Chemometrics and Intelligent Laboratory Systems 180 (2018) 64–69

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Prediction of bacteriophage proteins located in the host cell using hybrid features

Jing-Hui Cheng [a], Hui Yang [a], Meng-Lu Liu [a], Wei Su [a], Peng-Mian Feng [b], Hui Ding [a,**], Wei Chen [a,c,***], Hao Lin [a,*]

[a] Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China

[b] Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, School of Public Health, North China University of Science and Technology, Tangshan, China

[c] Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, 063000, China

ARTICLE INFO

Keywords:
Phage proteins
Host cell
Subcellular location
Feature selection
Webserver

ABSTRACT

The identification of bacteriophage proteins in the host subcellular localization could provide important clues for understanding the interaction between phage and host bacteria as well as antibacterial drug design. To date, computational methods have been reported to identify bacteriophage proteins located in the host cell. However, there is still space for improving the prediction accuracy. The existing methods considering the sequence order correlation and the physicochemical property of protein provide us insights to construct an integrated descriptor based on sequence for phage proteins. Meanwhile, we proposed a feature selection technique to obtain the optimal features. In the jackknife test, the prediction accuracies are 86.7% and 97.9%, respectively for discrimination between PH proteins and non-PH proteins as well as PHM proteins and PHC proteins. Based on our model, we updated the web server PHPred to version 2.0 which can be freely accessed from http://lin-group.cn/server/PHPred2.0.

1. Introduction

Genome duplication is the most fundamental and orchestrated step. A bacteriophage, i.e. phage, is the virus that infects and proliferates within a bacterium. It can also kill host bacterium. In recent years, as more and more bacteria display the multi-drug-resistance, phages can be used as antibacterial agents [1].

Like other viruses, bacteriophage is parasitic to the host cell by injecting viral genetic materials (RNA or DNA) into the bacterial cell [2]. Based on the physiological process in the infected bacteria, there are two types of phages: temperate phage and intemperate phage. The former integrates its DNA (RNA) to the chromosome of host cell to replicate prophages, which is called lysogenic cycle. The later can produce daughter phages by controlling the expression system of bacterium and kill the host to infect other bacteria, which is called lytic cycle [3]. But the temperate phage could turn to lytic cycle induced by physicochemical and biological factors [4].

Phage proteins located in the host cell (PH proteins) play a key role in physiological processes. Thus, it is important to identify whether a phage protein locates in host bacterial cellular or not. In facts, the subcellular location of PH proteins in host cell often correlates with its special function. Specifically, phage proteins located in the host cell membrane (PHM proteins) may be the enzymes of lysis, such as hydrolases and lyases [5], which is pivotal for daughter phage to depart from the host bacterium [6]. And phage proteins located in the host cell cytoplasm (PHC proteins) may be the capsulate proteins [7] or the regulators [8] of the gene expression. Therefore, it is necessary to identify the subcellular location of PH proteins in host bacterial cell.

Previous researchers have successfully developed many computational methods dealing with phages and phage proteins, such as identifying the prophages [9], classifying the viral structural proteins [10], predicting the phage virion proteins [11,12]. The research about PH

* Corresponding author.

** Corresponding author.

*** Corresponding author. Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China.

E-mail addresses: hding@uestc.edu.cn (H. Ding), chenweiimu@gmail.com (W. Chen), hlin@uestc.edu.cn (H. Lin).

https://doi.org/10.1016/j.chemolab.2018.07.006
Received 8 June 2018; Received in revised form 7 July 2018; Accepted 13 July 2018
Available online 17 July 2018
0169-7439/© 2018 Elsevier B.V. All rights reserved.

proteins was first developed by Ding et al. [13], in which they proposed a *g*-gap dipeptide composition descriptor and obtained an encouraging result [13]. Later on, Shatabda et al. proposed a new descriptor based on the structural and evolutionary information [14]. Although high accuracies were obtained, the evaluated results were not objective because of independent structural and evolutionary information. Thus, the correct way to design a powerful predictor is only based on sequence. However, to the best of our knowledge, no such descriptor based on sequence information can reach a wonderful prediction result for identifying PH proteins.

In this paper, we introduced an integrated descriptor based on the sequence composition and the basic property of amino acid to identifying PH proteins and their locations in host cell. The feature selection technique was used to obtain the optimal features. For the convenience of experimental scientists, an online web server called PHPred 2.0 was developed according to the proposed method.

2. Materials and methods

This work comprises four major steps: (i) constructing the benchmark dataset, (ii) formulating protein samples with feature extraction methods, (iii) selecting and obtaining optimal features, (iv) constructing and evaluating the model. The workflow diagram for constructing the prediction model can be found in Fig. 1.

2.1. Benchmark dataset

Ding's dataset [13], which could be obtained from http://lin-group.cn/server/PHPr/data, was used in this work. According to the description in Ding et al.'s work [13], the phage proteins in the benchmark dataset was extracted from the UniProt [15] database according to the following steps:

Firstly, only phage proteins whose subcellular locations are experimentally confirmed were selected. Secondly, only phage proteins which are not the fragments of other proteins were selected. Thirdly, only phage proteins whose sequences do not contain nonstandard letters ('B', 'U', 'X' or 'Z') were selected. Finally, phage proteins with sequence identity greater than 0.3 were removed by using the software CD-HIT [16]. After performing these rules, they obtained 278 phage proteins, of which 144 were located in host cell, 134 were not located in host cell. Based on these proteins, a benchmark dataset $\mathbb{S}$ is formulated as:

$$\mathbb{S} = \mathbb{S}_{\mathrm{PH}} \cup \mathbb{S}_{\mathrm{non-PH}} \tag{1}$$

where $\mathbb{S}_{\mathrm{PH}}$ contains 144 proteins located in host cell (PH proteins), $\mathbb{S}_{\mathrm{non-PH}}$ contains 134 proteins that do not locate in host cell (non-PH proteins). The PH proteins can be further classified into two classes, i.e. the phage proteins that located in membrane of host cell (PHM proteins) and the phage proteins that located in host cell cytoplasm (PHC proteins), which can be described as:

$$\mathbb{S}_{\mathrm{PH}} = \mathbb{S}_{\mathrm{PHM}} \cup \mathbb{S}_{\mathrm{PHC}} \tag{2}$$

where $\mathbb{S}_{\mathrm{PHM}}$ contains the 68 PHM proteins and $\mathbb{S}_{\mathrm{PHC}}$ contains the 76 PHC proteins, respectively.

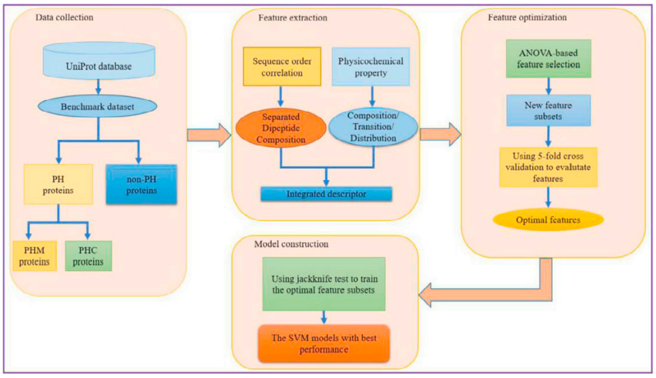

Data collection
UniProt database
Benchmark dataset
PH proteins
non-PH proteins
PHM proteins
PHC proteins
Feature extraction
Sequence order correlation
Physicochemical property
Separated Dipeptide Composition
Composition/ Transition/ Distribution
Integrated descriptor
Feature optimization
ANOVA-based feature selection
New feature subsets
Using 5-fold cross validation to evalutate features
Optimal features
Model construction
Using jackknife test to train the optimal feature subsets
The SVM models with best performance

Fig. 1. The workflow of this work.

2.2. Feature vector construction

After constructing the objective and strict benchmark dataset, we should formulate each protein sample with a mathematical descriptor. However, the lengths of proteins are different. Therefore, it's necessary to convert them to vectors that can be handled by the existing machine-learning algorithms. In fact, many efficient descriptors have been proposed and applied for this aim, such as the amino acid composition (AAC) [17] and dipeptide composition (DC) [18]. To consider both sequence order correlation and amino acid composition of the protein, Chou proposed a pseudo amino acid composition (PseAAC) [19] to formulate proteins. Here, we applied three kinds of higher dimensional descriptors described as follows.

(I) *g*-gap dipeptide composition (*g*-gap DC)

Suppose a protein sequence **P** with the length of L, denoted as follows:

$$P = R_1R_2R_3R_4\cdots R_i\cdots R_{L-1}R_L \tag{3}$$

where R_i means the i-th residue of the protein **P**.

In order to contain the long-range correlation information of residues, the interval of *g*-gap residues extended from dipeptide composition [11] was used in work. Then, the protein **P** can be expressed as:

$$P = [f_1^g, f_2^g, \cdots, f_i^g, \cdots, f_{400}^g]^{\mathrm{T}} \tag{4}$$

where f_i^g is the normalized frequency of the *i*-th ($i = 1, 2, \cdots, 400$) *g*-gap dipeptide [13] and is calculated by

$$f_i^g = \frac{n_i^g}{\sum_{k=1}^{400} n_k^g} = \frac{n_i^g}{L-g-1} \tag{5}$$

where n_i^g means the occurrence number of the *i*-th *g*-gap dipeptide, L denotes the length of protein **P**.

(II) Separated dipeptide compositions (SDC)

With the avalanche of protein sequences generated in the post-genomic era, the lengths of protein sequences vary widely. To extract important information from one protein, some researchers have spited a sequence to different fragments [20,21], which could highlight the properties of head or tail or special part of protein. Considering the lengths of proteins in the benchmark dataset are from 32 to 1825 residues, we segmented each protein sequence into two parts: the first 30 residues and the rest part. Then we calculated the *g*-gap dipeptide composition for each part based on Eq. (4) and Eq. (5). Finally, the SDC was obtained by combining the *g*-gap dipeptide composition of the two parts. Thus, the protein P can be expressed as a 800-D vector.

(III) composition/Transition/distribution (CTD)

Although SDCs contain much more sequence order correlation, the physicochemical properties are still lost [22]. The previous researchers have successfully developed some reasonable approaches [19,23] to extract the physicochemical property. Here, we chose the lower dimensional but more integrative physicochemical property descriptor—CTD (Composition/Transition/Distribution) to encode protein sequences.

CTD was first proposed to predict the protein folding class by Dubchak et al. [24] and has also been used to predict other protein cellular attributes. This paper also used the CTD to formulate protein samples. In the CTD feature, C represents the global composition of the given property in a protein sequence, T denotes the frequencies of the property changed along the protein sequence, and D is the distribution pattern for each class of the property, respectively. The details about how to calculate CTD features can be referred from Refs. [25,26].

2.3. Support vector machine (SVM)

SVM is a widely-used machine learning algorithm for classification and regression [27]. It is also popular in bioinformatics because of its remarkable performance [28,29]. Here, we used a free package LibSVM [30] which can be freely downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm to train and test our model. SVM provides four types of kernel functions: linear function, polynomial function, radial basis function (RBF) and sigmoid function. We chose the RBF as kernel function because it is suitable for nonlinear classification. The kernel parameter γ and regularization parameter C were optimized by the grid search with the searching space $[2^{-11}, 2^{1}]$ for γ and $[2^{-2}, 2^{9}]$ for C.

2.4. Feature selection technique

The features of the input data will limit the ceiling of improvement for the same classification algorithm. To avoid the noise-increasing from enlarging the dimension of the descriptor, it is necessary to select optimal features which could produce the best accuracy. Some efficient feature selection strategies such as Analysis Of Variance (ANOVA) [11], Maximum-Relevance-Maximum-Distance (MRMD) [31], and minimal Redundancy Maximal Relevance (mRMR) [32] have been applied in protein prediction [22,33–36]. In this work, the ANOVA was used to select optimal features because of its speed. The detail statistical explanation for ANOVA has been reported in previous work [11]. Here, we only provide its brief description.

ANOVA can measure the difference of each feature between positive and negative data according to the score F which is defined as:

$$F(i) = \frac{n_+\left(\overline{f}_i^{(+)} - \overline{f}_i\right)^2 + n_-\left(\overline{f}_i^{(-)} - \overline{f}_i\right)^2}{\frac{1}{n_+ + n_- - 2}\left[\sum_{k=1}^{n_+}\left(f_{i,k}^{(+)} - \overline{f}_i^{(+)}\right)^2 + \sum_{k=1}^{n_-}\left(f_{i,k}^{(-)} - \overline{f}_i^{(-)}\right)^2\right]} \tag{6}$$

where n_+ and n_- are, respectively the number of positive samples and negative samples, $\overline{f}_i$, $\overline{f}_i^{(+)}$, $\overline{f}_i^{(-)}$ is the mean frequencies of the i-th feature in the whole, positive and negative datasets, respectively. $f_{i,k}^{(+)}$ or $f_{i,k}^{(-)}$ is the frequency of the i-th feature of the k-th sample in the positive or negative datasets. The numerator and denominator represent the sample variance between groups and the sample variance within groups, respectively.

It is obvious that the larger the F-score is, the more discriminate power the feature has. Therefore, by using incremental feature selection (IFS) [37–39], the best feature subset can be obtained.

2.5. Evaluation metrics

The prediction quality of a classifier should be estimated by the cross-validation ordinarily. Three cross-validation methods including independent dataset test, subsampling test and jackknife test (leave-one-out cross-validation) are often used in protein prediction [13,40–43]. Among them, jackknife test could always yield a unique result for a given benchmark dataset. Hence, the jackknife test has been widely adopted to assess the prediction quality of various predictors [44–46]. Accordingly, we also used jackknife test to evaluate the prediction capability of our method. Considering the time-consuming property of the jackknife test, we first used the 5-fold cross-validation with grid search to figure out the optimal parameters γ and C, then we executed jackknife test to perform final evaluation.

The sensitivity (Sn), specificity (Sp) and overall accuracy (Acc) were used to evaluate the performance of the proposed method, and their definitions are described as follows [44–46]:

$$\begin{cases} Sn = 1 - \dfrac{\mathrm{N}_-^+}{\mathrm{N}^+} & 0 \le Sn \le 1 \\[2ex] Sp = 1 - \dfrac{\mathrm{N}_+^-}{\mathrm{N}^-} & 0 \le Sp \le 1 \\[2ex] Acc = 1 - \dfrac{\mathrm{N}_-^+ + \mathrm{N}_+^-}{\mathrm{N}^+ + \mathrm{N}^-} & 0 \le Acc \le 1 \end{cases} \tag{7}$$

where N^+ and N^- are the total numbers of the positive and negative samples, respectively, N_-^+ and N_+^- respectively denote the number of positive samples incorrectly predicted as negative samples (false negatives) and the number of negative samples incorrectly predicted as positive samples (false positives).

Furthermore, we introduced the Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC) to display the performance of our predictive model. The vertical coordinate of ROC curve indicates the true positive rate (sensitivity) and the horizontal coordinate indicates the false positive rate (1-specificity). AUC is an indicator of the performance quality, a value of 0.5 is equivalent to random prediction and a value of 1 represents a perfect prediction.

3. Results

At first, we should determine the parameter g in g-gap DC and SDC. Although long-range correlation between two residues also plays important role in protein structure and function, the short-range correlation still dominates in information storage. Thus, we utilized $g = 0$ in the two kinds of features. Subsequently, we used ANOVA combined with IFS technique to optimize the two feature sets. Results were recorded in Table 1. From the table, one may notice that the performance by using optimal SDC is superior to that by using optimal DC.

Subsequently, we examined the performance of CTD. Here, the five kinds of physicochemical properties that are Hydrophobicity, Hydrophilicity, pk1, pk2, and PI [22] were used to build the CTD descriptor, respectively. The ANOVA combined with IFS was used to find out the best features which can produce maximum prediction accuracy. Results were listed in Table 1. Although the accuracies of CTD were also lower than those of DC and SDC, the numbers of optimal features of CTD were dramatically less than those of DC and SDC.

Furthermore, we investigated the performance of different combinations of these features. Because SDC usually contains the information of DC, we just performed the following two feature integrations. The first is to combine SDC with five properties-based CTDs. Then the ANOVA-based feature selection technique was utilized to pick out the best feature subset. Table 1 showed that overall accuracies of 83.5% and 95.1% were obtained respectively for PH vs non-PH proteins classification and PHM vs PHC proteins classification, which were lower than those obtained by optimal SDCs, suggesting that noise or redundant information prevents from the improvement of prediction accuracy. The

Table 1
The performance by using different descriptors in jackknife cross-validation.

Descriptor Name	PH proteins vs non-PH proteins		PHM proteins vs PHC proteins	
	Acc(%)	Feature number	Acc(%)	Feature number
Optimal DCs	79.9	104	88.2	100
Optimal SDCs	86.0	131	97.2	97
Optimal (five properties-based CTDs) features	75.2	22	86.8	81
Optimal (SDCs + five properties-based CTDs) features	83.5	279	95.1	130
Optimal (SDCs + single property-based CTDs) features	86.7	144	97.9	122

second is to use ANOVA with IFS to optimize the five combinations between SDC and five kinds of CTDs respectively. By comparing these accuracies, we found that the combination of SDC and PI-based CTD can produce the best accuracy (86.7%) for discriminating PH proteins from non-PH proteins. For distinguishing between PHM proteins and PHC proteins, the maximum accuracy (97.9%) was achieved by combining SDC with hydrophobicity-based CTD. Meanwhile, the feature dimensions were not increased dramatically, suggesting that the model is reliable.

To further demonstrate the performance of our proposed method, we compared it with the previous pioneering work [13]. The performances of different methods were listed in Table 2.

As shown in Table 2, based on the same dataset, our method has improved the accuracies from 84.2% to 86.7% for the classification between PH proteins and non-PH proteins, and from 92.4% to 97.9% for discriminating PHM proteins from PHC proteins, respectively. Especially, in terms of classifying PH proteins and non-PH proteins, although the Sn obtained in this paper (82.6%) was lower than previous published results, the Sp was dramatically improved from 83.6% to 91.0%. These results demonstrated that the method proposed in this paper is powerful. In addition, ROC curves for the two classifications were plotted in Fig. 2 and the AUCs are 0.988 and 0.911, respectively. These results demonstrate again that our predictive model is powerful. The *p*-value was also provided to investigate the statistical significance of AUC between two prediction methods according to Hanley et al. [47].

Establish a webserver [48–55] or database [56–63] could provide conveniences to scientific community. Thus, based on the method proposed in this work, the freely web-server PHPred was updated to version 2.0. The new webserver PHPred2.0 can be freely accessed at http://lin-group.cn/server/PHPred2.0.

In order to improve the prediction efficiency of the webserver and provide the convenience to users, a brief guidance was provided as follows. Users can access the homepage of the webserver at the website (http://lin.uestc.edu.cn/server/PHPred2.0) as shown in Fig. 3. After clicking the **Web server** button, users can see a submitted page. Users can click the **Download** button to obtain the benchmark datasets. The **Help** page provides a detailed guidance for users. The **Citation** page provides the relevant papers. Users could find the research interesting and email from **About us** and **Contact** pages.

4. Discussions

The aim of this work is to extract enough information from phage protein sequence to develop a smart model for classifying bacteriophage proteins. In fact, the feature is main difference between this work and the previous work [13]. Different from the previous work [13] which used g-gap dipeptide composition, this work used a hybrid feature-based method by combining separated dipeptide composition with CTD. From the results in Table 1, it is easy to find that optimal SDCs could produce quite encouraging accuracies, but it cannot burden enough information to represent protein samples because the sequence composition cannot totally replace the physicochemical properties of protein to quantitatively characterize a protein [19]. Thus, it is possible to further improve the accuracy by including more information.

From above results, we should note that not all feature integrations can make positive effect. Let's use our examination as an example that when five properties-based CTDs were combined with SDCs, the jackknife cross-validated accuracies were reduced. In other words, noise or redundant information was included in model when we enlarged the dimension of descriptor to describe the protein comprehensively. Generally, high dimension features contain more information to describe samples. However, using more features in model does not always produce better results because the high-dimensional features may cause three problems: over-fitting, information redundancy and dimension disaster [37,39,64].

Table 2
The comparison with the previous work.

Classification	Ref.	Acc(%)	Sn(%)	Sp(%)	AUC	*P*-value (AUC)
PH proteins vs non-PH proteins	[13]	84.2	84.7	83.6	0.872	0.081
	This paper	86.7	82.6	91.0	0.911	
PHM proteins vs PHC proteins	[13]	92.4	89.7	94.7	0.970	0.155
	This paper	97.9	95.6	100.0	0.988	

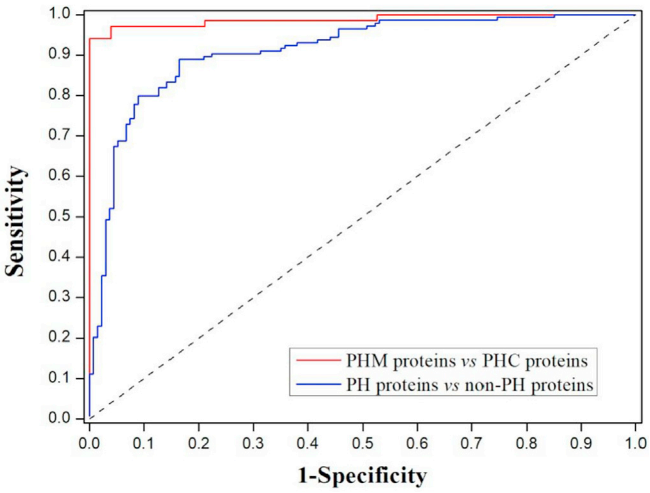
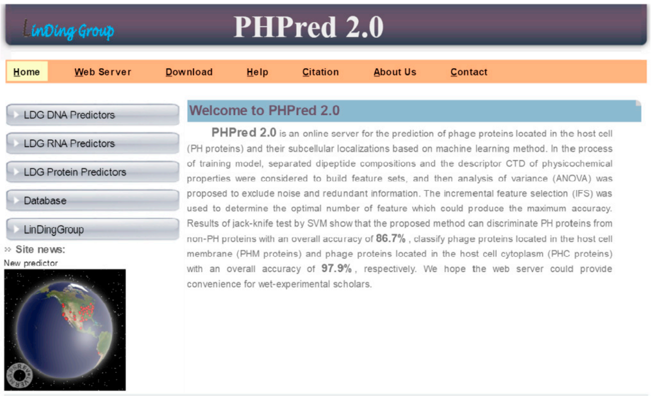


Fig. 2. The ROC curves for two predictions.



Welcome to PHPred 2.0

PHPred 2.0 is an online server for the prediction of phage proteins located in the host cell (PH proteins) and their subcellular localizations based on machine learning method. In the process of training model, separated dipeptide compositions and the descriptor CTD of physicochemical properties were considered to build feature sets, and then analysis of variance (ANOVA) was proposed to exclude noise and redundant information. The incremental feature selection (IFS) was used to determine the optimal number of feature which could produce the maximum accuracy. Results of jack-knife test by SVM show that the proposed method can discriminate PH proteins from non-PH proteins with an overall accuracy of **86.7%** , classify phage proteins located in the host cell membrane (PHM proteins) and phage proteins located in the host cell cytoplasm (PHC proteins) with an overall accuracy of **97.9%** , respectively. We hope the web server could provide convenience for wet-experimental scholars.

Fig. 3. A semi-screen shot for the top-page of the PHPred2.0 web server.

Thus, a reasonable strategy is to select the optimal features. Theoretically, the maximum accuracy could be found by examining all possible feature combinations. However, it is not possible to perform these examinations because more than $\binom{a}{1} + \binom{a}{2} + \cdots + \binom{a}{a} = 2^a - 1$ feature combinations for an a-dimension feature vector should be examined. Thus, we proposed using ANOVA-based feature selection technique, and finally total of 144 and 122 best features were obtained for the two classifications.

Recently, high accuracies were obtained by using evolutionary information and structure information [14,65]. However, it is unfair to compare our results reported in this paper with the published results because evolution information was generated by searching a non-redundant (NR) database using PSI-Blast. Generally, the NR database contains the phage protein samples, which results in a lack of objectivity of cross-validation and then overestimation of the method.

5. Conclusions

Bacteriophage proteins play a key role in the proliferation of phage and the death of bacterium. Thus, designing computational method to identify phage functional proteins will provide more clues to design anti-bacterial drugs. In this paper, based on protein sequence, an integrated descriptor composed of SDCs and CTD was proposed by progressively approximating the information of protein. By using ANOVA, we captured the best features to build our predictors. Promising results were obtained in cross-validation. To make it more convenient for the experimental researchers to study phage function-unknown proteins, a user friendly web server PHPred2.0 was established and can be freely used. We hope the tool will provide novel insights into the study of phage-related problems. The rapid application of deep learning in bioinformatics [66–68] provide us a chance to develop a more powerful tool for phage protein prediction.

Conflicts of interest

The authors declare that there is no conflict of interests.

Author contributions

Hui Ding, Wei Chen and Hao Lin designed the experiments. Jing-Hui Chen, Meng-Lu Liu and Wei Su performed most of the computational tests. Hui Yang established webserver. Jing-Hui Chen, Peng-Mian Feng, Hui Ding, Wei Chen and Hao Lin wrote the paper.

Acknowledgments

This work was supported by the National Nature Scientific Foundation of China (61772119, 31771471), the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015Z006, ZYGX2016J125, ZYGX2016J118, ZYGX2016J223), Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244), the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028).

References

[1] E.C. Keen, Phage therapy: concept to cure, Front. Microbiol. 3 (2012) 238.

[2] J.T. Chang, M.F. Schmid, C. Haase-Pettingell, P.R. Weigele, J.A. King, W. Chiu, Visualizing the structural changes of bacteriophage Epsilon15 and its Salmonella host during infection, J. Mol. Biol. 402 (2010) 731–740.

[3] H. Ding, L.-F. Luo, H. Lin, Entropy production rate changes in lysogeny/lysis switch regulation of bacteriophage lambda, Commun. Theor. Phys. 55 (2011) 371.

[4] G. Ofir, R. Sorek, Contemporary phage biology: from classic models to new insights, Cell 172 (2018) 1260–1270.

[5] H. Ding, W. Yang, H. Tang, P.M. Feng, J. Huang, W. Chen, H. Lin, PHYPred: a tool for identifying bacteriophage enzymes and hydrolases, Virol. Sin. 31 (2016) 350–352.

[6] A. Leo-Macias, G. Katz, H. Wei, A. Alimova, A. Katz, W.J. Rice, R. Diaz-Avalos, G.-B. Hu, D.L. Stokes, P. Gottlieb, Toroidal surface complexes of bacteriophage ϕ12 are responsible for host-cell attachment, Virology 414 (2011) 103–109.

[7] D.A. Marvin, M.F. Symmons, S.K. Straus, Structure and assembly of filamentous bacteriophages, Prog. Biophys. Mol. Biol. 114 (2014) 80–122.

[8] B. Nejman-Falenczyk, S. Bloch, K. Licznerska, A. Felczykowska, A. Dydecka, A. Wegrzyn, G. Wegrzyn, Small regulatory RNAs in lambdoid bacteriophages and phage-derived plasmids: not only antisense, Plasmid 78 (2015) 71–78.

[9] Y. Zhou, Y. Liang, K.H. Lynch, J.J. Dennis, D.S. Wishart, PHAST: a fast phage search tool, Nucleic Acids Res. 39 (2011) W347–W352.

[10] V. Seguritan, N. Alves Jr., M. Arnoult, A. Raymond, D. Lorimer, A.B. Burgin Jr., P. Salamon, A.M. Segall, Artificial neural networks trained to detect viral and phage structural proteins, PLoS Comput. Biol. 8 (2012), e1002657.

[11] H. Ding, P.M. Feng, W. Chen, H. Lin, Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis, Mol. Biosyst. 10 (2014) 2229–2235.

[12] P.M. Feng, H. Ding, W. Chen, H. Lin, Naive Bayes classifier with feature selection to identify phage virion proteins, Comput. Math. Meth. Med. 2013 (2013) 530696.

[13] H. Ding, Z.Y. Liang, F.B. Guo, J. Huang, W. Chen, H. Lin, Predicting bacteriophage proteins located in host cell with feature selection technique, Comput. Biol. Med. 71 (2016) 156–161.

[14] S. Shatabda, S. Saha, A. Sharma, A. Dehzangi, iPHLoc-ES: identification of bacteriophage protein locations using evolutionary and structural features, J. Theor. Biol. 435 (2017) 229–237.

[15] U. Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212.

[16] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[17] H. Lin, W. Chen, Prediction of thermophilic proteins using feature selection technique, J. Microbiol. Meth. 84 (2011) 67–70.

[18] H. Tang, Z.D. Su, H.H. Wei, W. Chen, H. Lin, Prediction of cell-penetrating peptides with feature selection techniques, Biochem. Biophys. Res. Commun. 477 (2016) 150–154.

[19] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (2001) 246–255.

[20] Q. Xiang, B. Liao, X. Li, H. Xu, J. Chen, Z. Shi, Q. Dai, Y. Yao, Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine, Artif. Intell. Med. 78 (2017) 41–46.

[21] W.C. Wong, S. Maurer-Stroh, B. Eisenhaber, F. Eisenhaber, On the necessity of dissecting sequence similarity scores into segment-specific contributions for inferring protein homology, function prediction and annotation, BMC Bioinf. 15 (2014) 166.

[22] H. Tang, W. Chen, H. Lin, Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique, Mol. Biosyst. 12 (2016) 1269–1275.

[23] R. Sharma, A. Dehzangi, J. Lyons, K. Paliwal, T. Tsunoda, A. Sharma, Predict Gram-Positive, Gram-Negative Subcellular, Localization via incorporating evolutionary information and physicochemical features into Chou's general PseAAC, IEEE Trans. NanoBioscience 14 (2015) 915–926.

[24] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, Prediction of protein folding class using global description of amino acid sequence, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 8700–8704.

[25] V. Saravanan, P.T. Lakshmi, APSLAP: an adaptive boosting technique for predicting subcellular localization of apoptosis protein, Acta Biotheor. 61 (2013) 481–497.

[26] Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu, Z. Lin, An approach for identifying cytokines based on a novel ensemble classifier, BioMed Res. Int. 2013 (2013) 686090.

[27] B.E. Boser, I. Guyon, V. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the 5th Annual Workshop on Computational Learning Theory, ACM Press, 1992, pp. 144–152.

[28] Y.W. Zhao, Z.D. Su, W. Yang, H. Lin, W. Chen, H. Tang, IonchanPred 2.0: a tool to predict ion channels and their types, Int. J. Mol. Sci. 18 (2017) 1838.

[29] D. Li, Y. Ju, Q. Zou, Protein folds prediction with hierarchical structured SVM, Curr. Proteonomics 13 (2016) 79–85.

[30] C.C. Chang, C.J. Lin, in: Libsvm, ACM Transactions on Intelligent Systems and Technology, vol. 2, 2011, pp. 1–27.

[31] Q. Zou, J. Zeng, L. Cao, R. Ji, A novel features ranking metric with application to scalable visual and bioinformatics data classification, Neurocomputing 173 (2016) 346–354.

[32] H.C. Peng, F.H. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.

[33] H. Ding, S.H. Guo, E.Z. Deng, L.F. Yuan, F.B. Guo, J. Huang, N. Rao, W. Chen, H. Lin, Prediction of Golgi-resident protein types by using feature selection technique, Chemometr. Intell. Lab. Syst. 124 (2013) 9–13.

[34] X.X. Chen, H. Tang, W.C. Li, H. Wu, W. Chen, H. Ding, H. Lin, Identification of bacterial cell wall lyases via pseudo amino acid composition, BioMed Res. Int. 2016 (2016) 1654623.

[35] H. Yang, H. Tang, X.X. Chen, C.J. Zhang, P.P. Zhu, H. Ding, W. Chen, H. Lin, Identification of secretory proteins in Mycobacterium tuberculosis using pseudo amino acid composition, BioMed Res. Int. 2016 (2016) 5413903.

[36] Q. Zou, S. Wan, Y. Ju, J. Tang, X. Zeng, Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy, BMC Syst. Biol. 10 (2016) 114.

[37] H. Yang, W.R. Qiu, G.Q. Liu, F.B. Guo, W. Chen, K.C. Chou, H. Lin, iRSpot-Pse6NC: identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC, Int. J. Biol. Sci. 14 (2018) 883–891.

[38] H. Tang, Y.W. Zhao, P. Zou, C.M. Zhang, R. Chen, P. Huang, H. Lin, HBPred: a tool to identify growth hormone-binding proteins, Int. J. Biol. Sci. 14 (2018) 957–964.

[39] Z.D. Su, Y. Huang, Z.Y. Zhang, Y.W. Zhao, D. Wang, W. Chen, K.C. Chou, H. Lin, iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty508.

[40] P.M. Feng, W. Chen, H. Lin, K.C. Chou, iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, Anal. Biochem. 442 (2013) 118–125.

[41] W. Chen, P. Xing, Q. Zou, Detecting N(6)-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines, Sci. Rep. 7 (2017) 40242.

[42] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest, Front. Pharmacol. 9 (2018) 276.

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

[44] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, K.C. Chou, iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites, Mol. Ther. Nucleic Acids 11 (2018) 468–474.

[45] W. Chen, H. Yang, P. Feng, H. Ding, H. Lin, iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, Bioinformatics 33 (2017) 3518–3523.

[46] P.M. Feng, H. Lin, W. Chen, Identification of antioxidants from sequence information using naive Bayes, Comput. Math. Meth. Med. 2013 (2013) 567529.

[47] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.

[48] B. Manavalan, T.H. Shin, G. Lee, Pvp-svm: sequence-based prediction of phage virion proteins using a support vector machine, Front. Microbiol. 9 (2018) 476.

[49] J. Kang, Y. Fang, P. Yao, N. Li, Q. Tang, J. Huang, NeuroPP: a tool for the prediction of neuropeptide precursors based on optimal sequence composition, Interdiscipl. Sci. Comput. Life Sci. (2018), https://doi.org/10.1007/s12539-018-0287-2.

[50] B. Manavalan, S. Basith, T.H. Shin, S. Choi, M.O. Kim, G. Lee, MLACP: machine-learning-based prediction of anticancer peptides, Oncotarget 8 (2017) 77121–77136.

[51] N. Li, J. Kang, L. Jiang, B. He, H. Lin, J. Huang, PSBinder: a web service for predicting polystyrene surface-binding peptides, BioMed Res. Int. 2017 (2017) 5761517.

[52] F.Y. Dao, H. Yang, Z.D. Su, W. Yang, Y. Wu, D. Hui, W. Chen, H. Tang, H. Lin, Recent advances in conotoxin classification by using machine learning methods, Molecules 22 (2017) 1057.

[53] B. He, J. Kang, B. Ru, H. Ding, P. Zhou, J. Huang, SABinder: a web service for predicting streptavidin-binding peptides, BioMed Res. Int. 2016 (2016) 9175143.

[54] B. Manavalan, J. Lee, SVMQA: support-vector-machine-based protein single-model quality assessment, Bioinformatics 33 (2017) 2496–2503.

[55] B. Manavalan, T.H. Shin, G. Lee, DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest, Oncotarget 9 (2018) 1944–1956.

[56] B. He, L. Jiang, Y. Duan, G. Chai, Y. Fang, J. Kang, M. Yu, N. Li, Z. Tang, P. Yao, P. Wu, R. Derda, J. Huang, Biopanning data bank 2018: hugging next generation phage display, Database: J. Biol. Databases Curation 2018 (2018), https://doi.org/10.1093/database/bay032.

[57] T. Zhang, P. Tan, L. Wang, N. Jin, Y. Li, L. Zhang, H. Yang, Z. Hu, L. Zhang, C. Hu, C. Li, K. Qian, C. Zhang, Y. Huang, K. Li, H. Lin, D. Wang, RNALocate: a resource for RNA subcellular localizations, Nucleic Acids Res. 45 (2017) D135–D138.

[58] Z.Y. Liang, H.Y. Lai, H. Yang, C.J. Zhang, H. Yang, H.H. Wei, X.X. Chen, Y.W. Zhao, Z.D. Su, W.C. Li, E.Z. Deng, H. Tang, W. Chen, H. Lin, Pro54DB: a database for experimentally verified sigma-54 promoters, Bioinformatics 33 (2017) 467–469.

[59] P. Feng, H. Ding, H. Lin, W. Chen, AOD: the antioxidant protein database, Sci. Rep. 7 (2017) 7449.

[60] B. He, G. Chai, Y. Duan, Z. Yan, L. Qiu, H. Zhang, Z. Liu, Q. He, K. Han, B. Ru, F.B. Guo, H. Ding, H. Lin, X. Wang, N. Rao, P. Zhou, J. Huang, BDB: biopanning data bank, Nucleic Acids Res. 44 (2016) D1127–D1132.

[61] J. Huang, B. Ru, P. Zhu, F. Nie, J. Yang, X. Wang, P. Dai, H. Lin, F.B. Guo, N. Rao, MimoDB 2.0: a mimotope database and beyond, Nucleic Acids Res. 40 (2012) D271–D277.

[62] Y. Yi, Y. Zhao, C. Li, L. Zhang, H. Huang, Y. Li, L. Liu, P. Hou, T. Cui, P. Tan, Y. Hu, T. Zhang, Y. Huang, X. Li, J. Yu, D. Wang, RAID v2.0: an updated resource of RNA-associated interactions across organisms, Nucleic Acids Res. 45 (2017) D115–D118.

[63] T. Cui, L. Zhang, Y. Huang, Y. Yi, P. Tan, Y. Zhao, Y. Hu, L. Xu, E. Li, D. Wang, MNDR v2.0: an updated resource of ncRNA-disease associations in mammals, Nucleic Acids Res. 46 (2018) D371–D374.

[64] H.Y. Lai, X.X. Chen, W. Chen, H. Tang, H. Lin, Sequence-based predictive modeling to identify cancerlectins, Oncotarget 8 (2017) 28169–28175.

[65] L. Wei, M. Liao, X. Gao, Q. Zou, An improved protein structural prediction method by incorporating both sequence and structure information, IEEE Trans. NanoBioscience 14 (2015) 339–349.

[66] R. Cao, C. Freitas, L. Chan, M. Sun, H. Jiang, Z. Chen, ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network, Molecules 22 (2017) 1732.

[67] R.Z. Cao, D. Bhattacharya, J. Hou, J.L. Cheng, DeepQA: improving the estimation of single protein model quality with deep belief networks, BMC Bioinf. 17 (2016) 495.

[68] L. Wei, Y. Ding, R. Su, J. Tang, Q. Zou, Prediction of human protein subcellular localization using deep learning, J. Parallel Distr. Comput. 117 (2018) 212–217.