frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 24 March 2020
doi: 10.3389/fbioe.2020.00183

Predicting Bacteriophage Enzymes and Hydrolases by Using Combined Features

*Hong-Fei Li[1,2], Xian-Fang Wang[2] and Hua Tang[1]**

[1] Department of Pathophysiology, Key Laboratory of Medical Electrophysiology, Ministry of Education, Southwest Medical University, Luzhou, China, [2] School of Computer and Information Engineering, Henan Normal University, Henan, China

Bacteriophage is a type of virus that could infect the host bacteria. They have been applied in the treatment of pathogenic bacterial infection. Phage enzymes and hydrolases play the most important role in the destruction of bacterial cells. Correctly identifying the hydrolases coded by phage is not only beneficial to their function study, but also conducive to antibacteria drug discovery. Thus, this work aims to recognize the enzymes and hydrolases in phage. A combination of different features was used to represent samples of phage and hydrolase. A feature selection technique called analysis of variance was developed to optimize features. The classification was performed by using support vector machine (SVM). The prediction process includes two steps. The first step is to identify phage enzymes. The second step is to determine whether a phage enzyme is hydrolase or not. The jackknife cross-validated results showed that our method could produce overall accuracies of 85.1 and 94.3%, respectively, for the two predictions, demonstrating that the proposed method is promising.

Keywords: bacteriophage enzymes, hydrolase, analysis of variance, sequence feature, classification

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Balachandran Manavalan,
Ajou University, South Korea
Yongchun Zuo,
Inner Mongolia University, China

***Correspondence:**
Hua Tang
huatang@swmu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *02 February 2020*
Accepted: *24 February 2020*
Published: *24 March 2020*

Citation:
Li H-F, Wang X-F and Tang H (2020) Predicting Bacteriophage Enzymes and Hydrolases by Using Combined Features. Front. Bioeng. Biotechnol. 8:183. doi: 10.3389/fbioe.2020.00183

INTRODUCTION

Bacteriophage, as safe agent, can lyse and infect specific bacteria without destroying natural beneficial microflora (Parmar et al., 2018). Hydrolytic enzymes encoded by phages are key ingredients of lysis, which is helpful to fighting bacterial pathogens, especially those that cannot be killed by antibiotics and chemicals. In fact, in some countries, they have been used therapeutically to treat bacterial infections that do not respond to antibiotics (Thiel, 2004; Parfitt, 2005; Keen, 2012). They have also been used as a food safety tool to reduce bacterial contamination (Pirisi, 2000). Hence, rapid detection of bacteriophage and hydrolase responsible for antibacterial drugs is a growing necessity for public health.

Because of abuse of antibiotics, certain resistant viruses cannot be effectively controlled. This problem can be resolved by therapy of phage hydrolytic that disintegrates host viruses during releasing progeny phage. Therefore, the identification of hydrolases encoded by phages has become an important research topic. It not only has been studied in chemistry and physics through

Abbreviations: CTD, composition transition and distribution; SVM, support vector machine; RF, random forest; MLP, multilayer perceptron; KNN, k-nearest neighbors; Sn, sensitivity; Sp, specificity; Ac, accuracy; MCC, matthew correlation coefficient; PseAAC, pseudo–amino acid composition; GTPC, grouped tripeptide composition; ROC, receiver operating characteristic; AUC, area under receiver operating characteristic (ROC) curve; GGDC, g-gap dipeptide composition; ANOVA, analysis of variance; RBF, Radial Basis Function; ORFs, Open Reading Frames.

experimental methods, but also achieved good results in theory through recently popular machine learning algorithms. Some experiments have been performed to study the function of phage hydrolase (Kimura and Itoh, 2003; Rodriguez-Rubio et al., 2013). In addition, in the study of host cell lysis by hydrolytic enzyme activation, Kovalenko et al. (2019) found that the calcium could regulate phage-induced bacterial lysis. Although those biochemical-based methods can accurately recognize phage hydrolases and clearly elucidate the functional mechanism of the enzyme, it is time-consuming and expensive. Additionally, biochemical experiments always need rigorous experimental conditions, which will prevent most of scholars from doing more in-depth studies. Computational methods provide another chance to study phage hydrolase without the disadvantage of biochemical-based methods. Phylogenetic analysis or similarity search could find relative conservation of motifs among related species (Lin and Li, 2011; Liu et al., 2019). However, it is extremely diverse for phage Open Reading Frames (ORFs), of which more than 70% of them cannot find out similar genes with annotated functions in GenBank (Seguritan et al., 2012). Moreover, it is also time-consuming.

With the accumulation of more and more postgenomic data, some computational methods have been proposed to study the function of phage proteins. Riede and his colleagues (Riede et al., 1987) have proposed a model to predict tail-fiber proteins' three-dimensional structure of T-even–type phages. The results are consistent with electron microscopic data. Subsequently, a computer program was developed to identify DNA-binding regulatory proteins in bacteriophage T7 (White, 1987; Song et al., 2014; Zou et al., 2016a; Qu et al., 2019). Recently, the virion proteins encoded by phages were studied by using naive Bayes combined with primary sequence information (Feng et al., 2013). The proposed model could yield the overall accuracy (Ac) of 79.15%. By using feature selection technique, the overall Ac was improved to 85.02% (Ding et al., 2014). A free webserver called PVPred (Ding et al., 2014) was constructed for predicting phage virion proteins.

The success of previous works on the prediction of phage functional proteins (Feng et al., 2013; Ding et al., 2014) and enzyme prediction (Zuo et al., 2014; Ding H. et al., 2016) provided good strategy to discriminate hydrolases encoded by phages by transforming protein sequences into digital features and further establishing machine learning-based models. Thus, this work aims to develop a powerful computational model to recognize phage hydrolase by combining feature selection and expression of multiple features. The entire experiment was divided into two steps. First is to discriminate phage enzymes from phage nonenzymes and then to identify phage hydrolases from phage enzymes. In this model, the support vector machine (SVM) was applied as the algorithm to perform the classification. Different features were proposed to formulate protein samples and then inputted into SVM. The best features that can achieve the maximum accuracies were discovered by using analysis of variance (ANOVA). The model's performance was estimated by using jackknife cross-validation.

MATERIALS AND METHODS

Benchmark Dataset

Constructing a reliable benchmark dataset could guarantee the reliability of the proposed computational model (Ma et al., 2014; Liang et al., 2017; Yang et al., 2017; Wang et al., 2018; Cheng et al., 2019; Hu et al., 2019; Zheng et al., 2019). In this work, samples were gained from Ding H. et al. (2016), which were rigorously screened through the following three steps: (1) phage proteins have been annotated by standard operating procedure for UniProt manual curation (Swiss-Prot); (2) protein sequences samples containing illegal characters were deleted; (3) sequence identity in the dataset must be less than 30%, which was implemented by CD-HIT (Fu et al., 2012) software. Consequently, the definitive benchmark dataset contains 255 phage proteins, of which 124 proteins belong to phage enzymes (positive samples of set 1), and the remaining 131 are phage nonenzymes (negative samples of set 1). Furthermore, 124 phage enzymes are divided into 69 hydrolases (positive samples of set 2) and 55 nonhydrolases (negative samples of set 2), respectively. The following calculations are all based on these data.

Protein Feature Extraction

The perfect expression of protein sequences by digital features can dramatically increase the Ac and robust of computing models (Wang et al., 2008, 2010; Song et al., 2010, 2018; Zuo et al., 2017; Basith et al., 2018; Chen W. et al., 2018; Wei et al., 2018b; Boopathi et al., 2019; Ding et al., 2019; Manavalan et al., 2019b; Shen et al., 2019; Tan et al., 2019; Zhang and Liu, 2019; Zhu et al., 2019). The specific order of residues in the peptide sequence dictates the protein to fold up into a special three-dimensional structure. Thus, the interaction between two residues in a protein is a main factor to characterize the protein. In the past 20 years, scholars have developed dipeptide composition to formulate peptide samples (Tang et al., 2016). However, the feature can only describe the short-range interaction between two residues. In fact, there are lots of long-range interaction for a protein in three-dimensional space. For example, the secondary structures (α helix and β sheet) were formed by the interaction of two nonadjoining residues. Hence, it will be more reasonable to investigate the performance of other kinds of correlations.

Based on the above analysis and other peer works (Ding and Li, 2015), in this work, the g-gap dipeptide composition (GGDC), which is extended from general dipeptide composition, is used as the main feature to denote the residues' correlation in the original peptide sequence. For the perfect expression of the sample, the combination of GGDC, pseudo–amino acid composition (PseAAC), grouped tripeptide composition (GTPC), and composition transition and distribution (CTD) is used as the final feature vector. Pseudo–amino acid composition provides the correlation of physical and chemical properties between two residues (Chen et al., 2016; Yang et al., 2016). Grouped tripeptide composition provides tripeptide information (Tan et al., 2019). CTD provides distribution patterns of a specific structural property for residues (Cheng et al., 2018) and indirectly

contains information about 20 amino acid residues, so PseAAC, in our work, does not contain amino acid information.

G-Gap Dipeptide Composition

The GGDC proposed by Ding et al. (2014) is the extension of the proximate dipeptide composition, because proteins contain deep correlation of residues relating with hydrogen bonding in secondary structure. For different g, the protein sequence P with L residues is expressed by a 400-dimensional GGDC as follows:

$$P = [f_1^g, f_2^g, ..., f_\varepsilon^g, ..., f_{400}^g]^T \tag{1}$$

where T is called the transposing operator, the f_ε^g can be calculated by:

$$f_\varepsilon^g = n_\varepsilon^g/(L - g - 1) \tag{2}$$

where the n_ε^g denotes the absolute occurrence number of the GGDC in a protein. Since previous studies (Ding H. et al., 2016) have shown that $g = 2$ has the best prediction effect, only 2-gap was used in our experiments.

Pseudo–Amino Acid Composition

Hydrophobicity, hydrophilicity, and other physicochemical properties are important characteristics of amino acids. In order to incorporate these properties with amino acid composition, two types of PseAAC were used. In our work, motived by PseAAC, the protein sample, can be expressed as follows:

$$\begin{cases} \tau_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{k,k+1}^1 \\ \tau_2 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{k,k+1}^2 \\ \cdots \\ \tau_n = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{k,k+1}^n \\ \tau_{n+1} = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{k,k+2}^1, (l < L) \\ \tau_{n+2} = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{k,k+2}^2 \\ \cdots \\ \tau_{\lambda n} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{k,k+\lambda}^n \end{cases} \tag{3}$$

$H_{k,k+\lambda}^n$ th residue and the $(k+\lambda)$-th residue; L is length of sample. After experimental comparison, we selected 10 physical and chemical properties containing hydrophobicity, hydrophilicity, amino acid side chain group mass, -COOH group dissociation constant, -NH3 group dissociation constant, isoelectric point at 25°C, rigidity, flexibility, irreplaceability, and polarity. We used $\lambda = 15$.

GTPC and CTD

iFeature is a comprehensive Python-based toolkit that contains four major functions: feature representation, dimensionality reduction algorithms, feature selection algorithms, and feature clustering algorithms (Chen Z. et al., 2018). In our study, we have used GTPC and CTD provided by iFeature (Chen Z. et al., 2018) to extract numerical descriptors from samples. Grouped tripeptide composition converts protein sequences into 125-dimensional digital features expressed as follows:

$$f(r, s, t) = \frac{N_{rst}}{N-1}, \quad r, s \in \{g1, g2, g3, g4, g5\} \tag{4}$$

where N_{rst} denotes the number of tripeptides in groups r, s, and t (Chen Z. et al., 2018). N is the length of a protein.

CTD converts protein sequences into 39-dimensional digital features defined as follows:

$$C(r) = \frac{N_r}{N}, \quad r\{polar, neutral, hydrophoic\} \tag{5}$$

where $N(r)$ represents the number of residue type r in the peptide sequence (Chen Z. et al., 2018). Thus, samples are transformed into 164 -dimensional features.

Support Vector Machine

Support vector machine is a classical machine learning algorithm and has been widely adopted in computational biology (Jiang et al., 2013; Zhao et al., 2015, 2017; Ding H. et al., 2016; Ding et al., 2016a,b; Dao et al., 2018; Feng et al., 2018; Manavalan et al., 2018a,b; Zhang et al., 2018; Chao et al., 2019; Chen et al., 2019a; Wang et al., 2019; Basith et al., 2020). For nonlinear samples, its projects inputted data into high-dimensional spare by a kernel function. There are four kernel functions including Sigmoid function, Gaussian function, line function, and polynomial function, among which Gaussian function is most commonly used. C and g are the most important parameters to adjust performance of Gaussian function. The value of g is related to the partitioning of samples, and the value of C determines the tolerance of the model. In our work, SVC functions in Scikit-learn (Swami and Jain, 2012), based on Python, are used to build models, and Gaussian functions are used as kernel functions, because the Gaussian function can efficiently map small samples with fewer features to high-dimensional space and distinguish positive and negative samples with high Ac. In addition, the GridSearchCV function in Scikit-learn was used to optimize the parameters C and g.

Feature Selection Method

Because one type of feature does not fully represent the characteristics of a protein sequence, the combination of features is a good approach to perform classifications. The combined features could also cause a lot of inconvenience, such as noise, dimension disaster, and so on. Analysis of variance (Feng et al., 2013; Tang et al., 2017; Xianfang et al., 2019), principal component analysis (Dong et al., 2015), minimal redundancy maximal relevance (Ding et al., 2013), maximum relevance maximum distance (Zou et al., 2016b), and increment of diversity (Zuo and Li, 2009; Zhao et al., 2010; Fan and Li, 2012) can solve these problems. In our study, ANOVA is used to screen the best feature set; the idea is to calculate the ratio of the categories to sample variance. Obviously, features with larger ratios are more suitable for classification. The details can be referred from Feng et al. (2013), Tang et al. (2017) and Xianfang et al. (2019).

Performance Evaluation

In statistical prediction, the performance of the model needs to be measured by some methods and parameters (Chen et al., 2017, 2019b; Ding et al., 2017; Tang et al., 2018; Yang et al., 2018). The cross-validation test has been widely used to evaluate methods (Yang et al., 2019; Zhu et al., 2019). To provide a fair comparison, we used the jackknife test in this study. The four parameters, namely, sensitivity (Sn), specificity (Sp), Ac, and Matthew correlation coefficient (MCC), are used to evaluate the performance of the model (Liu et al., 2018; Manavalan et al., 2018c, 2019a,c; Basith et al., 2019), which are defined as follows:

$$\begin{cases} S_n = \frac{TP}{TP+FN} \\ S_p = \frac{TP}{TP+FN} \\ A_c = \frac{TP+FN}{TP+FP+TN+FN} \\ MCC = \frac{(TP\times TN)+(FP\times FN)}{(TP+FN)(TN+FP)(TP+FP)(TN+FN)} \end{cases} \quad (6)$$

where TP and TN are the number of the correctly identified positive samples and the number of the correctly identified negative samples; FP indicates the number of negative samples recognized as positive samples; FN indicates the number of positive samples recognized as negative samples. Also, the area under receiver operating characteristic (ROC) curve (AUC) is often used to evaluate the performance of binary classification models.

RESULTS

Discriminating Phage Enzymes From Nonenzymes

For a new sequenced phage protein, we first need to judge whether the phage protein is an enzyme. Thus, the predictive performances of three combined vectors were investigated by using SVM with jackknife test. First, samples are expressed by three kinds of combinations: GGDC combined with PseAAC, GTPC combined with CTD, and all features. Prediction results are listed in **Table 1**. We observed that all features cannot achieve the best Ac. The reason is maybe noise or redundant information. Thus, we performed feature selection for three feature combinations to discover the best feature subsets. The results are also shown in **Table 1**. After feature selection, the highest Ac was obtained by using 191 features, which was based on all features. **Figure 1A** was drawn to show the *F*-value for all features. The above results implied that the information of phage enzymes requires multiple types of

TABLE 1 | The results by using different features for phage enzymes prediction.

Combined vector features	Original feature		Optimal features	
	Accuracy	Dimensions	Accuracy	Dimensions
GGDC + PseAAC	74.5%	550	83.1%	154
GTPC + CTD	67.8%	164	77.6%	35
GGDC + PseAAC + GTPC + CTD	72.9%	714	85.1%	191

GGDC, g-gap dipeptide composition; CTD, composition transition and distribution; PseAAC, pseudo–amino acid composition; GTPC, grouped tripeptide composition.

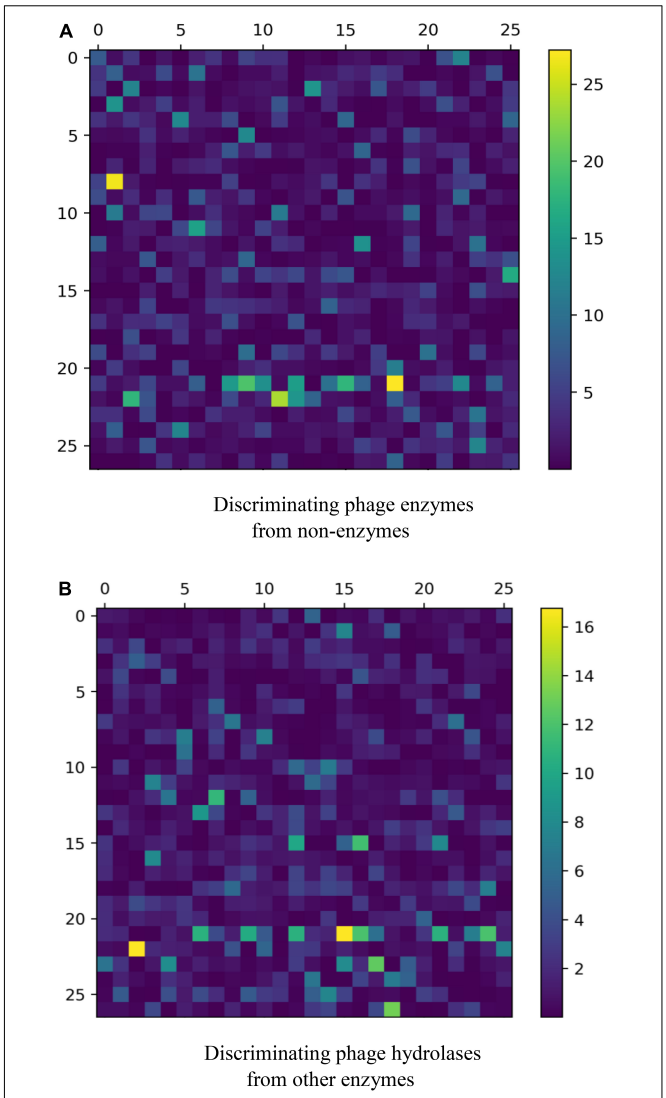

FIGURE 1 | A plot showing the *F*-values for **(A)** discriminating phage enzymes from nonenzymes and **(B)** discriminating phage hydrolases from other enzymes.

TABLE 2 | The comparison of different classifiers for predicting phage enzymes.

Classifier	Sn	Sp	Ac	MCC	AUC
KNN	0.98	0.16	0.702	0.232	0.664
RF	0.73	0.76	0.752	0.490	0.798
SVM	0.83	0.88	0.851	0.703	0.897
MLP	0.77	0.84	0.812	0.610	0.858

SVM, support vector machine; RF, random forest; MLP, multilayer perceptron; KNN, k-nearest neighbors; Sn, sensitivity; Sp, specificity; Ac, accuracy; MCC, Matthew correlation coefficient; AUC, area under receiver operating characteristic (ROC) curve.

TABLE 3 | The results by using different feature for discriminating phage hydrolases from other enzymes.

Combined vector features	Original features		Optimal features	
	Accuracy	Dimensions	Accuracy	Dimensions
GGDC + PseAAC	75.8	550	94.3%	61
GTPC + CTD	76.6%	164	86.4%	37
GGDC + PseAAC + GTPC + CTD	75.8%	714	92.7%	89

GGDC, g-gap dipeptide composition; CTD, composition transition and distribution; PseAAC, pseudo–amino acid composition; GTPC, grouped tripeptide composition.

TABLE 4 | The comparison of different classifiers for discriminating phage hydrolases from other enzymes.

Classifier	Sn	Sp	Ac	MCC	AUC
KNN	0.70	0.89	0.814	0.588	0.863
RF	0.91	0.80	0.86	0.722	0.898
SVM	0.96	0.93	0.943	0.886	0.961
MLP	0.93	0.91	0.927	0.837	0.948

SVM, support vector machine; RF, random forest; MLP, multilayer perceptron; KNN, k-nearest neighbors; Sn, sensitivity; Sp, specificity; Ac, accuracy; MCC, Matthew correlation coefficient; AUC, area under receiver operating characteristic (ROC) curve.

feature expressions. However, noises or redundant information may be results in the poor predictive capabilities of other groups, and the combining vectors of the first and second groups cannot fully express the peculiarity of the samples, which lead to its poor prediction effect. Subsequently, we investigated the performance of four classifiers, including random forest (RF), multilayer perceptron (MLP), k-nearest neighbor (KNN), and SVM, whose input features are the third set of 191-D optimal features. The result parameters of four classifiers have been exhibited in **Table 2**. We found the highest Ac of 85.1% and MCC of 70.3%. The AUC reaches to 89.3% by using SVM. k-Nearest neighbor has achieved the highest Sn of 98% with the lowest Sp of 16%. Moreover, performance of RF has an Sn of 73%, Sp of 76%, Ac of 75.2%, MCC of 0.490, and AUC of 0.798, respectively. Similarity, MLP obtained 77, 84, 81.2, 0.61, and 0.858%, respectively, for Sn, Sp, Ac, MCC, and AUC. These data indicate that SVM is the most suitable for distinguishing phage enzymes.

Discriminating Phage Hydrolases From Other Enzymes

When a phage protein is predicted as a phage enzyme, it is necessary to immediately judge whether the enzyme is a hydrolase. Like phage enzyme prediction, the performances of three combined vectors on phage hydrolase prediction were also examined by using SVM with jackknife cross-validation. As shown in **Table 3**, the three combined vectors were also processed by the feature selection algorithm, which not only improves the Ac but also greatly reduces the dimensions. Obviously, ANOVA can remove redundant information from features. It should be noticed that the optimal features (61-D) obtained from GGDC combined with PseAAC could produce the maximum Ac of 94.3%. This phenomenon indicates that features with a large *F*-value in the second group are not suitable for expressing hydrolases. The heat map for the features is also drawn in **Figure 1B**. Similarly, we compared the performances of different classifiers. In **Table 4**, KNN has yielded Ac of 81.4%, whereas KNN has obtained Ac of 84.64%. The performance of MLP is 93% Sn, 91% Sp, 92.7% Ac, 0.837 MCC, and 0.948 AUC. Support vector machine with Radial Basis Function (RBF) as kernel function gained the best prediction performance (94.3% Ac).

Performance Comparison With Existing Methods

In order to prove that our proposed model performs better than the model by Ding H. et al. (2016), who first used computational methods to predict hydrolases, the performance indexes of the two models were recorded in **Table 5**. In discriminating phage enzymes from nonenzymes, our model is better in Ac and Sp that are 85.1 and 88.0%, respectively. In discriminating phage hydrolases from other enzymes, all the evaluated indexes of our proposed model are better than those of Ding H. et al. (2016). Indeed, hydrolyzing enzymes adopt two types of features to encode samples. Compared with Ding and colleagues' experiment, we have selected more kinds of features in the sample expression, which makes the digital features of the sample more informative.

DISCUSSION

The purpose of this study is to establish a predictive model to predict phage enzymes and hydrolases. In fact, similarity search could be used to perform sequence analysis and function

TABLE 5 | Comparison of predictive performance with exist method.

		Ac	Sp	Sn
Discriminating phage enzymes from nonenzymes	(Ding H. et al., 2016)	84.3%	81.7%	87.1%
	This study	85.1%	88.0%	83.0%
Discriminating phage hydrolases from other enzymes	(Ding H. et al., 2016)	93.5%	92.8%	94.5%
	This study	94.3%	93.0%	96.0%

Sn, sensitivity; Sp, specificity; Ac, accuracy.

prediction. However, the strategy cannot work well on low-similar sequences. Especially, the phage genes display the extreme diversity. Protein functions are inextricably linked to correlation of nucleotides or residues, physicochemical properties, spatial structure, and other information. Therefore, we used multiple characteristics to represent phage and hydrolase, but this method has some problems that multiple features contain too much redundant information; different types of features are suitable for different samples. On the basis of the feature selection technique, promising results for phage enzymes and hydrolases prediction were achieved. In the future, we will pay more attention on deep learning, which has solved several protein prediction problems (Peng et al., 2018; Wei et al., 2018a, 2019; Yu et al., 2018; Lv et al., 2019) and may get well performance on this topic. Moreover, we will establish a free webserver that facilitates users to download data and predict phage hydrolases.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the http://lin-group.cn/server/PHYPred.

AUTHOR CONTRIBUTIONS

H-FL and HT designed the study. H-FL carried out all data collection and drafted the manuscript. X-FW and HT revised the manuscript. All authors approved the final manuscript.

FUNDING

This work has been supported by the National Nature Scientific Foundation of China (61702430).

REFERENCES

Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020). Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med. Res. Rev.* [Epub ahead of print].

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2018). iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.* 16, 412–420. doi: 10.1016/j.csbj.2018.10.007

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2019). SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids* 18, 131–141. doi: 10.1016/j.omtn.2019.08.011

Boopathi, V., Subramaniyam, S., Malik, A., Lee, G., Manavalan, B., and Yang, D. C. (2019). mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.* 20: E1964.

Chao, L., Jin, S., Wang, L., Guo, F., and Zou, Q. (2019). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.* 7:224. doi: 10.3389/fbioe.2019.00224

Chen, W., Feng, P., Liu, T., and Jin, D. (2018). Recent advances in machine learning methods for predicting heat shock proteins. *Curr. Drug Metab.* 20, 224–228. doi: 10.2174/1389200219666181031105916

Chen, W., Feng, P., Song, X., Lv, H., and Lin, H. (2019a). iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. *Mol. Ther. Nucleic Acids* 18, 269–274. doi: 10.1016/j.omtn.2019.08.022

Chen, W., Lv, H., Nie, F., and Lin, H. (2019b). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 35, 2796–2800. doi: 10.1093/bioinformatics/btz015

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Biomed Res. Int.* 2016:1654623. doi: 10.1155/2016/1654623

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

Cheng, J. H., Yang, H., Liu, M. L., Su, W., Feng, P. M., Ding, H., et al. (2018). Prediction of bacteriophage proteins located in the host cell using hybrid features. *Chemometr. Intell. Lab. Syst.* 180, 64–69. doi: 10.1016/j.chemolab.2018.07.006

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144. doi: 10.1093/nar/gky1051

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2018). Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics* 35, 2075–2083. doi: 10.1093/bioinformatics/bty943

Ding, H., and Li, D. M. (2015). Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids* 47, 329–333. doi: 10.1007/s00726-014-1862-4

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, H., Guo, S. H., Deng, E. Z., Yuan, L. F., Guo, F. B., Huang, J., et al. (2013). Prediction of Golgi-resident protein types by using feature selection technique. *Chemometr. Intell. Lab.* 124, 9–13. doi: 10.1016/j.chemolab.2013.03.005

Ding, H., Yang, W., Tang, H., Feng, P. M., Huang, J., Chen, W., et al. (2016). PHYPred: a tool for identifying bacteriophage enzymes and hydrolases. *Virol. Sin.* 31, 350–352. doi: 10.1007/s12250-016-3740-6

Ding, Y., Tang, J., and Guo, F. (2016a). Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information. *Int. J. Mol. Sci.* 17:1623. doi: 10.3390/ijms17101623

Ding, Y., Tang, J., and Guo, F. (2016b). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinformatics* 17:398.

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inform. Sci.* 41, 546–560. doi: 10.1016/j.ins.2017.08.045

Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

Dong, W., Han, S., Qu, X., Bao, W., Chen, Y., Fan, Y., et al. (2015). "A novel feature fusion method for predicting protein subcellular localization with multiple sites," in *Proceedings of the International Conference on Informative & Cybernetics for Computational Social Systems 2015*, (Piscataway, NJ: IEEE).

Fan, G.-L., and Li, Q.-Z. (2012). Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* 304, 88–95. doi: 10.1016/j.jtbi.2012.03.017

Feng, C. Q., Zhang, Z. Y., Zhu, X. J., Lin, Y., Chen, W., Tang, H., et al. (2018). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 35, 1469–1477. doi: 10.1093/bioinformatics/bty827

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.* 2013:530696. doi: 10.1155/2013/530696

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Hu, B., Zheng, L., Long, C., Song, M., Li, T., Yang, L., et al. (2019). EmExplorer: a database for exploring time activation of gene expression in mammalian embryos. *Open Biol.* 9:190054. doi: 10.1098/rsob.190054

Jiang, Q. H., Wang, G. H., Jin, S. L., Li, Y., and Wang, Y. D. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* 8, 282–293.

Keen, E. C. (2012). Phage therapy: concept to cure. *Front. Microbiol.* 3:238. doi: 10.3389/fmicb.2012.00238

Kimura, K., and Itoh, Y. (2003). Characterization of poly-gamma-glutamate hydrolase encoded by a bacteriophage genome: possible role in phage infection of *Bacillus subtilis* encapsulated with poly-gamma-glutamate. *Appl. Environ. Microbiol.* 69, 2491–2497. doi: 10.1128/aem.69.5.2491-2497.2003

Kovalenko, A. O., Chernyshov, S. V., Kutyshenko, V. P., Molochkov, N. V., and Mikoulinskaia, G. V. (2019). Investigation of the calcium-induced activation of the bacteriophage T5 peptidoglycan hydrolase promoting the host cell lysis. *Metallomics* 11, 799–809. doi: 10.1039/c9mt00020h

Liang, Z. Y., Lai, H. Y., Yang, H., Zhang, C. J., Yang, H., Wei, H. H., et al. (2017). Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics* 33, 467–469. doi: 10.1093/bioinformatics/btw630

Lin, H., and Li, Q. Z. (2011). Eukaryotic and prokaryotic promoter prediction using hybrid approach. *Theory Biosci.* 130, 91–100. doi: 10.1007/s12064-010-0114-8

Liu, B., Han, L., Liu, X., Wu, J., and Ma, Q. (2018). Computational prediction of sigma-54 promoters in bacterial genomes by integrating motif finding and machine learning strategies. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1211–1218. doi: 10.1109/tcbb.2018.2816032

Liu, D., Li, G., and Zuo, Y. (2019). Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Brief. Bioinform.* 20, 1826–1835. doi: 10.1093/bib/bby053

Lv, Z. B., Ao, C. Y., and Zou, Q. (2019). Protein function prediction: from traditional classifier to deep learning. *Proteomics* 19:2.

Ma, Q., Zhang, H., Mao, X., Zhou, C., Liu, B., Chen, X., et al. (2014). DMINDA: an integrated web server for DNA motif identification and analyses. *Nucleic Acids Res.* 42, W12–W19.

Manavalan, B., Basith, S., Shin, T. H., Lee, D. Y., Wei, L., and Lee, G. (2019a). 4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse genome. *Cells* 8:1332. doi: 10.3390/cells8111332

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019b). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35, 2757–2765. doi: 10.1093/bioinformatics/bty1047

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019c). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Manavalan, B., Shin, T. H., and Lee, G. (2018a). DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 9, 1944–1956. doi: 10.18632/oncotarget.23099

Manavalan, B., Shin, T. H., and Lee, G. (2018b). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

Manavalan, B., Subramaniyam, S., Shin, T. H., Kim, M. O., and Lee, G. (2018c). Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.* 17, 2715–2726. doi: 10.1021/acs.jproteome.8b00148

Parfitt, T. (2005). Georgia: an unlikely stronghold for bacteriophage therapy. *Lancet* 365, 2166–2167. doi: 10.1016/s0140-6736(05)66759-1

Parmar, K. M., Dafale, N. A., Tikariha, H., and Purohit, H. J. (2018). Genomic characterization of key bacteriophages to formulate the potential biocontrol agent to combat enteric pathogenic bacteria. *Arch. Microbiol.* 200, 1–12. doi: 10.1007/s00203-017-1471-1

Peng, L., Peng, M. M., Liao, B., Huang, G. H., Li, W. B., and Xie, D. F. (2018). The advances and challenges of deep learning application in biological big data processing. *Curr. Bioinform.* 13, 352–359. doi: 10.2174/1574893612666170707095707

Pirisi, A. (2000). Phage therapy–advantages over antibiotics? *Lancet* 356:1418. doi: 10.1016/s0140-6736(05)74059-9

Qu, K., Wei, L., and Zou, Q. (2019). A review of DNA-binding proteins prediction methods. *Curr. Bioinform.* 14, 246–254. doi: 10.2174/1574893614666181212102030

Riede, I., Schwarz, H., and Jahnig, F. (1987). Predicted structure of tail-fiber proteins of T-even type phages. *FEBS Lett.* 215, 145–150. doi: 10.1016/0014-5793(87)80130-8

Rodriguez-Rubio, L., Quiles-Puchalt, N., Martinez, B., Rodriguez, A., Penades, J. R., and Garcia, P. (2013). The peptidoglycan hydrolase of Staphylococcus aureus bacteriophage 11 plays a structural role in the viral particle. *Appl. Environ. Microbiol.* 79, 6187–6190. doi: 10.1128/AEM.01388-13

Seguritan, V., Alves, N. Jr., Arnoult, M., Raymond, A., Lorimer, D., Burgin, A. B. Jr., et al. (2012). Artificial neural networks trained to detect viral and phage structural proteins. *PLoS Comput. Biol.* 8:e1002657. doi: 10.1371/journal.pcbi.1002657

Shen, C., Jiang, L., Ding, Y., Tang, J., and Guo, F. (2019). LPI-KTASLP: prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* 7, 13486–13496. doi: 10.1109/access.2019.2894225

Song, J. N., Tan, H., Shen, H. B., Mahmood, K., Boyd, S. E., Webb, G. I., et al. (2010). Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* 26, 752–760. doi: 10.1093/bioinformatics/btq043

Song, J., Li, F., Leier, A., Marquez-Lago, T. T., Akutsu, T., Haffari, G., et al. (2018). PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* 34, 684–687. doi: 10.1093/bioinformatics/btx670

Song, L., Li, D., Zeng, X., Wu, Y., Guo, L., and Zou, Q. (2014). nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinformatics* 15:298. doi: 10.1186/1471-2105-15-298

Swami, A., and Jain, R. (2012). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480.

Tang, H., Su, Z. D., Wei, H. H., Chen, W., and Lin, H. (2016). Prediction of cell-penetrating peptides with feature selection techniques. *Biochem. Biophys. Res. Commun.* 477, 150–154. doi: 10.1016/j.bbrc.2016.06.035

Tang, H., Zhang, C. M., Chen, R., Huang, P., Duan, C. G., and Zou, P. (2017). Identification of secretory proteins of malaria parasite by feature selection technique. *Lett. Org. Chem.* 14, 621–624.

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Thiel, K. (2004). Old dogma, new tricks–21st century phage therapy. *Nat. Biotechnol.* 22, 31–36. doi: 10.1038/nbt0104-31

Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151. doi: 10.1093/nar/gkx1096

Wang, G., Wang, Y., Feng, W., Wang, X., Yang, J. Y., Zhao, Y., et al. (2008). Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genomics* 9(Suppl. 2):S22. doi: 10.1186/1471-2164-9-S2-S22

Wang, G., Wang, Y., Teng, M., Zhang, D., Li, L., and Liu, Y. (2010). Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon gamma-stimulated HeLa cells. *PLoS One* 5:e11794. doi: 10.1371/journal.pone.0011794

Wang, Y., Shi, F. Q., Cao, L. Y., Dey, N., Wu, Q., Ashour, A. S., et al. (2019). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinform.* 14, 282–294. doi: 10.2174/1574893614666190304125221

Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018a). Prediction of human protein subcellular localization using deep learning. *J. Parall. Distrib. Comput.* 117, 212–217.

Wei, L., Su, R., Wang, B., Li, X., Zou, Q., and Gao, X. (2019). Integration of deep feature representations and handcrafted features to improve the prediction of

N 6-methyladenosine sites. *Neurocomputing* 324, 3–9. doi: 10.1016/j.neucom.2018.04.082

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018b). ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016.

White, S. W. (1987). Prediction of DNA-binding regulatory proteins in bacteriophage T7. *Protein Eng.* 1, 373–376. doi: 10.1093/protein/1.5.373

Xianfang, W., Hongfei, L., Peng, G., Yifeng, L., and Wenjing, Z. (2019). Combining support vector machine with dual g-gap dipeptides to discriminate between acidic and alkaline enzymes. *Lett. Org. Chem.* 16, 325–331. doi: 10.2174/1570178615666180925125912

Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens. *J. Comput. Biol.* 25, 1266–1277. doi: 10.1089/cmb.2018.0004

Yang, H., Tang, H., Chen, X. X., Zhang, C. J., Zhu, P. P., Ding, H., et al. (2016). Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition. *Biomed Res. Int.* 2016:5413903. doi: 10.1155/2016/5413903

Yang, J., Chen, X., McDermaid, A., and Ma, Q. (2017). DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics* 33, 2586–2588. doi: 10.1093/bioinformatics/btx223

Yang, W., Zhu, X. J., Huang, J., Ding, H., and Lin, H. (2019). A brief survey of machine learning methods in protein sub-Golgi localization. *Curr. Bioinform.* 14, 234–240. doi: 10.2174/1574893613666181113131415

Yu, L., Sun, X., Tian, S. W., Shi, X. Y., and Yan, Y. L. (2018). Drug and nondrug classification based on deep learning with various feature selection strategies. *Curr. Bioinform.* 13, 253–259. doi: 10.2174/1574893612666170125124538

Zhang, J., and Liu, B. (2019). A review on the recent developments of sequence-based protein feature extraction methods. *Curr. Bioinform.* 14, 190–199. doi: 10.2174/1574893614666181212102749

Zhang, N., Sa, Y., Guo, Y., Lin, W., Wang, P., and Feng, Y. M. (2018). Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinform.* 13, 50–56. doi: 10.2174/1574893611666160608102537

Zhao, X., Pei, Z., Liu, J., Qin, S., and Cai, L. (2010). Prediction of nucleosome DNA formation potential and nucleosome positioning using increment of diversity combined with quadratic discriminant analysis. *Chromosome Res.* 18, 777–785. doi: 10.1007/s10577-010-9160-9

Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA promoter identification in arabidopsis using multiple histone markers. *Biomed. Res. Int.* 2015:861402. doi: 10.1155/2015/861402

Zhao, Y., Wang, F., Chen, S., Wan, J., and Wang, G. (2017). Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network. *Biomed. Res. Int.* 2017:7049406. doi: 10.1155/2017/7049406

Zheng, L., Huang, S., Mu, N., Zhang, H., Zhang, J., Chang, Y., et al. (2019). RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. *Database* 2019:190054.

Zhu, X. J., Feng, C. Q., Lai, H. Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016a). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* 10:114. doi: 10.1186/s12918-016-0353-5

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016b). 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

Zuo, Y. C., and Li, Q. Z. (2009). Using reduced amino acid composition to predict defensin family and subfamily: integrating similarity measure and structural alphabet. *Peptides* 30, 1788–1793. doi: 10.1016/j.peptides.2009.06.032

Zuo, Y. C., Peng, Y., Liu, L., Chen, W., Yang, L., and Fan, G. L. (2014). Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns. *Anal. Biochem.* 458, 14–19. doi: 10.1016/j.ab.2014.04.032

Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* 33, 122–124. doi: 10.1093/bioinformatics/btw564

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.