes. Deep-learning tools such as T4SEfinder are anticipated to support rapidly escalating demands of the discovery of disease-associated secreted protein factors across diverse bacterial pathogens.

Key Points

- The T4SE training dataset from the newly updated SecReT4 database and other previous studies was integrated to form the benchmark dataset.
- The pre-trained language model of protein sequences is used for protein sequence embedding.
- The developed TAPEBert_MLP model achieved a better trade-off between the sensitivity and precision on the independent test dataset.
- The HybridBiLSTM model that incorporated the TAPEBert embedding and local PSSM features achieved the highest accuracy and MCC.
- A publicly available web server is publicly available for the genome-scale identification of T4SEs.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Data availability

Training and independent datasets used in this study are available from the corresponding authors upon reasonable request.

Availability

The online version of T4SEfinder is freely accessible at https://tool2-mml.sjtu.edu.cn/T4SEfinder_TAPE/.

Funding

Science and Technology Commission of Shanghai Municipality (19JC1413000 and 19430750600), National Natural Science Foundation of China (32070572), Medicine and Engineering Interdisciplinary Research Fund of Shanghai Jiao Tong University [19X190020171].

References

1. Grohmann E, Christie PJ, Waksman G, *et al*. Type IV secretion in gram-negative and gram-positive bacteria. *Mol Microbiol* 2018;**107**:455–71.
2. Cascales E, Christie PJ. The versatile bacterial type IV secretion systems. *Nat Rev Microbiol* 2003;**1**:137–49.
3. Wozniak RAF, Waldor MK. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat Rev Microbiol* 2010;**8**:552–63.
4. Alvarez-Martinez CE, Christie PJ. Biological diversity of prokaryotic type IV secretion systems. *Microbiol Mol Biol Rev* 2009;**73**:775–808.
5. Wallden K, Rivera-Calzada A, Waksman G. Type IV secretion systems: versatility and diversity in function. *Cell Microbiol* 2010;**12**:1203–12.
6. Personnic N, Bärlocher K, Finsel I, *et al*. Subversion of retrograde trafficking by translocated pathogen effectors. *Trends Microbiol* 2016;**24**:450–62.
7. Sherwood RK, Roy CR. Autophagy evasion and endoplasmic reticulum subversion: the yin and Yang of legionella intracellular infection. *Annu Rev Microbiol* 2016;**70**:413–33.
8. Lee YW, Wang J, Newton HJ, *et al*. Mapping bacterial effector arsenals: in vivo and in silico approaches to defining the protein features dictating effector secretion by bacteria. *Curr Opin Microbiol* 2020;**57**:13–21.
9. Zou L, Nan C, Hu F. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* 2013;**29**:3135–42.
10. Wang Y, Wei X, Bao H, *et al*. Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics* 2014;**15**:50.
11. An Y, Wang J, Li C, *et al*. Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI. *Brief Bioinform* 2018;**19**:148–61.
12. Xue L, Tang B, Chen W, *et al*. A deep learning framework for sequence-based bacteria type IV secreted effectors prediction. *Chemometrics Intellig Lab Syst* 2018;**183**:134–9.
13. Xiong Y, Wang Q, Yang J, *et al*. PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front Microbiol* 2018;**9**:2571.
14. Esna Ashari Z, Brayton KA, Broschat SL. Prediction of T4SS effector proteins for Anaplasma phagocytophilum using OPT4e. *A New Software Tool Front Microbiol* 2019;**10**:1391.
15. Wang J, Yang B, An Y, *et al*. Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. *Brief Bioinform* 2019;**20**:931–51.
16. Hong J, Luo Y, Mou M, *et al*. Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief Bioinform* 2020;**21**:1825–36.
17. Chen T, Wang X, Chu Y, *et al*. T4SE-XGB: interpretable sequence-based prediction of type IV secreted effectors using eXtreme gradient boosting algorithm. *Front Microbiol* 2020;**11**:580382.
18. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
19. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.
20. Breiman L. Random forests, Random forests. *Mach Learn* 2001;**45**:5–32.
21. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015;**61**:85–117.
22. Suzek BE, Huang H, McGarvey P, *et al*. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 2007;**23**:1282–8.
23. Lv Z, Ao C, Zou Q. Protein function prediction: from traditional classifier to deep learning. *Proteomics* 2019;**19**:e1900119.
24. Bellegarda JR. Statistical language model adaptation: review and perspectives. *Speech Commun* 2004;**42**:93–108.
25. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *arXiv preprint arXiv* 2017;**1706**:03762.
26. Devlin J, Chang MW, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv* 2018;**1810**:04805.
27. Zhu J, Xia Y, Wu L, *et al*. Incorporating BERT into neural machine translation. *arXiv preprint arXiv* 2020;**2002**:06823.
28. Lan Z, Chen M, Goodman S, *et al*. ALBERT: a Lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv* 2019;**1909**:11942.
29. Rao R, Bhattacharya N, Thomas N, *et al*. Evaluating protein transfer learning with TAPE. *Adv Neural Inf Process Syst* 2019;**32**:9689–701.
30. Min S, Park S, Kim S, *et al*. Pre-training of deep bidirectional protein sequence representations with structural information. *arXiv preprint arXiv* 2019;**1912**:05625.
31. Elnaggar A, Heinzinger M, Dallago C, *et al*. ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv* 2020;**2007**:06225.
32. Rao R, Meier J, Sercu T *et al*. Transformer protein language models are unsupervised structure learners. *bioRxiv* 2020. doi: 10.1101/2020.12.15.422761.
33. Rives A, Meier J, Sercu T, *et al*. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**:e2016239118.
34. Rao R, Liu J, Verkuil R, *et al*. MSA transformer. *bioRxiv* 2021. doi: 10.1101/2021.02.12.430858.
35. Bi D, Liu L, Tai C, *et al*. SecReT4: a web-based bacterial type IV secretion system resource. *Nucleic Acids Res* 2013;**41**:D660–5.
36. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.

37. Wang Y, Guo Y, Pu X, *et al*. Effective prediction of bacterial type IV secreted effectors by combined features of both C-termini and N-termini. *J Comput Aided Mol Des* 2017;**31**:1029–38.
38. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15.
39. Meyer DF, Noroy C, Moumène A, *et al*. Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. *Nucleic Acids Res* 2013;**41**:9218–29.
40. Makino K, Oshima K, Kurokawa K, *et al*. Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae. *Lancet* 2003;**361**:743–9.
41. El-Gebali S, Mistry J, Bateman A, *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;**47**:D427–32.
42. Ruck DW, Rogers SK, Kabrisky M, *et al*. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Trans Neural Netw* 1990;**1**:296–8.
43. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 2005;**18**:602–10.
44. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *Proceedings of the 32nd International Conference on Machine Learning* 2015;**37**:448–56.
45. van der Maaten L. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
46. Ninio S, Celli J, Roy CR. A legionella pneumophila effector protein encoded in a region of genomic plasticity binds to dot/Icm-modified vacuoles. *PLoS Pathog* 2009;**5**:e1000278.
47. Beare PA, Gilk SD, Larson CL, *et al*. Dot/Icm type IVB secretion system requirements for Coxiella burnetii growth in human macrophages. *MBio* 2011;**2**:e00175–11.
48. Myeni S, Child R, Ng TW, *et al*. Brucella modulates secretory trafficking via multiple type IV secretion effector proteins. *PLoS Pathog* 2013;**9**:e1003556.
49. Grant JR, Stothard P. The CGView server: a comparative genomics tool for circular genomes. *Nucleic Acids Res* 2008;**36**:W181–4.
50. Vig J, Madani A, Varshney LR, *et al*. BERTology meets biology: interpreting attention in protein language models. *arXiv preprint arXiv* 2020;**2006**:15222.