Computational and Structural Biotechnology Journal 23 (2024) 801–812

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

T4SEpp: A pipeline integrating protein language models to predict bacterial type IV secreted effectors

Yueming Hu [a,1], Yejun Wang [b,c,1], Xiaotian Hu [a], Haoyu Chao [a], Sida Li [a], Qinyang Ni [a], Yanyan Zhu [a], Yixue Hu [b], Ziyi Zhao [b], Ming Chen [a,d,*]

[a] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, China
[b] Youth Innovation Team of Medical Bioinformatics, Shenzhen University Medical School, Shenzhen, China
[c] Department of Cell Biology and Genetics, College of Basic Medicine, Shenzhen University Medical School, Shenzhen, China
[d] Institute of Hematology, Zhejiang University School of Medicine, The First Affiliated Hospital, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

Keywords:
T4SEpp
T4SE Prediction
T4SS
Deep learning
Protein language model
Helicobacter pylori T4SEs

ABSTRACT

Many pathogenic bacteria use type IV secretion systems (T4SSs) to deliver effectors (T4SEs) into the cytoplasm of eukaryotic cells, causing diseases. The identification of effectors is a crucial step in understanding the mechanisms of bacterial pathogenicity, but this remains a major challenge. In this study, we used the full-length embedding features generated by six pre-trained protein language models to train classifiers predicting T4SEs and compared their performance. We integrated three modules into a model called T4SEpp. The first module searched for full-length homologs of known T4SEs, signal sequences, and effector domains; the second module fine-tuned a machine learning model using data for a signal sequence feature; and the third module used the three best-performing pre-trained protein language models. T4SEpp outperformed other state-of-the-art (SOTA) software tools, achieving ~0.98 accuracy at a high specificity of ~0.99, based on the assessment of an independent validation dataset. T4SEpp predicted 13 T4SEs from *Helicobacter pylori*, including the well-known CagA and 12 other potential ones, among which eleven could potentially interact with human proteins. This suggests that these potential T4SEs may be associated with the pathogenicity of *H. pylori*. Overall, T4SEpp provides a better solution to assist in the identification of bacterial T4SEs and facilitates studies of bacterial pathogenicity. T4SEpp is freely accessible at https://bis.zju.edu.cn/T4SEpp.

1. Introduction

Gram-negative bacteria employ more than one dozen secretion systems to transport proteins out of the cell envelope [1,2]. Among them, the type IV secretion system (T4SS) is a complex molecular machine spanning both the inner and outer membranes, which translocates substrate proteins into eukaryotic host cells in only one step [3–9]. Protein-translocating T4SSs can be divided into two major families according to the composition of component elements: type IVA, exemplified by the *A. tumfaciens* VirB/VirD4 T4SS and *H. pylori* Cag T4SS, and type IVB, exemplified by *Legionella* Dot/Icm T4SS [9]. Substrate proteins translocated by T4SSs, also called effectors, play important roles in bacterial infections and pathogenicity [1,10,11].

Effectors of T4SSs (T4SEs) are transported directly or as complexes with DNA in many pathogenic bacteria, such as *Helicobacter pylori*, *Legionella pneumophila*, *Bordetella pertussis*, *Coxiella*, *Brucella*, and *Bartonella* [12–17]. T4SS-mediated entry of effector proteins into recipient cells is contact-dependent [18]. Once they enter the eukaryotic host cytoplasm, they disrupt signal transduction and cause various host diseases. Identifying these effectors is crucial for understanding the mechanisms of infection and pathogenicity caused by these bacteria. However, because the composition and sequences vary significantly, it is challenging to identify new T4SEs experimentally. Although many T4SEs have been identified and characterized in a few model organisms [19–22], the exact mechanism remains unclear.

Since 2009, when the first machine-learning algorithms were introduced [23], tens of computational models have been developed to predict T4SEs [2,23,24]. Early algorithms were mainly species-specific, such as those predicting T4SEs in *Legionella pneumophila*, *Anaplasma marginale*, and *Anaplasma phagocytophilum* [23,25–27]. In another study,

* Corresponding author at: Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, China.
E-mail address: mchen@zju.edu.cn (M. Chen).
[1] Y.H. and Y.W. contributed equally to this study.

https://doi.org/10.1016/j.csbj.2024.01.015
Received 21 September 2023; Received in revised form 20 January 2024; Accepted 20 January 2024
Available online 23 January 2024

Wang *et al.* developed an SVM-based model, T4SEpre, which exhibited good overall and cross-species performance [28]. However, T4SEpre only considers the features buried in the C-terminal 100 amino acids [28]. Notably, Esner Ashari *et al.* conducted a thorough study, generating four datasets for *L. pneumophila*, *Coxiella burnetii*, *Brucella spp*, and *Bartonella spp*, comprising T4SE and non-T4SE sequences [29]. They computed 51 features for all protein sequences in each dataset, encompassing amino acid composition, position-specific scoring matrix (PSSM), and binary peptide composition [29]. Their findings underscored that the optimal feature set for the four pathogens included vector features such as PSSM, amino acid composition, and dipeptide composition [29]. Several other studies, especially ensemble models recently developed with multi-aspect features, found features from full-length proteins to improve performance [30,31]. Deep learning algorithms have also been applied for the prediction of T4SEs. For example, CNN-T4SE integrated three convolutional neural network (CNN) models to learn the features of amino acid composition, solvent accessibility, and secondary structure of full-length T4SEs [32]. T4SEfinder is a multi-layer perception (MLP) model that learns the features generated by a pre-trained BERT model [33], which can predict T4SEs accurately [34]. Notably, BERT is a natural language processing (NLP) model that is appealing in biology and other fields [35–40]. NLP models have been successfully applied to the prediction of protein subcellular localization [36,37], secondary structure [37,38,40], and others [39]. Besides T4SEfinder, NLP-based pre-trained transformers have also been used for the prediction of bacterial type III secreted effectors and Sec/Tat substrates, both achieving superior performance [41,42].

Although machine learning strategies have achieved some success in the identification of T4SEs [2,23,28], the high false-positive rate has been a major challenge. To reduce the false-positive rate in predicting type III effectors, Hui *et al.* proposed a strategy to combine machine learning models with homology searching and integrate multiple modules, considering the multi-aspect biological features of the effector genes [43]. To improve model performance, other models have also considered multiple features and a combination of homology-based strategies in the prediction of type III effectors [44–46]. For T4SE prediction, homology searching has also been applied independently. For example, S4TE integrates 13 sequence homology-based features, including homology to known effectors, homology to eukaryotic domains, presence of subcellular localization signals, and secretion signals, and develops a scoring scheme to predict T4SEs mainly from α- and γ-proteobacteria [47]. A recent iteration, S4TE 2.0, introduced a new module dedicated to locating phosphorylation (EPIYA-like) domains. Notably, S4TE 2.0 boasts enhanced homology search efficiency [48]. Despite the high precision, the sensitivity could be influenced by the large diversity of T4SE composition and sequences. Therefore, it could be a better solution to take advantage of both machine learning approaches, especially ensembles, and homology-based methods, designing an integrated T4SE prediction pipeline that combines various models and comprehensively considers various characteristics of effector sequences.

In this study, we proposed a hybrid strategy for predicting T4SEs. First, a homology searching strategy scanned both the global homology of full-length proteins and the local homology of domains to known effectors. Additionally, we retrained a machine learning module T4SEpre [28] with updated T4SE data and hand-crafted amino acid composition features in the C-termini. Furthermore, a group of transfer learning models was developed based on the features generated by various pretrained transformers. For the transfer learning models, we utilized the deep context protein language models ESM-1b, ProtBert, ProtT5-XL, and ProtAlbert to represent protein sequence features [37,38]. These features can characterize the intrinsic but unclear properties of protein sequences and the interactions between positions. Based on these feature representations, application models were developed to classify T4SEs using a deep neural network architecture with an attention mechanism. Finally, we integrated the homology-based modules, machine learning models based on traditional handcrafted features, and transfer learning models with transformer-generated features into a pipeline, namely T4SEpp, which assembles the individual modules to generate a prediction score reflecting the likelihood of a protein to be a T4SE. A web application for T4SEpp is also available via the link: https://bis.zju.edu.cn/T4SEpp.

2. Results

2.1. Sequence homology among verified effectors and the integrated prediction framework

Experimentally verified effectors were collected from literature and databases, and 644 proteins were obtained after removing redundant sequences, representing the latest and most comprehensive list of experimentally verified T4SEs [31,49] (see Materials and Methods). We conducted an analysis of amino acid occurrences at each position in the C terminus of T4SS effectors. Utilizing the kpLogo program [50], we identified 100 C-terminal positions of T4SEs and control proteins (non-T4SEs), examining differences in amino acid preferences (Supplementary Fig. S3). Remarkably, we observed substantial residue enrichment at the C-terminus of T4SEs, particularly concentrated in the terminal 50 amino acid residues. Notably, the − 15 to − 9 region of the T4SEs C-terminus exhibited a significant enrichment of glutamate residues, consistent with prior studies [28,31].

We proceeded to perform pairwise sequence alignments of full-length (FL) effector proteins or their C-terminal peptides of 100 or 50 amino acids (C100 or C50, respectively) (Fig. 1A). For the FL proteins, 472 non-homologous clusters were identified after homology filtering for the proteins with $>$ 30% identity and $>$ 70% length coverage of the pair of proteins (FL_70%_30%_ID) (Supplementary Fig. S4). However, for the C100 sequences, 247 were homologous to others with an identity of $>$ 30%, and 469 non-redundant clusters were retained from these sequences after homology filtering (C100_30%_ID) (Supplementary Fig. S4). The reduction in the number of clusters indicated that the C-terminal 100 amino acids showed more homology than the full-length effector proteins, but there were no significant differences between them (C100_30%_ID, 469/629 vs., FL_70%_30%_ID, 472/644, EBT P = 0.609). The C50 sequences further reflected the typical C-terminal homology between effectors. A total of 339 peptides were found to have homology with the others, while 398 clusters remained for these peptides after homology filtering (C50_30%_ID, 398/644 vs. C100_30%_ID, 469/629, EBT P = 6.50e-04) (Supplementary Fig. S4). Rigorous homology filtering is a prerequisite for the application of machine learning to sequence analysis and effector identification. Sequence homology is often measured using similarity (SIM) rather than identity, with a cut-off of $\leq$ 30% for proteins. Therefore, we also employed a loose measure of homology, defined as $>$ 30% similarity, to examine sequence similarity between validated effectors. Surprisingly, the homology network involved all the 629 C100 peptides (C100_30%_SIM) (Supplementary Fig. S4). The results demonstrated that the validated T4SEs showed unexpectedly significant homology, especially for the C-terminus.

Taking full advantage of the fragmental similarity between T4SEs, combined with machine learning techniques, a comprehensive prediction pipeline (T4SEpp) was designed (Figs. 1B, C, and D). Several homology searching modules have been developed to detect full-length (flBlast), effector domain (effectHMM), and C-terminal signal region (sigHMM) homologs of known T4SEs. The previous machine learning model, T4SEpre, predicted T4SEs based on manually crafted C-terminal features [28] and was retrained with an updated dataset. Using the generative features from pre-trained transformers, we also developed a deep learning module, T4attention, incorporated with the Bi-Conv attention mechanism. Fig. 1E shows the framework of T4SEpp, combining the prediction scores of the homology search module (flBlast, effectHMM, and sigHMM), T4SEpre, and T4attention into a model to

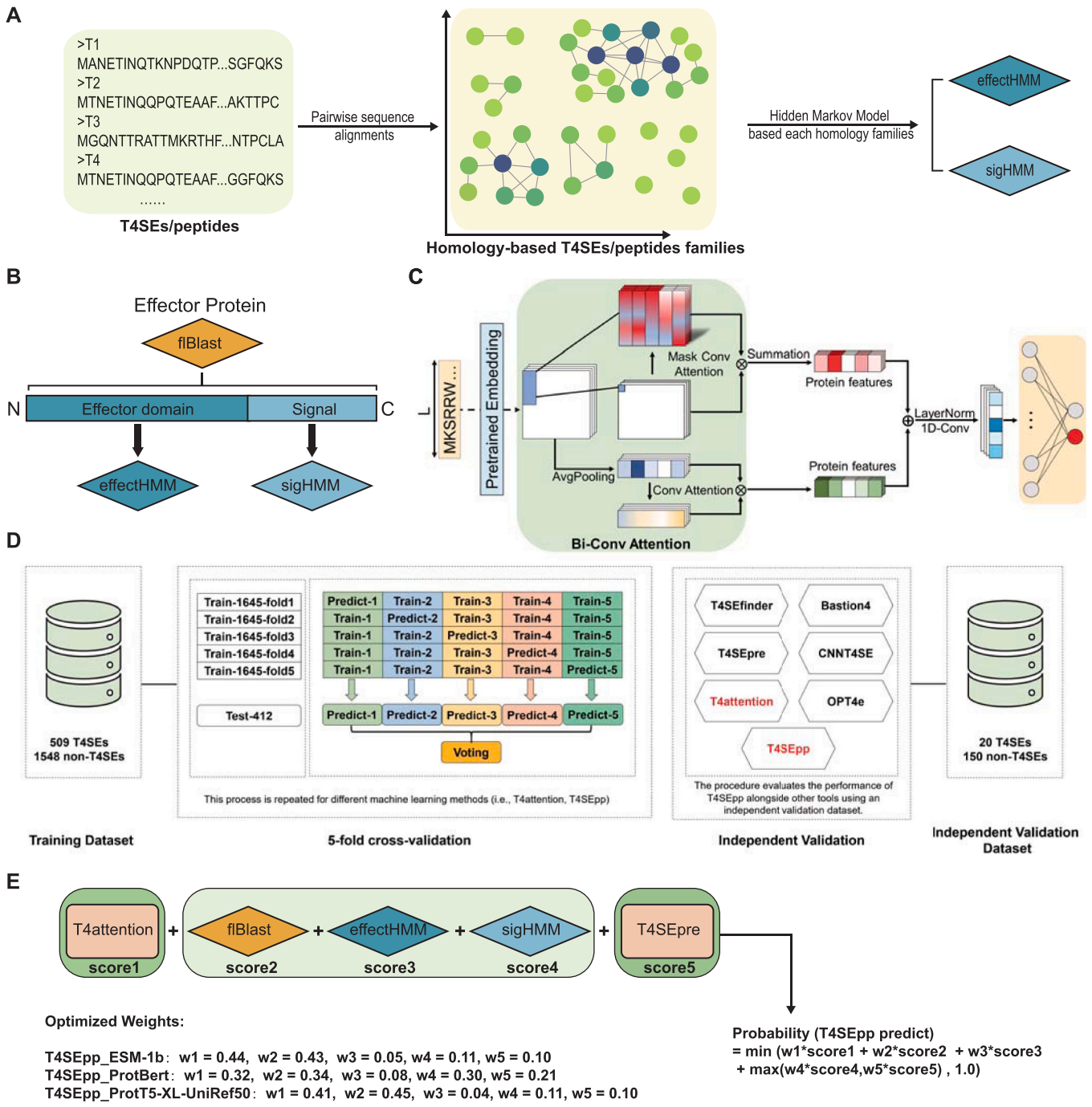

Fig. 1. Sequence homology among T4S effectors and an integrated prediction framework. (A) The workflow of homology sequence family identification and the construction of corresponding Hidden Markov Models (HMMs) for full-length (FL) effector proteins or their peptide fragments. Homologous family HMM models are constructed for C-terminal signal sequence (sigHMM), and effector (effectHMM) domains. (B) Homology-based modules developed for T4SEpp, based on the full-length effector proteins (flBlast) or signal sequence (sigHMM), and effector (effectHMM) domains. (C) T4attention is a deep learning framework incorporating Bi-Conv Attention. It utilizes a pretrained protein language model as input for feature extraction. (D) Procedure and datasets used for training and evaluation of the models. (E) Flowchart of the T4SEpp prediction program. The weighted sum of the prediction scores from each individual module is incorporated into the probability that a protein is a T4SE. For T4SEpre, we retrain using the updated T4SE dataset.

generate the final score, which reflects the likelihood of an input protein to be an effector.

2.2. T4SE families of signal sequences and functional domains

According to the homology of the C50 peptides, the effectors could be clustered into 398 signal sequence families, including 93 multi-member and 305 singlet families (Supplementary Table S3). After the signal sequences (C50) were removed, 635 effectors with a length of $\geq$ 30 amino acids remained, of which 268 were classified into 105 multi-member families and 367 represented singlet families (Supplementary Table S4). The sequences within each multi-component family showed striking similarity, and multiple positions appeared conserved, as shown for one example, sigFAM49 (Fig. 2A). The amino acid composition (AAC) showed apparent preference in multiple positions, e.g., leucine in positions 9, 24, and 37, serine in position 18, 30, and 64, and asparagine

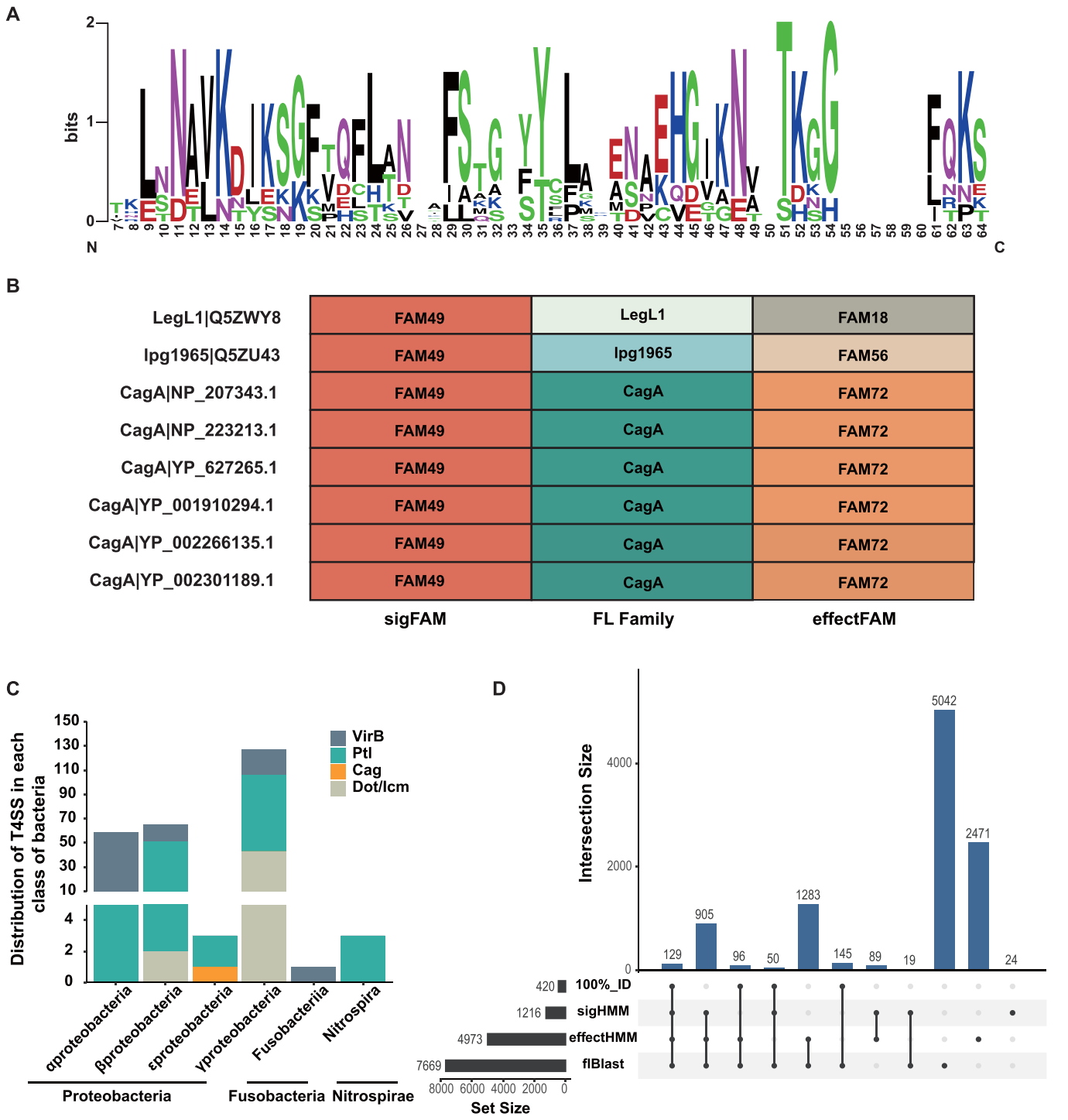

Fig. 2. Search for T4SS and effectors in the UniProt reference proteome based on sequence homology. (A) Multiple-sequence alignment (MSA) of a homologous cluster (i.e., sigFAM50) of T4SE signal sequences. Then, utilize the sequence logo of position-specific Amino Acid Compositions (AAC) corresponding to the alignment. The height of the amino acid in each position indicated the AAC preference. (B) Family clustering of the corresponding full-length effectors (FL Family) and effector domain (effectFAM) of sigFAM50 members. (C) Using the core protein components of T4SS to construct a Hidden Markov Model (HMM) to predict the distribution of T4SS in the UniProt reference proteome. (D) Three homologous units (sigHMM, effectHMM, and flBlast) were used to predict the potential T4SE in the UniProt reference proteome containing T4SS, respectively, where 100%_ID represents a known verified T4SE.

in position 11, 26, and 48, of sigFAM49 (Fig. 2A). Effectors of the same signal sequence family may belong to different effector functional domain families and vice versa. For example, six cytotoxin-associated gene A (CagA) effectors and two *Legionella* proteins contained the signal sequences of the same family (sigFAM49, Fig. 2B; Supplementary Table S3), but they also fell into three different effector functional domain families (effectFAM72 for all the CagAs, and effectFAM18 and effectFAM56 for the other two proteins; Fig. 2B; Supplementary Table S4). This could be related to frequent domain reshuffling events that have been reported in *Legionella* [51].

Furthermore, we searched for homologs of known T4SEs from the representative bacterial genomes downloaded from UniProt (8761 genomes; Supplementary Table S5). In total, 258 protein-translocating T4SSs were detected from 227 bacterial strains distributed in their

phyla (*Proteobacteria*, *Fusobacteria* and *Nitrospirae*), six classes (*Alphaproteobacteria*, *Betaproteobacteria*, *Epsilonproteobacteria*, *Gammaproteobacteria Fusobacteriia*, and *Nitrospira*), 117 genera and 227 species (Fig. 2C, Supplementary Table S6). In these strains with T4SSs, 10,253 proteins were detected with full-length or local homology to the known T4SEs using the individual homology searching units, and 1034 were identified by all the three units (Fig. 2D, Supplementary Table S7).

2.3. Prediction of T4SEs with pre-trained transformer-based models

Recently, protein language models have been successfully applied for structural prediction and sequence classification. In this research, we used six pre-trained models, ESM-1b, ProtAlbert, ProtBert-BFD, ProtBert-UniRef100, ProtT5-XL-BFD, and ProtT5-XL-UniRef50, to generate features; based on this, we developed deep learning models (T4attention) based on Bi-Conv attention to classify T4SEs and non-T4SEs. The T4attention models based on different sequence embedding features were compared for performance based on a five-fold cross-validation strategy (Table 1). Generally, T4attention_ESM-1b performed the best, followed by T4attention_ProtT5-XL-UniRef50, and T4attention_ProtAlbert showed the poorest performance, according to Matthew's correlation coefficient (MCC) and F1-score (Table 1). T4attention_ESM-1b not only reached the highest MCC and F1-score (0.828 and 0.869, respectively), but required the lowest computational resources (Supplementary Fig. S5). It was also noted that, for the same protein language model architecture, ProtBert or ProtT5-XL, for example, the generation of features from models pre-trained from various volumes of protein databases required similar computational resources, but the smaller database-based pre-trained models always generated features for subsequent T4attention models with better performance (MCC of T4attention_ProtBert vs. T4attention_ProtBert-BFD, 0.817 vs. 0.793; T4attention_ProtT5-XL-UniRef50 vs. ProtT5-XL-BFD, 0.812 vs. 0.804) (Table 1, Supplementary Fig. S5). The redundancy of protein sequences in the BFD dataset might lead to biases in model training and further compromise the performance of models addressing downstream tasks.

We also evaluated the performance and generalization abilities of these models on an independent validation dataset. T4attention_ProtBert showed the overall best performance, for which the MCC, F1-score, and accuracy reached 0.887, 0.900, and 0.976, respectively (Table 2). T4attention_ESM-1b unexpectedly showed poor performance (Table 2). Consistent with the cross-validation results, the ProtBert and ProtT5-XL models, based on the features generated by transformers pre-trained from a smaller database (UniRef100/UniRef50), showed better performance (Table 2, Supplementary Fig. S6).

Considering the performance of models based on both cross-validation results and the independent validation dataset, as well as the requirement of computational resources, we integrated three models, T4attention_ESM-1b, T4attention_ProtBert, and T4attention_ProtT5-XL-UniRef50, into the pipeline to predict T4SEs.

2.4. An integrated pipeline predicting T4SEs with largely improved performance

In addition to the models based on the features generated by the transformer, we tested traditional machine learning models based on hand-crafted features. To this end, we fine-tuned two models of T4SEpre models (T4SEpre_psAac and T4SEpre_bpbAac) to learn the amino acid composition features in the C-termini of T4SEs [28]. Both models showed a reasonable performance in the prediction of T4SEs according to the cross-validation results or the independent validation dataset, although they were not comparable to the T4attention models (Tables 1 and 2).

To further improve the accuracy and reduce the false-positive rate for T4SE prediction, we assembled a unified pipeline, T4SEpp, integrating the homology searching modules, machine learning models based on hand-crafted features, and models based on transformer-generated features (Fig. 1). The integrated pipeline showed strikingly better performance than the individual models, with MCC values of 0.917, 0.914, and 0.913 for T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50 based on the cross-validation evaluation and 0.883, 0.913, and 0.942 for the validation dataset, respectively (Tables 1 and 2).

T4SEpp was also compared to other state-of-the-art (SOTA) T4SE prediction models, such as Bastion4 [31], OPT4e [26], CNNT4SE [32], and T4SEfinder [34]. Among these other models, Bastion4 showed the best performance, which was close to that of the T4attention models but was far inferior to the integrated T4SEpp (Table 2).

Furthermore, we extracted effector proteins from the positive dataset, specifically those associated with the *L. pneumophila Philadelphia-1*, resulting in a total of 297 T4SEs. Employing T4SEpp for predicting the *L. pneumophila Philadelphia-1*, we compared its performance with candidate effector proteins predicted by two earlier studies that were focused on the same bacterial strain [27,48]. Notably, T4SEpp identified a minimum of 98.99% of experimentally confirmed T4SEs in the *L. pneumophila Philadelphia-1* (98.99% for T4SEpp_ESM-1b, 99.66% for T4SEpp_ProtBert, and 100% for T4SEpp_ProtT5-XL-UniRef50)

Table 1
Performance comparison of the models in T4SEpp on 5-fold cross-validation dataset.

Method	ACC	SN	SP	PR	F1	MCC	rocAUC	AUPRC
T4attention_ESM-1b	**0.937 ± 0.004**	0.854 ± 0.018	**0.964 ± 0.008**	**0.884 ± 0.022**	**0.869 ± 0.007**	**0.828 ± 0.010**	0.953 ± 0.013	0.895 ± 0.013
T4attention_ProtBert	0.932 ± 0.011	**0.862 ± 0.032**	0.955 ± 0.006	0.862 ± 0.019	0.862 ± 0.024	0.817 ± 0.031	**0.955 ± 0.013**	0.895 ± 0.020
T4attention_ProtBert-BFD	0.922 ± 0.018	0.858 ± 0.029	0.950 ± 0.009	0.843 ± 0.016	0.832 ± 0.043	0.793 ± 0.046	0.944 ± 0.012	0.874 ± 0.026
T4attention_ProtT5-XL-UniRef50	0.931 ± 0.009	0.852 ± 0.030	0.956 ± 0.010	0.863 ± 0.025	0.857 ± 0.018	0.812 ± 0.024	0.949 ± 0.014	0.896 ± 0.010
T4attention_ProtT5-XL-BFD	0.928 ± 0.004	0.850 ± 0.016	0.953 ± 0.002	0.851 ± 0.065	0.854 ± 0.006	0.804 ± 0.013	0.945 ± 0.016	0.885 ± 0.030
T4attention_ProtAlbert	0.922 ± 0.013	0.858 ± 0.038	0.942 ± 0.007	0.827 ± 0.021	0.842 ± 0.027	0.790 ± 0.035	0.951 ± 0.017	0.871 ± 0.053
T4SEpre_psAac[a]	0.843 ± 0.020	0.816 ± 0.030	0.871 ± 0.028	0.864 ± 0.027	0.839 ± 0.023	0.688 ± 0.043	0.911 ± 0.015	0.886 ± 0.010
T4SEpre_bpbAac[a]	0.854 ± 0.024	0.818 ± 0.042	0.891 ± 0.010	0.883 ± 0.020	0.849 ± 0.028	0.712 ± 0.046	0.925 ± 0.020	**0.896 ± 0.017**
T4SEpp_ESM-1b	**0.970 ± 0.006**	**0.905 ± 0.016**	**0.991 ± 0.005**	**0.969 ± 0.015**	**0.936 ± 0.013**	**0.917 ± 0.017**	0.993 ± 0.004	0.957 ± 0.041
T4SEpp_ProtBert	0.969 ± 0.007	**0.905 ± 0.016**	0.986 ± 0.006	0.965 ± 0.017	0.934 ± 0.014	0.914 ± 0.018	**0.994 ± 0.003**	**0.964 ± 0.027**
T4SEpp_ProtT5-XL-UniRef50	0.968 ± 0.003	**0.905 ± 0.016**	0.989 ± 0.003	0.963 ± 0.007	0.933 ± 0.008	0.913 ± 0.010	0.993 ± 0.003	0.946 ± 0.042

ACC, Accuracy; SN, sensitivity; SP, specificity; PR, precision; F1, F1-score; MCC, Matthews correlation coefficient; rocAUC, area under the receiver operating characteristic curve; AUPRC, precision recall rate curve; a, re-trained the model; data in table are represented as mean ± standard deviation (SD). The upper and lower horizontal lines represent the single prediction model and the integrated prediction model respectively.

Table 2
Performance comparison of the models in T4SEpp and other tools on the independent dataset.

Method	ACC	SN	SP	PR	F1	MCC	rocAUC	AUPRC
T4attention_ESM-1b	0.929	0.850	0.940	0.654	0.739	0.707	0.955	0.907
T4attention_ProtBert	0.976	0.900	0.987	0.900	0.900	0.887	**0.994**	**0.973**
T4attention_ProtBert-BFD	0.935	**0.950**	0.933	0.655	0.776	0.757	0.979	0.937
T4attention_ProtT5-XLUniRef50	0.959	0.900	0.967	0.783	0.837	0.816	0.971	0.889
T4attention_ProtT5-XL-BFD	0.941	0.900	0.947	0.692	0.783	0.758	0.977	0.941
T4attention_ProtAlbert	0.959	0.850	0.973	0.810	0.829	0.806	0.979	0.926
T4SEpp_ESM-1b	0.976	0.850	0.993	0.944	0.894	0.883	0.953	0.928
T4SEpp_ProtBert	0.982	0.900	0.993	0.947	0.923	0.913	0.965	0.950
T4SEpp_ProtT5-XL-UniRef50	**0.988**	0.900	**1.000**	**1.000**	**0.947**	**0.942**	0.963	0.932
T4SEfinder-TAPEBert_MLP	0.958	0.850	0.973	0.810	0.829	0.806	0.959	0.805
T4SEfinder-hybridbilstm	0.941	0.800	0.960	0.727	0.762	0.730	0.945	0.852
T4SEfinder-pssm_cnn	0.906	0.800	0.920	0.571	0.667	0.625	0.923	0.759
Bastion4	0.965	0.900	0.973	0.818	0.857	0.838	-	-
CNNT4SE	0.953	0.700	0.987	0.875	0.778	0.758	0.943	0.860
OPT4e	0.865	0.200	0.953	0.363	0.258	0.201	-	-
T4SEpre_psAac[a]	0.888	0.700	0.913	0.519	0.596	0.541	0.921	0.740
T4SEpre_bpbAac[a]	0.829	0.700	0.847	0.378	0.491	0.427	0.895	0.730

ACC, Accuracy; SN, sensitivity; SP, specificity; PR, precision; F1, F1-score; MCC, Matthews correlation coefficient; rocAUC, area under the receiver operating characteristic curve; AUPRC, precision recall rate curve; a, re-trained the model. The upper and lower horizontal lines respectively represent the prediction model constructed in this article and the prediction model constructed by previous researchers.

(Supplementary Fig. S7). In contrast, S4TE v2 managed to identify only 94.61% of T4SEs [48], and Esna Ashari *et al.* exhibited the lowest proportion of experimentally verified T4SEs, at 93.60% (Supplementary Fig. S7) [27]. These findings strongly suggest that T4SEpp excels in predicting effector proteins of the *L. pneumophila Philadelphia-1*, positioning it with significant potential in the realm of single-strain prediction.

2.5. Genome-wide screening of T4SEs in Helicobacter pylori and other bacteria

H. pylori is a Gram-negative, spiral-shaped bacterium that colonizes the stomach in approximately half of the world's population [52]. Although most individuals do not experience any adverse health outcomes attributable to *H. pylori*, the presence of these bacteria in the stomach increases the risk of developing gastric diseases [53–57]. *H. pylori* infection is also the strongest known risk factor for gastric cancer, the third leading cause of cancer-related death worldwide [58]. T4SS plays an important role in *H. pylori* [54–57]. However, to date, only one T4SE, CagA, has been identified for the T4SS in *H. pylori* [59]. Here, we applied T4SEpp to screen T4SE candidates from the proteins derived from the genome of *H. pylori* 26695, a model *H. pylori* strain (NCBI accession number: NC_000915.1). The three T4SEpp integrated models, T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50, predicted 56, 27, and 44 T4SE candidates, respectively, and 13 were shared by the prediction results of all the three models (Fig. 3A-B; Supplementary Table S6, S8). The 13 potential effector genes were scattered throughout the genome (Fig. 3B). Notably, *HP_RS02695,* which encodes the only known effector CagA, was among the 13 candidates (Fig. 3B).

Gene co-expression was analyzed for the 13 T4SE candidate genes in *H. pylori 26695* using an RNA-seq dataset sampled from the strain collected under 12 different conditions [60]. Except for HP_RS06290, HP_RS03730, HP_RS06295, HP_RS00300, and HP_RS02820, the remaining seven genes showed a strong expression correlation with *cagA* expression (Fig. 3C). The genes co-expressed with *cagA* also showed a significant correlation with the expression of the core component genes of the Cag T4SS (Fig. 3C). Furthermore, we annotated 12 human proteins that showed experimentally verified interactions with CagA by literature search, including ASPP2, c-Abl, c-Met, Crk, E-cadherin, GSK-3, PAR1, PRK2, SHP-1, SHP-2, TAK1, and ZO-1 [61–72]. The interaction network between the 13 potential *H. pylori* 26695 T4SEs and the 12 human proteins was inferred (Fig. 3D). Eleven of the candidate T4SEs showed potential interaction with at least one of the human proteins (Fig. 3D). Similar to CagA, HP_RS02225, HP_RS02255, HP_RS00300, HP_RS06295, and HP_RS03730 interacted with all 12 human proteins (Fig. 3D). Taken together, the proteins predicted by T4SEpp could potentially represented new T4SEs or may be closely related to the pathogenicity of *H. pylori 26695*.

We also used T4SEpp to screen the T4SE candidates from the genomes of 227 bacterial strains bearing T4SSs. T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50 detected 16,729, 19,781, and 18,176 T4SE candidates respectively, with 12,553 common candidates co-predicted by all three T4SEpp models (Supplementary Table S9, Supplementary Fig. S8).

2.6. Web server and implementation of T4SEpp

To facilitate the implementation of T4SEpp, we developed a user-friendly web application (https://bis.zju.edu.cn/T4SEpp). The three T4SEpp integrated models, T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50 can be chosen and implemented by users. Both the overall prediction results and the results of the individual modules are displayed in table format, which can be downloaded and filtered easily.

3. Discussion

T4SS plays a crucial role in bacterial pathogenicity by secreting effectors into host cells. *L. pneumophila* can translocate more than 300 known effectors into human cells via the Dot/Icm T4SS system, causing legionellosis [74,75]. In *H. pylori*, CagA is the only known T4SE that can hijack multiple signaling pathways in gastric epithelial cells, leading to gastritis, gastric ulcer, and even gastric cancer [76,77]. Identifying the full repertoire of T4SEs in a pathogen is important to understand its pathogenic mechanisms. Computational methods can assist with the effective identification of new effectors [24]. However, the currently available T4SE prediction tools still show high false-positive rates [2]. To address this issue, we developed a unified T4SE prediction pipeline, T4SEpp, which includes homologous search modules, traditional machine learning modules, and natural language processing-based modules. T4SEpp outperformed other SOTA methods for predicting T4SEs, with improved sensitivity and specificity. Furthermore, we created a web server that conveniently implements the T4SEpp pipeline, providing the prediction results for each module.

Although the component modules of T4SEpp can be used for T4SE prediction, they often show higher false-positive rates when used alone. This could be related to the low power of the individual dimensions of

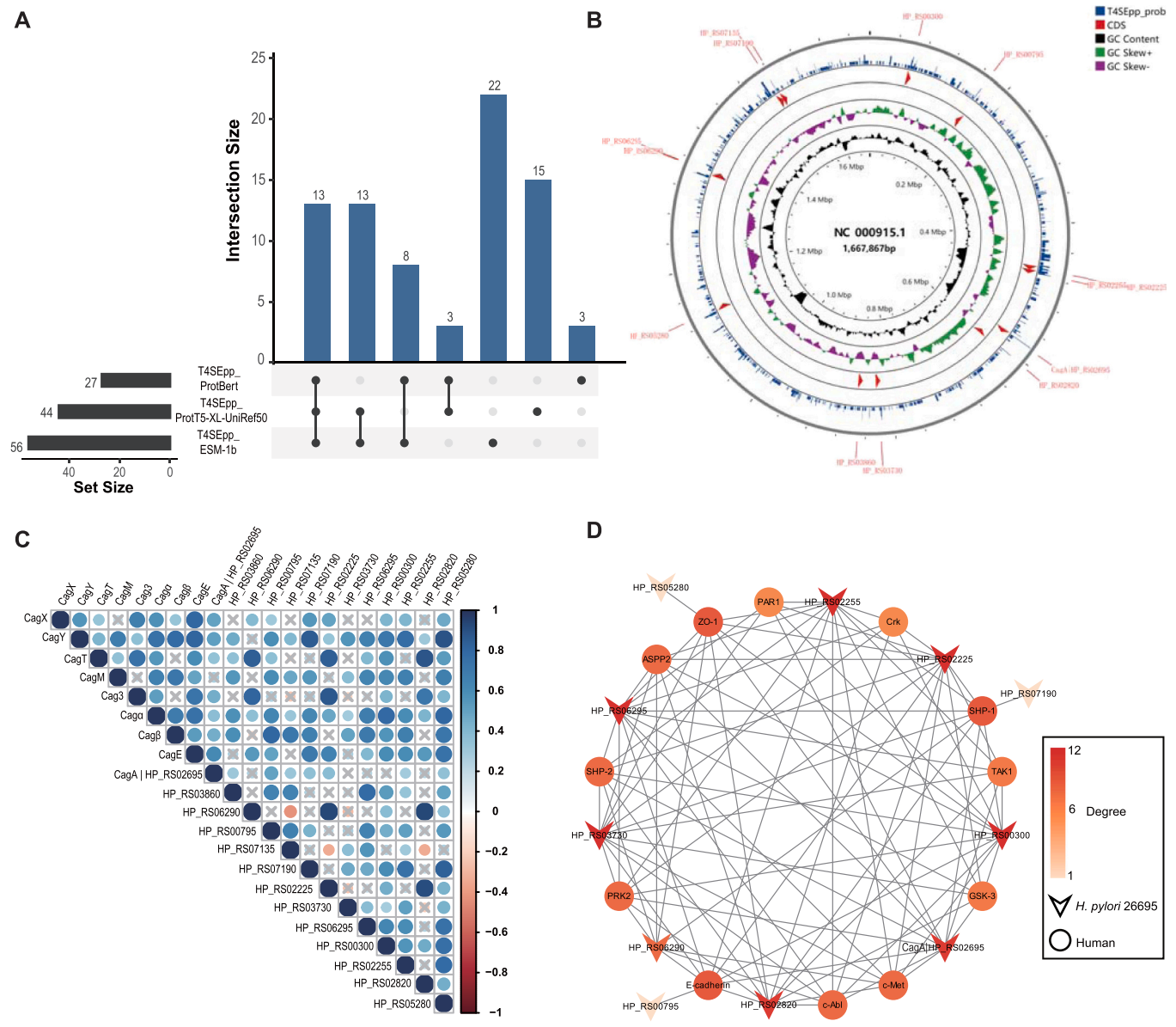

Fig. 3. Whole-proteome detection for T4SEs in pathogenic bacteria (*H. pylori* 26695). (A) Prediction of potential T4SEs in the *H. pylori 26695* proteome using three T4SEpp models. (B) Use the circos diagram to show the distribution of potential T4SEs predicted by the three T4SEpp models on the *H. pylori 26695* chromosome (NC_000915.1), where T4SEpp_prob represents the mean value of the prediction results of the three T4SEpp models, and the outer circle of the circos diagram represents the three T4SEpp model predictions were all positive. (C) Under 12 different expression conditions of *H. pylori 26695*, the expression correlation of Cag T4SS core components with 12 potential T4SEs and CagA (HP_RS02695) predicted by three T4SEpp models were positive. (D) Prediction of potential interactions between 12 potential T4SEs in H. pylori 26695 and 12 human proteins using DeepHPI [73]. These 12 human proteins are known to interact with CagA (HP_RS02695). Nodes represent *H. pylori 26695* (V-shape) or human (ellipse-shape) proteins. Edges represent protein–protein interactions. The nodes are colored according to their degrees.

the features. Specifically, T4SE signal sequences were considered to contain important common features guiding T4SE secretion and translocation, which were used for effective T4SE prediction using tools such as T4SEpre [28]. However, the computational models based only on the signal sequences showed performance inferior to other models based on multiple-aspect features extracted from full-length proteins [31]. In this study, we discovered high sequence similarity in the C-terminal signal region among the proteins, without apparent homology to full-length effectors. Such undetected homology could have introduced bias and led to overfitting of various established machine learning algorithms and the discrepancy between the reported and actual accuracy of these methods. However, the C-terminal homology could also suggest the independent evolution of the signal sequences, and it could potentially be applied to facilitate the identification of new effectors [47].

In this study, three types of modules were integrated to predict T4SEs. Homology searching-based modules provide more accurate results, but they also show a lower capacity to detect new effectors with or without remote homology. The re-trained T4SEpre modules focused on the important features of the C-terminal signal sequences of T4SEs. T4attention learns from the full-length effector proteins the features generated by protein language models (pLMs) pre-trained with large-scale protein databases. These pLM-based models can learn new, previous unknown features that may involve position-position interactions and have demonstrated outstanding performance in the prediction of proteins with various biological functions, such as subcellular localization and secondary structure. We used multiple pLMs to build transfer learning models, most of which exhibited excellent performance in T4SE prediction. Interestingly, we noticed that the pre-trained pLMs based on

the larger datasets did not generate better prediction performance. pLMs pre-trained on smaller datasets are more efficient. Therefore, the transfer models were trained with the pLMs based on smaller non-redundant protein datasets. T4SEpp, which integrated all three types of modules, significantly outperformed both individual modules and other similar applications.

Considering the comprehensive performance of T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50 across various aspects, we offer these three models on our web server for user selection. T4SEpp_ESM-1b excelled in 5-fold cross-validation, while T4SEpp_ProtBert and T4SEpp_ProtT5-XL-UniRef50 demonstrated excellent prediction performance in independent datasets (Tables 1 and 2). This balanced approach for overall model performance ensures our tool's ability to provide users with reliable T4SE prediction results under diverse conditions. Additionally, the flexibility of model selection is provided to cater to different researchers' preferences based on their research questions and dataset characteristics, aiming to ensure that T4SEpp meets a wide range of research needs and exhibits maximum applicability in practical scenarios.

Using T4SEpp, we analyzed the potential new T4SEs in both *H. pylori* and other strains bearing a T4SS. We identified 12 new T4SEs in *H. pylori*. We also identified 12,138 new T4SEs and 420 known T4SEs from 227 strains bearing a T4SS. The results suggested that there are many new effectors yet to be identified.

Despite the significant performance improvement of T4SEpp, there remains a need to further improve the prediction of T4SEs. Other features that have been known to contribute to the recognition of T4SEs, such as the GC content of genomic loci, phylogenetic profiles, amino acid composition, dipeptide composition, consensus regulatory motifs in promoters, physicochemical properties, secondary structures, hydropathy, homology to eukaryotic domains, and organelle-targeting signals, have not been integrated into the current version of the model [24,29]. Novel features that could be further integrated to improve the model performance remain to be disclosed. The different types (IVA and IVB) of effectors, chaperone-dependent or chaperone-independent effectors, or species-specific effectors can also be modeled and predicted separately to make more accurate prediction [24].

4. Conclusion

In this study, we introduced T4SEpp, an advanced tool for identifying bacterial Type IV Secretion System effectors (T4SEs). T4SEs play a critical role in bacterial infections, and T4SEpp significantly outperforms existing methods. Achieving a sensitivity of ~0.85 with ~0.99 specificity on an independent dataset, it marks a major advancement. Additionally, our comprehensive search across bacterial species unveiled 227 species spanning 3 phyla and 117 genera with T4SSs. T4SEpp identified 12,138 new putative T4SEs, enriching our understanding of bacterial virulence mechanisms. While T4SEpp is a remarkable achievement, future work could involve integrating more features and creating specialized models for different effectors. Overall, T4SEpp empowers researchers in the fight against infectious diseases and microbial-host interactions.

5. Materials and methods

5.1. Datasets

The 390 T4SEs used by Bastion4 as the positive training dataset [31] and 540 T4SEs annotated in SecReT4 v2.0 [49] were collected and merged. Subsequently, we filtered out fragment proteins, proteins originating from Gram-positive bacteria, and certain inaccurately annotated T4SEs. In total we got 644 non-identical, validated T4SEs. CD-HIT [78] was used to filter homology-redundant proteins with sequence identity $\geq$ 60%, generating 509 non-redundant T4SEs, which were used as the positive training dataset (Supplementary Fig. S1A, Fig. 1D). For the negative training dataset, we collected 1112 and 1548 non-T4SE protein sequences from Bastion4 [31] and PredT4SE-stack [79], respectively. The same procedure was used to eliminate the sequence redundancy among the non-T4SEs and between the non-T4SEs and T4SEs in the positive training dataset, generating 1590 non-redundant non-T4SEs (Supplementary Fig. S1A, Fig. 1D). An independent validation dataset was also prepared, for which the T4SEs were collected from the testing dataset of Bastion4 (30) and others (74) annotated from literature published recently (Supplementary Table S1), and the 150 testing non-T4SEs of Bastion4 were also used as negative ones. CD-HIT was used to filter the redundant proteins with $\geq$ 60% sequence identity to the training proteins and among proteins in the independent validation dataset, resulting in 20 non-redundant T4SEs and 150 non-redundant non-T4SEs (Supplementary Fig. S1B, Fig. 1D).

5.2. Genome-wide screening of protein-translocation T4SSs

The conserved core component proteins were collected from four representative protein-translocation T4SSs, including the *Agrobacterium tumefaciens* VirB/VirD4 T4SS (inner membrane complex proteins VirB3, VirB6, VirB8, the N-terminal region of VirB10, and VirD4, and outer membrane complex proteins VirB7, VirB9, and a C-terminal domain of VirB10) [16], the *Bordetella pertussis* Ptl T4SS (inner membrane complex proteins PtlB, PtlE and PtlH, and outer membrane complex proteins PtlF and PtlG) [80], the *Helicobacter pylori* Cag T4SS (inner membrane complex proteins Cagα, Cagβ, and CagE, and outer membrane complex proteins CagX, CagY, CagT, CagM, and Cag3) [18], the *Legionella pneumophila* Dot/Icm T4SS (inner membrane complex proteins IcmB, IcmG, and DotB, and outer membrane complex proteins DotC, DotD, DotG, and IcmK) [16]. Hidden Markov Model (HMM) profiles were built using HMMER 3.1 for the T4SS component protein families [81]. Protein sequences derived from the 8761 reference bacterial genomes curated in UniProt were scanned with HMMER and the HMM profiles to determine the distribution of homologs of T4SS core component proteins (Supplementary Table S5).

5.3. Homology networks of the T4SE peptide sequences

The sequences of 653 non-identical verified T4SE proteins were used to construct the homology networks. JAligner implemented the Smith-Waterman algorithm to determine the similarity between any pair of full-length effectors or peptide fragments of designated length (http://jaligner.sourceforge.net/). The identity and similarity percentages between any pair of sequences were used as measures to determine the homology level [43]. Networks were built and visualized using Cytoscope v3.9.1 [82].

5.4. Homology-based T4SE detection modules

Diamond blastp was used to determine the homology and cluster the full-length effector proteins [83] and to screen new full-length homologs (flBlast). Two proteins showing $\geq$ 30% similarity for $\geq$ 70% of the full length of either protein were considered to be full-length homologs [43,84]. The C-terminal 50-aa signal sequences of the verified effectors were clustered according to homology networks with 30% identity for 70% length aligned by JAligner. HMM profiles were built for each signal sequence family, and a sigHMM module was developed to screen for proteins with C-terminal sequences homologous to the profiles of known T4SE signal sequence families. The homology cutoff for HMM searching was optimized for each family, ensuring that all or most of the known effectors recalled and maintained a higher specificity. For effectHMM, we removed the C-terminal 50-aa signal from each known effector sequence, and the remaining peptide fragment with $>$ 30-aa length was used for domain clustering. Pairwise alignment was repeatedly performed with BLAST between the domain sequences, and the cutoff for homology was optimized based on the average coverage of the aligned

length multiplied by the identity, that is, ≥ 10 [43]. The HMM profiles were built for the effector domain families, and effectHMM was developed using a similar procedure as sigHMM to screen the proteins with homologous T4SE effector-domains. We used EBT to compare general homology between proteins [43,85].

5.5. Re-trained T4SEpre models with updated datasets

We retrained the T4SEpre models (T4SEpre_psAac and T4SEpre_bpbAac) using the new training datasets of T4SEs and non-T4SEs. The original T4SEpre procedure was followed for feature representation, parameter optimization, and model training [28]. Briefly, sequential amino acid, bi-residue and motif composition features, and position-specific amino acid composition profiles for the positive training dataset were represented for each C-terminal 100-aa sequence for the psAac model. For the bpbAac model, position-specific amino acid composition profiles of both the positive and the negative training datasets (Bi-Profile Bayesian features) were represented for each C-terminial 100-aa sequence. Support vector machine (SVM) models were trained for feature matrices. The kernel functions, that is, linear, polynomial, sigmoid, and radial base function (RBF), and corresponding parameters (cost and gamma) were optimized using a 5-fold cross-validation grid search strategy. sklearn v1.0.1 was used for implementing SVM model training and kernel/parameter optimization.

5.6. The deep learning architecture of T4attention based on pre-trained protein language models

5.6.1. Input embeddings

Frozen embeddings were extracted directly from protein language models (pLMs) without fine-tuning the training data. Four different basic LMs were used in this study, and six different pLMs were pre-trained with different datasets. The basic LMs include: (i) “ESM-1b” [38], which is a Transformer model, (ii) "ProtBert" [37], which is a BERT-based encoder model [35], generating two pLMs pre-trained on BFD [86] and UniRef100 [87] data, respectively, (iii) ProtT5-XL [37], which is an encoder model based on T5 [88], generating two pLMs pre-trained on BFD and UniRef50, respectively, and (iv) ProtAlbert [37], which is an encoder model based on Albert [89] and pre-trained only with UniRef100.

5.6.2. Optimization strategy

We used a BERT-like optimizer AdamW and a Cosine Warm-up strategy [35] to optimize the loss of the learning model. The initial learning rate was set to 0.0001, the batch size was set to 18, and the warm-up steps were set to 10. An early stopping strategy was applied to monitor the validation ACC with 30 epochs to prevent overfitting. To address the challenges of imbalanced positive and negative samples and the difficulty of training individual samples in deep learning model training, we adopted the Focal Loss method to mitigate the issue of gradient descent difficulty [90]. Focal Loss adjusts the hyperparameter γ (default γ = 2) based on the weighted cross-entropy loss, influencing the shape of the curve.

$$\boldsymbol{FL}(\boldsymbol{p}_t) = -\boldsymbol{\alpha}_t(\mathbf{1}-\boldsymbol{p}_t)^{\gamma}\mathbf{log}(\boldsymbol{p}_t)$$

$\boldsymbol{\alpha}_t$: Weight of the sample t,

$\boldsymbol{p}_t$: Binary cross entropy loss.

Consequently, as p_t approaches 1, the loss decreases, diminishing the impact of easily classifiable samples, whereas as p_t approaches 0, the loss increases, amplifying the influence of hard-to-classify samples. In essence, Focal Loss diminishes the emphasis on easily classifiable samples and amplifies the significance of challenging samples, enhancing the model’s capability to handle difficult instances.

5.6.3. T4attention model

The input to T4attention (Fig. 1C, Supplement Fig. S2) is a protein embedding $\boldsymbol{E_0} \in \mathbb{R}^{n\times d_0}$, where **n** is the sequence length and $\mathbf{d_0}$ is the size of the embedding (depending on the feature extraction model). T4attention is a model based on Bi-Conv attention. In the protein embedding direction, average pooling is performed directly, and the input is transformed by two separate 1D convolutions, where the 1D convolution serves as the attention coefficient **e** and value **v** for computing the embedding dimension, $\boldsymbol{e}, \boldsymbol{v} \in \mathbb{R}^{d_1}$. Thus, we obtained the feature representation of the embedding dimension $\boldsymbol{x} = \boldsymbol{softmax}(\boldsymbol{e}) \times \boldsymbol{v}$. In the direction of the protein sequence, we randomly intercept the length of **m** in the length direction of the protein-embedding sequence such that the protein embedding becomes $\boldsymbol{E_1} \in \mathbb{R}^{m\times d_0}$. Similar to the convolutional attention calculation in the protein embedding direction, the attention coefficient **e’** and value **v’** are obtained, $\boldsymbol{e}', \boldsymbol{v}' \in \mathbb{R}^{m\times d_1}$. The difference is that the direction of the convolution is in the direction of the sequence length, so that we can obtain the feature representation of the protein sequence direction and converge according to the sequence length direction by $\boldsymbol{x}' = \sum_i^m \boldsymbol{softmax}(\boldsymbol{e}') \times \boldsymbol{v}'$. The convolution attention results of the embedding direction and the protein sequence direction are merged and passed through the LayerNorm and the residual 1D convolution, and the class probabilities are obtained through the two-layer multi-layer perceptron (MLP), $\boldsymbol{p}(\mathbf{c}|\boldsymbol{x}) = \boldsymbol{softmax}(\boldsymbol{MLP}(\boldsymbol{Conv}(\boldsymbol{x}+\boldsymbol{x}')+(\boldsymbol{x}+\boldsymbol{x}')))$, where **c** indicates the category of the output (i.e., T4SE or nonT4SE).

T4attention was developed using PyTorch v1.10.1. The models were trained and evaluated with 24-GB of memory and an NVIDIA GeForce RTX 3090 GPU for acceleration.

5.7. Integrated T4SE prediction model

T4SEpp is a non-linear model that integrates multiple prediction modules developed or re-trained in this study, including homology-searching modules for full-length or fragmented effector proteins, traditional machine-learning modules with hand-crafted features, and the attention-based transfer learning modules using the features generated by pre-trained protein language models. For any prediction module, the factor was set to 1.0 if there was a positive prediction result, and 0 otherwise. We assigned initial weights to each module within the range of 0 to 0.5 based on empirical considerations. To determine the optimal combination of weights, we employed a grid search strategy during 5-fold cross-validation (Fig. 1D). The grid search involved systematically exploring different weight combinations within predefined ranges, allowing us to evaluate the performance of the model across various configurations. This iterative process helped identify the set of weights that maximized the predictive accuracy of the T4SEpp integrated model. The early stopping strategy was similar to that used for T4attention. The final optimal parameters are shown in Fig. 1E.

5.8. Assessment of model performance

Measures including accuracy (ACC), sensitivity (SN), specificity (SP), precision (PR), F1-score, Matthew’s correlation coefficient (MCC), the area under the receiver operating characteristic curve (rocAUC), and the precision recall rate curve (AUPRC) were calculated to evaluate and compare the performance of models predicting T4SEs. Some of these measures are defined as follows:

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

$$\boldsymbol{SN} = \frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FN}}$$

$$\boldsymbol{SP} = \frac{\boldsymbol{TN}}{\boldsymbol{TN}+\boldsymbol{FP}}$$

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

$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN}$$

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

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

5.9. RNA-seq analysis

RNA-seq datasets of *H. pylori* 26695 under different conditions were downloaded from the NCBI GEO DataSets database with accessions GSE165055 and GSE165056 [60]. After removing the adapters and low-quality sequences with Trimmomatic v0.39 [91], the cleaned reads were mapped to the *H. pylori* 26695 reference genome (NC_000915.1) using READemption (Version 2.0.0) [92]. The annotated genes were then quantified and analyzed. Protein-Protein Interaction (PPI) networks were built and visualized using Cytoscope v3.9.1 [82].

Availability

The online version of the T4SEpp is freely accessible at https://bis.zju.edu.cn/T4SEpp. The standalone version of the T4SEpp model and the individual modules are also available at https://github.com/yuemhu/T4SEpp. RNA-seq data are publicly available in the NCBI GEO DataSets database with accession numbers GSE165055 and GSE165056.

Authors' contribution

MC conceived and supervised the project. YH, MC, and YW coordinated the project. YH, YZ, YH, and ZZ dataset collection. YH provided codes, models and software tools. YH, XH, and HC developed the website. YH and YW performed model comparison and RNA-seq data analyses. YH, XH, HC, SL, QN, YW, and MC wrote the first draft of this manuscript. YH, YW, and MC revised the manuscript accordingly.

Funding

This work was supported by the National Key Research and Development Program of China (2016YFA0501704, 2023YFE0112300), the National Natural Sciences Foundation of China (31771477, 32070677), the Science and Technology Innovation Leading Scientist (2022 R52035), and the 151 talent project of Zhejiang Province (first level), Collaborative Innovation Center for Modern Crop Production co-sponsored by province and ministry, and the Natural Science Fund of Shenzhen (JCYJ20190808165205582).

Author statement

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.01.015.

References

[1] Costa TR, Felisberto-Rodrigues C, Meir A, Prevost MS, Redzej A, et al. Secretion systems in Gram-negative bacteria: structural and mechanistic insights. Nat Rev Microbiol 2015;13(6):343–59. https://doi.org/10.1038/nrmicro3456.
[2] Hui X, Chen Z, Zhang J, Lu M, Cai X, et al. Computational prediction of secreted proteins in gram-negative bacteria. Comput Struct Biotechnol J 2021;19:1806–28. https://doi.org/10.1016/j.csbj.2021.03.019.
[3] Grohmann E, Christie PJ, Waksman G, Backert S. Type IV secretion in Gram-negative and Gram-positive bacteria. Mol Microbiol 2018;107(4):455–71. https://doi.org/10.1111/mmi.13896.
[4] Galan JE, Waksman G. Protein-injection machines in bacteria. Cell 2018;172(6):1306–18. https://doi.org/10.1016/j.cell.2018.01.034.
[5] Waksman G. From conjugation to T4S systems in Gram-negative bacteria: a mechanistic biology perspective. EMBO Rep 2019;20(2). https://doi.org/10.15252/embr.201847012.
[6] Li YG, Hu B, Christie PJ. Biological and structural diversity of Type IV secretion systems. Microbiol Spectr 2019;7(2). https://doi.org/10.1128/microbiolspec.PSIB-0012-2018.
[7] Gonzalez-Rivera C, Bhatty M, Christie PJ. Mechanism and function of type IV secretion during infection of the human host. Microbiol Spectr 2016;4(3). https://doi.org/10.1128/microbiolspec.VMBF-0024-2015.
[8] Christie PJ. The mosaic type IV secretion systems. EcoSal 2016;7(1). https://doi.org/10.1128/ecosalplus.ESP-0020-2015.
[9] Christie PJ, Gomez Valero L, Buchrieser C. Biological diversity and evolution of type IV secretion systems. Curr Top Microbiol Immunol 2017;413:1–30. https://doi.org/10.1007/978-3-319-75241-9_1.
[10] Chandran Darbari V, Waksman G. Structural biology of bacterial type IV secretion systems. Annu Rev Biochem 2015;84:603–29. https://doi.org/10.1146/annurev-biochem-062911-102821.
[11] Sheedlo MJ, Ohi MD, Lacy DB, Cover TL. Molecular architecture of bacterial type IV secretion systems. PLoS Pathog 2022;18(8):e1010720. https://doi.org/10.1371/journal.ppat.1010720.
[12] Ansari S, Yamaoka Y. Helicobacter pylori virulence factor cytotoxin-associated gene A (CagA)-mediated gastric pathogenicity. Int J Mol Sci 2020;21(19). https://doi.org/10.3390/ijms21197430.
[13] Hubber A, Roy CR. Modulation of host cell function by Legionella pneumophila type IV effectors. Annu Rev Cell Dev Biol 2010;26:261–83. https://doi.org/10.1146/annurev-cellbio-100109-104034.
[14] Wozniak RA, Waldor MK. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol 2010;8(8):552–63. https://doi.org/10.1038/nrmicro2382.
[15] Wallden K, Rivera-Calzada A, Waksman G. Type IV secretion systems: versatility and diversity in function. Cell Microbiol 2010;12(9):1203–12. https://doi.org/10.1111/j.1462-5822.2010.01499.x.
[16] Costa TRD, Harb L, Khara P, Zeng L, Hu B, et al. Type IV secretion systems: advances in structure, function, and activation. Mol Microbiol 2021;115(3):436–52. https://doi.org/10.1111/mmi.14670.
[17] Burns DL. Type IV transporters of pathogenic bacteria. Curr Opin Microbiol 2003;6(1):29–34. https://doi.org/10.1016/s1369-5274(02)00006-1.
[18] Cover TL, Lacy DB, Ohi MD. The Helicobacter pylori cag type IV secretion system. Trends Microbiol 2020;28(8):682–95. https://doi.org/10.1016/j.tim.2020.02.004.
[19] Ward DV, Zambryski PC. The six functions of Agrobacterium VirE2. Proc Natl Acad Sci USA 2001;98(2):385–6. https://doi.org/10.1073/pnas.98.2.385.
[20] Schrammeijer B, den Dulk-Ras A, Vergunst AC, Jurado Jacome E, Hooykaas PJ. Analysis of Vir protein translocation from Agrobacterium tumefaciens using Saccharomyces cerevisiae as a model: evidence for transport of a novel effector protein VirE3. Nucleic Acids Res 2003;31(3):860–8. https://doi.org/10.1093/nar/gkg179.
[21] Hofreuter D, Odenbreit S, Puls J, Schwan D, Haas R. Genetic competence in Helicobacter pylori: mechanisms and biological implications. Res Microbiol 2000;151(6):487–91. https://doi.org/10.1016/s0923-2508(00)00164-9.

[22] Lee YW, Wang J, Newton HJ, Lithgow T. 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. https://doi.org/10.1016/j.mib.2020.04.002.

[23] Burstein D, Zusman T, Degtyar E, Viner R, Segal G, et al. Genome-scale identification of Legionella pneumophila effectors using a machine learning approach. PLoS Pathog 2009;5(7):e1000508. https://doi.org/10.1371/journal.ppat.1000508.

[24] Zhao Z, Hu Y, Hu Y, White AP, Wang Y. Features and algorithms: facilitating investigation of secreted effectors in Gram-negative bacteria. Trends Microbiol 2023;31(11):1162–78. https://doi.org/10.1016/j.tim.2023.05.011.

[25] Lockwood S, Voth DE, Brayton KA, Beare PA, Brown WC, et al. Identification of Anaplasma marginale type IV secretion system effector proteins. Plos One 2011;6(11). https://doi.org/10.1371/journal.pone.0027724.

[26] 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. https://doi.org/10.3389/fmicb.2019.01391.

[27] Esna Ashari Z, Brayton KA, Broschat SL. Using an optimal set of features with a machine learning-based approach to predict effector proteins for Legionella pneumophila. PLoS One 2019;14(1):e0202312. https://doi.org/10.1371/journal.pone.0202312.

[28] Wang Y, Wei X, Bao H, Liu SL. Prediction of bacterial type IV secreted effectors by C-terminal features. BMC Genom 2014;15:50. https://doi.org/10.1186/1471-2164-15-50.

[29] Esna Ashari Z, Dasgupta N, Brayton KA, Broschat SL. An optimal set of features for predicting type IV secretion system effector proteins for a subset of species based on a multi-level feature selection approach. PLoS One 2018;13(5):e0197041. https://doi.org/10.1371/journal.pone.0197041.

[30] Zou L, Nan C, Hu F. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. Bioinformatics 2013;29(24):3135–42. https://doi.org/10.1093/bioinformatics/btt554.

[31] Wang J, Yang B, An Y, Marquez-Lago T, Leier A, et al. Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. Brief Bioinform 2019;20(3):931–51. https://doi.org/10.1093/bib/bbx164.

[32] Hong J, Luo Y, Mou M, Fu J, Zhang Y, 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(5):1825–36. https://doi.org/10.1093/bib/bbz120.

[33] Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, et al. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst 2019;32:9689–701.

[34] Zhang Y, Zhang Y, Xiong Y, Wang H, Deng Z, et al. T4SEfinder: a bioinformatics tool for genome-scale prediction of bacterial type IV secreted effectors using pre-trained protein language model. Brief Bioinform 2022;23(1). https://doi.org/10.1093/bib/bbab420.

[35] Devlin J., Chang M.-W., Lee K., Toutanova K., editors. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding2019 June; Minneapolis, Minnesota: Association for Computational Linguistics.

[36] Stärk H, Dallago C, Heinzinger M, Rost B. Light attention predicts protein location from the language of life. Bioinform Adv 2021;1(1):vbab035. https://doi.org/10.1093/bioadv/vbab035.

[37] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, et al. ProtTrans: toward Understanding the Language of Life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2022;44(10):7112–27. https://doi.org/10.1109/TPAMI.2021.3095381.

[38] Rives A, Meier J, Sercu T, Goyal S, Lin Z, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 2021;118(15). https://doi.org/10.1073/pnas.2016239118.

[39] Lee J, Yoon W, Kim S, Kim D, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40. https://doi.org/10.1093/bioinformatics/btz682.

[40] Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinforma 2019;20(1):723. https://doi.org/10.1186/s12859-019-3220-8.

[41] Wagner N, Alburquerque M, Ecker N, Dotan E, Zerah B, et al. Natural language processing approach to model the secretion signal of type III effectors. Front Plant Sci 2022;13:1024405. https://doi.org/10.3389/fpls.2022.1024405.

[42] Teufel F, Almagro Armenteros JJ, Johansen AR, Gislason MH, Pihl SI, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 2022;40(7):1023–5. https://doi.org/10.1038/s41587-021-01156-3.

[43] Hui X, Chen Z, Lin M, Zhang J, Hu Y, et al. T3SEpp: an integrated prediction pipeline for bacterial type III secreted effectors. mSystems 2020;5(4). https://doi.org/10.1128/mSystems.00288-20.

[44] Dong X, Lu X, Zhang Z. BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. Database 2015;2015:bav064. https://doi.org/10.1093/database/bav064.

[45] Goldberg T, Rost B, Bromberg Y. Computational prediction shines light on type III secretion origins. Sci Rep 2016;6:34516. https://doi.org/10.1038/srep34516.

[46] Wagner N, Avram O, Gold-Binshtok D, Zerah B, Teper D, et al. Effectidor: an automated machine-learning-based web server for the prediction of type-III secretion system effectors. Bioinformatics 2022;38(8):2341–3. https://doi.org/10.1093/bioinformatics/btac087.

[47] Meyer DF, Noroy C, Moumene A, Raffaele S, Albina E, 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(20):9218–29. https://doi.org/10.1093/nar/gkt718.

[48] Noroy C, Lefrançois T, Meyer DF. Searching algorithm for Type IV effector proteins (S4TE) 2.0: improved tools for Type IV effector prediction, analysis and comparison in proteobacteria. Plos Comput Biol 2019;15(3). https://doi.org/10.1371/journal.pcbi.1006847.

[49] Bi D, Liu L, Tai C, Deng Z, Rajakumar K, et al. SecReT4: a web-based bacterial type IV secretion system resource. Nucleic Acids Res 2013;41(Database issue):D660–5. https://doi.org/10.1093/nar/gks1248.

[50] Wu X, Bartel DP. kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. Nucleic Acids Res 2017;45(W1):W534–8. https://doi.org/10.1093/nar/gkx323.

[51] Burstein D, Amaro F, Zusman T, Lifshitz Z, Cohen O, et al. Genomic analysis of 38 Legionella species identifies large and diverse effector repertoires. Nat Genet 2016;48(2):167–75. https://doi.org/10.1038/ng.3481.

[52] Hooi JKY, Lai WY, Ng WK, Suen MMY, Underwood FE, et al. Global prevalence of helicobacter pylori infection: systematic review and meta-analysis. Gastroenterology 2017;153(2):420–9. https://doi.org/10.1053/j.gastro.2017.04.022.

[53] Cover TL, Blaser MJ. Helicobacter pylori in health and disease. Gastroenterology 2009;136(6):1863–73. https://doi.org/10.1053/j.gastro.2009.01.073.

[54] Blaser MJ, Perez-Perez GI, Kleanthous H, Cover TL, Peek RM, et al. Infection with Helicobacter pylori strains possessing cagA is associated with an increased risk of developing adenocarcinoma of the stomach. Cancer Res 1995;55(10):2111–5.

[55] Figueiredo C, Machado JC, Pharoah P, Seruca R, Sousa S, et al. Helicobacter pylori and interleukin 1 genotyping: an opportunity to identify high-risk individuals for gastric carcinoma. J Natl Cancer Inst 2002;94(22):1680–7. https://doi.org/10.1093/jnci/94.22.1680.

[56] Plummer M, van Doorn LJ, Franceschi S, Kleter B, Canzian F, et al. Helicobacter pylori cytotoxin-associated genotype and gastric precancerous lesions. J Natl Cancer Inst 2007;99(17):1328–34. https://doi.org/10.1093/jnci/djm120.

[57] Cover TL. Helicobacter pylori diversity and gastric cancer risk. e01869-01815 mBio 2016;7(1). https://doi.org/10.1128/mBio.01869-15.

[58] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424. https://doi.org/10.3322/caac.21492.

[59] Knorr J, Ricci V, Hatakeyama M, Backert S. Classification of helicobacter pylori virulence factors: Is CagA a toxin or not? Trends Microbiol 2019;27(9):731–8. https://doi.org/10.1016/j.tim.2019.04.010.

[60] Loh JT, Shum MV, Jossart SDR, Campbell AM, Sawhney N, et al. Delineation of the pH-Responsive regulon controlled by the helicobacter pylori ArsRS two-component system. Infect Immun 2021;89(4). https://doi.org/10.1128/IAI.00597-20.

[61] Nesic D, Buti L, Lu X, Stebbins CE. Structure of the helicobacter pylori CagA oncoprotein bound to the human tumor suppressor ASPP2. Proc Natl Acad Sci 2014;111(4):1562–7. https://doi.org/10.1073/pnas.1320631111.

[62] Poppe M, Feller SM, Romer G, Wessler S. Phosphorylation of Helicobacter pylori CagA by c-Abl leads to cell motility. Oncogene 2007;26(24):3462–72. https://doi.org/10.1038/sj.onc.1210139.

[63] Churin Y, Al-Ghoul L, Kepp O, Meyer TF, Birchmeier W, et al. Helicobacter pylori CagA protein targets the c-Met receptor and enhances the motogenic response. J Cell Biol 2003;161(2):249–55. https://doi.org/10.1083/jcb.200208039.

[64] Suzuki M, Mimuro H, Suzuki T, Park M, Yamamoto T, et al. Interaction of CagA with Crk plays an important role in Helicobacter pylori-induced loss of gastric epithelial cell adhesion. J Exp Med 2005;202(9):1235–47. https://doi.org/10.1084/jem.20051027.

[65] Murata-Kamiya N, Kurashima Y, Teishikata Y, Yamahashi Y, Saito Y, et al. Helicobacter pylori CagA interacts with E-cadherin and deregulates the beta-catenin signal that promotes intestinal transdifferentiation in gastric epithelial cells. Oncogene 2007;26(32):4617–26. https://doi.org/10.1038/sj.onc.1210251.

[66] Lee DG, Kim HS, Lee YS, Kim S, Cha SY, et al. Helicobacter pylori CagA promotes Snail-mediated epithelial-mesenchymal transition by reducing GSK-3 activity. Nat Commun 2014;5:4423. https://doi.org/10.1038/ncomms5423.

[67] Saadat I, Higashi H, Obuse C, Umeda M, Murata-Kamiya N, et al. Helicobacter pylori CagA targets PAR1/MARK kinase to disrupt epithelial cell polarity. Nature 2007;447(7142):330–3. https://doi.org/10.1038/nature05765.

[68] Mishra JP, Cohen D, Zamperone A, Nesic D, Muesch A, et al. CagA of Helicobacter pylori interacts with and inhibits the serine-threonine kinase PRK2. Cell Microbiol 2015;17(11):1670–82. https://doi.org/10.1111/cmi.12464.

[69] Saju P, Murata-Kamiya N, Hayashi T, Senda Y, Nagase L, et al. Host SHP1 phosphatase antagonizes Helicobacter pylori CagA and can be downregulated by Epstein-Barr virus. Nat Microbiol 2016;1:16026. https://doi.org/10.1038/nmicrobiol.2016.26.

[70] Higashi H, Tsutsumi R, Muto S, Sugiyama T, Azuma T, et al. SHP-2 tyrosine phosphatase as an intracellular target of Helicobacter pylori CagA protein. Science 2002;295(5555):683–6. https://doi.org/10.1126/science.1067147.

[71] Lamb A, Yang XD, Tsang YH, Li JD, Higashi H, et al. Helicobacter pylori CagA activates NF-kappaB by targeting TAK1 for TRAF6-mediated Lys 63 ubiquitination. EMBO Rep 2009;10(11):1242–9. https://doi.org/10.1038/embor.2009.210.

[72] Amieva MR, Vogelmann R, Covacci A, Tompkins LS, Nelson WJ, et al. Disruption of the epithelial apical-junctional complex by Helicobacter pylori CagA. Science 2003;300(5624):1430–4. https://doi.org/10.1126/science.1081919.

[73] Kaundal R, Loaiza CD, Duhan N, Flann N. deepHPI: a comprehensive deep learning platform for accurate prediction and visualization of host-pathogen protein-protein interactions. Brief Bioinform 2022;23(3). https://doi.org/10.1093/bib/bbac125.

[74] Goncalves IG, Simoes LC, Simoes M. Legionella pneumophila. Trends Microbiol 2021;29(9):860–1. https://doi.org/10.1016/j.tim.2021.04.005.

[75] Mondino S, Schmidt S, Rolando M, Escoll P, Gomez-Valero L, et al. Legionnaires' Disease: state of the art knowledge of pathogenesis mechanisms of Legionella. Annu Rev Pathol 2020;15:439–66. https://doi.org/10.1146/annurev-pathmechdis-012419-032742.
[76] Hatakeyama M. Oncogenic mechanisms of the Helicobacter pylori CagA protein. Nat Rev Cancer 2004;4(9):688–94. https://doi.org/10.1038/nrc1433.
[77] Hatakeyama M. Helicobacter pylori CagA and gastric cancer: a paradigm for hit-and-run carcinogenesis. Cell Host Microbe 2014;15(3):306–16. https://doi.org/10.1016/j.chom.2014.02.008.
[78] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13):1658–9. https://doi.org/10.1093/bioinformatics/btl158.
[79] Xiong Y, Wang Q, Yang J, Zhu X, Wei DQ. PredT4SE-Stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. Front Microbiol 2018;9:2571. https://doi.org/10.3389/fmicb.2018.02571.
[80] O'Callaghan D, Cazevieille C, Allardet-Servent A, Boschiroli ML, Bourg G, et al. A homologue of the Agrobacterium tumefaciens VirB and Bordetella pertussis Ptl type IV secretion systems is essential for intracellular survival of Brucella suis. Mol Microbiol 1999;33(6):1210–20. https://doi.org/10.1046/j.1365-2958.1999.01569.x.
[81] Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14(9):755–63. https://doi.org/10.1093/bioinformatics/14.9.755.
[82] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–504. https://doi.org/10.1101/gr.1239303.
[83] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;12(1):59–60. https://doi.org/10.1038/nmeth.3176.
[84] Hu Y, Huang H, Cheng X, Shu X, White AP, et al. A global survey of bacterial type III secretion systems and their effectors. Environ Microbiol 2017;19(10):3879–95. https://doi.org/10.1111/1462-2920.13755.
[85] Hui X, Hu Y, Sun MA, Shu X, Han R, et al. EBT: a statistic test identifying moderate size of significant features with balanced power and precision for genome-wide rate comparisons. Bioinformatics 2017;33(17):2631–41. https://doi.org/10.1093/bioinformatics/btx294.
[86] Steinegger M, Soding J. Clustering huge protein sequence sets in linear time. Nat Commun 2018;9(1):2542. https://doi.org/10.1038/s41467-018-04964-5.
[87] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31(6):926–32. https://doi.org/10.1093/bioinformatics/btu739.
[88] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21:1–67. https://doi.org/10.48550/arXiv.1910.10683.
[89] Lan Z., Chen M., Goodman. S., Gimpel. K., Sharma. P., et al. (2019) ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. International Conference on Learning Representations 2019; 20 Dec. doi: 10.48550/arXiv.1909.11942.
[90] Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42(2):318–27. https://doi.org/10.1109/TPAMI.2018.2858826.
[91] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
[92] Forstner KU, Vogel J, Sharma CM. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. Bioinformatics 2014;30(23):3421–3. https://doi.org/10.1093/bioinformatics/btu533.