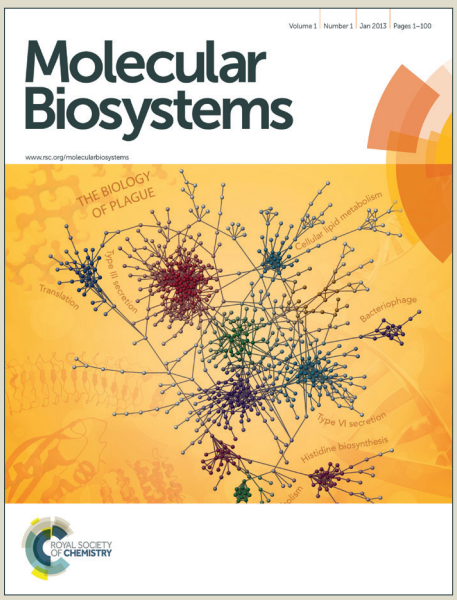
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Identification of bacteriophage virion proteins with the ANOVA feature selection and analysis

Hui Ding[1*], Peng-Mian Feng[2], Wei Chen[3*], Hao Lin[1*]

[1] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China;

[2] School of Public Health, Hebei United University, Tangshan 063000, China

[3] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China;

***Corresponding authors**

Hui Ling: hding@uestc.edu.cn

Wei Chen: greatchen@heuu.edu.cn

Hao Lin: hlin@uestc.edu.cn

Corresponding author's mail addresses

School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054 China;

Tel: +86-28-8320-2351

Fax: +86-28-8320-8238

Abstract:

The bacteriophage virion proteins play extremely important roles in the fate of host bacterial cells. Accurate identification of bacteriophage virion proteins is very important for understanding their functions and clarifying the lysis mechanism of bacterial cells. In this study, a new sequence-based method was developed to identify phage virion proteins. In the new method, the protein sequences were initially formulated by the *g*-gap dipeptide compositions. Subsequently, the analysis of variance (ANOVA) with incremental feature selection (IFS) was used to search for the optimal feature set. It was observed that, in jackknife cross-validation, the optimal feature set including 160 optimized features can produce the maximum accuracy of 85.02%. By performing feature analysis, we found that the correlation between two amino acids with one gap was more important than other correlations for phage virion protein prediction and that some of the 1-gap dipeptides were important and mainly contributed to the virion protein prediction. This analysis will provide novel insights into the function of phage virion proteins. On the basis of the proposed method, an online web-server, **PVPred**, was established and can be freely accessed from the website (http://lin.uestc.edu.cn/server/PVPred). We believe that the PVPred will become a powerful tool to study phage virion proteins and to guide the related experimental validations.

Keywords: phage virion protein; feature selection; g-gap dipeptide; feature analysis

1. Introduction

The bacteriophage, also called phage, is a group of viruses that can infect bacteria and be replicated in bacteria. They have been the essential tools in the development of bacterial genetics for the construction of genetic manipulation tools [1]. Phage virion is a complete fully infectious extracellular phage virus particle which mainly consists of two or three parts: genetic material, a protein coat and an envelope for some phages [2]. The genetic material is either DNA or RNA and is protected by the protein coat. In some case, the lipids envelope surrounding the protein coat can protect the phages when they are outside the cells. After the virion binds to the surface of a specific host bacteria cell, its DNA or RNA is injected into the host cell. In the lytic state, the virions are replicated with eventual break of the host cell. Subsequently, the offspring virions spread and infect other host cells.

Phage proteins are of great significance to understand the mechanism of interaction between the phage and its host bacteria and to develop new antibacterial drugs. Due to the relatively limited experimental data, it is difficult to determine the function of phage virion proteins from sequence information [3]. Machine learning approaches have been proved to be quite powerful and efficient in dealing with various biological problems. Thus, it may be feasible to predict the functions of phage proteins with machine learning approaches.

Actually, Segall et al. [3] have developed an Artificial Neural Network (ANN)-based method to classify viral structural proteins by using amino acid composition and protein isoelectric points. Recently, our group proposed a Naïve Bayes classifier with feature selection to identify phage virion proteins by using primary sequence information [4]. Although the aforementioned methods could yield encouraging results, the accuracies of these methods were still far from satisfactory. Furthermore, no web-server was provided for these methods. Hence, their usage is quite limited, particularly for the broad experimental scientists.

The present study was devoted to enhance the prediction performance and quality for identifying the phage virion proteins. Firstly, we introduced a universal *g*-gap dipeptide composition to formulate the protein samples. Secondly, a powerful feature selection technique, analysis of various (ANOVA), was proposed to optimize the features. Finally, the support vector machine (SVM) was used to perform virion protein prediction. The jackknife cross-validation was performed to objectively evaluate the anticipated accuracy of the predictor. Prediction results demonstrate that

the proposed method is reliable. To deeply understand the composition of phage virion proteins, the feature analysis was performed. For the convenience of most experimental scientists, a user-friendly web server was constructed based on the proposed method.

2. Materials and methods

2.1. Benchmark Dataset

The original positive and negative dataset used in this study was obtained from the Universal Protein Resource (Uniprot) [5]. The following steps were performed to guarantee a high quality dataset. Firstly, the phage proteins whose subcellular location is virion were regarded as positive samples (virion proteins), and the phage proteins whose subcellular location is not virion were considered as negative samples (non-virion proteins). Secondly, the protein sequences which are fragments of other proteins were dislodged. Thirdly, if the protein sequences contain nonstandard letters, such as 'B', 'U', 'X' or 'Z', these proteins were excluded because their meanings are ambiguous. As a result, a total of 121 phage virion and 231 phage non-virion proteins were obtained.

Generally, if a predictor is trained and tested by a benchmark dataset with high homologous sequences, misleading results with overestimated accuracy will be obtained [6, 7]. To get rid of redundancy and avoid bias, the CD-HIT software [8] was used. Thus, a cutoff threshold of 40% was imposed to exclude those proteins that are no less than 40% sequence identity to any other in a same subset [9]. Finally, we obtained a strict and objective benchmark dataset as formulated by

$$S = S_{\text{virion}} \bigcup S_{\text{non-virion}} \quad (1)$$

where the S_{virion} contains 99 phage virion proteins and the $S_{\text{non-virion}}$ contains 208 phage non-virion proteins. The detailed sequences can be freely downloaded from http://lin.uestc.edu.cn/server/PVP/data.

2.2. The g-gap dipeptide composition

In the development of a sequence-based predictor of phage virion proteins, it is important to formulate its sequence with an effective mathematical expression that can truly reflect the correlation between the intrinsic features of the sequence and the

protein types to be predicted. The most straightforward formulation method is to formulate the sample of protein **P** with *L* residues with its entire amino acid sequence as:

$$\mathbf{P}=R_1R_2R_3R_4\ldots R_L \quad (2)$$

where R_1 represents the 1st residue of the protein **P**, R_2 represents the 2nd residue of the protein **P**, and so forth.

Another common strategy is to formulate protein sequences with amino acid composition (AAC). To obtain the sequence-order information, the simple AAC was replaced by the adjoining dipeptide composition to represent the sample of a protein [4, 10, 11]. However, the adjoining dipeptide composition can only reflect the correlation between two adjoining amino acids. In fact, two amino acids with the interval of *g*-gap residues are maybe adjacent in three dimension space. Especially, in some regular secondary structures, such as alpha helices and beta-structural form, two non-adjoining residues are connected by hydrogen bonds. To search for the important correlation, we proposed a universal dipeptide composition, namely, *g*-gap dipeptide composition which was extended from the adjoining dipeptide composition [12, 13]. Thus, protein **P** can be formulated by

$$\mathbf{P}=[f_1^g, f_2^g, \cdots, f_\xi^g \cdots f_{400}^g]^{\mathrm{T}} \quad (3)$$

where the f_ξ^g is the frequency of the ξ-th (ξ=1, 2, …, 400) *g*-gap dipeptide and calculated by

$$f_\xi^g = n_\xi^g / \sum_{\xi=1}^{400} n_\xi^g = n_\xi^g / (L-g-1) \quad (4)$$

where n_ξ^g denotes the occurrence number of the ξ-th *g*-gap dipeptide and *L* is the length of the protein **P**. *g*=0 is the adjoining dipeptide and indicates the correlation of two proximate residues; *g* =1 describes the correlation between two residues with one residue interval; *g* =2 indicates the correlation between two residues with the interval of two residues, and so forth.

2.3. The analysis of variance (ANOVA)

For economizing run-time and computational resource, a wise strategy is to use algorithm to find the optimal features and eventually improve the prediction quality. In the present study, we performed feature selection through the analysis of variance (ANOVA). ANOVA is a very simple and powerful method to test the difference in

means between groups. Due to the following advantages, the ANOVA has been used for feature selection [6, 13, 14]. Firstly, it is robust to most violations of its assumptions. Secondly, it is more intuitive for us to analyze the interaction of the two variables. Thirdly, it can be used effectively even when the number of observations is different in each group. Finally, it is easily generalized to more than two groups without increasing the Type 1 error.

According to the principle of ANOVA [6, 13, 14], the score (F) of the ξ-th g-gap dipeptide in benchmark dataset can be defined by

$$F(\xi) = \frac{s_B^2(\xi)}{s_W^2(\xi)} \tag{5}$$

where $s_B^2(\lambda)$ and $s_W^2(\lambda)$ denote the sample variance between groups (also called Means Square Between, MSB) and sample variance within groups (also called Mean Square Within, MSW), respectively, and are calculated by

$$s_B^2(\xi) = \sum_{i=1}^{K} m_i \left(\frac{\sum_{j=1}^{m_i} f_\xi^g(i,j)}{m_i} - \frac{\sum_{i=1}^{K}\sum_{j=1}^{m_i} f_\xi^g(i,j)}{\sum_{i=1}^{K} m_i} \right)^2 / df_B \tag{6}$$

$$s_W^2(\xi) = \sum_{i=1}^{K}\sum_{j=1}^{m_i} \left(f_\xi^g(i,j) - \frac{\sum_{j=1}^{m_i} f_\xi^g(i,j)}{m_i} \right)^2 / df_W \tag{7}$$

where $df_B = K - 1$ and $df_W = M - K$ are the degrees of freedom for MSB and MSW, respectively. K and M represent the number of groups (here K=2) and total number of samples (here M=307), respectively. $f_\xi^g(i,j)$ denotes the frequency of the ξ-th g-gap dipeptide of the j-th sample in the i-th group; m_i denotes the number of samples in the i-th group (here m_1=99, m_2=208).

Obviously, a large value of $F(\xi)$ means that the ξ-th feature has a better discriminative capability. Hence, all features can be ranked according to their F values. Subsequently, the incremental feature selection (IFS) [15, 16] was used to determine the optimal number of features as described below. Firstly, the feature subset started from a feature with the highest F value in the ranked feature set. Secondly, a new feature subset was produced when the feature with the second highest F value was added. This process was repeated from the higher F to the lower F value until all candidate features were added. Thus, for any gap g, the 400 feature subsets will be produced. The ε-th feature subset is composed of ε ranked g-gap dipeptides and can be

expressed as:

$$\mathbf{P}_{\varepsilon}^{g} = \left[f_1^g, f_2^g, \cdots, f_{\varepsilon}^g \right]^{\mathrm{T}} \qquad 1 \le \varepsilon \le 400 \quad , \quad g \ge 0 \tag{8}$$

For each of the 400 feature sets, the prediction accuracy of the proposed method was examined on the benchmark dataset by using jackknife cross-validation. Then we obtained an IFS curve in a 2D Cartesian coordinate system with index ε (the number of features) as its abscissa (or *X*-coordinate) and the overall accuracy as its ordinate (or *Y*-coordinate). If g varies from 0 to g_θ, there are g_θ+1 IFS curves. The peak (the maximum accuracy) can be observed in these curves. Then the optimal feature subset with parameters ε_0 and g_ϕ can be determined and expressed as:

$$\mathbf{P}_{\varepsilon_0}^{g_\phi} = \left[f_1^{g_\phi}, f_2^{g_\phi}, \cdots, f_{\varepsilon_0}^{g_\phi} \right]^{\mathrm{T}} \qquad (1 \le \varepsilon_0 \le 400; \;\; 0 \le g_\phi \le g_\theta \;) \tag{9}$$

where ε_0 is the number of optimal g_ϕ-gap dipeptides.

Based on above processes, the high-dimensional data will be projected into a low-dimensional space. The final classifier model was built based on the optimal feature subset.

2.4. Support vector machine (SVM)

SVM has been widely used in bioinformatics [17-27] and was adopted as the classification algorithm in this work. Its basic principle is to transform the input vector into a high-dimension Hilbert space and seek a separating hyperplane with the maximal margin in this space by using the decision function:

$$f(\vec{X}) = \mathrm{sgn}(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\vec{X}, \vec{X}_i) + b) \tag{10}$$

where α_i is Lagrange multipliers; b is the offset; $\vec{X}_i$ is the i-th training vector; y_i represents the type of the i-th training vector; $K(\vec{X}, \vec{X}_i)$ is a kernel function which defines an inner product in a high dimensional feature space; sgn is sign function. The radial basis kernel function (RBF) $K(\vec{X}_i, \vec{X}_j) = \exp(-\gamma \| \vec{X}_i - \vec{X}_j \|^2)$ was used in the current work due to its effectiveness and speed in nonlinear classification process. The software toolbox (LibSVM) for implementing SVM can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. A grid search method was used to optimize the regularization parameter C and kernel parameter γ through 5-fold cross-validation. The search spaces for C and γ are [2^{15}, 2^{-5}] and [2^{-5}, 2^{-15}] with the steps of 2^{-1} and 2,

respectively. The probability estimates from LibSVM were calculated by the Bradley-Terry model [28]. The guide on how to obtain probability can be obtained from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/.

2.5. Performance Evaluation

In statistical prediction, three cross-validation methods, namely independent dataset test, sub-sampling (e.g., 2, 5 or 10-fold cross-validation) test, and jackknife test are often used to evaluate the performance of the predicted methods in practical application [13, 14]. Among the three test methods, as elucidated in [9] and demonstrated by Eq.50 of [29], the jackknife test was deemed as the most objective one that can always yield a unique result for a given benchmark dataset. Therefore, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods (see, e.g. [30-44]). Thus, the jackknife cross-validation was used in this study to examine the anticipated success rates of the predictor. Furthermore, to reduce the computational time, the 5-fold cross-validation was used to select the parameters C and γ in SVM.

To provide a simple method to measure the prediction quality, the following three metrics: sensitivity (*Sn*), specificity (*Sp*) and accuracy (*Acc*) were used and expressed as

$$Sn = \frac{n^+}{N^+} \tag{11}$$

$$Sp = \frac{n^-}{N^-} \tag{12}$$

$$Acc = \frac{n^+ + n^-}{N^+ + N^-} \tag{13}$$

where N^+ and N^- denote the number of phage virion proteins and the number of phage non-virion proteins, respectively; n^+ and n^- are the number of the correctly recognized phage virion proteins and the number of the correctly recognized phage non-virion proteins, respectively.

To describe the performance of models across the entire range of SVM decision values, the receiver operating characteristic (ROC) curves were also provided. The quality of the proposed method can be objectively evaluated by measuring the area under the receiver operating characteristic curve (auROC).

3. Results and discussion

3.1. Feature selection for improving accuracy

According to the aforementioned *g*-gap dipeptide composition in Eqs. (3-4), for each *g* parameter, a 400-dimension vector will be produced. The feature dimension is much larger than the number of samples (307 samples). Generally, the high dimension features would not only lead to the over-fitting problem, but also bring about information redundancy or noise [4, 6, 10, 13, 15, 16]. These would result in low capability in the generalization of a predictor and the poor prediction in cross-validation. For example, the 400 1-gap dipeptides can only yield the Acc of 77.85% in 5-fold cross-validation. Although the low dimension feature can improve the robust of a predictor, the number of the selected features is too small to afford enough information, thus resulting in the poor predictive accuracy. For example, in 5-fold cross-validation, the Acc of 76.9% was obtained by 10 selected 0-gap dipeptides.

To overcome these disadvantages and deal with the high-dimension disaster, it is necessary to pick out informative parameters. This will not only gain a deeper insight into the intrinsic properties of phage virion proteins, but also improve the understandability and the quality of the predictor. Obviously, the best feature combination can be found by examining the performance of all feature sets. However, the computation time is so long that it is impossible to investigate the performance of all feature sets. Taking the amino acid composition containing 20-dimension feature vector as an example, the number of all possible combinations for 20-D vector is $C_{20}^{1}+C_{20}^{2}+\cdots+C_{20}^{19}+C_{20}^{20}=1,048,575$. For a 400-dimension vector, the number of all possible combinations will be greater than 2.58×10^{120}.

In order to economize the computational time and source, the ANOVA with IFS process was used to search for the optimal feature set with the maximum accuracy. We controlled gap *g* to vary from 0 to 9 and investigated the performances of 10×400=4,000 feature subsets. And the 10 IFS curves were plotted in **Fig. 1**. The *Acc* reached its peak (85.02%) when the top ranked 160 1-gap dipeptides ($P<10^{-5}$) were used (**Fig. 1**). With the top ranked 160 1-gap dipeptides as the input parameters of SVM, 75.75% phage virion proteins and 89.42% phage non-virion proteins can be correctly predicted. The ROC curve was plotted in **Fig.2** for investigating the performance of the model across the entire range of SVM decision values. The auROC reaches 0.899. It should be noted that the number of features (160) are about a

half of the number of samples (307), suggesting that the proposed method is reliable and efficient. The high accuracy obtained by cross-validation demonstrates that the method is robust.

3.2. Comparison with other methods

It would be instructive to make a comparison between the proposed method and other published methods. The comparative results of different methods on the same benchmark dataset are listed in **Table 1**. As we can see from **Table 1**, although the *Sn* obtained in this paper is equal to that of our previously proposed Naïve Bayes model, the *Sp*, *Acc* and auROC of the proposed method are all dramatically higher than those of Naïve Bayes [4]. Furthermore, we calculated the *Acc* achieved by completely random guess (CRG) [45]. Obviously, the *Acc* achieved by CRG is 50.00%. If considering the weight or prior probability, the *Acc* is [99×(99/307)+208×(208/307)]/307=56.30%. These results demonstrate that our method is superior to the published method and random guess.

In addition, we investigated the performances of four state-of-the-art classifiers (BayesNet, RBFNetwork, Random Forest and Naïve Bayes) on the same benchmark dataset using same feature selection technique. Firstly, we repeated the feature selection process in which ANOVA was adopted to optimize *g*-gap dipeptides. Secondly, each feature set was input into the four algorithms. Finally, the maximum accuracies of three algorithms were selected for comparison. Comparison in **Table 1** demonstrates that the SVM is the best one among all classifiers for phage virion predictions.

Furthermore, we checked the performance of proposed method for low identity datasets using jackknife cross-validation. By use of 25% sequence identity as the cutoff, we obtained 278 phage proteins included 87 virion proteins and 191 non-virion proteins. The sensitivities of virion proteins and non-virion proteins were 68.97% and 88.48%, respectively. The overall accuracy of 82.37% with average accuracy of 78.72% was achieved. The overall accuracy just decreased by 2.65% with the sequence identity decreasing from 40 to 25%. These results demonstrated that the proposed model is robust.

Finally, for demonstrating the prediction capability of the proposed model, we built an independent dataset which contained 11 phage virion proteins and 19 phage

non-virion proteins. Our model can correctly identify the 9 virion proteins and 17 non-virion proteins.

3.2. Feature analysis

The results in **Fig.1** also reveal that the correlation between two residues with one residue interval (g=1) is more important than other correlations in phage virion protein sequences. It is sure that some important 1-gap dipeptides contribute to the recognition of phage virion proteins. To provide an overall and intuitive view, the following normalized function was introduced to scale the $F(\xi)$ of the ξ-th 1-gap dipeptide as follows

$$F^0(\xi) = \frac{F(\xi) - F_{\min}}{F_{\max} - F_{\min}} \times \mathrm{sgn}\left[\overline{f^1_{\xi,virion}} - \overline{f^1_{\xi,non-virion}}\right] \quad (14)$$

where $F_{\min}$ and $F_{\max}$ are the minimum and maximum F values of all the 400 1-gap dipeptides. The $\overline{f^1_{\xi,virion}}$ and $\overline{f^1_{\xi,non-virion}}$ are the average frequencies of the ξ-th 1-gap dipeptide in virion proteins and non-virion proteins, respectively; sgn is the sign function. Thus, we obtained $F^0(\xi) \in (-1,1)$. If $F^0(\xi) < 0$, the ξ-th 1-gap dipeptide prefers phage virion proteins, otherwise it prefers phage non-virion proteins.

To analyze the contributions of different 1-gap dipeptides to the prediction model, a heat map was drawn in **Fig.3**. In **Fig.3**, the column and row of the heat map represent the first residue and the second residue of 1-gap dipeptides, respectively. Each element in the heat map represents a 1-gap dipeptide and is colorized according to its $F^0(\xi)$. It is observed that the majority of 1-gap dipeptides have very small absolute value of $F^0(\xi)$ (green), indicating that these features are irrelevant with the phage virion protein prediction. We also found that the amino acids A, G, P, S, T and V (red) as well as their 1-gap correlations often appear in phage virion proteins, whereas the amino acids E, K, L and R (blue) as well as their 1-gap correlations are not preferred in phage virion proteins. Ala, Gly, Pro, Ser, Thr and Val are small amino acids and their side-chain masses are 15.0, 1.0, 42.0, 31.0, 45.0 and 43.0, respectively. Glu, Lys, Leu and Arg are big amino acids and their side-chain masses are 73.0, 73.0, 57.0 and 101.0, respectively [46].

Small amino acids are prone to lead to the conformation transformation, implying that small amino acids play important roles in the function of phage virion proteins. Coia et al. [47] have found that the small amino acid (such as Gly-Lys-Arg) usually

occurred in the flanking potential cleavage site in virion proteins. The flexibility of the small side chain amino acids is required to accommodate the variation observed in the cleavage sites [48]. Kuzmicheva et al. [49] have demonstrated that, to ensure a high constant of binding (low dissociation constant), the domain C of phage major coat protein must include predominantly small amino-acid residues which possess a diminutive positively charged surface and a low energy of the higher occupied molecular orbital. Our statistical results coincide with these findings.

As we can see from **Fig.3**, the colors of some 1-gap dipeptides are dramatically different from that of other 1-gap dipeptides. We cautiously picked out 17 1-gap dipeptides (A*G, A*T, A*P, S*T, S*A, V*A, T*S, V*T, G*A, G*G, S*G, V*G, V*I, E*L, K*L, K*E, E*E) according to the criteria that the absolute value of $F^0(\xi)$ is larger than 0.55. Among the 17 features, 13 features in **Fig.3** are marked in red, indicating that the occurrence frequencies of these features in virion proteins are dramatically larger than that in non-virion proteins. Only 4 1-gap dipeptides in **Fig.3** marked in blue prefer to non-virion proteins. The reason of this phenomenon is that non-virion proteins are consisted of phage-secreted proteins and other non-structural proteins. The features of different non-virion proteins are annihilated each other. Thus, according to the strategy in the promoter prediction[50], it is better to use multi negative sets, in which each negative set has its given type, to train and test the model. However, in this study, the currently available data do not allow the strategy. Otherwise, the number of proteins for some subsets would be too few to implement statistical significance. However, these 17 features do play important roles in virion protein prediction and yield the Acc of 78.18% in 5-fold cross-validation, suggesting that the ANOVA-based feature selection technique is powerful.

3.3. Web-Server Guide

Establishing a user-friendly web-server will improve the efficiency and avoid repeating a complicated mathematics and program for studying phage virion proteins. The predictor established via aforementioned procedures is called **PVPred**. For the convenience of the vast majority of experimental scientists, we provided a guide to help experimental scientists to use the web-server to get the desired results.

Firstly, open the web server at http://lin.uestc.edu.cn/server/PVPred and you will see the top page of **PVPred** on your computer screen, as shown in **Fig.4**. Click on the

Read Me button to see a brief introduction about the predictor and the caveat when using it. Click on the Data button to download the benchmark datasets used to train and test the **PVPred** predictor. Click on the Citation button to find the relevant papers that document the detailed development and algorithm of **PVPred**. Secondly, either type 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. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box. Thirdly, click on the Submit button to see the predicted result. It should be noted that each of the input query sequences should exclude all illegal characters: such as 'B', 'X', 'U', 'Z'.

4. Conclusion

The available evidence indicates that the bacteriophage is a new way to fight against bacterial infections. The knowledge for phage virion proteins is conductive to the development of antibacterial drugs. Thus, we proposed a feature selection technique based on AVONA to discriminate phage virion proteins from phage non-virion proteins by using SVM. A high accuracy model was obtained. Results demonstrate that the ANOVA can accurately pick out informative features and efficiently improve predictive performance. Based on this model, an online predictor **PVPred** was established for identifying phage proteins. We hope that this predictor will become a useful tool for phage virion protein analysis and further experimental research. Moreover, the method proposed in this study can be generalized to the prediction of other proteomics.

Conflict of interest:

The authors declare that there is no conflict of interests.

Acknowledgements

This work was supported by the National Nature Scientific Foundation of China (Nos. 61202256, 61301260 and 61100092), the Nature Scientific Foundation of Hebei Province (No.C2013209105), and the Fundamental Research Funds for the Central Universities (Nos. ZYGX2012J113, ZYGX2013J102).

References:

1. E. J. Stella, J. J. Franceschelli, S. E. Tasselli and H. R. Morbidoni, *PloS one*, 2013, **8**, e56384.
2. W. Gibson, *Intervirology*, 1996, **39**, 389-400.
3. V. Seguritan, N. Alves, Jr., M. Arnoult, A. Raymond, D. Lorimer, A. B. Burgin, Jr., P. Salamon and A. M. Segall, *PLoS computational biology*, 2012, **8**, e1002657.
4. P. M. Feng, H. Ding, W. Chen and H. Lin, *Computational and mathematical methods in medicine*, 2013, **2013**, 530696.
5. C. UniProt, *Nucleic acids research*, 2013, **41**, D43-47.
6. H. Ding, S. H. Guo, E. Z. Deng, L. F. Yuan, F. B. Guo, J. Huang, N. N. Rao, W. Chen and H. Lin, *Chemometrics Intell. Lab. Syst.*, 2013, **124**, 9-13.
7. H. Ding, L. Liu, F. B. Guo, J. Huang and H. Lin, *Protein and peptide letters*, 2011, **18**, 58-63.
8. L. Fu, B. Niu, Z. Zhu, S. Wu and W. Li, *Bioinformatics*, 2012, **28**, 3150-3152.
9. K. C. Chou and H. B. Shen, *PloS one*, 2010, **5**, e11335.
10. L. F. Yuan, C. Ding, S. H. Guo, H. Ding, W. Chen and H. Lin, *Toxicology in vitro : an international journal published in association with BIBRA*, 2013, **27**, 852-856.
11. H. Lin and H. Ding, *Journal of theoretical biology*, 2011, **269**, 64-69.
12. H. Lin, *Journal of theoretical biology*, 2008, **252**, 350-356.
13. H. Lin, W. Chen and H. Ding, *PloS one*, 2013, **8**, e75726.
14. H. Lin and W. Chen, *Journal of microbiological methods*, 2011, **84**, 67-70.
15. W. Chen, H. Lin, P. M. Feng, C. Ding, Y. C. Zuo and K. C. Chou, *PloS one*, 2012, **7**, e47843.
16. B. Q. Li, L. L. Hu, S. Niu, Y. D. Cai and K. C. Chou, *Journal of proteomics*, 2012, **75**, 1654-1665.
17. Z. Chen, Y. Zhou, J. Song and Z. Zhang, *Biochimica et biophysica acta*, 2013, **1834**, 1461-1467.
18. S. Nakariyakul, Z. P. Liu and L. Chen, *Biochimica et biophysica acta*, 2014, **1844**, 165-170.
19. B. Niu, Y. Zhang, J. Ding, Y. Lu, M. Wang, W. Lu, X. Yuan and J. Yin, *Biochimica et biophysica acta*, 2014, **1844**, 214-223.
20. S. Nakariyakul, Z. P. Liu and L. Chen, *Amino acids*, 2012, **42**, 1947-1953.
21. C. Sun, X. M. Zhao, W. Tang and L. Chen, *BMC systems biology*, 2010, **4 Suppl 2**, S12.
22. J. F. Xia, X. M. Zhao, J. Song and D. S. Huang, *BMC bioinformatics*, 2010, **11**, 174.
23. J. Song and K. Burrage, *BMC bioinformatics*, 2006, **7**, 425.
24. Y. Cai, J. He, X. Li, K. Feng, L. Lu, K. Feng, X. Kong and W. Lu, *Protein and peptide letters*, 2010, **17**, 464-472.
25. Y. Cai, J. He, X. Li, L. Lu, X. Yang, K. Feng, W. Lu and X. Kong, *Journal of proteome research*, 2009, **8**, 999-1003.
26. P. Du and Y. Li, *BMC bioinformatics*, 2006, **7**, 518.
27. P. Du and Y. Yu, *BioMed research international*, 2013, **2013**, 263829.
28. T. K. Huang, R. C. Weng and C. J. Lin, *J Mach Learn Res*, 2006, **7**, 85-115.
29. K. C. Chou and H. B. Shen, *Analytical biochemistry*, 2007, **370**, 1-16.

30. K. C. Chou, *Proteins*, 2001, **43**, 246-255.
31. K. C. Chou, *Journal of theoretical biology*, 2011, **273**, 236-247.
32. P. Du, S. Cao and Y. Li, *Journal of theoretical biology*, 2009, **261**, 330-335.
33. P. Du, Y. Tian and Y. Yan, *Journal of theoretical biology*, 2012, **313**, 61-67.
34. P. Du and L. Wang, *PloS one*, 2014, **9**, e86879.
35. W. Z. Lin, J. A. Fang, X. Xiao and K. C. Chou, *Molecular bioSystems*, 2013, **9**, 634-644.
36. J. L. Min, X. Xiao and K. C. Chou, *BioMed research international*, 2013.
37. L. Nanni and A. Lumini, *Amino acids*, 2008, **34**, 635-641.
38. L. Nanni and A. Lumini, *Amino acids*, 2008, **34**, 653-660.
39. L. Nanni and A. Lumini, *Amino acids*, 2009, **36**, 167-175.
40. Z. C. Wu, X. Xiao and K. C. Chou, *Protein and peptide letters*, 2012, **19**, 4-14.
41. X. Xiao, J. L. Min, P. Wang and K. C. Chou, *Journal of theoretical biology*, 2013, **337**, 71-79.
42. X. Xiao, P. Wang and K. C. Chou, *PloS one*, 2012, **7**, e30869.
43. X. Xiao, P. Wang, W. Z. Lin, J. H. Jia and K. C. Chou, *Analytical biochemistry*, 2013, **436**, 168-177.
44. X. Xiao, Z. C. Wu and K. C. Chou, *Journal of theoretical biology*, 2011, **284**, 42-51.
45. K. C. Chou, *Molecular bioSystems*, 2013, **9**, 1092-1100.
46. H. B. Shen and K. C. Chou, *Analytical biochemistry*, 2008, **373**, 386-388.
47. G. Coia, M. D. Parker, G. Speight, M. E. Byrne and E. G. Westaway, *The Journal of general virology*, 1988, **69 (Pt 1)**, 1-21.
48. G. Speight, G. Coia, M. D. Parker and E. G. Westaway, *The Journal of general virology*, 1988, **69 (Pt 1)**, 23-34.
49. G. A. Kuzmicheva, P. K. Jayanna, A. M. Eroshkin, M. A. Grishina, E. S. Pereyaslavskaya, V. A. Potemkin and V. A. Petrenko, *Protein engineering, design & selection : PEDS*, 2009, **22**, 631-639.
50. H. Lin and Q. Z. Li, *Theory in biosciences = Theorie in den Biowissenschaften*, 2011, **130**, 91-100.

TABLE

Table 1 Comparing the proposed method with other methods

Classifier	Sn(%)	Sp(%)	Acc(%)	auROC
Naïve Bayes(38-D)[4]	75.76	80.77	79.15	0.855
BayeNet(148-D)	56.57	80.29	72.64	0.776
RBFNetwork(166-D)	70.71	84.62	80.13	0.806
Random Forest(160-D)	45.45	93.27	77.85	0.798
Naïve Bayes(105-D)	75.76	86.06	82.74	0.862
SVM(160-D)	**75.76**	**89.42**	**85.02**	**0.899**

Figure captions

Fig. 1. A plot to show the IFS procedure. When the top160 1-gap dipeptides were used to perform prediction, the overall success rate reaches IFS peak of 85.0%.

Fig. 2. The ROC curve for the prediction of phage virion proteins by using 160 optimal 1-gap dipeptides. The auROC of 0.899 was obtained in jackknife cross-validation. The diagonal dot line denotes a random guess with the auROC of 0.5.

Fig. 3. A chromaticity diagram for the 400 $F^0(\xi)$ of the 1-gap dipeptides.

The red elements indicate $\overline{f^1_{\xi}}_{,virion} > \overline{f^1_{\xi}}_{,non-virion}$, whereas the blue elements indicate $\overline{f^1_{\xi}}_{,virion} < \overline{f^1_{\xi}}_{,non-virion}$

Fig. 4. A semi-screenshot to show the top page of the PVPred web-server. Its website address is at http://lin.uestc.edu.cn/server/PVPred.

Figures

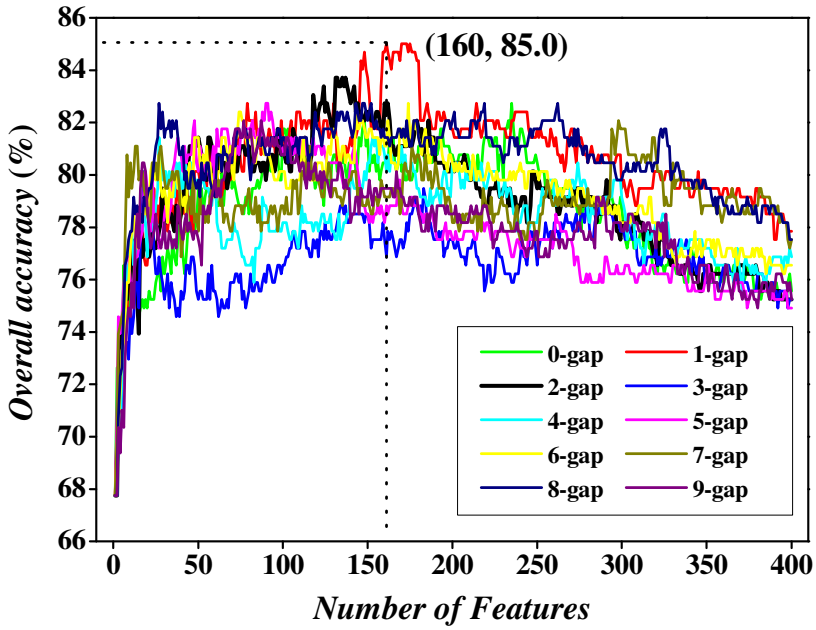


Fig. 1

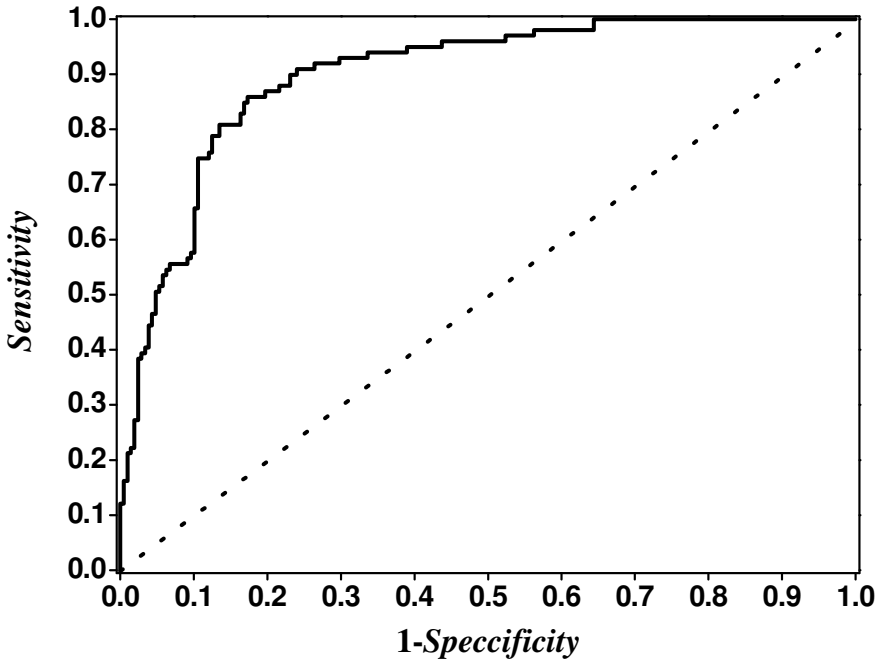


Fig. 2

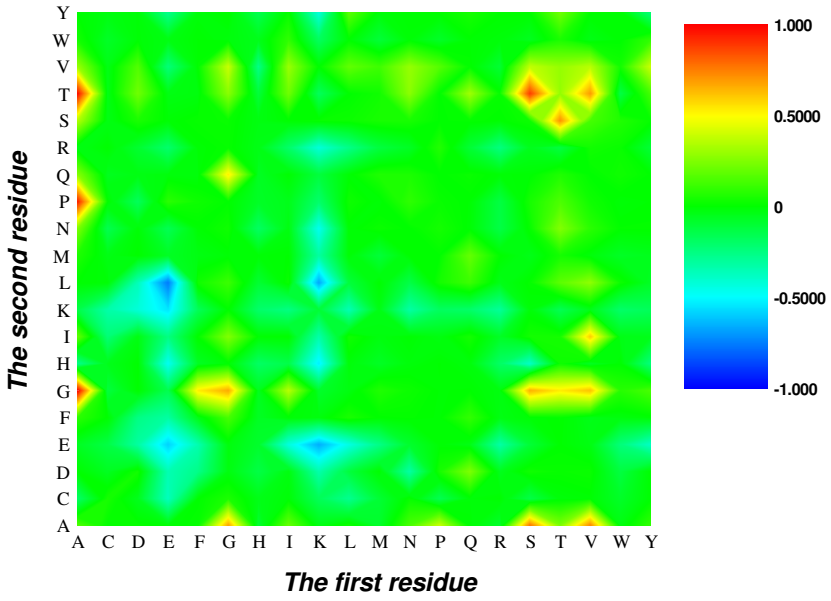


Fig. 3

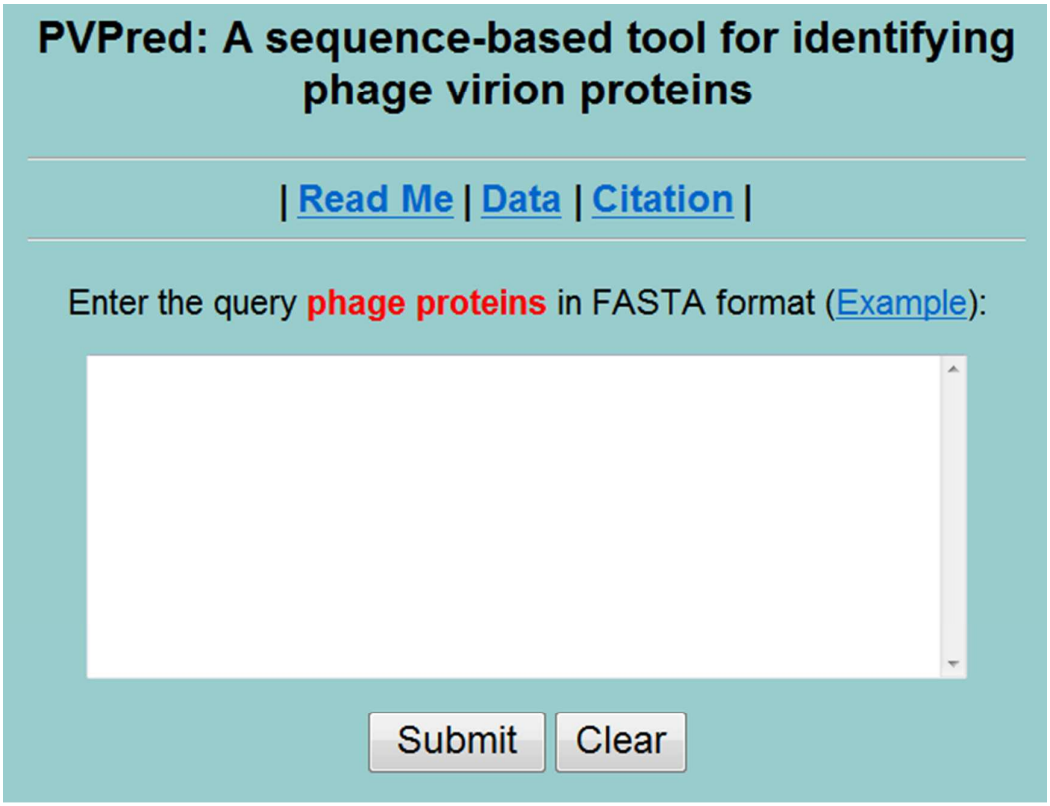


Fig. 4