frontiers | Frontiers in Microbiology

METHODS
published: 15 July 2022
doi: 10.3389/fmicb.2022.932661

Ensemble Learning-Based Feature Selection for Phage Protein Prediction

Songbo Liu[1], *Chengmin Cui*[2], *Huipeng Chen*[1]* *and Tong Liu*[1]

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] Beijing Institute of Control Engineering, China Academy of Space Technology, Beijing, China

Phage has high specificity for its host recognition. As a natural enemy of bacteria, it has been used to treat super bacteria many times. Identifying phage proteins from the original sequence is very important for understanding the relationship between phage and host bacteria and developing new antimicrobial agents. However, traditional experimental methods are both expensive and time-consuming. In this study, an ensemble learning-based feature selection method is proposed to find important features for phage protein identification. The method uses four types of protein sequence-derived features, quantifies the importance of each feature by adding perturbations to the features to influence the results, and finally splices the important features among the four types of features. In addition, we analyzed the selected features and their biological significance.

Keywords: machine learning, ensemble learning, feature selection, phage, protein classification

OPEN ACCESS

Edited by:
Jian Huang,
University of Electronic Science and Technology of China, China

Reviewed by:
Zhiwei Ji,
Nanjing Agricultural University, China
Yi Xiong,
Shanghai Jiao Tong University, China

***Correspondence:**
Huipeng Chen
chp@ir.hit.edu.cn

Specialty section:
This article was submitted to Phage Biology, a section of the journal Frontiers in Microbiology

Received: *30 April 2022*
Accepted: *14 June 2022*
Published: *15 July 2022*

Citation:
Liu S, Cui C, Chen H and Liu T (2022) Ensemble Learning-Based Feature Selection for Phage Protein Prediction. Front. Microbiol. 13:932661. doi: 10.3389/fmicb.2022.932661

INTRODUCTION

Phages, which are the most abundant and widespread organisms on the Earth, can replicate within and destroy the host cell. Phages play an important role in microbial physiology, population dynamics, evolution, and therapy (Clokie et al., 2011), affecting biochemical systems worldwide (Jahn et al., 2019).

Phages also influence the development of anti-cancer drugs. The use of phages to target cancer cells for a specific binding for therapeutic purposes has been applied in clinical trials for cancer treatment due to differences in surface markers of tumor cells from normal cells (Yu et al., 2021). The identification of phage proteins is important for understanding the relationship between phages and host bacteria and for developing novel drugs or antibiotics (Lekunberri et al., 2017), and therefore, thorough investigations must be performed to identify the specific components recognized by phages.

Traditional physical experimental methods such as mass spectrometry (MS), sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), and protein arrays (Lavigne et al., 2009; Yuan and Gao, 2016; Jara-Acevedo et al., 2018), which have been used to identify phage viral proteins, are expensive and often time-consuming. Traditional biological methods such as cell separation, electron microscopy, and fluorescence microscopy are less feasible for analyzing large-scale biological data (Mei, 2012; Li et al., 2015). Computational models can not only analyze large amounts of biological data but also make preliminary predictions of unknown protein sequences, which is an excellent complement to traditional experimental methods.

In recent years, protein function prediction has been a hot topic in the field of computational biology (Ding et al., 2020; Fu et al., 2020; Guo et al., 2020). With the increasing amount of protein data, the techniques of applying machine learning and data mining to protein function prediction have gradually matured (Liu et al., 2019; Zhao et al., 2021). Several researchers have used machine learning to predict the function of protein sequences through sequence analysis (Chou, 2009; Cui et al., 2019; Jin et al., 2021), position-specific scoring matrix (PSSM) (Jones, 1999), various physicochemical and biochemical properties of amino acids, sequence conservation, amino acid composition, domain interactions, and geometrical and biophysical properties (Kawashima and Kanehisa, 2000; Cai et al., 2003; Zulfiqar et al., 2021). Feng et al. (2013) developed a Naive Bayes-based model for protein classification, which used amino acid composition (AAC) and dipeptide combination (DPC) as input features. Ding et al. (2014) developed a support vector machine (SVM) prediction model. In this method, analysis of variance was applied to select significant features from the g-gap DPC. Recently, Zhang et al. (2015) developed a random forest classification method to distinguish phage virus protein (PVP) from non-PVP. A novel feature extraction method with a two-layered structure is proposed (Xiong et al., 2018; Jiang et al., 2021). First, the features irrelevant to the results are removed by the filter or wrapper method, and then, the results of the previous step are used in the model for classification and prediction.

Since each feature extraction method extracts only part of the protein sequence information, the method of combining multiple protein information for classification is proposed in the absence of a clear sequence or structural information. Jiao and Du (2017) proposed the functional domain enrichment score with position-specific physicochemical properties (PSPCP). Li et al. (2015) proposed to fuse the position-specific scoring matrix (PSSM) and gene ontology to extract feature sequences.

Protein sequences often have high feature dimensionality and contain a large amount of redundant information, which reduces the prediction performance of a model. Dimensionality reduction of feature vectors with high-dimensional data is performed to eliminate unnecessary features. The cmmonly used methods are principal component analysis (PCA) (Ahmad et al., 2016), information gain (Wen et al., 2016), maximum correlation and minimum redundancy (MRMR) (Khan et al., 2017), maximum correlation maximum distance (MRMD) (Zou et al., 2016), singular value decomposition (SVD) (Silvério-Machado et al., 2015), local linear discriminant analysis (Yu et al., 2018), and dipeptide composition (DPC) (Ahmad et al., 2016). Xie et al. (2021) proposed a method for k-size optimal parsimony features based on the rough set theory, which found the effective features by fixing the parsimony size and dynamic weighting strategy. NMFBFS reduced the dimensionality of the data by decomposing the non-negative matrix of the data (Ji et al., 2015). Despite the specific advantages of existing methods for PVP prediction, there is a need to improve the accuracy and transferability of predictive models.

In the present study, we propose an ensemble learning-based feature selection method for phage virus protein classification that uses a four-step pipeline for protein prediction, (I) extracting the amino acid composition content (AAC), physicochemical properties CTD, dipeptide composition CKSAAP, and reduced position specificity scoring matrix (RPSSM) of the protein; (II) using ensemble learning to measure the importance of each feature component from each type of feature; (III) using an incremental strategy to select the most important feature subset; and (IV) combining the optimal feature subset derived from each type of feature to retrain the data to be filtered again and finally applying the obtained optimal feature subset to predict the protein type. Instead of the PCA-like feature dimension reduction method, our method can directly obtain important features for further biological analysis. Experimental results demonstrate the effectiveness of our method.

MATERIALS AND METHODS

Data

In this study, we used the dataset constructed by Ding et al. (2014). This dataset was processed by UniProt in the following ways. First of all, the phage proteins whose subcellular location is a virion were considered positive sample and *vice versa*. The sequences containing unknown amino acids such as "B," "J," "O," "U," "X," or "Z" were removed. To eliminate the influence of homologous sequences, more than 40% homologous sequences were removed by CD-HIT. Eventually, 99 phage proteins and 208 non-phage proteins were obtained.

The data from various literature studies (Feng et al., 2013; Ding et al., 2014; Zhang et al., 2015) were processed in the same way to build an independent test set. Besides, more than 40% homologous sequences with the training set were removed.

Feature Extraction

The functions of a protein consists of the amino acid type, quantity, arrangement order of the peptide chain, and the spatial structure of the protein. Therefore, the main description methods of a protein can be divided into the global description of the protein and the description of the amino acid level (Xu et al., 2020). The global description of the protein includes the first-class features and spatial features of the protein. However, the acquisition of spatial features is expensive. Under certain conditions, the primary structure of the protein can determine the secondary, tertiary, and quaternary structures. Therefore, in this study, we only extracted the primary features of the protein and also extracted simple spatial structural features from the primary structure.

For the convenience of discussion, we define a protein sequence P as follows:

$$P = p_1 p_2 p_3 \ldots p_L, p_i \in \{A,\ C,\ D,\ \ldots Y\} \tag{1}$$

where p_i is an amino acid, i is the position of the amino acid in the sequence, and L is the length of the amino acid sequence.

AAC

The Amino Acid Composition (AAC) is to calculate the content of each amino acid in a protein sequence. The AAC feature of an

amino acid sequence is as follows:

$$AAC_i = \frac{N\left(p_i\right)}{L},\ 0 < i \leq 20 \tag{2}$$

where AAC_i represents the proportion of p_i in a protein sequence, $N\left(p_i\right)$ is the number of the amino acid p_i in a protein sequence, and L is the number of amino acids in a sample.

CTD

Protein-related chemical reactions commonly occur in the cell and tissue fluids, so the physicochemical properties of a protein are closely related to the function of the protein. There are eight types commonly used with the following physicochemical properties: hydrophobicity, polarity, surface tension, polarizability, charge, van der Waals force, secondary structure, and solubility (Cai et al., 2003).

For each physicochemical property, the amino acids are divided into three groups (positive, neural, and negative), and then, the three values of Composition (C), Transition (T), and Distribution (D) of each property are calculated. C is the percentage of the composition of a certain physicochemical property. T describes the following three types of residue pairs: a negative residue followed by a neural residue; a positive residue followed by a negative residue; and a positive followed by a neural residue. The percentage of the amino acids of a particular property located at the first, 25, 50, 75, and 100% is measured as the distribution of the protein (D). Finally, 168 physicochemical features are obtained.

CKSAAP

The composition of k-space amino acid pairs (CKSAAP) is encoded by the proportion of amino acid pairs separated by any k residues.

$$\begin{aligned} &CKSAAP = \frac{N_{p_i p_j}}{L}, p_i, p_j \in \{A,\ C,\ D,\ \ldots,\ Y\} \\ &j = i + k + 1,\ i,\ j \leq L, \end{aligned} \tag{3}$$

where $N_{p_i p_j}$ is the proportion of residue pairs p_i and p_j. According to Ding et al. (2014), the best results of extracting phage protein features for classification are obtained when $k = 1$.

RPSSM

In the process of biological evolution, amino acid sequences mutate corresponding to common changes, including deletion, substitution, and insertion of amino acid residues. However, these changed protein sequences still have similar structures and functions (Ding et al., 2014). The position-specific scoring matrix (PSSM) can describe this change.

For the convenience of description, we defined the sequence P_A as follows:

$$P_A = (P_{1A}, P_{2A},\ \ldots\ , P_{LA})^T \tag{4}$$

where P_{iA} indicates the score of amino acid mutation to amino acid A at the *ith* position of the protein sequence.

According to the similarity of amino acids, Li et al. (2003) found that 10 residues can construct a set with the smallest reasonable folding model, i.e., {F, Y, W}, {M, L}, {I, V}, {A, T, S}, {N, H}, {Q, E, D}, {R, K}, {P}, {C}, and {G}. We combined the same categories based on the similarity of amino acids as follows:

$$\begin{cases} P_1 = \frac{(P_F + P_Y + P_W)}{3}, \\ P_2 = \frac{(P_M + P_L)}{2}, \\ \ldots, \\ P_{10} = P_G. \end{cases} \tag{5}$$

The reduced position-specific scoring matrix (RPSSM) is represented as follows:

$$PSSM_S = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,j} & \cdots & P_{1,10} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,j} & \cdots & P_{2,10} \\ \vdots & \vdots & & \vdots & & \\ P_{i,1} & P_{i,2} & \cdots & P_{i,j} & \cdots & P_{i,10} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{L,1} & P_{L,2} & \cdots & P_{L,j} & \cdots & P_{L,10} \end{bmatrix}. \tag{6}$$

The variance of each column is calculated to get the feature D_s,

$$D_S = \frac{1}{L}\sum_{i=1}^{L}\left(p_{is} - \bar{p}_s\right)^2,\ s \leq 10,\ i \leq L \tag{7}$$

However, D_S does not contain the amino acid order information. To get the information on the local sequence order effect, the dipeptide composition of the protein sequence is extended to PSSM. Assuming that p_i is mutated into p_s and that p_{i+1} is mutated into p_t in the sequence, then

$$\begin{aligned} D_{i,i+1} &= \left(p_{i,s} - \frac{p_{i,s} + p_{i+1,t}}{2}\right)^2 + \left(p_{i+1,t} - \frac{p_{i,s} + p_{i+1,t}}{2}\right)^2 \\ &= \frac{\left(p_{i,s} - p_{i+1,t}\right)^2}{2} \end{aligned} \tag{8}$$

where $s, t = 1, 2, \ldots 10,\ i = 1, 2, \ldots\ L\ 1$, and $D_{i,i+1}$ represents the difference between the average values of $p_{i,s}$ and $p_{i+1,t}$, while $D_{i,i+1}$ in the protein sequence is expressed as follows:

$$D_{s,t} = \frac{1}{L}\sum_{i-1}^{L-1}\frac{\left(p_{i,s} - p_{i+1,t}\right)^2}{2},\ s,\ t = 1, 2, \ldots,\ 10 \tag{9}$$

Finally, 110 features were obtained by splicing $D_{s,t}$ and D_s.

Feature Selection

The goal of the feature-selection module is to select as few features as possible with guaranteed classification accuracy so that the model does not degrade significantly when the model is learned using only the subset of features and the learning results are close to or higher than the learning results using the full set of features.

The relationship between the sequence structure and the function of protein is not entirely clear. Therefore, the features based on knowledge extraction are not necessarily related to the function of the protein or are even some irrelevant features.

Redundancy will also affect the fitting of the classifier to protein data and will interfere with the prediction. Therefore, we cannot select features only in a knowledge-driven way. It is necessary to further screen the extracted features in a data-driven way so as to screen out effective feature subsets beneficial to the learning algorithm. This can not only reduce the difficulty of learning tasks but also improve the efficiency of the model. We assume that only some of all the features play a decisive role in model fitting. If this feature is modified, it will have a greater impact on the results. To quantify this effect, we proposed a feature selection model. The specific steps we followed were as follows.

Importance Accumulation

First, the features from four types of protein sequences were extracted by a feature extraction module in **Figure 1A**. Then, the importance score of each feature component was calculated. According to the feature selection module shown in **Figure 1B**, we assumed that all the features extracted are valid for classification. Then, 70% of the training samples were randomly selected to train the classifier, and the classification accuracy rate was obtained by testing the data out-of-bag, which is recorded as $score_1$. If a feature component was very relevant to the protein function, just changing it would have a great impact on the result. According to this idea, we only shuffled the values of a feature component randomly by a permutation way on the testing dataset in order to maintain the data distribution unchanged before and after this modification, then used the model trained on the training set to predict the shuffled data, and obtained the classification accuracy, which is recorded as $score_2$. The importance of this feature component is defined as:

$$imp_i = score_1 - score_2 \tag{10}$$

Each feature component is shuffled to get its important score. To reduce the error, it repeats n rounds to get the average importance scores of each feature.

Incremental Feature Selection

To find the optimal feature subset, we added each feature component incrementally according to its score in descending order, trained the classification model, and calculated the classification accuracy and finally showed the result in **Figure 1C**. In this way, a feature subset that maintains a comparative classification result equivalent to the feature is selected and considered an important feature. Then, the classifier was used in a feature subset to train all the training samples, and the independent test set was used to predict the phage protein. The final processed was shown in **Figure 1D**.

Ensemble Classifier

Due to the unclear sample distribution and various classification boundaries, only using a single classifier may not fit the data well and may not get a good classification result. Therefore, we used seven common classifiers with default parameters as the base classifier in scikit-learn (Pedregosa et al., 2011), linear discriminant analysis (LDA), decision tree classifier (DTC), k-neighbors classifier (KNN), support vector machine (SVM), logistic regression (LR), Gaussian Naive Bayes classifier (GNB), and multilayer perceptron (MLP) using 1,000 iterations. At the same time, seven classifiers were trained in each sampling, and the classifier with the highest accuracy was selected. We believed that this classifier can best fit the distribution of the data. This classifier was used as the best classifier in this round to classify or calculate feature importance. Although the ensemble learning method can be insensitive to the distribution of data, the time

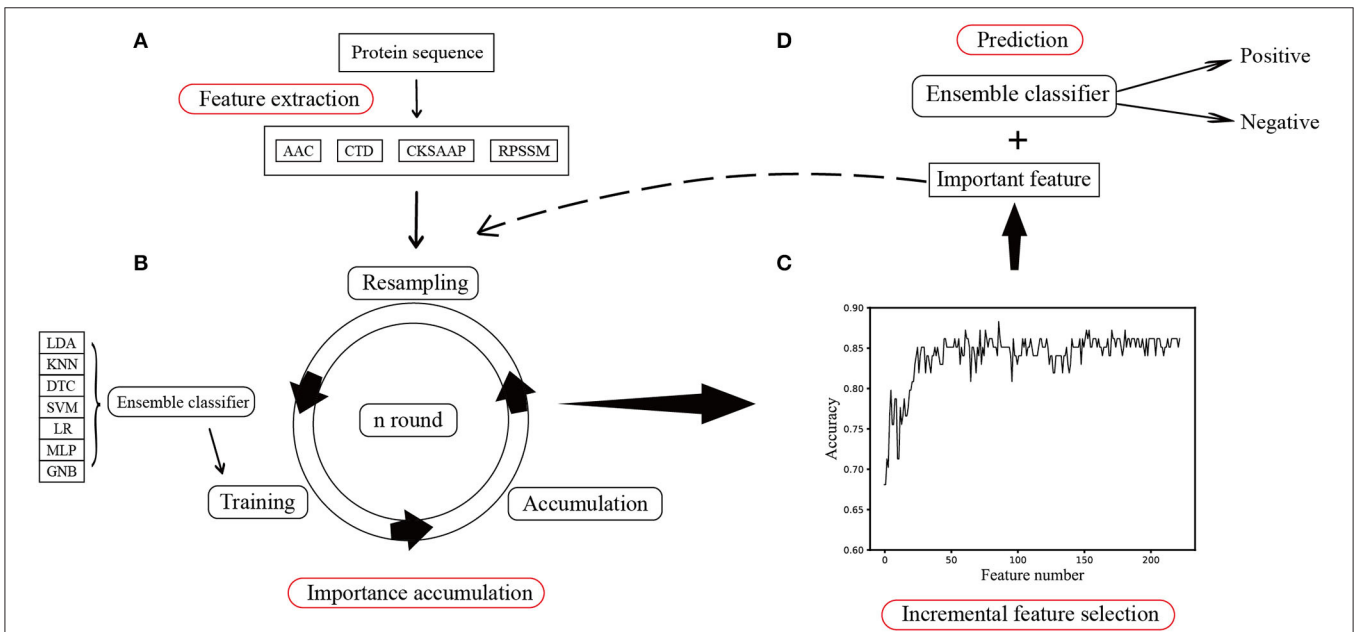

FIGURE 1 | Flowchart of the proposed method. **(A)** Feature extraction method includes extracting AAC, CTD, CKSAAP-1GAP, and RPSSM from the original protein sequences. **(B)** Resampling all samples and then calculating the importance of each feature component using the ensemble classification. **(C)** Selecting a subset of important features using the incremental method. **(D)** Predicting phage proteins on the testing set using the important feature components.

cost will increase due to the use of multiple classifiers. The time complexity of the proposed method $O(rounds * \sum_i classifier_i)$ was mainly to calculate the importance score of features, where rounds refer to the number of iterations. Therefore, if the time complexity needs to be reduced, the preferred method is to choose a classifier with smaller time complexity.

Evaluation Criteria

To evaluate our model comprehensively, we used the common measures, i.e., ACC, SN, and SP. These methods are defined as follows:

$$ACC = \frac{TN + TP}{TP + FN + FN + FP} \tag{11}$$

$$SN = \frac{TP}{TP + FN} \tag{12}$$

$$SP = \frac{TN}{TN + FP} \tag{13}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

RESULTS

The function of the protein consists of amino acid composition, arrangement order, and spatial structure. In this study, we selected amino acid content (AAC), physicochemical properties (CTD), dipeptide (CKSAAP-1GAP), and PSSM matrix of protein (RPSSM) as the features of protein data. The classification accuracy of each feature was calculated using the ensemble classifier to get the importance of each feature component. According to the importance of feature components, the effective feature subset of each feature of different species was selected in an incremental way. Finally, all feature subsets were spliced and predicted on the independent test set and compared with other methods.

Performance Comparison of Individual Types of Features on Training Dataset Results Using Independent Features

For features of different types, each classification model in the ensemble classifier was trained using the training dataset, and

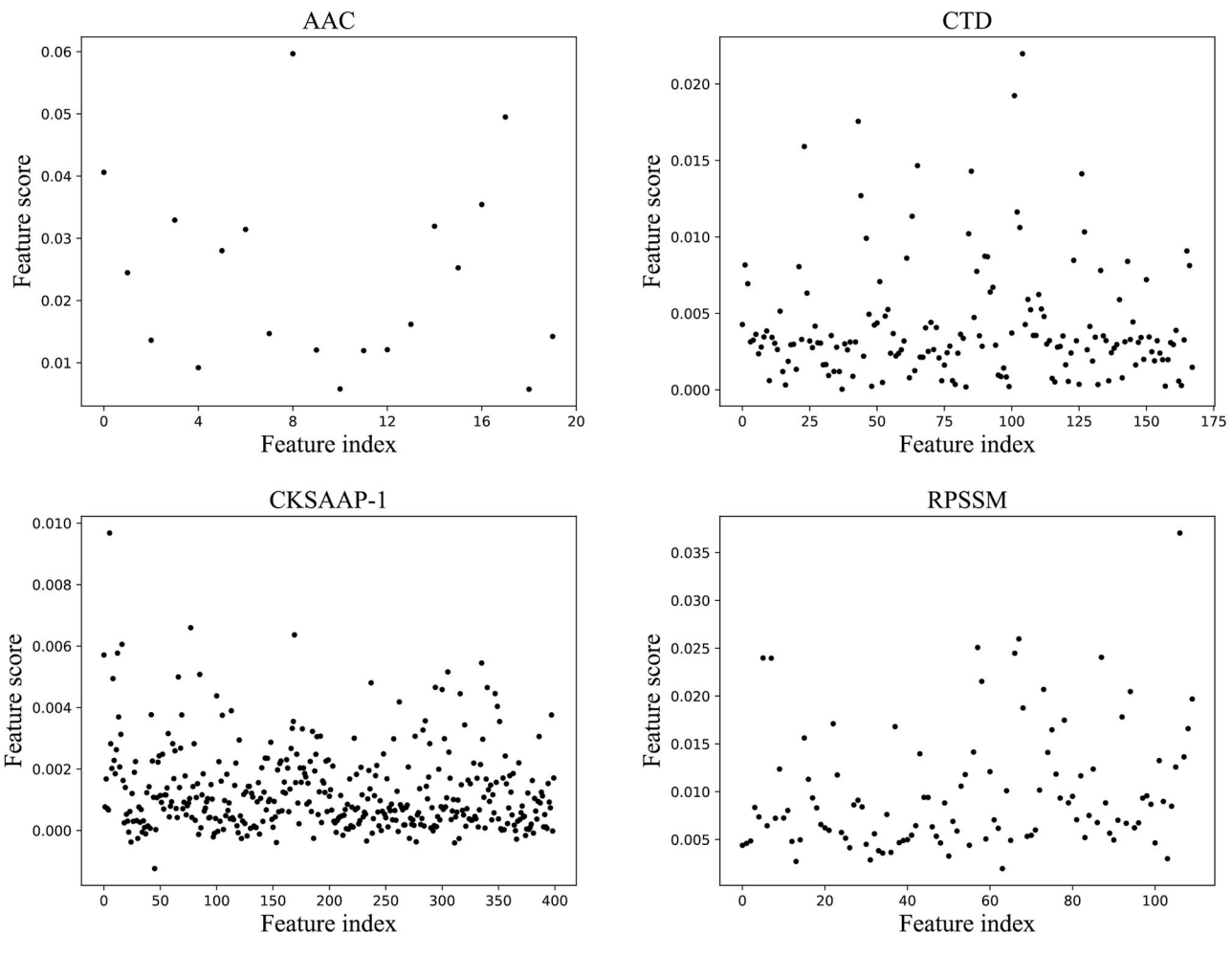

FIGURE 2 | The scatterplots of feature importance using different feature extraction methods.

the model with the best fitting effect was selected by ACC in the ensemble classifier. The feature importance score was calculated using a feature selection module shown in **Figure 1B**. **Figure 2** shows the scatterplots of the accumulated scores according to feature components. It can be found that AAC can affect the results by 6% at the highest and the physicochemical properties (CTD) by 2%, but dipeptide can only affect the results by 1% at the highest. RPSSM has an impact of 3.5%. Some features have no effect on the results after modification.

To get rid of the ineffective feature components, we used the incremental method according to the rank of feature importance scores. After stacking feature components according to their importance, we trained the ensemble classifier and obtained the classification accuracies. The ACC curve was obtained using a 10-fold cross-validation method, as seen in **Figure 3**.

It can be found that the classification accuracy of the model in AAC has shown an increasing trend with the superposition of feature components, but it is around 82%. The accuracy of physicochemical properties (CTD) is more stable when features are superimposed. It can be found that the classification accuracy is the highest using only the first 50 important features, and the classification accuracy does not increase when feature components are superimposing. The classification accuracy of the first 103 feature components in CKSAAP-1GAP is also good with the highest classification accuracy, and the classification accuracy even decreases when new feature components are added. RPSSM is always more stable, and the first 50 feature components with the highest classification accuracy are selected as the optimal feature subset of the features.

Performance Comparison of Concatenated Features on Training Data Set Results Using Integrated Features

We selected the subset of features with the best classification result and the least number of features among different kinds of features and spliced the obtained subset of features to form an important feature with multiple types of superpositions.

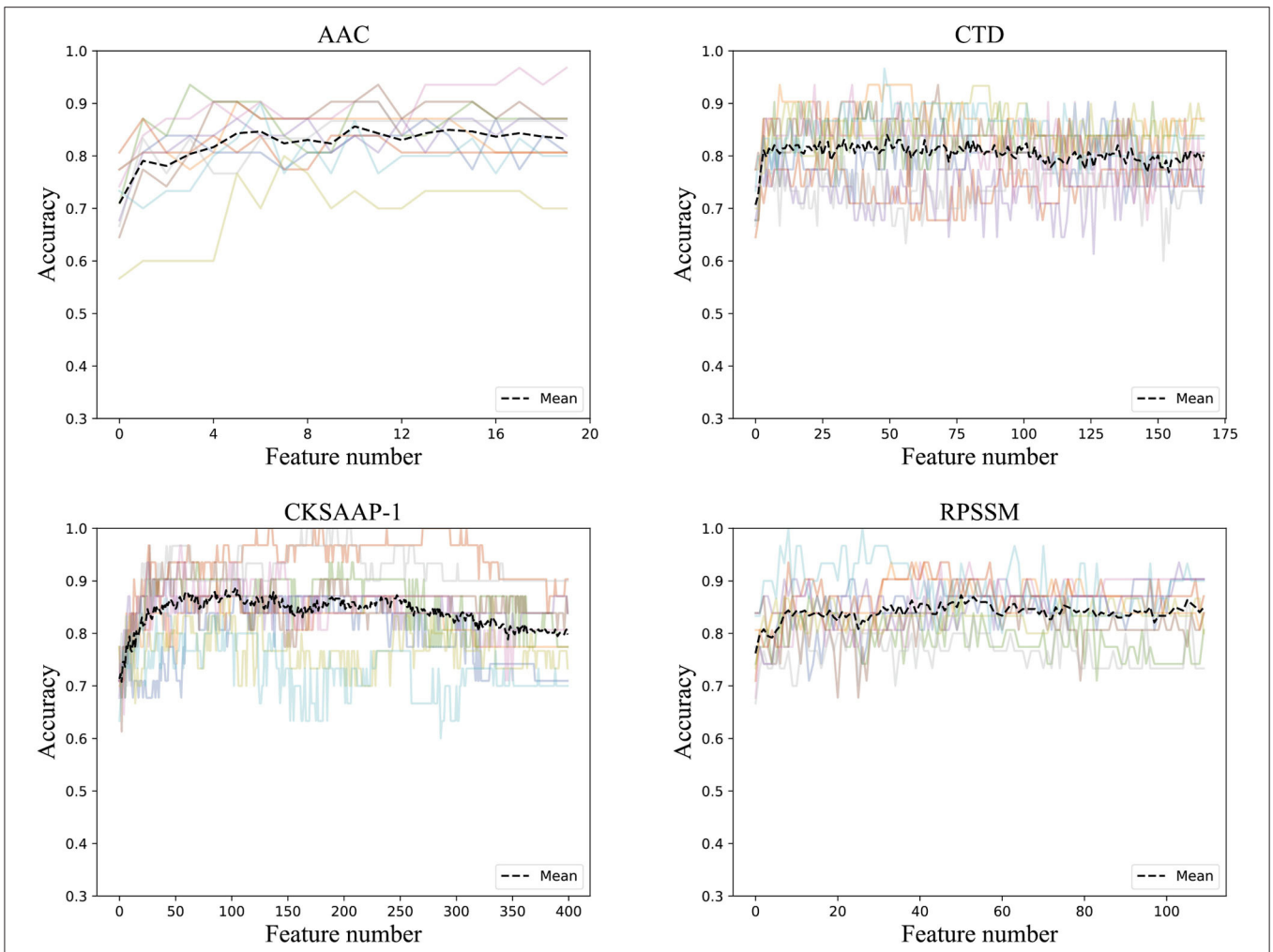

FIGURE 3 | Accuracies derived from the incremental strategy using different feature extraction methods (The colored line shows the accuracy results of the 10-fold data).

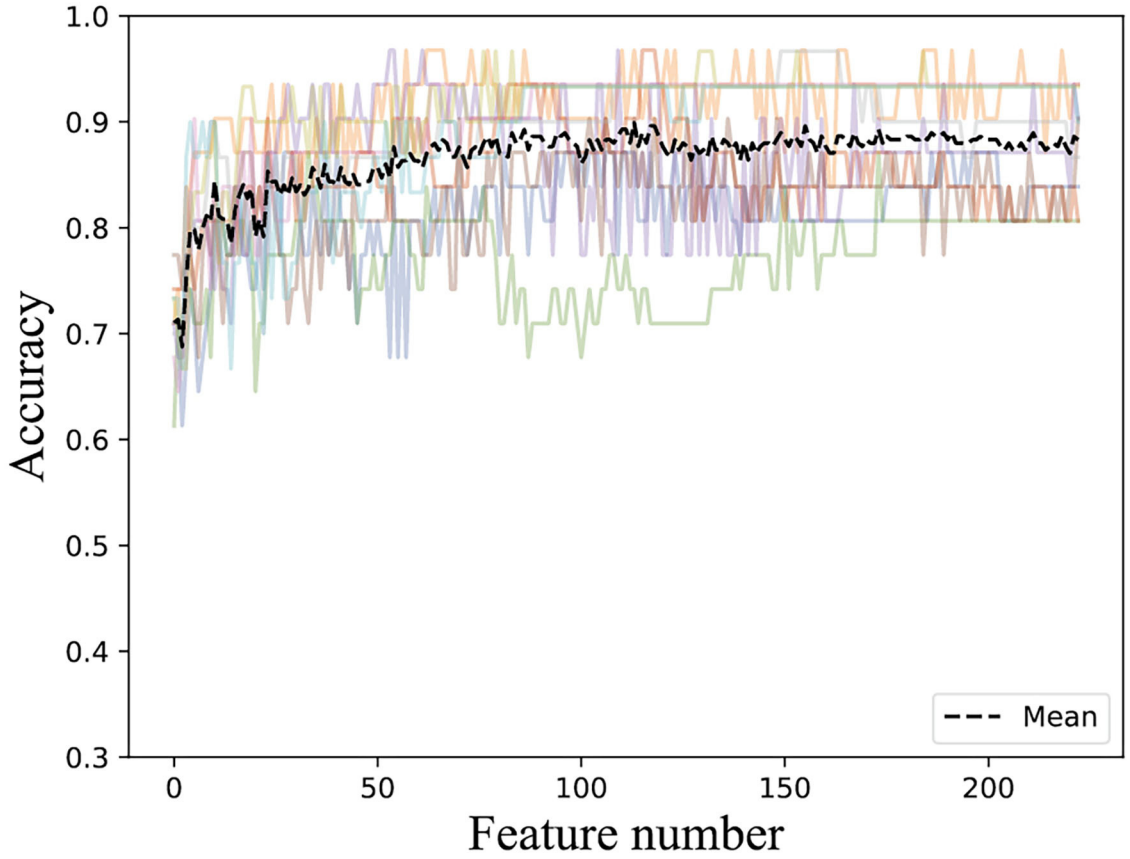

FIGURE 4 | Accuracies derived from the incremental strategy using integrated features.

TABLE 1 | Average accuracy of a 10-fold cross-validation on the training set using different features.

Classifier	Features	SN (%)	SP (%)	ACC (%)
DTC, GNB, LR, MLP	AAC (20D)	71.89	85.02	80.81
MLP, KNN, DTC, LR	CTD (168D)	69.00	86.12	80.41
KNN, MLP, LR, GNB	CKSAAP_1gap (400D)	70.78	80.71	77.51
GNB, LDA, LR, MLP	RPSSM (110D)	82.78	79.81	80.81
MLP, GNB, LR, DTC, KNN	Concatenation (698D)	56.67	90.63	79.79
GNB, MLP, KNN, LR	**CTD (50D)**	**73.88**	**89.40**	**84.29**
GNB, KNN	**CKSAAP_1gap (103D)**	**82.06**	**88.07**	**86.04**
LR, MLP, DTC, GNB, LDA	**RPSSM (50D)**	**81.78**	**87.05**	**85.37**
DTC, LR, MLP, LDA, KNN	**Concatenation (38D)**	**83.00**	**85.05**	**86.00**
GNB, LDA, LR	**Concatenation (87D)**	**89.00**	**86.55**	**89.28**

Bold indicates the result of the processing of the features.

TABLE 2 | Accuracy on independent test sets using different kinds of features.

Classifier	Features	SN (%)	SP (%)	ACC (%)
MLP	AAC (20D)	66.67	85.94	79.79
MLP	CTD (168D)	70.00	79.69	76.60
GNB	CKSAAP_1gap (400D)	63.33	85.94	78.72
GNB	RPSSM (110D)	73.33	76.56	75.53
GNB	**Concatenation (698D)**	56.67	90.63	79.79
LDA	**CTD (50D)**	**70.00**	**82.81**	**78.72**
SVM	**CKSAAP_1gap (103D)**	**46.67**	**92.19**	**77.66**
MLP	**RPSSM (50D)**	**60.00**	**92.19**	**81.92**
MLP	**Concatenation (38D)**	**73.33**	**90.63**	**85.11**
MLP	**Concatenation (87D)**	**73.33**	**93.75**	**87.23**

Bold indicates the result of the processing of the features.

According to the important feature subset selected from the AAC sequence feature, the physicochemical feature, dipeptide 1-gap content, which has been indicated to be the best (Ding et al., 2014), and RPSSM feature, we obtained 223 feature components totally. Accordingly, the average classification accuracy was calculated in turn on the training set using a 10-fold cross-validation. The qualitative results are shown in **Figure 4**. The quantitative results compared with other methods are listed in **Tables 1**, **2**.

It can be seen from the 10-fold results that the classification accuracy of only the first 20 feature components of AAC is better than that of the others. On the independent test set, the classification accuracy rate is 79.79%. It shows that the amino acid content of phage protein is quite different from that of the

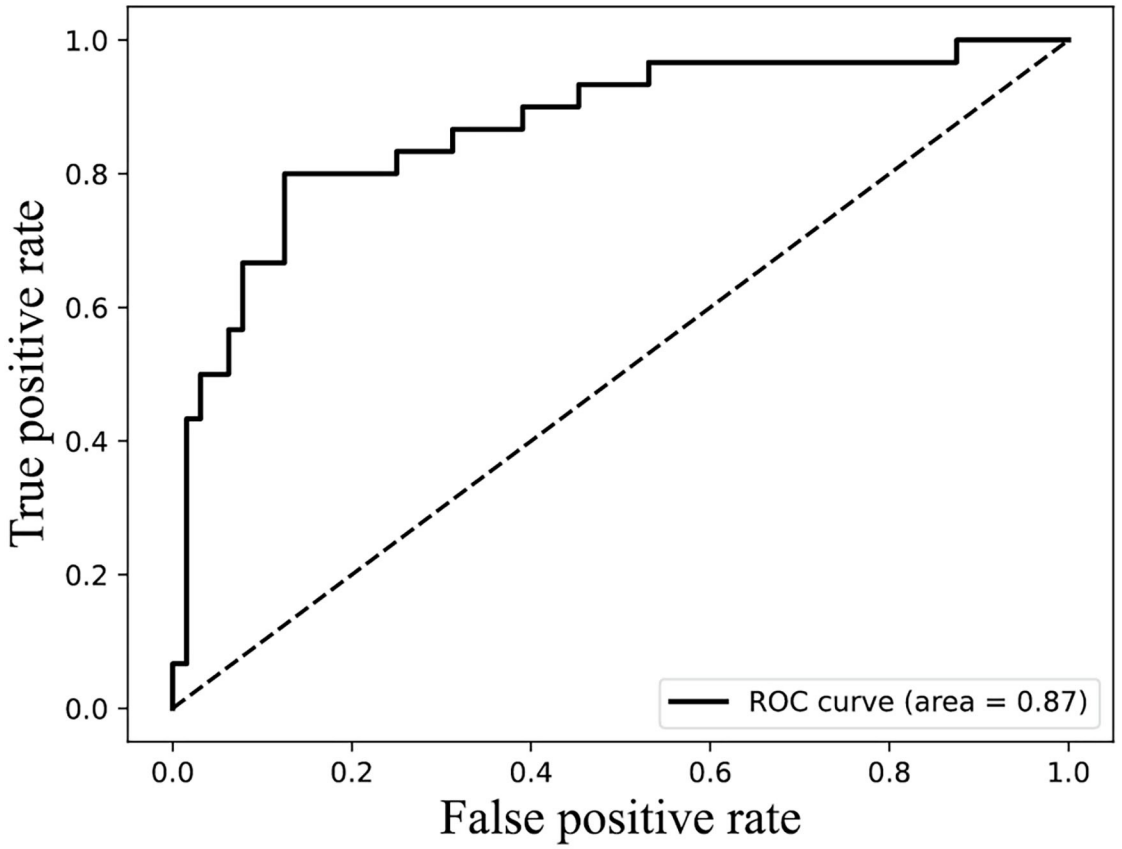

FIGURE 5 | ROC curves of 87 features on independent test data.

non-bacterial protein in these data, especially lysine and valine are the most important residues. The classification accuracy of CKSAAP-1GAP is not as good as that of AAC. Because the number of the feature subsets is 30% more than the total number of training samples, the classification accuracy rate is increased to 86.04% when 103 important feature components are selected, but the classification accuracy rate is only 77.66% when only CKSAAP-1GAP is used in the independent test set. However, the classification results of other individual feature subsets also show that the classification accuracy of only using any individual feature subsets is always low, so we decided to splice different types of feature subsets. To prevent the dimension from being too high, we only selected important features to combine.

For the training set of 223 dimensional features after splicing, we calculated the feature importance again using the feature selection modules shown in **Figure 1B** to calculate the influence of each feature on the results when different feature components were spliced. The features were then further filtered in the training data set using the incremental approach and cross-validation at 10-fold.

The result of **Figure 4** shows that 85% classification accuracy, which is the same classification accuracy as stated in the original article (Ding et al., 2014), can be achieved on the independent test set with 55 feature components selected, and 89.28% on the training set and 86.17% on the independent test with 87 feature components selected. Although the accuracy is 0.65% higher than 87 at the 113-dimension level, the number of feature components is greatly increased. Therefore, we chose 87 features for prediction on the independent test set, and the ROC curve is shown in **Figure 5**.

TABLE 3 | Performance comparison of the different features in independent test sets.

Classifier	Features	SN (%)	SP (%)	ACC (%)
Naïve Bayes	Ding et al., 2014 (38D)	75.76	80.77	79.15
SVM	Ding et al., 2014 (160D)	75.76	89.42	85.02
Bin et al., 2020	Nine feature groups (8D)	50.00	92.19	78.72
MLP	**Concatenation (38D)**	**73.33**	**90.63**	**85.11**
MLP	**Concatenation (87D)**	**73.33**	**93.75**	**87.23**

Bold indicates the result of the processing of the features.

From **Table 1**, it can be seen that the classification accuracy is low using only one type of all features and that using feature selection is significantly higher. As can be seen from **Table 3**, we can achieve the same results with fewer feature components compared to the original article (Ding et al., 2014).

For the 17 selected 1-gap dipeptides (A*G, A*T, A*P, S*T, S*A, V*A, T*S, V*T, G*A, G*G, S*G, V*G, V*I, E*L, K*L, K*E, and

E*E) in the original article (Ding et al., 2014), the order of feature importance is (1, 4, 5, 17, 15, 14, 7, 61, 18, 25, 8, 161, 16, 23, 3, 87, and 54) when determined individually. It can be found that there is a significant overlap between the feature we selected and the original article (Ding et al., 2014). Lin et al. analyzed the amino acid composition of filamentous bacterial virus xf (*Xanthomonas oryzae*) coat protein, which showed that His, Cys, and Phe were absent from the xf protein. This indicates that these three amino acids are not important in phage classification. In AAC features, the rankings of these three amino acids are 7, 9, and 17. It also occupies 1, 0, and 1 amino acids in the top 20 CKSAAPs.

DISCUSSION AND CONCLUSION

By analyzing the selected feature, it can be found that physicochemical properties are important for phage protein identification. In fact, 40 components representing physicochemical properties appear in the 87 features spliced. Of the first 15 features, 12 refer to the physicochemical properties. The most important feature comes from the physicochemical properties. The charge property is the most important, followed by polarity and polarization rate. Besides, the effective physicochemical properties are derived from different feature extraction methods. The contribution of CKSAAP-1GAP to the 87-dimensional feature is limited. Only 9 of CKSAAP-1GAP features are selected, while the secondary structure composed of the physicochemical properties derived from CTD occupies 20. Thus, it can be concluded that CTD makes a greater contribution than CKSAAP-1GAP on the selected feature for the classification of phage protein. As to AAC, its most important components are proline and leucine, which are highly ranked in the 87 features spliced. For RPSSM, its selected features are not top ranked in the 87-dimensional feature, but 30 features appear in the 87 features spliced. The extracted secondary structures using CTD and RPSSM after classifying amino acids in advance according to their properties are highly ranked in the 87 feature components, while the ones derived from CKSAAP-1GAP are not.

In this study, a feature selection framework is proposed for the protein classification of phages. The model improves the classification accuracy of the data by overlaying different types of features. To prevent overfitting caused by high feature dimensionality, the feature importance was quantified and the important features with high scores were selected as the final feature for classification. We performed ensemble learning using different classifiers, which are insensitive to the distribution of the original data, quantified the importance of each feature, and then performed feature selection on it. Finally, only an 87-dimensional feature was used to achieve a high classification accuracy. Compared with the original article (Ding et al., 2014) and PredNeuroP (Bin et al., 2020), a new method for recognizing phage protein, the model can achieve the same classification accuracy using only 38 feature components. The classification accuracy reaches 87.23% when the optimal 87-dimensional feature is used. It also shows that, when the relationship between the structure and function of phage proteins is not fully understood, the knowledge-driven approach to feature extraction alone does not necessarily lead to better prediction results. In contrast, the prediction of phage proteins by a combination of knowledge-driven and data-driven is more accurate, and the functions of phage proteins can be further investigated by analyzing the selected feature.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SL and CC proposed this research topic. TL completed the experiment and wrote the manuscript. HC supervised the experimental process and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Ahmad, K., Waris, M., and Hayat, M. (2016). Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition. *J. Membr. Biol.* 249, 293–304. doi: 10.1007/s00232-015-9868-8

Bin, Y., Zhang, W., Tang, W., Dai, R., Li, M., Zhu, Q., et al. (2020). Prediction of neuropeptides from sequence information using ensemble classifier and hybrid features. *J. Proteome Res.* 19, 3732–3740. doi: 10.1021/acs.jproteome.0c00276

Cai, C., Han, L., Ji, Z. L., Chen, X., and Chen, Y. Z. (2003). SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31, 3692–3697. doi: 10.1093/nar/gkg600

Chou, K.-C. (2009). Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics* 6, 262–274. doi: 10.2174/157016409789973707

Clokie, M. R., Millard, A. D., Letarov, A. V., and Heaphy, S. (2011). Phages in nature. *Bacteriophage* 1, 31–45. doi: 10.4161/bact.1.1.14942

Cui, X., Yu, Z., Yu, B., Wang, M., Tian, B., and Ma, Q. (2019). UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components. *Chemometr. Intell. Lab. Syst.* 184, 28–43. doi: 10.1016/j.chemolab.2018.11.012

Ding, H., Feng, P.-M., Chen, W., and Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol. Biosyst.* 10, 2229–2235. doi: 10.1039/C4MB00316K

Ding, Y., Tang, J., and Guo, F. (2020). Human protein subcellular localization identification via fuzzy model on kernelized neighborhood representation. *Appl. Soft Comput.* 96, 106596. doi: 10.1016/j.asoc.2020.106596

Feng, P.-M., Ding, H., Chen, W., and Lin, H. (2013). Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.* 2013. doi: 10.1155/2013/530696

Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. *Bioinformatics* 36, 3028–3034. doi: 10.1093/bioinformatics/btaa131

Guo, Z., Wang, P., Liu, Z., and Zhao, Y. (2020). Discrimination of thermophilic proteins and non-thermophilic proteins using feature dimension reduction. *Front. Bioeng. Biotechnol.* 8, 584807. doi: 10.3389/fbioe.2020.584807

Jahn, M. T., Arkhipova, K., Markert, S. M., Stigloher, C., Lachnit, T., Pita, L., et al. (2019). A phage protein aids bacterial symbionts in eukaryote immune

evasion. *Cell Host Microbe* 26, 542–550.e545. doi: 10.1016/j.chom.2019.08.019

Jara-Acevedo, R., Díez, P., González-González, M., Dégano, R. M., Ibarrola, N., Góngora, R., et al. (2018). "Screening phage-display antibody libraries using protein arrays," in *Phage Display*. (Totowa, NJ: Springer), 365–380. doi: 10.1007/978-1-4939-7447-4_20

Ji, Z., Meng, G., Huang, D., Yue, X., and Wang, B. (2015). NMFBFS: a NMF-based feature selection method in identifying pivotal clinical symptoms of hepatocellular carcinoma. *Comput. Math. Methods Med.* 2015, 846942. doi: 10.1155/2015/846942

Jiang, M., Zhao, B., Luo, S., Wang, Q., Chu, Y., Chen, T., et al. (2021). NeuroPpred-Fuse: an interpretable stacking model for prediction of neuropeptides by fusing sequence information and feature selection methods. *Brief. Bioinformatics* 22, bbab310. doi: 10.1093/bib/bbab310

Jiao, Y.-S., and Du, P.-F. (2017). Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions. *J. Theor. Biol.* 416, 81–87. doi: 10.1016/j.jtbi.2016.12.026

Jin, S., Zeng, X., Xia, F., Huang, W., and Liu, X. (2021). Application of deep learning methods in biological networks. *Brief. Bioinformatics* 22, 1902–1917. doi: 10.1093/bib/bbaa043

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292, 195–202. doi: 10.1006/jmbi.1999.3091

Kawashima, S., and Kanehisa, M. (2000). AAindex: amino acid index database. *Nucleic Acids Res.* 28, 374–374. doi: 10.1093/nar/28.1.374

Khan, M., Hayat, M., Khan, S. A., Ahmad, S., and Iqbal, N. (2017). Bi-PSSM: position specific scoring matrix based intelligent computational model for identification of mycobacterial membrane proteins. *J. Theor. Biol.* 435, 116–124. doi: 10.1016/j.jtbi.2017.09.013

Lavigne, R., Ceyssens, P.-J., and Robben, J. (2009). "Phage proteomics: applications of mass spectrometry," in *Bacteriophages*. Totowa, NJ: Humana Press, 239–251. doi: 10.1007/978-1-60327-565-1_14

Lekunberri, I., Subirats, J., Borrego, C. M., and Balcázar, J. L. (2017). Exploring the contribution of bacteriophages to antibiotic resistance. *Environ. Pollut.* 220, 981–984. doi: 10.1016/j.envpol.2016.11.059

Li, L., Yu, S., Xiao, W., Li, Y., Hu, W., Huang, L., et al. (2015). Protein submitochondrial localization from integrated sequence representation and SVM-based backward feature extraction. *Mol. Biosyst.* 11, 170–177. doi: 10.1039/C4MB00340C

Li, T., Fan, K., Wang, J., and Wang, W. (2003). Reduction of protein sequence complexity by residue grouping. *Protein Eng.* 16, 323–330. doi: 10.1093/protein/gzg044

Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2. 0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* 47, e127–e127. doi: 10.1093/nar/gkz740

Mei, S. (2012). Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. *J. Theor. Biol.* 293, 121–130. doi: 10.1016/j.jtbi.2011.10.015

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.5555/1953048.2078195

Silvério-Machado, R., Couto, B. R., and Dos Santos, M. A. (2015). Retrieval of Enterobacteriaceae drug targets using singular value decomposition. *Bioinformatics* 31, 1267–1273. doi: 10.1093/bioinformatics/btu792

Wen, P.-P., Shi, S.-P., Xu, H.-D., Wang, L.-N., and Qiu, J.-D. (2016). Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization. *Bioinformatics* 32, 3107–3115. doi: 10.1093/bioinformatics/btw377

Xie, X., Gu, X., Li, Y., and Ji, Z. (2021). K-size partial reduct: positive region optimization for attribute reduction. *Knowl. Based Syst.* 228, 107253. doi: 10.1016/j.knosys.2021.107253

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D.-Q. (2018). PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol.* 9, 2571. doi: 10.3389/fmicb.2018.02571

Xu, Y., Verma, D., Sheridan, R. P., Liaw, A., Ma, J., Marshall, N. M., et al. (2020). Deep dive into machine learning models for protein engineering. *J. Chem. Inf. Model.* 60, 2773–2790. doi: 10.1021/acs.jcim.0c00073

Yu, B., Li, S., Qiu, W., Wang, M., Du, J., Zhang, Y., et al. (2018). Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. *BMC Genomics* 19, 1–17. doi: 10.1186/s12864-018-4849-9

Yu, L., Wang, M., Yang, Y., Xu, F., Zhang, X., Xie, F., et al. (2021). Predicting therapeutic drugs for hepatocellular carcinoma based on tissue-specific pathways. *PLoS Comput. Biol.* 17, e1008696. doi: 10.1371/journal.pcbi.1008696

Yuan, Y., and Gao, M. (2016). Proteomic analysis of a novel bacillus jumbo phage revealing glycoside hydrolase as structural component. *Front. Microbiol.* 7, 745. doi: 10.3389/fmicb.2016.00745

Zhang, L., Zhang, C., Gao, R., and Yang, R. (2015). An ensemble method to distinguish bacteriophage virion from non-virion proteins based on protein sequence characteristics. *Int. J. Mol. Sci.* 16, 21734–21758. doi: 10.3390/ijms160921734

Zhao, X., Wang, H., Li, H., Wu, Y., and Wang, G. (2021). Identifying plant pentatricopeptide repeat proteins using a variable selection method. *Front. Plant Sci.* 12, 298. doi: 10.3389/fpls.2021.506681

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Sun, Z.-J., Dao, F.-Y., Yu, X.-L., et al. (2021). Identification of cyclin protein using gradient boost decision tree algorithm. *Comput. Struct. Biotechnol. J.* 19, 4123–4131. doi: 10.1016/j.csbj.2021.07.013

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.