Computers in Biology and Medicine 71 (2016) 156–161

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# Predicting bacteriophage proteins located in host cell with feature selection technique

Hui Ding [a,*], Zhi-Yong Liang [a], Feng-Biao Guo [a], Jian Huang [a], Wei Chen [a,b,*], Hao Lin [a,**]

[a] *Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology and Center for Information in Biomedicine, University of Electronic Science and Technology of China, Chengdu 610054, China*

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

ARTICLE INFO

*Article history:*
Received 16 December 2015
Accepted 18 February 2016

*Keywords:*
Bacteriophage proteins
Analysis of variance
g-gap dipeptide
Feature analysis

ABSTRACT

A bacteriophage is a virus that can infect a bacterium. The fate of an infected bacterium is determined by the bacteriophage proteins located in the host cell. Thus, reliably identifying bacteriophage proteins located in the host cell is extremely important to understand their functions and discover potential anti-bacterial drugs. Thus, in this paper, a computational method was developed to recognize bacteriophage proteins located in host cells based only on their amino acid sequences. The analysis of variance (ANOVA) combined with incremental feature selection (IFS) was proposed to optimize the feature set. Using a jackknife cross-validation, our method can discriminate between bacteriophage proteins located in a host cell and the bacteriophage proteins not located in a host cell with a maximum overall accuracy of 84.2%, and can further classify bacteriophage proteins located in host cell cytoplasm and in host cell membranes with a maximum overall accuracy of 92.4%. To enhance the value of the practical applications of the method, we built a web server called PHPred (〈http://lin.uestc.edu.cn/server/PHPred〉). We believe that the PHPred will become a powerful tool to study bacteriophage proteins located in host cells and to guide related drug discovery.

## 1. Introduction

Bacteriophages (phages) are viruses that can attack and kill bacteria. Thus, phages have become important potential resources of the development of anti-bacterial drugs [1].

In the infection, bacteriophage initially attaches tightly to the bacterial surface via a specific receptor [2]. Subsequently, the genetic material of the phage is injected into the bacterial cell [3]. According to the type of phage, one of two life cycles (the lysis or lysogeny) will occur after infection [4, 5]. In the lytic cycle, the phage will produce daughter phage nucleic acids and proteins by using the bacteria's genetic mechanism. The phage proteins produced within the bacterial cell will make the cell wall to lyse, further releasing the offspring phages to infect other bacteria. During the lysogenic cycle, the phage DNA is integrated with the bacterial chromosome to create the prophage. The prophage will replicate along with the reproduction of bacterial host. The host cell is not destroyed. Thus, the offspring bacterial cells also contain the prophage which has capability to produce new phages. However, some adverse conditions such as UV or mutagenic chemicals can trigger the termination of the lysogenic state by changing the concentration of phage proteins [6]. The infected system will switch from lysogeny to lysis. Thus, the phage proteins located in host cell play a key role in destroying host cell membrane and killing bacteria [7].

In fact, the phage proteins in host cell (abbreviated as phage host proteins or PH proteins) are mainly distributed in two subcellular locations of host. The first location is host cell membrane, in which the phage proteins are hydrolases or lyases and in charge of destroying host cell membrane [8,9]. The second location is host cell cytoplasm, in which the phage proteins are used to regulate the transcription of phage genes [10], assemble protease and procapsid [11], and mediate ssDNA packaging into virion [12].

Accurate identification of the PH proteins and their host subcellular locations are of great importance for the exploration of the mechanism of host cell lysis and the development of potential antibacterial drugs. However, due to the limited experimental

* Corresponding authors.

** Corresponding author at: Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology and Center for Information in Biomedicine, University of Electronic Science and Technology of China, Chengdu 610054, China.

*E-mail addresses:* hding@uestc.edu.cn (H. Ding), greatchen@heuu.edu.cn (W. Chen), hlin@uestc.edu.cn (H. Lin).

http://dx.doi.org/10.1016/j.compbiomed.2016.02.012
0010-4825/© 2016 Elsevier Ltd. All rights reserved.

resource and expensive cost, it is not realistic to investigate all phage proteins. Computational methods pave a new way to study various biological problems. In fact, bioinformatics approaches are quite powerful and efficient in dealing with phage protein problems. Seguritan and Alves initially developed a computational method to classify viral structural proteins [13]. Since phage virion proteins have important functions, some methods have been proposed to identify them by using residue sequence information [14,15]. A free web-server PVPred has been constructed for phage virion proteins prediction [14]. However, to the best of our knowledge, no computational system was developed for the prediction of PH proteins or their host subcellular locations.

This paper aims to develop a novel sequence-based method to identify PH proteins and their host subcellular locations.

According to a recent review [16] and previous results [17–25], in order to establish a useful statistical predictor for identifying PH proteins and their host subcellular locations, we constructed an objective and strict benchmark dataset to train and test the proposed model, proposed *g*-gap dipeptide composition which could truly reflect the intrinsic correlation between two amino acids to formulate the protein samples, developed the analysis of various (ANOVA) based technique to perform feature selection and adopted the support vector machine (SVM) to perform the prediction. The anticipated accuracy was evaluated by using jackknife cross-validation. In addition, for the convenience of most experimental scientists, an online web server called *PHPred* was constructed based on the proposed method.

## 2. Materials and methods

### 2.1. Benchmark dataset

To develop a sequence-based predictor for PH proteins, it is necessary to firstly construct a valid benchmark dataset to train and test the predictor. In this study, all data were collected from the Universal Protein Resource (Uniprot) [26]. To construct a high quality dataset, we firstly selected the phage proteins whose subcellular localizations had been experimentally confirmed. Secondly, the phage proteins which are the fragments of other phage proteins were dislodged because their information is incomplete. Thirdly, we excluded the proteins whose sequences contain nonstandard letters, such as ‘B’, ‘U’, ‘X’ or ‘Z’ because their meanings are ambiguous. Finally, the redundancy of the dataset was reduced by excluding those proteins with the sequence identity of less than 30% to any other protein in the dataset by using the CD-HIT software [27]. As a result, a total of 278 phage proteins with subcellular annotation were obtained and formulated by

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

where the $\mathbb{S}_{\mathrm{PH}}$ contains 144 phage proteins located in host cell (PH proteins) and the $\mathbb{S}_{\mathrm{non-PH}}$ contains 134 phage proteins which are not located in host cell (non-PH proteins). The PH proteins can be further classified into two types: the phage proteins located in host cell membrane (PHM proteins) and the phage proteins located in host cell cytoplasm (PHC proteins) and expressed as:

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

where the $\mathbb{S}_{\mathrm{PHM}}$ contains 68 PHM proteins and the $\mathbb{S}_{\mathrm{PHC}}$ contains 76 PHC proteins. The detailed sequences can be freely downloaded from the website (〈http://lin.uestc.edu.cn/server/PHP/data〉).

### 2.2. The g-gap dipeptide composition

After obtaining a standard dataset, we can formulate protein sequence samples with a mathematics descriptor. According to the most straightforward formulation method, a protein **P** can be expressed with the amino acid sequence as follows:

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\ldots\mathrm{R}_L \tag{3}$$

where $\mathrm{R}_1$, $\mathrm{R}_2$ and $\mathrm{R}_\mathrm{L}$, respectively denote the 1st, 2nd and *L*-th residues of the protein **P**. However, due to the length difference among protein sequences, it can only be used in some similarity searching programs such as BLAST and FASTA.

For most of machine learning methods, it is required that the samples are all denoted by the vector with the same dimension. Amino acid composition (AAC) including 20-D features is the first strategy to formulate protein samples [28]. However, the residue-order information of sequence is lost. To overcome this disadvantage, dipeptide composition was used because it could reflect the correlation between two adjoining residues [29,30]. However, the long-range correlation information was still not considered in dipeptide. The pseudo amino acid composition was proposed to improve the formulation of protein samples [31–33]. It not only included the AAC information, but also contained the correlation of physicochemical properties between two residues. However, the direct correlation between two residues with the interval of *g*-gap residues was omitted. In fact, it is possible that two amino acids are adjacent in three-dimensional space, but far from each other in primary sequences. For example, in alpha helix, two non-adjoining residues are connected by hydrogen bonds. Thus, to incorporate the correlation of more residues as possible, the *g*-gap dipeptide composition was proposed in this work to formulate PH protein samples. Then a protein **P** can be expressed as

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

where the $f_\varepsilon^g$ is the frequency of the $\varepsilon$-th ($\varepsilon$=1, 2, …, 400) *g*-gap dipeptide and calculated by

$$f_\varepsilon^g = \frac{n_\varepsilon^g}{\sum_{\varepsilon=1}^{400} n_\varepsilon^g} = \frac{n_\varepsilon^g}{L-g-1} \tag{5}$$

where $n_\varepsilon^g$ is the occurrence number of the $\varepsilon$-th *g*-gap dipeptide. *L* denotes the length of the protein **P**.

### 2.3. Support vector machine (SVM)

SVM was used as the classification algorithm in this work because of its powerful performance in the field of bioinformatics [18,21,29,34–38]. The software (LibSVM) (〈https://www.csie.ntu.edu.tw/∼cjlin/libsvm/〉) was selected to implement SVM. The radial basis kernel function (RBF) was selected as the kernel function. A grid search method was applied in the selection of the regularization parameter *C* and kernel parameter $\gamma$ through jackknife cross-validation. The search spaces for *C* and $\gamma$ are, respectively $[2^{15}, 2^{-5}]$ and $[2^{-5}, 2^{-15}]$ with the steps of $2^{-1}$ and 2.

### 2.4. Performance evaluation

Three testing methods including independent dataset test, subsampling test, and jackknife test, can be used to evaluate the performance of the proposed predictors [16,39]. The jackknife test has been widely adopted to evaluate the performance of their methods [34,35,40–46] as it can yield a unique result for a given benchmark dataset. Therefore, we also used the jackknife cross-validation test to estimate the anticipated success rates of our method. In order to reduce the computational time, the 5-fold cross-validation with a grid search was used to select the parameters $C$ and $\gamma$ in SVM.

The following set of three metrics was adopted [44]. The sensitivity (Sn), specificity (Sp) and overall accuracy (Acc) can be expressed as

$$\begin{cases} \mathrm{Sn} = 1 - \frac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \frac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \end{cases} \tag{6}$$

where $N^+$ and $N^-$ are, respectively the total numbers of the positive samples and negative samples. $N_-^+$ and $N_+^-$ are, respectively the number of positive samples incorrectly predicted as negative samples and the number of negative samples incorrectly predicted as positive samples.

The receiver operating characteristic (ROC) curves were also plotted to show the predictive capability of our method. The area under the receiver operating characteristic curve (auROC) was calculated to quantitatively and objectively measure the performance of the proposed method.

### 2.5. Analysis of variance (ANOVA)-based technique

To obtain a deeper insight into the intrinsic properties of PH protein sequences and improve the understandability, scalability and accuracy of the prediction model, the feature selection technique was utilized to optimize the features [47]. Here, the ANOVA was used to perform feature selection.

Based on the statistical theory [14,18,20], the ANOVA is used to test the difference between two or more means. In this study, we faced the binary classification problems. Thus, the difference of the $\varepsilon$-th $g$-gap dipeptide between two groups can be measured by the ANOVA score ($F$) which can be expressed as

$$F(\varepsilon) = \frac{m^{(+)} \cdot \left( \overline{f_\varepsilon^g}^{(+)} - \overline{f_\varepsilon^g} \right)^2 + m^{(-)} \cdot \left( \overline{f_\varepsilon^g}^{(-)} - \overline{f_\varepsilon^g} \right)^2}{\frac{1}{m^{(+)} + m^{(-)} - 2} \left[ \sum_{k=1}^{m^{(+)}} \left( f_{\varepsilon,k}^{g(+)} - \overline{f_\varepsilon^g}^{(+)} \right)^2 + \sum_{k=1}^{m^{(-)}} \left( f_{\varepsilon,k}^{g(-)} - \overline{f_\varepsilon^g}^{(-)} \right)^2 \right]} \tag{7}$$

where $m^{(+)}$ and $m^{(-)}$ are, respectively the total numbers of the positive samples and the negative samples. $\overline{f_\varepsilon^g}^{(+)}$, $\overline{f_\varepsilon^g}^{(-)}$ and $\overline{f_\varepsilon^g}$ are, respectively the mean values of the $\varepsilon$-th $g$-gap dipeptide in the entire positive samples, the entire negative samples and the total samples. $f_{\varepsilon,k}^{g(+)}$ and $f_{\varepsilon,k}^{g(-)}$, respectively represent the $\varepsilon$-th $g$-gap dipeptide of the $k$-th sample in the positive data set and the negative data set.

On the basis of the principle of ANOVA, the $F(\varepsilon)$ obeys the $F$ sampling distribution under the null hypothesis. Obviously, the$F(\varepsilon)$ indicates the degree that the $\varepsilon$-th $g$-gap dipeptide is correlated to the group variables. The stronger correlation degree between the $g$-gap dipeptide and the group variable means its larger contribution to the classification. Thus, a large $F(\varepsilon)$ means that the $\varepsilon$-th $g$-gap dipeptide contributes to the better prediction.

Based on the above analysis, we employed the incremental feature selection (IFS) [14] to determine the optimal number of features according to the following procedures. Firstly, all features were ranked according to their $F$ values. Then, the feature with the highest $F$ value was selected as the first feature subset and its prediction accuracy was measured by the SVM-based model. Thirdly, the second feature subset was produced by adding the second feature with the second highest $F$ value. The accuracy of this feature subset was also investigated. We repeated the above procedure until the accuracies of all candidate feature subsets were evaluated. To save the computational time, the 5-fold cross-validation test was proposed to examine the performance of each feature subset. If $g$ varied from 0 to $g_\alpha$, the accuracies of $(g_\alpha+1) \times 400$ feature subsets should be investigated. The best feature subset was composed of the features which could achieve the maximum accuracy. Thus, the final classifier model was built according to this feature subset.

## 3. Results

### 3.1. Discrimination between PH proteins and non-PH proteins

PH proteins are important in destroying host cell membrane and killing bacteria. Therefore, we initially performed the classification between 144 PH proteins and 134 non-PH proteins. According to the aforementioned $g$-gap dipeptide composition in (Eqs. (4)–5), for each gap $g$, we obtained a 400-dimension vector which was much larger than the number of samples ($144+134=278$ samples). The ANOVA-based technique was used to perform feature selection.

In general, the high-dimensional features cause two outcomes: the information redundancy and the over-fitting problem. The two outcomes result in the low prediction accuracy in cross-validation and the poor generalization ability of the proposed predictor. Taking the 400 0-gap dipeptides as an example, in jackknife cross-validation, the Acc is only 72.7%. On the contrary, if a model is built on a low-dimensional feature subset, the robustness of the model can be improved. However, the dimension of feature subset is so low that the features cannot afford enough information, thus resulting in the poor performance of the model. An evidence is that 10 0-gap dipeptides with high $F$ values can only generate the overall accuracy of 70.1% in jackknife cross-validation.

To find out the best feature subset which can produce the maximum accuracy, we plotted the IFS curves in Fig. 1 by selecting the feature number as $X$-coordinate and the overall accuracy as $Y$-coordinate. In this work, the gap $g$ varied from 0 to 9. Thus, there are 10 IFS curves in Fig. 1. The accuracies of $10 \times 400 = 4000$ feature subsets were investigated. As shown in Fig. 1, the maximum Acc reaches 84.2% when the feature subset contains 69 7-gap dipeptides. In total, 84.7% PH proteins and 83.6% non-PH proteins can be correctly identified in jackknife cross-validation. To further investigate the whole performance of the model, we plotted the ROC curve in Fig. 2. The auROC is 0.872.

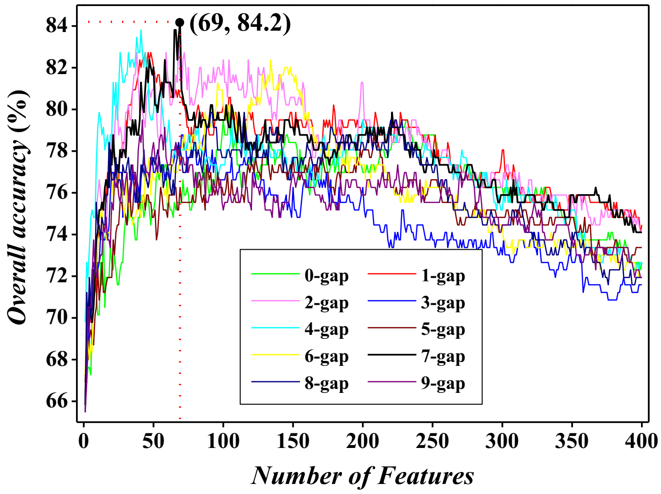


**Fig. 1.** A plot showing the IFS procedure for discriminating between PH proteins and non-PH proteins. When the top 69 7-gap dipeptides were used to perform prediction, the overall success rate reaches an IFS peak of 84.2%.

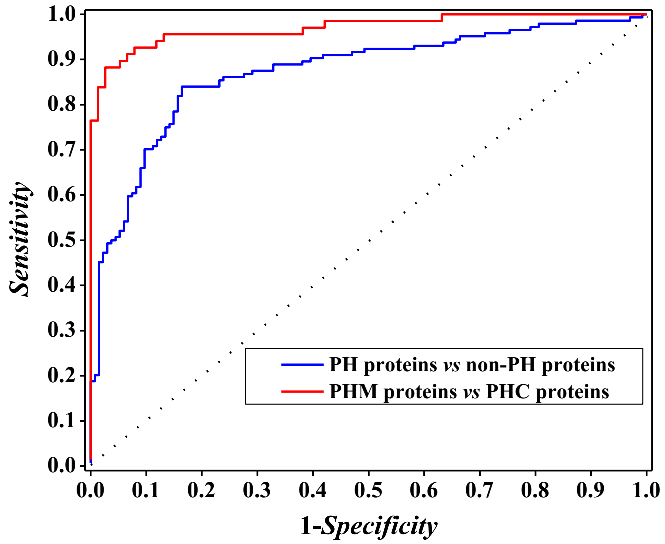


**Fig. 2.** The ROC curves for predicting PH proteins and non-PH proteins. The auROCs of 0.872 and 0.970 were obtained in jackknife cross-validation. The diagonal dot line denotes a random guess with the auROC of 0.5.

The number of features (69) is dramatically less than the number of samples (278), suggesting that the model is robust and reliable. The high accuracy proves that the proposed method is powerful and efficient.

### 3.2. Discrimination between PHM proteins and PHC proteins

PH proteins are distributed in the host cell membrane or the host cell cytoplasm. Thus, the section aims to construct a model to discriminate the 68 PHM proteins from 76 PHMC proteins.

To build a powerful model, we still used the ANOVA-based technique to optimize feature subset. The feature selection process was repeated as described above. As shown in the 10 IFS curves (Fig. 3 shows), when the top 49 3-gap dipeptides are used, the maximum Acc of 92.4% is achieved in jackknife cross-validation. At the same time, the best model can correctly identify 89.7% PHM proteins and 94.7% PHC proteins. The ROC curve for this prediction is also drawn in Fig. 2. The auROC is 0.970. Likewise, the model is reliable and robust as the number of features (49) is about one-third of the number of samples (144).

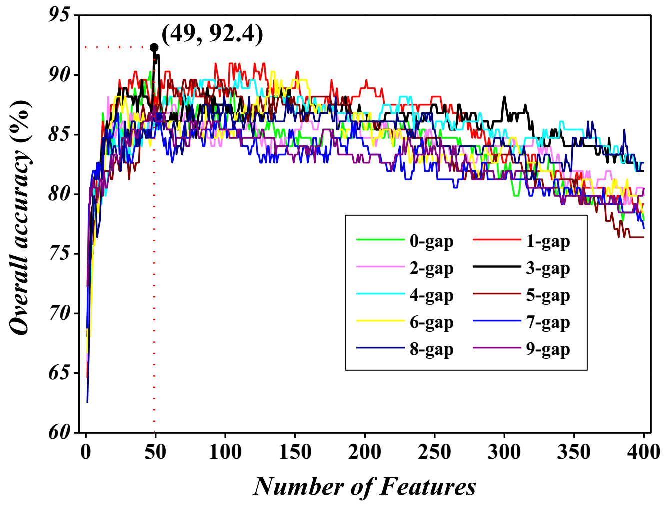
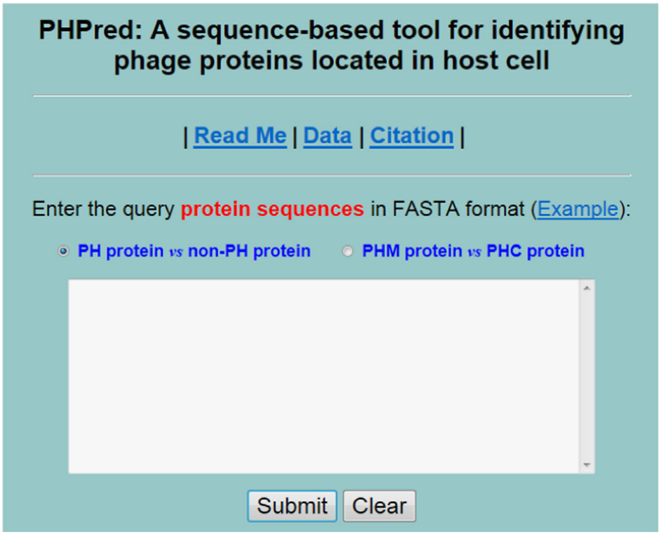


**Fig. 3.** A plot showing the IFS procedure for discriminating between PHM proteins and PHC proteins. When the top 49 3-gap dipeptides were used to perform prediction, the overall success rate reaches an IFS peak of 92.4%.



**PHPred: A sequence-based tool for identifying phage proteins located in host cell**

| Read Me | Data | Citation |

Enter the query protein sequences in FASTA format (Example):

● PH protein vs non-PH protein ○ PHM protein vs PHC protein

**Fig. 4.** A semi-screen shot to show the top page of the PHPred web-server. Its website address is 〈http://lin.uestc.edu.cn/server/PHPred〉.

### 3.3. Web-server construction

In order to improve the prediction efficiency of PH proteins, 2e constructed a web-server called *PHPred* via aforementioned procedures. The web interface for entry browse and submitted window are coded in PHP. The server is freely accessible at the website (〈http://lin.uestc.edu.cn/server/PHPred〉). For the convenience of users, the guidance was provided as follows.

Firstly, users can access the web server at the website (〈http://lin.uestc.edu.cn/server/PHPred〉) and will see the top page of PHPred on the computer screen, as shown in Fig. 4. After clicking the *Read Me* button, users can see a brief introduction about the predictor and the caveat. Users can click the *Data* button to download the benchmark datasets. The *Citation* button provides the relevant papers.

Secondly, either input or copy/paste the query phage peptide sequences into the input box at the center of Fig. 4. The input sequence should be in the FASTA format. To get the anticipated prediction results, the discrimination button ("PH protein *vs* non-PH protein" and "PHM protein *vs* PHC protein") should be checked. After clicking the "Submit" button, users may obtain the predicted results.

Thirdly, the webserver is just focused on bacteriophage proteins. Thus, it should be noted that it will give ridiculous results when submitting non-bacteriophage proteins.

## 4. Discussion

The study aims to build powerful models to predict PH proteins and their distribution in host cell. The above calculations have demonstrated that our proposed method is efficient. Here, we further discuss the prediction from the following aspects.

The first is the objectivity of benchmark dataset. It is well-known that the prediction results will be overestimated if the proposed method was trained and tested by a benchmark dataset with high homologous sequences [21,48,49]. Generally, if two protein sequences share the sequence similarity of 40% or above, they are homologs. The most predictors in previous studies [50,51] always achieved good accuracies when the model was trained and tested on a dataset with the sequence identity of 40% or above. However, those predictors could not provide satisfactory accuracies when the dataset with low sequence identity (30% or less) was used

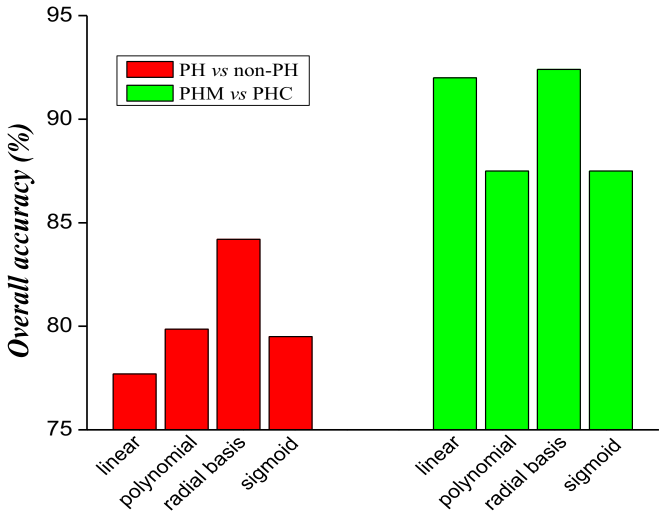


**Fig. 5.** The performance comparison of the four kernel functions for the two discriminations.

in training and test [52]. To reduce redundancy and avoid bias, we constructed an objective and strict benchmark dataset by excluding those proteins which had the sequence identity of 30% and below to any other protein in the same data subset. Such standard data would guarantee the reliability of the prediction model.

The second is the importance of feature selection in prediction. In the era of big data, it is inevitable to produce high-dimensional disaster. High-dimensional features will bring about three problems [14,17,22,53,54]: over-fitting, information redundancy and dimension disaster. Thus, feature selection techniques can not only overcome these disadvantages, but also pick out useful features for understanding the intrinsic properties of PH proteins. Therefore, the accuracies of proposed model can be improved.

Obtaining the best features is important for prediction. After examining the performance of all feature combinations, the best feature set will be found. However, we cannot complete it because the computation time is too long. For example, for the 400 dipeptides, the number of all possible combinations for 400-D vector is $C_{400}^{1}+C_{400}^{2}+\cdots+C_{400}^{399}+C_{400}^{400}>2.58\times10^{120}$. Thus, we used a feature selection technique (ANOVA) to obtain the best feature subsets and eventually improved the prediction quality.

The results showed that the optimal features did improve the prediction accuracy. The Accs reached 84.2% and 92.4% in jackknife cross-validation for two discriminations. Moreover, the feature selection process also provided us an important clue to study the correlation between two residues. For example, for the prediction of PH protein, before we performed feature selection, the 1-gap dipeptides can produce the maximum overall accuracy of 75.2%. After performing the feature selection, the two residues with the interval of 7 residues is the main correlation mode. In summary, the ANOVA-based technique not only improved the performance of predictor, but also provided a great deal of quantitative insight to residues correlation.

It is important to compare the proposed method with other published methods. However, it is not reality because no computational system was developed for the prediction of PH proteins or their host subcellular locations. Thus, we just investigated the Acc achieved by random guess [55]. If the weight or prior probability is considered, the Accs for discriminating PH proteins from non-PH proteins and PHM proteins from PHC proteins are, respectively only $[134\times(134/278)+144\times(144/278)]/278=50.1\%$ and $[68\times(68/144)+76\times(76/144)]/144=50.2\%$. Thus, the proposed prediction method has capability to perform the two predictions.

In fact, there are four kernel functions (linear function, polynomial function, radial basis function and sigmoid function) in LibSVM. Thus, we should also compare the performances of different kernel function (Fig. 5). From the figure, we noticed that the RBF function is more suitable for the two classifications.

## 5. Conclusion

More and more evidences indicate that the PH proteins are important sources for antibacterial drug development. Thus, a classification method was proposed to predict PH proteins and their distribution in host cell. The proposed feature selection algorithm could dramatically improve prediction performance. On the basis of this model, we built an online predictor *PHPred* for the convenience of the experimental scholars. The predictor will become a powerful tool for PH protein analysis and research. Moreover, the feature selection technique proposed in this study can be generalized to other fields of computational biology.

## Conflicts of interest statement

The authors declare no conflict of interest.

## Acknowledgments

This work was supported by the National Nature Scientific Foundation of China (Nos. 61301260 and 61202256), the Applied Basic Research Program of Sichuan Province (No. 2015JY0100), the China Scholarship Council (No. 201406075035) and the Fundamental Research Funds for the Central Universities (Nos. ZYGX2015J144, ZYGX2015Z006).

## References

[1] I. Sorokulova, E. Olsen, V. Vodyanoy, Expert Rev. Med. Devices 11 (2014) 175–186.
[2] J.T. Chang, M.F. Schmid, C. Haase-Pettingell, P.R. Weigele, J.A. King, W. Chiu, J. Mol. Biol. 402 (2010) 731–740.
[3] Y. Choi, H. Shin, J.H. Lee, S. Ryu, Appl. Environ. Microbiol. 79 (2013) 4829–4837.
[4] H. Ding, L.F. Luo, Chin. Phys. Lett. 26 (2009).
[5] H. Ding, L.F. Luo, H. Lin, Commun. Theory Phys. 55 (2011) 371–375.
[6] B.A. Duerkop, K.L. Palmer, M.J. Horsburgh, Enterococcal bacteriophages and genome defense, in: M.S. Gilmore, D.B. Clewell, Y. Ike, N. Shankar (Eds.), Enterococci: From Commensals to Leading Causes of Drug Resistant Infection, Boston, 2014.
[7] 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, Virology 414 (2011) 103–109.
[8] M. Rajaure, J. Berry, R. Kongari, J. Cahill, R. Young, Proceedings of the National Academy of Sciences of the United States of America, 112 (2015) 5497–5502.
[9] Q. Zou, X.B. Li, Y. Jiang, Y.M. Zhao, G.H. Wang, Curr. Proteom. 10 (2013) 2–9.
[10] A. Chakraborty, B.D. Paul, V. Nagaraja, Protein Eng. Des. Sel. 20 (2007) 1–5.
[11] E.M. Medina, B.T. Andrews, E. Nakatani, C.E. Catalano, J. Mol. Biol. 412 (2011) 723–736.
[12] K.L. Brentlinger, S. Hafenstein, C.R. Novak, B.A. Fane, R. Borgon, R. McKenna, M. Agbandje-McKenna, J. Bacteriol. 184 (2002) 1089–1094.
[13] V. Seguritan, N. Alves Jr., M. Arnoult, A. Raymond, D. Lorimer, A.B. Burgin Jr., P. Salamon, A.M. Segall, PLoS Comput. Biol. 8 (2012) e1002657.
[14] H. Ding, P.M. Feng, W. Chen, H. Lin, Mol. BioSyst. 10 (2014) 2229–2235.
[15] P.M. Feng, H. Ding, W. Chen, H. Lin, Computational and mathematical methods in medicine, , 2013 (2013) 530696.
[16] K.C. Chou, J. Theor. Biol. 273 (2011) 236–247.
[17] C. Ding, L.F. Yuan, S.H. Guo, H. Lin, W. Chen, J. Proteom. 77 (2012) 321–328.
[18] H. Ding, S.H. Guo, E.Z. Deng, L.F. Yuan, F.B. Guo, J. Huang, N.N. Rao, W. Chen, H. Lin, Chemom. Intell Lab 124 (2013) 9–13.
[19] W.C. Li, E.Z. Deng, H. Ding, W. Chen, H. Lin, Chemom. Intell. Lab 141 (2015) 100–106.
[20] H. Lin, W. Chen, H. Ding, PloS One 8 (2013) e75726.
[21] H. Lin, W. Chen, L.F. Yuan, Z.Q. Li, H. Ding, Acta Biotheor. 61 (2013) 259–268.
[22] P.P. Zhu, W.C. Li, Z.J. Zhong, E.Z. Deng, H. Ding, W. Chen, H. Lin, Mol. BioSyst. 11 (2015) 558–563.

[23] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, K.C. Chou, PloS One 9 (2014) e106691.
[24] B. Liu, D.Y. Zhang, R.F. Xu, J.H. Xu, X.L. Wang, Q.C. Chen, Q.W. Dong, K.C. Chou, Bioinformatics 30 (2014) 472–479.
[25] Z. Wang, Q. Zou, Y. Jiang, Y. Ju, X.X. Zeng, Curr. Bioinform. 9 (2014) 331–342.
[26] C. UniProt, Nucleic Acids Res. 43 (2015) D204–D212.
[27] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, Bioinformatics 28 (2012) 3150–3152.
[28] A. Cornish-Bowden, J. Theor. Biol. 76 (1979) 369–386.
[29] M. Bhasin, G.P. Raghava, Nucleic Acids Res. 32 (2004) W414–W419.
[30] B. Liu, F.L. Liu, X.L. Wang, J.J. Chen, L.Y. Fang, K.C. Chou, Nucleic Acids Res. 43 (2015) W65–W71.
[31] K.C. Chou, Proteins 43 (2001) 246–255.
[32] B. Liu, S.Y. Wang, X.L. Wang, Sci. Rep. UK 5 (2015).
[33] B. Liu, X.L. Wang, Q. Zou, Q.W. Dong, Q.C. Chen, Mol. Inform. 32 (2013) 775–782.
[34] H. Ding, L. Luo, H. Lin, Protein Pept. Lett. 16 (2009) 351–355.
[35] V. Tripathi, D.K. Gupta, J. Biomol. Struct. Dyn. 32 (2014) 1575–1582.
[36] R. Cao, Z. Wang, Y. Wang, J. Cheng, BMC Bioinform. 15 (2014) 120.
[37] B. Liu, J. Chen, X. Wang, Bioinformatics 31 (2015) 3492–3498.
[38] B. Liu, L. Fang, R. Long, X. Lan, K.C. Chou, Bioinformatics (2015).
[39] B. Liu, L.Y. Fang, F.L. Liu, X.L. Wang, J.J. Chen, K.C. Chou, PloS One 10 (2015).
[40] S. Nakariyakul, Z.P. Liu, L. Chen, Amino Acids 42 (2012) 1947–1953.
[41] B. Liu, L.Y. Fang, J.J. Chen, F.L. Liu, X.L. Wang, Mol. BioSyst. 11 (2015) 1194–1204.
[42] H. Ding, D.M. Li, Amino Acids 47 (2015) 329–333.
[43] S.-H Guo, E.-Z Deng, L.-Q Xu, H Ding, H Lin, W Chen, K.-C Chou, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (11) (2014) 1522–1529, http://dx.doi.org/10.1093/bioinformatics/btu083.
[44] H. Lin, E.Z. Deng, H. Ding, W. Chen, K.C. Chou, Nucleic Acids Res. 42 (2014) 12961–12972.
[45] H. Lin, H. Ding, F.B. Guo, A.Y. Zhang, J. Huang, Protein Pept. Lett. 15 (2008) 739–744.
[46] H. Lin, W.X. Liu, J. He, X.H. Liu, H. Ding, W. Chen, Sci. Rep. U.K. 5 (2015).
[47] B. Liu, J.J. Chen, X.L. Wang, Mol. Genet. Genom. 290 (2015) 1919–1931.
[48] H. Nielsen, S. Brunak, G. von Heijne, Protein Eng. 12 (1999) 3–9.
[49] E.C. Su, H.S. Chiu, A. Lo, J.K. Hwang, T.Y. Sung, W.L. Hsu, BMC Bioinform. 8 (2007) 330.
[50] H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, H. Lin, W. Chen, K.C. Chou, BioMed Res. Int. 2014 (2014) 286419.
[51] Y.C. Wang, X.B. Wang, Z.X. Yang, N.Y. Deng, Protein Pept. Lett. 17 (2010) 1441–1449.
[52] K. Chen, L.A. Kurgan, J. Ruan, J. Comput. Chem. 29 (2008) 1596–1604.
[53] B. Liu, L.Y. Fang, S.Y. Wang, X.L. Wang, H.T. Li, K.C. Chou, J. Theor. Biol. 385 (2015) 153–159.
[54] B. Liu, F.L. Liu, L.Y. Fang, X.L. Wang, K.C. Chou, Bioinformatics 31 (2015) 1307–1309.
[55] K.C. Chou, Mol. BioSyst. 9 (2013) 1092–1100.