IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 13, 2020, accepted January 25, 2020, date of publication February 3, 2020, date of current version February 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2971091

SMOPredT4SE: An Effective Prediction of Bacterial Type IV Secreted Effectors Using SVM Training With SMO

ZIHAO YAN[1], DONG CHEN[2], ZHIXIA TENG[1], DONGHUA WANG[3], AND YANJUAN LI[1]

[1]School of information and computer engineering, Northeast Forestry University, Harbin 150040, China
[2]School of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin 150050, China
[3]Department of General Surgery, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin 150088, China

Corresponding authors: Donghua Wang (wangdonghua7885@163.com) and Yanjuan Li (liyanjuan@nefu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61901103, in part by the Fundamental Research Funds for the Central Universities under Grant 2572017CB33, and in part by the Natural Science Foundation of Heilongjiang Province under Grant LH2019F002.

ABSTRACT Various bacterial pathogens can deliver their secreted effectors to host cells via type IV secretion system (T4SS) and cause host diseases. Since T4SS secreted effectors (T4SEs) play important roles in the interaction between pathogens and host, identifying T4SEs is crucial to understanding of the pathogenic mechanism of T4SS. We established an effective predictor called SMOPredT4SE to identify T4SEs from protein sequences. SMOPredT4SE employed combination features of series correlation pseudo amino acid composition and position-specific scoring matrix to present protein sequences, and employed support vector machines (SVM) training with sequential minimal optimization (SMO) arithmetic to train the prediction model (To distinguish it from the traditional SVM, we will abbreviate it as SMO later). In the 5-fold cross-validation test, SMOPredT4SE's overall accuracy was 95.6%. Experiments on comparison with other feature, classifiers, and existing methods are conducted. Experimental results show the effectiveness of SMOPredT4SE in predicting T4SEs.

INDEX TERMS Machine learning, protein classification, sequential minimal optimization, type IV secreted effector.

I. INTRODUCTION

Gram-negative bacteria are generally classified as bacteria that become red with gram staining, many of which are common bacteria that cause hospital infections. For example, *Escherichia coli*, *Proteusbacillus vulgaris*, *Bordetella pertussis*, *Acinetobacter baumannii*, *Serratia*, *Enterobacter*, and *Pseudomonas* all fall into this category [1], [2]. 100 years after the discovery of a bacterial 'endotoxin', 50 years after the introduction of antibiotics, and 25 years after the routine use of intensive care units to support septic shock patients, gram-negative infections continue to account for significant morbidity and mortality [3]. According to their outer membrane secretion mechanisms, gram-negative bacteria have been identified into eight different secretion systems (type I to type VIII) [4], all of which show differences in evolution and function. Of the eight secretion systems, secretion pathways of type IV secretion systems (T4SSs) ancestrally related to bacterial conjugation systems [5] and they are widely distributed in a variety of bacteria, such as *B. pertussis*, *Helicobacter pylori*, and *Legionella pneumophila* [6]–[9]. Such pathogens use T4SSs to translocate macromolecular substrates directly into bacterial, plant, or human target cells [10]. T4SSs are medically important, contributing to virulence-gene spread, genome plasticity, and the alteration of host cellular processes during infection [10]. Contact-dependent translocation of effector proteins via the T4SSs allows pathogens to secrete type IV secreted effectors (T4SEs) across both bacterial membranes [11]. Accurate and reliable identification of T4SEs is an important step to understand the pathogenesis of gram-negative bacteria. Due to the importance of T4SEs in biology, many experimental methods have been developed to identify them, such as genetic complementation, reporter protein fusion, and secretion

The associate editor coordinating the review of this manuscript and approving it for publication was Dariusz Mrozek.

apparatus or chaperone interactions [12]–[15]. However, these biological experimental approaches are time and resource consuming. Prediction methods based on machine learning (ML) algorithm shows high efficiency and can be targeted at large-scale data.

At the beginning of the 21st century, the success of the international Human Genome Project (HGP) resulted in rapid growth of biological information, causing researchers to consider using mathematical calculation methods to reveal the hidden laws in various biological data, and to predict the structure and function of unknown proteins [16], [17]. In recent years, some prediction methods for T4SEs based on ML algorithms have gradually developed [18]. Burstein *et al.* [19] first measured sequence similarity to known effector proteins and eukaryotic proteomes, taxonomic distribution among bacteria and metazoa, genome organization, G+C content, C-terminal signal, and regulatory elements as the input features. And then they used naive Bayes, Bayesian networks, support vector machines (SVM), sequential minimal optimization (SMO), neural networks (multilayer perceptron) [20], [21], and a Voting algorithm [22], [23] based on these algorithms to predict T4SEs. By combining gene screening and bioinformatics analyses, Chen *et al.* [24] obtained a large number of candidate protein substrates for *Coxiella burnetii* Dot/Icm T4SEs. Lifshitz *et al.* [6] implemented a hidden semi-Markov model (HSMM) to characterize the amino acid composition (AAC) of the input signal to identify T4SEs in *L. pneumophila*, *Legionella longbeachae*, and *C. burnetii*. Zou *et al.* [25] calculated four types of distinctive features, namely AAC, dipeptide composition, position-specific scoring matrix (PSSM) composition, and auto covariance transformation of PSSM from primary sequences. And they developed a classifier, T4EffPred, using the SVM with these features and their different combinations for T4SEs prediction. Wang *et al.* [26] compared C-terminal sequence and position-specific amino acid compositions, possible motifs, and structural features of T4SEs from different bacteria. Then presented the inter-species prediction tool package, T4SEpre, to help find new pathogenic T4SEs efficiently in a variety of pathogenic bacteria. In an effort to improve predictive performance, An *et al.* [27] constructed three ensemble models based on ML algorithms by integrating the output of all individual predictors reviewed. Wang *et al.* [28] suggested it would be incomplete to use only the features of C-terminal residues for prediction, targeting information can also be encoded in the N-terminal region of at least some T4SEs. Therefore, they integrated 50 N-terminal and 100 C-terminal residues, and calculated their AAC composition, transition, distribution, and PSSM. Then, ranked the importance of 150 residues to T4SE based on the information gain, and selected 125 residues at different positions as the prediction model of SVM. Recently, Xiong *et al.* [29] proposed an ensemble classification method based on stack generalization to further improve the predictive performance. They used the same datasets as in the study by Wang *et al.* and took PSSM-composition as the input features. And then, they used eight ML algorithms to build the base-classifier, and used the output of the optimal combination as input to the meta-classifier. In this way, they got better prediction results.

As is evident, there are currently many different feature representation methods and ML algorithms to predict T4SEs using bioinformatics experiments. However, T4SEs are an extremely small fraction of the total cellular proteins, and as the number continues to grow, it will be necessary to develop a classifier with higher specificity and performance. In this study, we used a combination of series correlation pseudo amino acid composition (SC-PseAAC) and PSSM-composition, which shows great specificity in the classification task. The SVM training with SMO algorithm, which is faster in operation and more suitable for processing large-scale data, was selected as the classifier to build the representative model. Our new method is more efficient and concise, and achieves better results than previously published methods.

II. MATERIALS AND METHODS

A. DATASET

In this study, we used the benchmark dataset collected and collated by Wang *et al*. [28]. T4SEs in the dataset all came from the effector dataset in the SecRet4 database and other studies. Non-effectors were randomly selected from the same strains where the positive training sequences originated, followed by removal of known effectors and their homologs. The dataset contained 1,765 protein sequences, 380 of which were T4SEs and 1,385 were non-T4SEs. All protein sequences were divided into two groups: 915 sequences as a set of 5-fold cross-validation and 850 sequences as a set of independent-validation. The training dataset (Train-915) composed of 305 T4SEs and 610 non-T4SEs, all randomly selected from the set of positive and negative sequences, respectively. They were then further randomly divided into five sets for the input of 5-fold cross-validation. The independent testing dataset (Test-850) contained 75 positive and 775 negative sequences.

B. FEATURE EXTRACTION

Feature extraction is an important step in building a protein sequence training model [30]–[39]. A feature with good specificity and high identification can greatly improve the prediction performance of the model. In this study, we tried five features and different combinations of them. Finally, the best feature or combination of features was selected according to the experimental results.

1) PSEUDO AMINO ACID COMPOSITION

AAC is a powerful expression of the composition of a protein sequence, where the 20 naturally occurring amino acids are represented by letters A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y [40]. AAC, however, has a limitation. It cannot express the sequence order. Pseudo

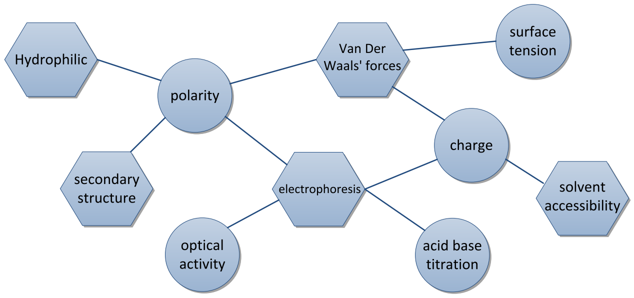


FIGURE 1. Expandable physical and chemical properties in SC-PseAAC.

amino acid composition (PseAAC) is the addition of a set of discrete sequence correlation factors based on the AAC, forming (20 + λ)-D vector discrete numbers to represent a protein [41], [42]:

$$X = [x_1 \cdots x_{20}, x_{20+1} \cdots x_{20+\lambda}]^T \tag{1}$$

where

$$x_i = \begin{cases} \dfrac{f_i}{\sum\limits_{j=1}^{20} f_j + \omega \sum\limits_{k=1}^{\lambda} \theta_k} & (1 \le i \le 20) \\ \dfrac{\omega\theta_{i-20}}{\sum\limits_{j=1}^{20} f_j + \omega \sum\limits_{k=1}^{\lambda} \theta_k} & (20+1 \le i \le 20+\lambda) \end{cases} \tag{2}$$

where f is the occurrence frequency of the 20 amino acids in the protein, θ is the sequence correlation factor, and w is the weight factor for the sequence order effect. The first 20 components reflect the effect of the amino acid composition, whereas components from 20+1 to 20 + λ reflect the effect of sequence order [43]. SC-PseAAC is a further improvement of PseAAC, with the following advantages: (a) It has a simple structure and a small number of features, which can greatly improve the operation speed for large-scale data; (b) the λ components can be defined by the user at will, introducing physical and chemical properties such as hydrophilicity, hydrophobicity, polarity, and charge properties; and (c) since its vectors are composed of discrete numbers, it is highly scalable without affecting the performance of other components. More extensive physical and chemical properties are shown in Figure 1.

2) POSITION-SPECIFIC SCORING MATRIX COMPOSITION

A given protein sequence S is represented as $S_1S_2 \cdots S_L$, where $S_i(1 \le i \le L)$ represents the amino acid residue appearing in the ith position of S, and L is the length of S [44], [45]. The so-called evolutionary profile of S is the PSSM, which has been applied to the prediction of protein sequences in numerous previous studies, achieving good results [28], [29], [46]–[49]. In this study, we performed three iterative searches of the Uniref50 database using PSI-BLAST, setting the e-value to 0.001 to generate the BLOSUM62 replacement matrix. The results are shown below:

$$PSSM = (S_1, S_2 \cdots S_{20}) \tag{3}$$

where $S_i(i = 1, 2 \cdots 20)$ is the column vector of amino acid type i in the matrix. We represented each of the column vectors as:

$$S_j = \left(s_{1,j}, s_{2,j}, \cdots s_{L,j}\right)^T \quad (j = 1, 2, \cdots 20) \tag{4}$$

The rows of the same amino acids in the PSSM matrix are then added together to obtain the 20∗20 feature of PSSM-composition (400-dimensional). The specific process is shown in figure 2.

3) 188-D FEATURE

188-D feature is a powerful representation of a protein sequence, containing basic amino acids and a variety of physical and chemical properties. Its reliability has also been demonstrated in a number of applications to the modeling of protein predictors [50], [51]. The first 20 dimensions are the number of amino acids in the sequence and the latter dimensions include specific physicochemical properties, such as hydrophilicity, hydrophobicity, Van der Waals forces, polarity, and conversion frequency.

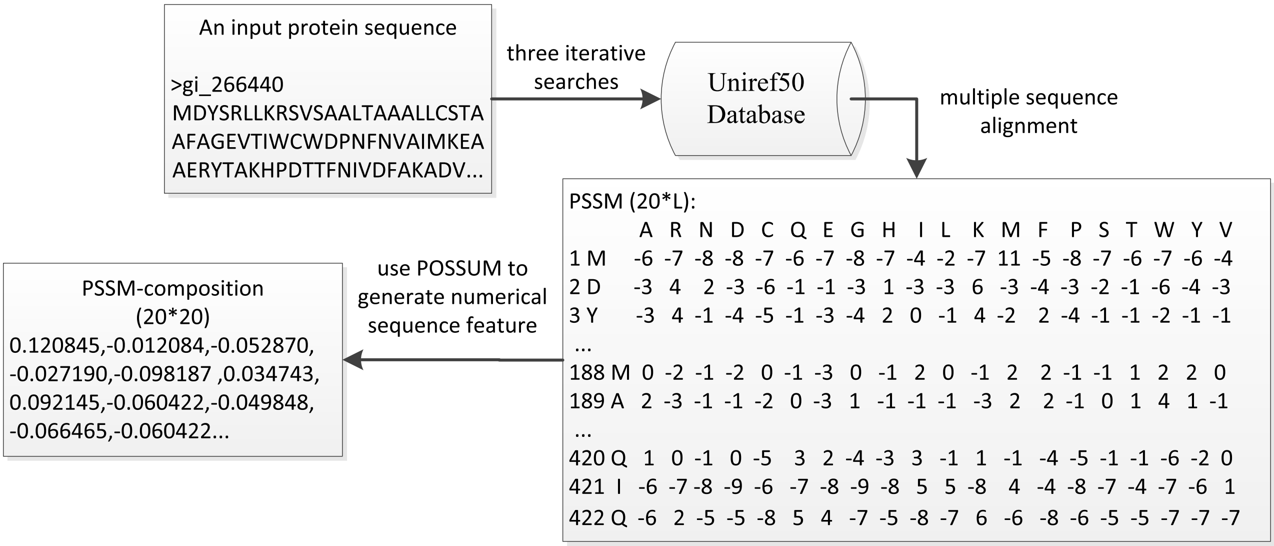
4) ADAPTIVE k-SKIP-n-GRAM FEATURES

Since the traditional n-gram feature is sparse in short amino acid sequences, Guthrie et al. proposed an improved feature that incorporates location information into it; the adaptive k-skip-n-gram (400-D) feature [52]. It obtains n-gram information by jumping a certain number of words or positions, which solves the problem of the sparsity of feature space to some extent [53].

C. CLASSIFICATION METHODS

Waikato Environment for Knowledge Analysis (Weka) is an open source ML and data mining software which assembled many ML algorithms capable of mining data [54]–[58]. Weka is a flexible tool in which we can integrate our own algorithms and even borrow his algorithms to implement visualization tools. In our research, we experimented with a variety of ML algorithms based on Weka platform such as SVM, Random Forest (RF), Naïve Bayes, k-Nearest Neighbor (kNN), Bagging, Stochastic gradient descent (SGD), LibD3C. Finally, we chose the SMO algorithm, and the results of the experiment will be presented in the next section.

Since SVMs were first proposed by Vladimir Vapnik [59], they have become very popular binary classification algorithms and have achieved good results in many classification tasks and regression problems, such as image recognition, text classification, and protein sequence classification [60]–[70]. SVM is a generalized linear classifier for binary classification of data by supervised learning, which is founded based on the statistical learning theory and structural risk minimization, its decision boundary is the maximum margin hyperplane to solve the learning sample. The classification problem is transformed into a convex quadratic programming problem to solve. Since the basic concept of classification learning is to find a partition hyperplane in the sample space to separate the samples of different categories,


three iterative searches
Uniref50 Database
multiple sequence alignment
use POSSUM to generate numerical sequence feature

An input protein sequence

>gi_266440
MDYSRLLKRSVSAALTAAALLCSTA
AFAGEVTIWCWDPNFNVAIMKEA
AERYTAKHPDTTFNIVDFAKADV...

PSSM (20*L):

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1 M	-6	-7	-8	-8	-7	-6	-7	-8	-7	-4	-2	-7	11	-5	-8	-7	-6	-7	-6	-4
2 D	-3	4	2	-3	-6	-1	-1	-3	1	-3	-3	6	-3	-4	-3	-2	-1	-6	-4	-3
3 Y	-3	4	-1	-4	-5	-1	-3	-4	2	0	-1	4	-2	2	-4	-1	-1	-2	-1	-1
...																				
188 M	0	-2	-1	-2	0	-1	-3	0	-1	2	0	-1	2	2	-1	-1	1	2	2	0
189 A	2	-3	-1	-1	-2	0	-3	1	-1	-1	-1	-3	2	2	-1	0	1	4	1	-1
...																				
420 Q	1	0	-1	0	-5	3	2	-4	-3	3	-1	1	-1	-4	-5	-1	-1	-6	-2	0
421 I	-6	-7	-8	-9	-6	-7	-8	-9	-8	5	5	-8	4	-4	-8	-7	-4	-7	-6	1
422 Q	-6	2	-5	-5	-8	5	4	-7	-5	-8	-7	6	-6	-8	-6	-5	-5	-7	-7	-7

PSSM-composition
(20*20)
0.120845,-0.012084,-0.052870,
-0.027190,-0.098187 ,0.034743,
0.092145,-0.060422,-0.049848,
-0.066465,-0.060422...

FIGURE 2. The process from protein sequence to PSSM-composition.

the goal of SVM is to find a hyperplane so that there can be a larger interval between the outlier points that are close to the hyperplane. In other words, it is not necessary to consider all sample points, but only to maximize the interval between the points that are close to the hyperplane. In the sample space, the partition hyperplane can be described by the following linear equation:

$$W^T x + b = 0 \tag{5}$$

where W is the normal vector, which determines the direction of the hyperplane, and b is the displacement, which determines the distance between the hyperplane and the origin. If the hyperplane satisfies the following formula for the training sample (x_i, y_i):

$$\begin{cases} W^T x_i + b \geq +1 & y_i = +1 \\ W^T x_i + b \leq -1 & y_i = -1 \end{cases} \tag{6}$$

this formula is called the maximum interval hypothesis, with $y_i = +1$ indicating that the sample is positive and $y_i = -1$ indicating that the sample is negative. The sample points closest to the hyperplane that satisfy $y_i(W^T x_i + b) = 1$ are called support vectors. The basic model of SVM is a convex quadratic programming problem, which can be solved by Lagrange multiplier method:

$$L(w, b, a) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i(w^T x_i + b)\right) \tag{7}$$

After the training is completed, most of the samples need not be retained, and the final model is only related to the support vector. Based on these strengths, SVM is a sparse and robust classifier which can be used for nonlinear classification by kernel method.

However, because the scale of the quadratic programming (QP) optimization problem SVM had to solve was enormous, it will be slow when processing large-scale data. SMO has a great advantage in solving the QP problem; by breaking the large QP problem into a number of smaller QP problems [71], [72]. Despite an increase in the number of problems, the computing speed of each small problem is greatly improved. Thus, avoiding the QP optimization problem as an internal cycle, greatly reducing processing time and improving computation speed [73]–[76]. The SMO algorithm requires the following steps in the calculation process. Firstly, SMO had to solve for the two Lagrange multipliers, compute the constraints for the two multipliers, and solve for the minimum constraint. Then, in order to speed convergence, SMO uses heuristics to choose which two Lagrange multipliers to jointly optimize [71]. Due to these advantages, the SVM training with SMO is chosen as the main classification method in this paper.

D. PERFORMANCE EVALUATION

To demonstrate the performance of our model, we introduced five evaluation indicators commonly used in bioinformatics [77]–[95]: accuracy (ACC), sensitivity (SE), specificity (SP), Matthew's correlation coefficient (MCC), and F-measure. These metrics are defined as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$SE = \frac{TP}{TP + FN} \tag{9}$$

$$SP = \frac{TN}{TN + FP} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FP)\times(TN+FN)}} \tag{11}$$

$$PR = \frac{TP}{TP + FP} \tag{12}$$

$$F - measure = \frac{2 \times SE \times PR}{SE + PR} \tag{13}$$

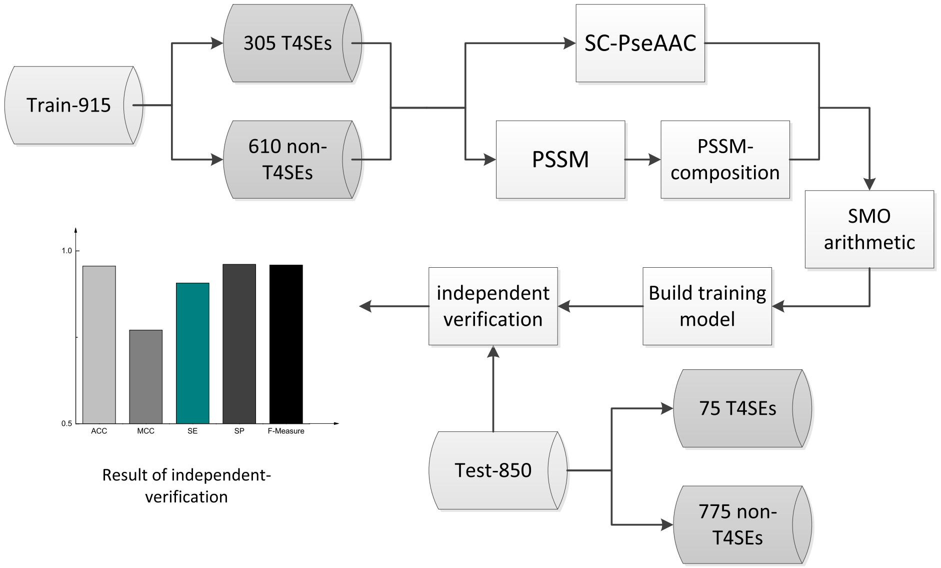


FIGURE 3. Training models and forecasting processes.

where TP (true position) is the number of correctly predicted T4SEs, FP (false position) is the number of wrongly predicted T4SEs as non-T4SEs, TN (true negation) is the number of correctly predicted non-T4SEs, and FN (false negation) is the number of wrongly predicted non-T4SEs as T4SEs. Figure 3 shows the entire process, including the detailed process of feature selection, modeling, and prediction

III. RESULTS AND DISCUSSION

A. CROSS-VALIDATION RESULTS OF TRAIN-915

In many experiments, we tried a variety of methods to extract highly recognizable features from protein sequences in the training set, and used several algorithms to train the model to achieve optimal accuracy. The experimental comparison results are as follows:

1) PERFORMANCE OF DIFFERENT FEATURES ON CROSS-VALIDATION

Using the SMO algorithm, we first tried the above-mentioned PSSM-composition, SC-PseAAC, 188-D, and 488-D features, and their combinations. Parameter c was set to 0.2. Table 1 lists the performance of the four single features and several combinations of features with good performance in 5-fold cross-validation. The results show that the combination of SC-PseAAC and PSSM-composition performed better than the other features and combinations. We also tried a combination of three or more features, with no significant improvement.

2) PERFORMANCE OF DIFFERENT CLASSIFIERS ON CROSS-VALIDATION

We compared SMO with SVM, RF, which performed well in previous studies, and algorithms that are suitable for binary classification problems. As shown in table 2, the SMO algorithm we chose was superior to the other classifiers in every index.

TABLE 1. The result of different features on Train-915.

Feature	ACC (%)	MCC	SE	SP	F-Measure
SC-PseAAC	83.5	0.618	0.659	0.921	0.831
400-D	83.8	0.629	0.708	0.903	0.836
188-D	85.4	0.661	0.672	0.944	0.848
PSSM	90.3	0.778	0.822	0.943	0.902
400-D+PSSM	91.6	0.808	0.819	0.964	0.914
188-D+SC-PseAAC+PSSM	91.9	0.815	0.836	0.961	0.918
188-D+PSSM	92.6	0.831	0.839	0.969	0.925
SC-PseAAC +PSSM	92.6	0.831	0.842	0.967	0.925

TABLE 2. The result of different classifiers on Train-915.

Classifier	ACC(%)	MCC	SE	SP	F-Measure
Naïve Bayes	72.1	0.472	0.839	0.662	0.729
kNN (k=1)	84.9	0.694	0.908	0.820	0.853
Bagging	87.0	0.700	0.707	0.951	0.866
SGD	88.3	0.739	0.845	0.902	0.884
SVM	88.4	0.734	0.750	0.951	0.881
LibD3C	88.6	0.739	0.750	0.954	0.883
RF	89.2	0.753	0.730	0.972	0.888
SMO	92.6	0.831	0.842	0.967	0.925

B. INDEPENDENT-VALIDATION RESULTS OF TEST-850

We used the model built with selected algorithms for independent validation on the Test-850 dataset to test its generalization performance. As shown in table 3 and 4, the combination of SC-PseAAC and PSSM-composition and SMO algorithm performed best. Although the combination of 188-D and PSSM performed as well as SC-PseAAC and PSSM-composition in cross-validation, it performed significantly worse in independent-validation. Therefore, the combination of SC-PseAAC and PSSM-composition with SMO algorithm was determined to have the highest specificity and best stability.

C. COMPARISON WITH OTHER PREDICTORS

Since the same data set as Wang *et al*. and Xiong *et al*. was used in our study, the results of our 5-fold cross-verification

TABLE 3. The result of different features on Test-850.

Feature	ACC(%)	MCC	SE	SP	F-Measure
400-D	87.5	0.481	0.767	0.888	0.892
SC-PseAAC	88.6	0.503	0.767	0.899	0.900
188-D	90.4	0.513	0.680	0.825	0.911
PSSM	93.3	0.676	0.867	0.939	0.939
188-D+PSSM	94.6	0.715	0.853	0.955	0.949
400-D+PSSM	94.6	0.732	0.880	0.951	0.950
188-D+SC-PseAAC+PSSM	94.8	0.728	0.893	0.950	0.951
SC-PseAAC +PSSM	95.6	0.771	0.907	0.961	0.959

TABLE 4. The result of different classifiers on Test-850.

Classifier	ACC (%)	MCC	SE	SP	F-Measure
Naïve Bayes	66.7	0.305	0.880	0.646	0.739
kNN (k=1)	86.6	0.558	0.947	0.858	0.889
SGD	91.5	0.631	0.880	0.919	0.925
SVM	92.7	0.632	0.800	0.939	0.933
Bagging	93.1	0.633	0.773	0.946	0.935
LibD3C	93.5	0.678	0.853	0.943	0.940
RF	94.0	0.685	0.827	0.951	0.944
SMO	95.6	0.771	0.907	0.961	0.959

TABLE 5. Comparison betweenSMOPredT4SE and other method on Train-915.

Method	ACC (%)	MCC	SE	SP	F-Measure
Wang Y. et al.'s	87.8	0.727	0.814	0.912	0.795
PredT4SE-Stack(SVM)	90.6	0.787	0.803	0.957	0.849
PredT4SE-Stack (NB)	90.9	0.795	0.823	0.952	0.857
PredT4SE-Stack (KNN)	91.0	0.797	0.820	0.956	0.857
PredT4SE-Stack (LR)	91.1	0.800	0.810	0.962	0.858
SMO	92.6	0.831	0.842	0.967	0.925

TABLE 6. Comparison betweenSMOPredT4SE and other method on Test-850.

Method	ACC (%)	MCC	SE	SP	F-Measure
Wang Y. et al.'s	85.3	0.518	0.907	0.848	0.521
PredT4SE-Stack(SVM,0.23)	87.5	0.556	0.907	0.872	0.562
PredT4SE-Stack (LR,0.11)	88.7	0.579	0.907	0.885	0.586
PredT4SE-Stack (LR,0.50)	94.4	0.715	0.880	0.950	0.733
PredT4SE-Stack (SVM,0.50)	94.5	0.715	0.867	0.952	0.734
SMOPredT4SE	95.6	0.771	0.907	0.961	0.959

and independent-verification results could be compared with these studies. To illustrate the differences between these methods by a more intuitive means, results comparing these methods are shown in table 5 and 6.

The results show that the new modeling method proposed in this paper is superior to the methods in the five aspects of ACC, SE, SP, MCC, and F-measure. In particular, there was a significant improvement in F-measure (from 0.734 to 0.959). Since F-measure is the weighted harmonic average of precision and recall, this suggests that the experimental method presented here is highly optimal. In a recent study by Xiong *et al.*, they used the stacked ensemble model to improve predictive performance. This increased the difficulty of use and reduced the efficiency of operation to some extent. In comparison, our modeling method is simple, feasible, and highly efficient. This also improves predictive performance in another way.

IV. CONCLUSION

In this work, we propose a simple, efficient, and reliable experimental method for predicting gram-negative bacteria T4SEs based on machine learning algorithms. After comparative analysis of several experiments, we selected the combination of SC-PseAAC and PSSM-composition, which showed high specificity and good stability. We then used SMO algorithms to build prediction models, obtaining excellent results in both training and test datasets. In terms of important indicators, our model yielded ACC of 92.6%, MCC of 0.831, and F-measure of 0.925 in 5-fold cross-validation based on the Train-915 dataset, and ACC of 95.6%, MCC of 0.771, and F-measure of 0.959 in independent-validation based on the Test-850 dataset. In conclusion, we believe that our new model provides a reliable and effective means of screening T4SEs from the huge number of protein sequences. In the future, we will pay more attentions on the deep learning classifiers [96]–[105] and evolutionary strategy [106]–[108].

REFERENCES

[1] R. J. Vivero, G. B. Mesa, S. M. Robledo, C. X. M. Herrera, and G. Cadavid-Restrepo, "Enzymatic, antimicrobial, and leishmanicidal bioactivity of gram-negative bacteria strains from the midgut of *Lutzomyia evansi*, an insect vector of leishmaniasis in Colombia," *Biotechnol. Rep.*, vol. 24, Dec. 2019, Art. no. e00379.

[2] M. Valenzuela-Valderrama, I. A. González, and C. E. Palavecino, "Photodynamic treatment for multidrug-resistant gram-negative bacteria: Perspectives for the treatment of *Klebsiella pneumoniae* infections," *Photodiagnosis Photodyn. Therapy*, vol. 28, pp. 256–264, Dec. 2019.

[3] D. Morrison, R. Danner, C. Dinarello, R. Munford, C. Natanson, M. Pollack, J. Spitzer, R. Ulevitch, S. Vogel, and E. Mcsweegan, "Bacterial endotoxins and pathogenesis of Gram-negative infections: Current status and future direction," *J. Endotoxin Res.*, vol. 1, no. 2, pp. 71–83, Jun. 1994.

[4] M. Desvaux, M. HÉbraud, R. Talon, and I. R. Henderson, "Secretion and subcellular localizations of bacterial proteins: A semantic awareness issue," *Trends Microbiol.*, vol. 17, no. 4, pp. 139–145, Apr. 2009.

[5] S. Backert and T. F. Meyer, "Type IV secretion systems and their effectors in bacterial pathogenesis," *Current Opinion Microbiol.*, vol. 9, no. 2, pp. 207–217, Apr. 2006.

[6] Z. Lifshitz, D. Burstein, M. Peeri, T. Zusman, K. Schwartz, H. A. Shuman, T. Pupko, and G. Segal, "Computational modeling and experimental validation of the legionella and coxiella virulence-related type-IVB secretion signal," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 8, pp. E707–E715, Feb. 2013.

[7] R. Fronzes, E. Schafer, L. Wang, H. R. Saibil, E. V. Orlova, and G. Waksman, "Structure of a type IV secretion system core complex," *Science*, vol. 323, no. 5911, pp. 266–268, Jan. 2009.

[8] V. Chandran, R. Fronzes, S. Duquerroy, N. Cronin, J. Navaza, and G. Waksman, "Structure of the outer membrane complex of a type IV secretion system," *Nature*, vol. 462, no. 7276, pp. 1011–1015, Dec. 2009.

[9] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, "GutMDisorder: A comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D554–D560, Jan. 2020.

[10] Z. Ding, K. Atmakuri, and P. J. Christie, "The outs and ins of bacterial type IV secretion substrates," *Trends Microbiol.*, vol. 11, no. 11, pp. 527–535, Nov. 2003.

[11] E. Cascales and P. J. Christie, "The versatile bacterial type IV secretion systems," *Nature Rev. Microbiol.*, vol. 1, no. 2, pp. 137–149, Nov. 2003.

[12] B. Schrammeijer, "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.*, vol. 31, no. 3, pp. 860–868, Feb. 2003.

[13] J. Coers, J. C. Kagan, M. Matthews, H. Nagai, D. M. Zuckman, and C. R. Roy, "Identification of Icm protein complexes that play distinct roles in the biogenesis of an organelle permissive for Legionella pneumophila intracellular growth," *Mol. Microbiol.*, vol. 38, no. 4, pp. 719–736, Nov. 2000.

[14] D. V. Ward and P. C. Zambryski, "The six functions of Agrobacterium VirE2," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, pp. 385–386, Jan. 2001.

[15] G. Schroder, S. Krause, E. L. Zechner, B. Traxler, H.-J. Yeo, R. Lurz, G. Waksman, and E. Lanka, "TraG-like proteins of DNA transfer systems and of the helicobacter pylori type IV secretion system: Inner membrane gate for exported substrates?" *J. Bacteriol.*, vol. 184, no. 10, pp. 2767–2779, May 2002.

[16] S. Yoshiyuki, "A Japanese history of the human genome project," *J. Proc. Jpn. Acad. B, Phys. Biol. Sci.*, vol. 95, no. 8, pp. 441–458, 2019.

[17] H. Zhuang, L. Cheng, Y. Wang, Y.-K. Zhang, M.-F. Zhao, G.-D. Liang, M.-C. Zhang, Y.-G. Li, Y.-B. Zhao, Y.-N. Gao, Y.-J. Zhou, and S.-L. Liu, "Dysbiosis of the gut microbiome in lung cancer," *Frontiers Cellular Infection Microbiol.*, vol. 9, p. 112, Apr. 2019.

[18] L. Xue, B. Tang, W. Chen, and J. Luo, "A deep learning framework for sequence-based bacteria type IV secreted effectors prediction," *Chemometric Intell. Lab. Syst.*, vol. 183, pp. 134–139, Dec. 2018.

[19] D. Burstein, T. Zusman, E. Degtyar, R. Viner, G. Segal, and T. Pupko, "Genome-scale identification of legionella pneumophila effectors using a machine learning approach," *PLoS Pathogens*, vol. 5, no. 7, Jul. 2009, Art. no. e1000508.

[20] X. Zeng, W. Wang, G. Deng, J. Bing, and Q. Zou, "Prediction of potential disease-associated MicroRNAs by using neural networks," *Mol. Therapy-Nucleic Acids*, vol. 16, pp. 566–575, Jun. 2019.

[21] L. Nie, L. Deng, C. Fan, W. Zhan, and Y. Tang, "Prediction of protein S-sulfenylation sites using a deep belief network," *Current Bioinf.*, vol. 13, no. 5, pp. 461–467, Sep. 2018.

[22] X. Ru, P. Cao, L. Li, and Q. Zou, "Selecting essential MicroRNAs using a novel voting method," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 16–23, Dec. 2019.

[23] S. Wan, Y. Duan, and Q. Zou, "HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source," *Proteomics*, vol. 17, nos. 17–18, Sep. 2017, Art. no. 1700262.

[24] C. Chen, S. Banga, K. Mertens, M. M. Weber, I. Gorbaslieva, Y. Tan, Z.-Q. Luo, and J. E. Samuel, "Large-scale identification and translocation of type IV secretion substrates by Coxiella burnetii," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 50, pp. 21755–21760, Dec. 2010.

[25] L. Zou, C. Nan, and F. Hu, "Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles," *Bioinformatics*, vol. 29, no. 24, pp. 3135–3142, Dec. 2013.

[26] Y. Wang, X. Wei, H. Bao, and S.-L. Liu, "Prediction of bacterial type IV secreted effectors by C-terminal features," *BMC Genomics*, vol. 15, no. 1, p. 50, 2014.

[27] Y. An, J. Wang, C. Li, A. Leier, T. Marquez-Lago, J. Wilksch, Y. Zhang, G. I. Webb, J. Song, and T. Lithgow, "Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI," *Briefings Bioinf.*, vol. 19, no. 1, pp. 148–161, 2016.

[28] Y. Wang, Y. Guo, X. Pu, and M. Li, "Effective prediction of bacterial type IV secreted effectors by combined features of both C-termini and N-termini," *J. Comput. Aided Mol. Des.*, vol. 31, no. 11, pp. 1029–1038, Nov. 2017.

[29] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D.-Q. Wei, "PredT4SE-stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method," *Frontiers Microbiol.*, vol. 9, p. 2571, Oct. 2018.

[30] W. Yang, X.-J. Zhu, J. Huang, H. Ding, and H. Lin, "A brief survey of machine learning methods in protein sub-Golgi localization," *Current Bioinf.*, vol. 14, no. 3, pp. 234–240, Mar. 2019.

[31] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier," *Artif. Intell. Med.*, vol. 83, pp. 67–74, Nov. 2017.

[32] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, 2018.

[33] G. Pan, L. Jiang, J. Tang, and F. Guo, "A novel computational method for detecting DNA methylation sites with DNA sequence information and physicochemical properties," *Int. J. Mol. Sci.*, vol. 19, no. 2, p. 511, Feb. 2018.

[34] L. Jiang, Y. Ding, J. Tang, and F. Guo, "MDA-SKF: Similarity kernel fusion for accurately discovering miRNA-disease association," *Frontiers Genet.*, vol. 9, p. 618, Dec. 2018.

[35] L. Yu, S. Yao, L. Gao, and Y. Zha, "Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments," *Frontiers Genet.*, vol. 9, p. 745, Jan. 2019.

[36] L. Yu, J. Huang, Z. Ma, J. Zhang, Y. Zou, and L. Gao, "Inferring drug-disease associations based on known protein complexes," *BMC Med. Genomics*, vol. 8, p. 13, Dec. 2015.

[37] P. Zhu, Q. Xu, Q. Hu, and C. Zhang, "Co-regularized unsupervised feature selection," *Neurocomputing*, vol. 275, pp. 2855–2863, Jan. 2018.

[38] P. Zhu, Q. Xu, Q. Hu, C. Zhang, and H. Zhao, "Multi-label feature selection with missing labels," *Pattern Recognit.*, vol. 74, pp. 488–502, Feb. 2018.

[39] P. Zhu, W. Zhu, Q. Hu, C. Zhang, and W. Zuo, "Subspace clustering guided unsupervised feature selection," *Pattern Recognit.*, vol. 66, pp. 364–374, Jun. 2017.

[40] Z. Lv, S. Jin, H. Ding, and Q. Zou, "A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 215, 2019.

[41] M. Awais, W. Hussain, Y. D. Khan, N. Rasool, S. A. Khan, and K.-C. Chou, "IPhosH-PseAAC: Identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[42] H. Tang, W. Chen, and H. Lin, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Mol. BioSyst.*, vol. 12, no. 4, pp. 1269–1275, Feb. 2016.

[43] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 44, no. 1, p. 60, Jul. 2001.

[44] L. Wei, J. Tang, and Q. Zou, "Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information," *Inf. Sci.*, vol. 384, pp. 135–144, Apr. 2017.

[45] X. J. Zhu, C. Q. Feng, H. Y. Lai, W. Chen, and L. Hao, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowl.-Based Syst.*, vol. 163, pp. 787–793, Jan. 2019.

[46] G. Liou, D. Li, Z. Li, S. Qiu, W. Li, C. C. Chao, N. Yang, H. Li, Z. Cheng, X. Song, and L. Cheng, "PSSMHCpan: A novel PSSM-based software for predicting class I peptide-HLA binding affinity," *Giga Sci.*, vol. 6, no. 5, 2017, Art. no. gix017.

[47] Y.-B. Wang, Z.-H. You, L.-P. Li, D.-S. Huang, F.-F. Zhou, and S. Yang, "Improving prediction of self-interacting proteins using stacked sparse auto-encoder with PSSM profiles," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 983–991, 2018.

[48] F. Ali, M. Arif, Z. U. Khan, M. Kabir, S. Ahmed, and D.-J. Yu, "SDBP-Pred: Prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM," *Anal. Biochem.*, vol. 589, Jan. 2020, Art. no. 113494.

[49] F. Yuan, G. Liu, X. Yang, S. Wang, and X. Wang, "Prediction of oxidoreductase subfamily classes based on RFE-SND-CC-PSSM and machine learning methods," *J. Bioinform. Comput. Biol.*, vol. 17, no. 04, Aug. 2019, Art. no. 1950029.

[50] Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu, and Z. Lin, "An approach for identifying cytokines based on a novel ensemble classifier," *BioMed. Res. Int.*, vol. 2013, pp. 1–11, 2013.

[51] Y. H. Li, J. Y. Xu, L. Tao, X. F. Li, S. Li, X. Zeng, S. Y. Chen, P. Zhang, C. Qin, C. Zhang, Z. Chen, F. Zhu, and Y. Z. Chen, "SVM-Prot 2016: A Web-server for machine learning prediction of protein functional families from sequence irrespective of similarity," *PLoS ONE*, vol. 11, no. 8, Aug. 2016, Art. no. e0155290.

[52] J. Argouarc'h, "Dependency, skip-grams, association tensors and machine learning for sentence completion," in *Proc. 14th Conf. Natural Lang. Process. (KONVENS)*, 2018, pp. 100–109.

[53] W. Zhu, H. Gong, J. Shen, C. Zhang, J. Shang, S. Bhat, and J. Han, "FUSE: Multi-faceted set expansion by coherent clustering of skip-grams," Oct. 2019, *arXiv:1910.04345*. [Online]. Available: https://arxiv.org/abs/1910.04345

[54] J. Jabez, S. Gowri, S. Vigneshwari, J. A. Mayan, and S. Srinivasulu, "Anomaly detection by using CFS subset and neural network with WEKA tools," in *Information and Communication Technology for Intelligent Systems* (Smart Innovation, Systems and Technologies), vol. 107, S. Satapathy and A. Joshi, Eds. Singapore: Springer, 2019.

[55] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, Oct. 2004.

[56] G. Holmes, A. Donkin, and I. H. Witten, "WEKA: A machine learning workbench," in *Proc. ANZIIS*, Nov./Dec. 1994, pp. 357–361.

[57] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with Java implementations," Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, Working paper 99/11, 1999.

[58] S. R. Garner, "WEKA: The waikato environment for knowledge analysis," in *Proc. New Zealand Comput. Sci. Res. Students Conf.*, 1995, pp. 1–8.

[59] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. Springer, 2006.

[60] S. V. M. Banadaki, S. Lattanzi, J. E. Feldman, S. Leonardi, H. Lynch, and V. Sharma, "Efficient similarity ranking for bipartite graphs," Google Patents 10 152 557, Dec. 11, 2018.

[61] H. Faris, M. A. Hassonah, A. M. Al-Zoubi, S. Mirjalili, and I. Aljarah, "A multi-verse optimizer approach for feature selection and optimizing SVM parameters based on a robust system architecture," *Neural Comput. Appl.*, vol. 30, no. 8, pp. 2355–2369, Oct. 2018.

[62] R. Varatharajan, G. Manogaran, and M. K. Priyan, "A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing," *Multimed. Tools Appl.*, vol. 77, no. 8, pp. 10195–10215, Apr. 2018.

[63] W. Deng, R. Yao, H. Zhao, X. Yang, and G. Li, "A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm," *Soft Comput.*, vol. 23, no. 7, pp. 2445–2462, Apr. 2019.

[64] B. Manavalan, T. H. Shin, and G. Lee, "PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine," *Frontiers Microbiol.*, vol. 9, p. 476, Dec. 2018.

[65] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via multiple information integration," *Inf. Sci.*, vols. 418–419, pp. 546–560, Dec. 2017.

[66] L. Yu, J. Zhao, and L. Gao, "Predicting potential drugs for breast cancer based on miRNA and tissue specificity," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 971–982, 2018.

[67] C. Jia, Y. Zuo, and Q. Zou, "O-GlcNAcPRED-II: An integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique," *Bioinformatics*, vol. 34, no. 12, pp. 2029–2036, Jun. 2018.

[68] L. Cheng, H. Yang, H. Zhao, X. Pei, H. Shi, J. Sun, Y. Zhang, Z. Wang, and M. Zhou, "MetSigDis: A manually curated resource for the metabolic signatures of diseases," *Briefings Bioinf.*, vol. 20, no. 1, pp. 203–209, Jan. 2019.

[69] L. Cheng, H. Zhao, P. Wang, W. Zhou, M. Luo, T. Li, J. Han, S. Liu, and Q. Jiang, "Computational methods for identifying similar diseases," *Mol. Therapy Nucleic Acids*, vol. 18, pp. 590–604, Dec. 2019.

[70] Y. Qiao, Y. Xiong, H. Gao, X. Zhu, and P. Chen, "Protein-protein interface hot spots prediction based on a hybrid feature selection strategy," *BMC Bioinf.*, vol. 19, no. 1, p. 14, 2018.

[71] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Tech. Rep. MSR-TR-98-14, 1998.

[72] K. M. Nakanishi, K. Fujii, and S. Todo, "Sequential minimal optimization for quantum-classical hybrid algorithms," Tech. Rep., 2019.

[73] D. Pir, "Functional-based acoustic group feature selection for automatic recognition of eating condition," in *Proc. Int. Conf. Multimodal Interact. (ICMI)*. New York, NY, USA: ACM, 2018, pp. 579–583.

[74] S. Maldonado, J. MerigÓ, and J. Miranda, "Redefining support vector machines with the ordered weighted average," *Knowl.-Based Syst.*, vol. 148, pp. 41–46, May 2018.

[75] V. Lishchuk, C. Lund, and Y. Ghorbani, "Evaluation and comparison of different machine-learning methods to integrate sparse process data into a spatial model in geometallurgy," *Minerals Eng.*, vol. 134, pp. 156–165, Apr. 2019.

[76] I. E. Livieris, K. Drakopoulou, V. T. Tampakas, T. A. Mikropoulos, and P. Pintelas, "Predicting secondary school students' performance utilizing a semi-supervised learning approach," *J. Educ. Comput. Res.*, vol. 57, no. 2, pp. 448–470, Apr. 2019.

[77] J.-X. Tan, S.-H. Li, Z.-M. Zhang, C.-X. Chen, W. Chen, H. Tang, and H. Lin, "Identification of hormone binding proteins based on machine learning methods," *Math. Biosci. Eng.*, vol. 16, no. 4, pp. 2466–2480, 2019.

[78] W. Chen, P. Feng, T. Liu, and D. Jin, "Recent advances in machine learning methods for predicting heat shock proteins," *Current Drug Metabolism*, vol. 20, no. 3, pp. 224–228, May 2019.

[79] L. Wei, P. Xing, R. Su, G. Shi, Z. S. Ma, and Q. Zou, "CPPred-RF: A sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency," *J. Proteome Res.*, vol. 16, no. 5, pp. 2044–2053, May 2017.

[80] L. Wei, P. Xing, J. Tang, and Q. Zou, "PhosPred-RF: A novel sequence-based predictor for phosphorylation sites using sequential information only," *IEEE Trans. Nanobiosci.*, vol. 16, no. 4, pp. 240–247, Jun. 2017.

[81] Y. Shen, J. Tang, and F. Guo, "Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC," *J. Theor. Biol.*, vol. 462, pp. 230–239, Feb. 2019.

[82] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, Jan. 2019.

[83] L. Yu, J. Zhao, and L. Gao, "Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome," *Artif. Intell. Med.*, vol. 77, pp. 53–63, Mar. 2017.

[84] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim scores," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 687–695, May 2017.

[85] M. Zhang, F. Li, T. T. Marquez-Lago, A. Leier, C. Fan, C. K. Kwoh, K.-C. Chou, J. Song, and C. Jia, "MULTiPly: A novel multi-layer predictor for discovering general and specific types of promoters," *Bioinformatics*, vol. 35, no. 17, pp. 2957–2965, Sep. 2019.

[86] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring MicroRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul. 2017.

[87] Z. Hong, X. Zeng, L. Wei, and X. Liu, "Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism," *Bioinformatics*, Sep. 2019, doi: 10.1093/bioinformatics/btz694.

[88] X. Zhang, Q. Zou, A. Rodriguez-Paton, and X. Zeng, "Meta-path methods for prioritizing candidate disease miRNAs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 1, pp. 283–291, Jan. 2019.

[89] L. Cheng, Y. Jiang, H. Ju, J. Sun, J. Peng, M. Zhou, and Y. Hu, "InfAcrOnt: Calculating cross-ontology term similarities using information flow by a random walk," *BMC Genomics*, vol. 19, p. 919, Jan. 2018.

[90] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, "DincRNA: A comprehensive Web-based bioinformatics toolkit for exploring disease associations and ncRNA function," *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, Jun. 2018.

[91] X. Zhu, J. He, S. Zhao, W. Tao, Y. Xiong, and S. Bi, "A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae," *Briefings Funct. Genomics*, vol. 18, no. 6, pp. 367–376, 2019.

[92] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, and Y. Xiong, "PseUI: Pseudouridine sites identification based on RNA sequence information," *BMC Bioinf.*, vol. 19, no. 1, p. 306, 2018.

[93] Q. Xu, Y. Xiong, H. Dai, K. M. Kumari, Q. Xu, H.-Y. Ou, and D.-Q. Wei, "PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm," *J. Theor. Biol.*, vol. 417, pp. 1–7, Mar. 2017.

[94] X. Shan, X. Wang, C.-D. Li, Y. Chu, Y. Zhang, Y. Xiong, and D.-Q. Wei, "Prediction of CYP450 enzyme–substrate selectivity based on the network-based label space division method," *J. Chem. Inf. Model.*, vol. 59, no. 11, pp. 4577–4586, Nov. 2019.

[95] T. Fang, Z. Zhang, R. Sun, L. Zhu, J. He, B. Huang, Y. Xiong, and X. Zhu, "RNAm5CPred: Prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 739–747, Dec. 2019.

[96] R. Su, X. Liu, L. Wei, and Q. Zou, "Deep-resp-forest: A deep forest model to predict anti-cancer drug response," *Methods*, vol. 166, pp. 91–102, Aug. 2019.

[97] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, "DeepDR: A network-based deep learning approach to in silico drug repositioning," *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, Dec. 2019, doi: 10.1093/bioinformatics/btz418.

[98] Z. Lv, C. Ao, and Q. Zou, "Protein function prediction: From traditional classifier to deep learning," *Proteomics*, vol. 19, no. 14, 2019, Art. no. 1900119.

[99] L. Yu, X. Sun, S. Tian, X. Shi, and Y. Yan, "Drug and nondrug classification based on deep learning with various feature selection strategies," *Current Bioinf.*, vol. 13, no. 3, pp. 253–259, May 2018.

[100] X. Zeng, X. Zhang, T. Song, and L. Pan, "Spiking neural P systems with thresholds," *Neural Comput.*, vol. 26, no. 7, pp. 1340–1361, Jul. 2014.

[101] L. Wei, Y. Ding, R. Su, J. Tang, and Q. Zou, "Prediction of human protein subcellular localization using deep learning," *J. Parallel Distrib. Comput.*, vol. 117, pp. 212–217, Jul. 2018.

[102] F. G. C. Cabarle, H. N. Adorna, M. Jiang, and X. Zeng, "Spiking neural P systems with scheduled synapses," *IEEE Trans. Nanobiosci.*, vol. 16, no. 8, pp. 792–801, Dec. 2017.

[103] L. Peng, M. Peng, B. Liao, G. Huang, W. Li, and D. Xie, "The advances and challenges of deep learning application in biological big data processing," *Current Bioinf.*, vol. 13, no. 4, pp. 352–359, Jul. 2018.

[104] Z. Zhang, Y. Zhao, X. Liao, W. Shi, K. Li, Q. Zou, and S. Peng, "Deep learning in omics: A survey and guideline," *Briefings Funct. Genomics*, vol. 18, no. 1, pp. 41–57, Feb. 2019.

[105] T. Song, A. Rodriguez-Paton, P. Zheng, and X. Zeng, "Spiking neural P systems with colored spikes," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 4, pp. 1106–1115, Dec. 2018.

[106] H. Xu, W. Zeng, D. Zhang, and X. Zeng, "MOEA/HD: A multiobjective evolutionary algorithm based on hierarchical decomposition," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 517–526, Feb. 2019.

[107] H. Xu, W. Zeng, X. Zeng, and G. G. Yen, "An evolutionary algorithm based on minkowski distance for many-objective optimization," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3968–3979, Nov. 2019.

[108] X. Zeng, W. Wang, C. Chen, and G. G. Yen, "A consensus community-based particle swarm optimization for dynamic community detection," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/tcyb.2019.2938895.

ZIHAO YAN was born in Harbin, Heilongjiang, China, in 1996. He received the B.S. degree in computer science and technology from Northeast Forestry University, Harbin, in 2018, where he is currently pursuing the degree. His main research interests include bioinformatics and machine learning.

DONG CHEN received the Ph.D. degree from the School of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, China, in 2009. His current research interests include lifting wavelet, data privacy, and deep learning.

ZHIXIA TENG received the Ph.D. degree in computer science and technology from the Harbin Institute of Technology, in 2016. She is currently a Lecturer with the College of Information and Computer Engineering, Northeast Forestry University, Harbin, China. Her current research interests include bioinformatics, computational system biology, and machine learning.

DONGHUA WANG received the bachelor's and master's degrees in medicine from the Harbin Medical School, in 1989 and 1992, respectively. From 2004 to 2019, he was the Chief Physician with the General Hospital of the Provincial Agricultural Reclamation. He is currently the Vice Chairman of the Association, the Executive Director of the Provincial Medical Association, the Executive Director of the Provincial Association of Doctors, and the Executive Director of the Association of Youth Prosperity Leaders of the Nongnongken. His research is mainly focused on general surgery. He is a General Member of the Provincial Medical Association, and a member of the Cancer Committee of the Provincial Medical Association.

YANJUAN LI received the Ph.D. degree in artificial intelligence and information processing from the Harbin Institute of Technology, Harbin, China, in 2012. She is currently an Associate Professor with the School of Information and Computer Engineering, Northeast Forestry University, Harbin. Her current research interests include machine learning, data privacy, inductive logic programming, and evolutionary algorithms.