molecules

MDPI

Article

# Identifying Phage Virion Proteins by Using Two-Step Feature Selection Methods

**Jiu-Xin Tan [1], Fu-Ying Dao [1], Hao Lv [1], Peng-Mian Feng [2,*] and Hui Ding [1,*]**

[1] 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; tjx0705@163.com (J.-X.T.); koyee_d@sina.com (F.-Y.D.); 13208188368@163.com (H.L.)

[2] Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, School of Public Health, North China University of Science and Technology, Tangshan 063000, China

* Correspondence: fengpengmian@gmail.com (P.-M.F.); hding@uestc.edu.cn (H.D.)

Received: 13 July 2018; Accepted: 8 August 2018; Published: 10 August 2018


**Abstract:** Accurate identification of phage virion protein is not only a key step for understanding the function of the phage virion protein but also helpful for further understanding the lysis mechanism of the bacterial cell. Since traditional experimental methods are time-consuming and costly for identifying phage virion proteins, it is extremely urgent to apply machine learning methods to accurately and efficiently identify phage virion proteins. In this work, a support vector machine (SVM) based method was proposed by mixing multiple sets of optimal g-gap dipeptide compositions. The analysis of variance (ANOVA) and the minimal-redundancy-maximal-relevance (mRMR) with an increment feature selection (IFS) were applied to single out the optimal feature set. In the five-fold cross-validation test, the proposed method achieved an overall accuracy of 87.95%. We believe that the proposed method will become an efficient and powerful method for scientists concerning phage virion proteins.

**Keywords:** phage virion protein; feature fusion; ANOVA; mRMR; machine learning

## 1. Introduction

The bacteriophage, which is also known as the phage, is a kind of virus that infects bacteria and can complete growth and reproduction in bacteria [1]. The bacteriophage consists of an outer protein coat and a single genetic material (DNA or RNA) inside [2] and is the most widely distributed group of the virus, which is usually found in places full of bacterial communities such as soil and animals' intestine.

Proteins are the major components of viruses including structural proteins (namely, phage virion proteins) and non-structural proteins (namely, phage non-virion proteins). Non-structural proteins are proteins encoded by viral genes that function in the process of viral gene expression, but they don't bind to viral particles. Structural proteins are the necessary proteins to form mature and infectious virus particles including shell proteins, envelope proteins, and virus particle enzymes, etc. Among them, the shell proteins are the proteins that constitute the structure of the virus capsid and their main function is protecting the viral nucleic acid, participating in the adsorption and invasion of bacteriophages, and more. Envelope proteins are the proteins that constitute the viral envelope structure and the main function is to act as a viral surface antigen, which maintains the viral structure, participates in virus budding, and more. Due to the clear differences in the function of structural proteins and non-structural proteins, the correct identification of them will be helpful to further understand the molecular mechanisms of bacteriophage genetics and the development of antimicrobial drugs.

The traditional method for the identification of the phage virion and non-virion proteins is mass spectrometry (MS) [3]. However, it has not kept pace with the explosive growth of protein sequences in the post-genome era. Therefore, it is necessary to adopt machine learning methods to identify phage virion proteins. In 2013, Feng et al. proposed a Naïve Bayes classifier to identify phage virion proteins [4]. Afterward, Ding et al. developed an SVM-based method to identify phage virion proteins in which the proteins were encoded using the optimal features obtained by using the ANOVA feature selection technique [5]. In 2015, Zhang et al. introduced an ensemble method for predicting phage virion proteins from phage protein sequences by combining the CTD, bi-profile Bayes, PseAAC, and PSSM [6]. Subsequently, Manavalan et al. proposed a method called PVP-SVM, which adopted the SVM classifier with multiple feature extraction methods [7]. More recently, Fan et al. proposed a novel method called PhagePred to predict phage virion proteins by using the Multinomial Naïve Bayes classifier combined with the g-gap features tree [8]. However, these current feature extraction methods do not describe the protein sequences completely. Therefore, it is necessary to apply new feature extraction and selection methods to investigate the identification of phage virion proteins further.

In this paper, by using a new feature extraction method and two feature selection methods (ANOVA and mRMR) to select optimal features, we proposed an SVM-based method to identify phage virion proteins. As shown in Figure 1, the rest of the paper is organized based on the following aspects: (1) collection of raw data sets and processing of raw data sets, (2) feature extraction, (3) feature selection, (4) classifier algorithm, and (5) performance evaluation.

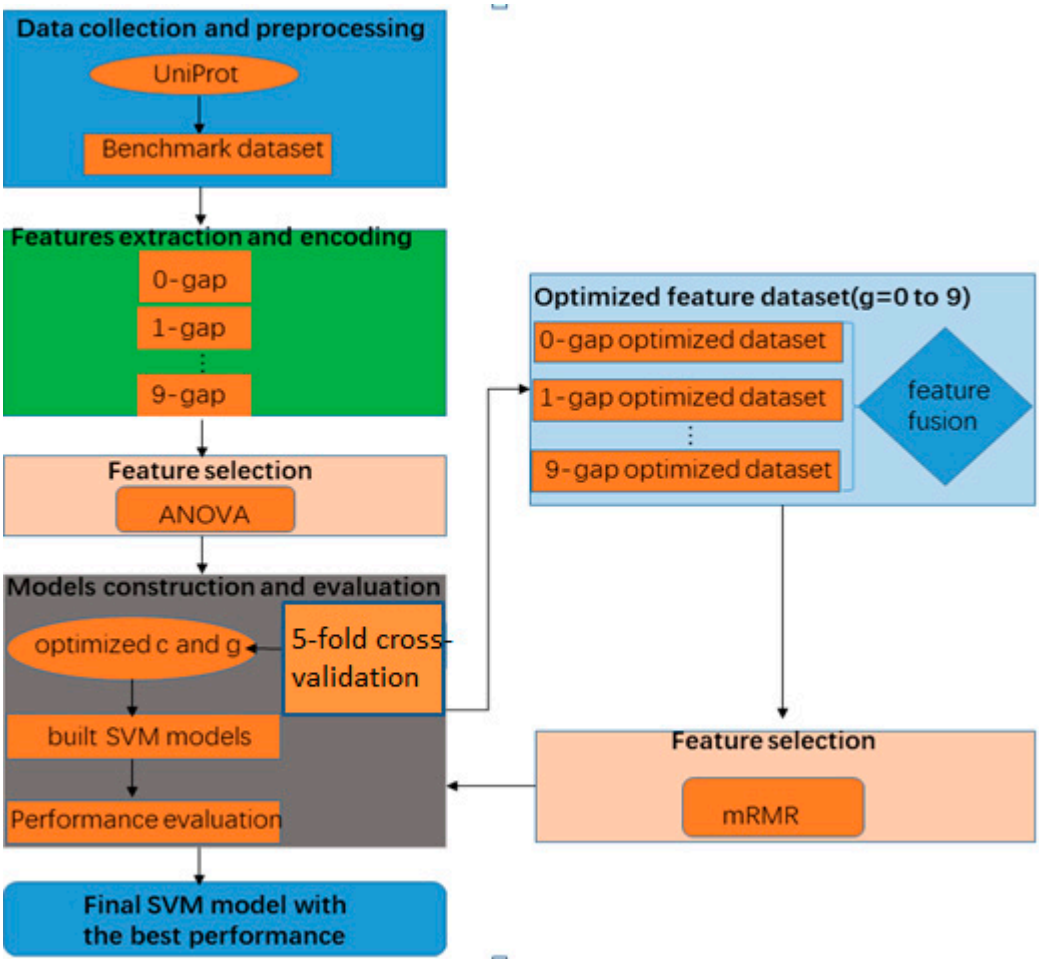


**Figure 1.** The framework of the proposed method.

## 2. Materials and Methods

### *2.1. Benchmark Dataset*

The raw positive and negative datasets adopted in this research were extracted from the Universal Protein Resource (UniProt) [9]. In order to obtain a high-quality and reliable benchmark dataset, the following steps were performed. First, only the phage virion proteins and phage non-virion proteins that have been experimentally confirmed can be included. Second, the protein sequences that are fragments of other proteins were excluded. Third, protein sequences containing nonstandard letters such as 'B,' 'U,' 'X,' or 'Z' were eliminated because their meanings are ambiguous. By following these three rigorous screening processes, a total of 121 phage virion protein sequences and 231 phage non-virion protein sequences were obtained. In order to obtain a high quality benchmark dataset, the CD-HIT [10] program was used by setting the cutoff threshold of the protein sequence identity to 40%. Lastly, we obtained 99 phage virion protein sequences and 208 phage non-virion protein sequences.

It is necessary to evaluate the proposed model by using an independent dataset to check whether the prediction model has a generalization capability. In this study, the independent dataset constructed by Manavalan et al. [7] was used, which can be downloaded from http://www.thegleelab.org/PVP-SVM/SVM-PVPData.html. It is a reliable independent dataset containing 30 phage virion protein sequences and 64 phage non-virion protein sequences.

### *2.2. The g-Gap Dipeptide Composition*

After the benchmark dataset was built, we needed to describe the protein sequences using a computer-readable form. Proteins are formed with the use of 20 amino acids, according to a certain order and space structure. The most common and simplest method is to formulate the sample protein **P** with $L$ residues with its entire amino acid sequence, which is shown below.

$$\mathbf{P} = A_1 A_2 A_3 \ldots A_L \tag{1}$$

where $A_1$ represents the first amino acid residue of the sample protein **P**, $A_2$ represents the second amino acid residue of the sample protein **P**, and so forth.

Another straightforward method to formulate the protein sequence is amino acid composition (AAC). In order to obtain the sequence-related information, the AAC is replaced by the adjacent dipeptide composition to represent the protein sequences [4]. However, the adjacent dipeptide composition can only express the correlation between two adjacent amino acid residues. In fact, in three-dimensional space, two amino acids with $g$-gap residues may be adjacent. In order to find the important correlation in protein sequences, we applied the $g$-gap dipeptide composition, which is extended from the adjacent dipeptides [8,11,12] and is a kind of mode of PseAAC [13–15]. By adopting this method, a sample protein **P** can be formulated by using the equation below.

$$\mathbf{P} = \left[ v_1^g, v_2^g, \ldots, v_i^g, \ldots v_{400}^g \right]^T \tag{2}$$

where the symbol $T$ represents the transposition of the vector while the $v_i^g$ represents the frequency of the $i$-th ($i$ = 1, 2, . . . , 400) $g$-gap dipeptide and can be formulated by using the equation below.

$$v_i^g = \frac{n_i^g}{L - g - 1} \tag{3}$$

where $n_i^g$ represents the number of the $i$-th $g$-gap dipeptide, $L$ represents the length of the protein **P**, $g$ represents the number of amino acid residues separated by two amino acid residues, $g$ = 1 indicates the correlation between two amino acid residues with the interval of one residue, $g$ = 2 indicates the correlation between two amino acid residues with the interval of two residues, and so forth.

### 2.3. The Analysis of Variance (ANOVA)

In general, if a model was built on a low-dimensional feature subset, the robustness of the model will be excellent. However, the low dimensionality of the feature subset will not provide enough information, which results in a poor performance of the model. On the contrary, a high-dimensional features subset can lead to information redundancy and overfitting problems. Both of these problems will lead to low accuracy of the cross-validation prediction and a poor predictor generalization ability. In order to overcome these shortcomings, the best way is to pick out the best feature subset, but it is time-consuming to investigate the performance of all feature subsets by the computer [16–19]. For example, if the amino acid composition contains a 400-dimension feature vector, the number of all possible combinations for the 400-D vector is $C_{400}^{1} + C_{400}^{2} + C_{400}^{3} + \ldots + C_{400}^{399} + C_{400}^{400} = 2.58 \times 10^{120}$. Therefore, the analysis of variance (ANOVA) method with the incremental feature selection (IFS) [20–22] process was applied to investigate the optimal feature set with the maximum accuracy.

The ANOVA method can score each feature according to a unified standard and then reduce the features according to their contribution. This will not only save the calculation time but will also improve the model's performance. According to the principle of ANOVA, the score (F) of the $i$-th $g$-gap dipeptide can be formulated by using the formula below.

$$F(i) = \frac{S_B^2(i)}{S_W^2(i)} \tag{4}$$

where $S_B^2(i)$ represents the variance between groups (MSB) of $i$-th feature in the sample and $S_W^2(i)$ represents the variance within groups (MSW) of $i$-th feature in the sample, which are calculated by using the equations below.

$$\begin{cases} S_B^2(i) = \frac{SS_B(i)}{df_B} \\ S_W^2(i) = \frac{SS_W(i)}{df_W} \end{cases} \tag{5}$$

where $df_B = K - 1$ and $df_W = M - K$ represent the degree of freedom for MSB and MSW, respectively. K and M represent the number of group (here $K$ = 2) and the number of samples (here $M$ = 307), respectively. $SS_B(i)$ and $SS_W(i)$ represent the sum of MSB and MSW, respectively, and are calculated by using the formula below.

$$\begin{cases} SS_B(i) = \sum_{j=1}^{2} m_j \left( \frac{\sum_{s=1}^{m_j} f_i^g(s,j)}{m_j} - \frac{\sum_{j=1}^{2}\sum_{s=1}^{m_j} f_i^g(s,j)}{\sum_{j=1}^{2} m_j} \right)^2 \\ SS_W(i) = \sum_{j=1}^{2}\sum_{s=1}^{m_j} \left( f_i^g(s,j) - \frac{\sum_{s=1}^{m_j} f_i^g(s,j)}{m_j} \right)^2 \end{cases} \tag{6}$$

where $f_i^g(s,j)$ represents the frequency of the $i$-th $g$-gap dipeptide of the $j$-th sample in the $s$-th group.

A feature with a high $F(i)$-value means that its ability to identify the sample is excellent, which is more conducive to building a highly robust model. Therefore, we ranked all features according to their $F(i)$-values from high to low and obtained new feature vectors, which are shown below.

$$P_g' = \left[ v_{1,g}', v_{2,g}', \ldots, v_{i,g}', \ldots v_{n,g}' \right]^T (0 \le g \le 9) \tag{7}$$

By using the ANOVA method, we have a clear understanding of each feature's capabilities for the model. Therefore, we don't have to exhaust all the feature subsets but instead selectively construct feature subsets according to their $F(i)$-values in this paper. The first feature subset contains the feature with the highest $F(i)$-value, $P_g' = \left[ v_{1,g}' \right]$. The second feature subset adds the second highest $F(i)$-value to the first subset, $P_g' = \left[ v_{1,g}', v_{2,g}' \right]$. The third feature subset adds the third highest $F(i)$-value to the

second subset, $P'_g = \left[v'_{1,g}, v'_{2,g}, v'_{3,g}\right]$. The procedure was repeated until the accuracy of the model no longer increased.

### 2.4. Minimal-Redundancy-Maximal-Relevance (mRMR)

The combination of some of the best-performing features does not mean that the best predictive effect can be achieved. The main reason for this phenomenon is that these features are likely to have a high degree of correlation, which leads to more redundant information in the feature vector. To solve this problem, Peng et al. proposed the mRMR algorithm [23]. MRMD [24] is another tool similar to mRMR. The main idea of the algorithm is to filter out some of the most relevant features in the subset to achieve the goal of minimizing information redundancy and then obtain the most 'concise' subset of features in theory. Therefore, when using the mRMR-ranked feature benchmark dataset with a smaller dimension, it can still effectively represent a dataset with a larger dimension, which can ensure that the feature dimension and the time of the training model are greatly reduced with almost no loss of effective information.

The mRMR algorithm is often used to select discretization features and continuity features. Based on the issues involved in this paper, the following is a description about discretization features. Given two random discrete variables $x$ and $y$, the mutual information $I(x, y)$ between them can be calculated by using the formula below.

$$I(x, y) = \iint p(x, y) log \frac{p(x, y)}{p(x)p(y)} dxdy \tag{8}$$

where the mutual information $I(x, y)$ is a measure of the degree of correlation between two random variables $x$ and $y$ and $p(x), p(y), p(x, y)$ denote the probabilistic density functions, respectively. The metrics of the mRMR algorithm are Max-Relevance and Min-Redundancy and they can be described by the equations below.

1. Max-Relevance:

$$max\ D(S, c), D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \tag{9}$$

2. Min-Redundancy:

$$min\ R(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \tag{10}$$

where $x_i$ represents the $i$-th feature attribute, $c$ represents the category variable, $S$ represents the feature subset, and $|S|$ represents the size of the feature subset. Furthermore, if the two metrics are considered equally important, we can abbreviate it by using the formula below.

$$\begin{cases} max(D_I - R_I) \\ min\left(\frac{D_I}{R_I}\right) \end{cases} \tag{11}$$

### 2.5. Support Vector Machine (SVM)

The support vector machine (SVM) is a widely used binary classification model, which has been widely used in bioinformatics [25–38]. It is a supervised machine learning method and its main idea is to map the input features from low-dimensional space to a high-dimensional space through nonlinear transformation and then find the optimal linear classification surface in this high-dimensional space. For convenience, SVM software packages LibSVM can be download from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. In the current study, the LibSVM package was adopted to investigate the performance of identifying the phage virion proteins. Since it has been widely adopted in bioinformatics, the radical basis function kernel was selected to perform predictions.

### 2.6. Performance Evaluation

Three cross-validation methods, which are called the independent dataset test, the sub-sampling test, and the jackknife test, are widely used to evaluate the predictive ability of a predictor in practical application [4,28,39–49]. Among these three methods, the jackknife test was considered to be the most rigorous one that can get a unique outcome in statistical prediction. This test has been widely used by investigators to assess the performance of the predictor [4,37,46,50–54]. In this paper, in order to save computational time, the five-fold cross-validation method was used to tune the parameters $C$ and $g$ in the SVM.

In this paper, we adopted five evaluation indexes to evaluate the model. Sensitivity ($S_n$) is used to evaluate the model's ability to predict positive samples. Specificity ($S_p$) is used to evaluate the model's ability to predict negative samples. Overall Accuracy ($Acc$) reflects the proportion of the entire benchmark dataset that can be correctly predicted. The Matthew correlation coefficient ($Mcc$) is an indicator used to reflect the reliability of the algorithm. Its value is between $-1$ and 1 and the high value of $Mcc$ indicates that the model has a good prediction performance. The four indexes are defined below.

$$\begin{cases} S_n = \frac{TP}{TP+FN} \\ S_p = \frac{TN}{TN+FP} \\ Acc = \frac{TP+TN}{TP+FN+TN+FP} \\ Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \quad (12)$$

where $TP$, $TN$, $FP$, and $FN$ represent the number of the correctly recognized phage virion proteins, the number of the correctly recognized phage non-virion proteins, the number of phage non-virion proteins recognized as phage virion proteins, and the number of phage virion proteins recognized as phage non-virion proteins, respectively.

The ROC (receiver operating characteristic) curve is a more intuitive way to demonstrate the performance of the model. Therefore, we plotted the ROC and calculated the area under the ROC curve (auROC). The high value of auROC indicates that the model has a good classification ability and deserves our trust.

## 3. Result and Discussion

### 3.1. Prediction of Phage Virion Proteins Based on Single Kind of g-Gap Dipeptides

The ANOVA method with the IFS process was applied to investigate the optimal feature set with the maximum accuracy. The details of the optimization process can be referred to Section 2.3. We chose residue parameter $g$ from 0 to 9 to estimate the performance for all the $10 \times 400 = 4000$ feature subsets using SVM until all the 4000 *Accs* (overall accuracies) were calculated. In addition, we plotted 10 curves by setting the overall $Acc$ as an ordinate and the number of features as abscissa shown in Figure 2. Additionally, the highest predictive accuracies for $g$-gap dipeptides are shown in Table 1.

**Table 1.** The maximum $Acc$ and the corresponding number of feature at different $g$ values.

| $g$ | Number of Feature | $Acc$ (%) |
|---|---|---|
| 0 | 107 | 83.06 |
| 1 | 213 | 84.69 |
| 2 | 135 | 83.39 |
| 3 | 87 | 81.76 |
| 4 | 42 | 80.78 |
| 5 | 89 | 84.04 |
| 6 | 70 | 82.41 |
| 7 | 174 | 82.73 |
| 8 | 255 | 82.41 |
| 9 | 94 | 83.06 |

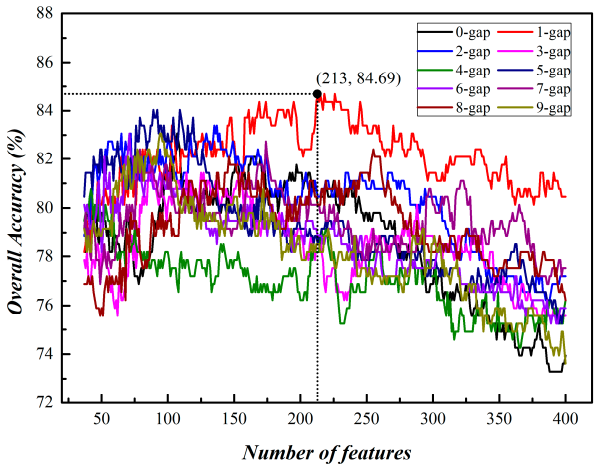


**Figure 2.** A plot showing the IFS curves for 0-gap to 9-gap.

### *3.2. Prediction of Phage Virion Proteins Based on Fusing Features*

For different g (from 0 to 9) values, each optimum subset represents the best characterization of proteins at different levels. If the 10 best feature subsets were fused together, each protein in the benchmark dataset can be encoded using a 107 + 213 + 135 + 87 + 42 + 89 + 70 + 174 + 255 + 94 = 1266-dimensional feature vector, which would be more comprehensive in representing protein sequences. We investigated the performance of identifying phage virion proteins based on the 1266 features. An *Acc* of 85.02% was achieved by adopting SVM in the five-fold cross-validation method. It is worthy of further investigation because the prediction performance was still far from satisfactory.

However, there exists noise in such a set of features. Therefore, we used the mRMR to score the 1266 features and ranked the features according to their scores. The details of the optimization process is outlined in Section 2.4. In order to show the contributions of each feature to the prediction, we made a heat map for the 1266 features based on their *Z*-scores, which is shown in Figure 3. The scores can be calculated with the help of the formula below.

$$Z_i = \frac{x_i - \mu}{\sigma} \ (i = 1, 2, \ldots, 1266) \tag{13}$$

where $x_i$ represents mRMR score for the $i$-th features, $\mu$ represents the overall average of 1266 mRMR scores, and $\sigma$ represents the standard deviation of 1266 mRMR scores.

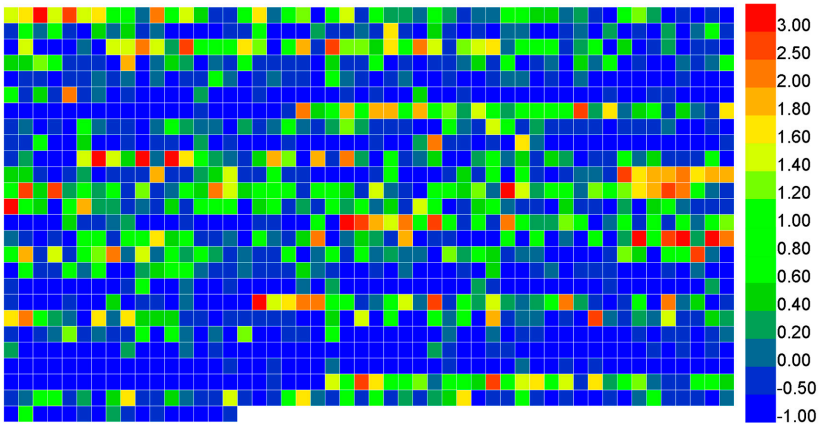


**Figure 3.** A heat map for 1266 features based on different Z-scores.

By using the IFS process to investigate the performance for all the subsets, the *Acc* reached its peak (87.95%) when the top ranked 368 dipeptides were used to build the model (Figure 4). In this case, the $S_n$, $S_p$, and *Mcc* are 83.84%, 89.90%, and 0.761%, respectively, with the auROC at 0.915 (Figure 5). This result indicates that the performance of the proposed model is smart and reliable for identifying phage virion proteins.

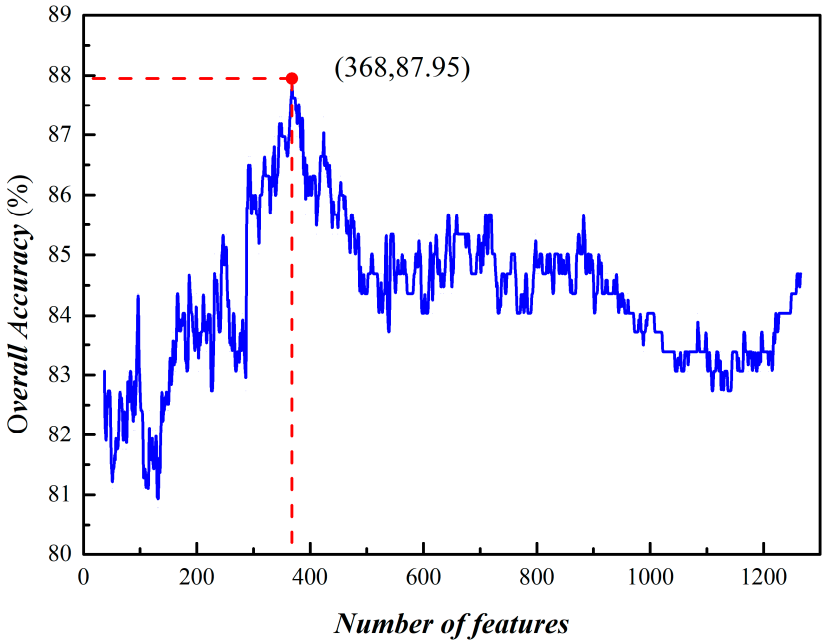


**Figure 4.** A plot showing the IFS curve by using mRMR.

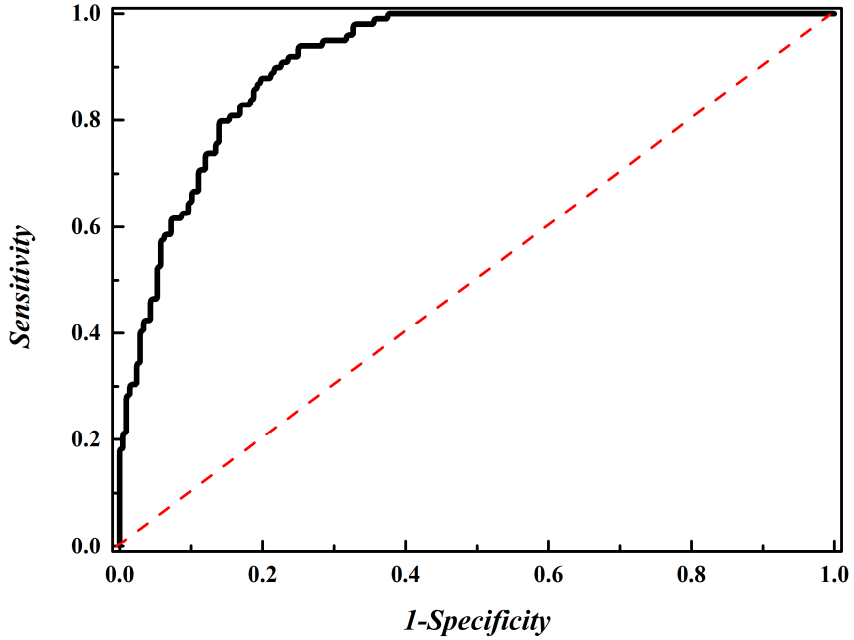


**Figure 5.** The ROC curve for the prediction of phage virion proteins by using 368 optimal features. The auROC of 0.915 was obtained in a five-fold cross-validation test. The diagonal dot line denotes a random guess with the auROC of 0.5.

### 3.3. Comparision with Other Published Methods

It is necessary to compare the methods used in this article with other published methods. Table 2 shows the detailed predictive results from the published papers. Based on the same benchmark dataset, Feng et al. proposed a Naïve Bayes-based method to predict bacteriophage virion proteins by using the amino acid composition and the dipeptide composition and obtained an overall accuracy of 79.15% [4]. Ding et al. adopted the ANOVA feature selection method to select optimal 1-gap dipeptides and obtained an overall accuracy of 85.02% [5]. Manavalan et al. proposed a novel method called 'PVP-SVM' in which the AAC, ATC, CTD, DPC, and PCP were used to represent the protein sequences and got an overall accuracy of 87.0% in the jackknife test [7]. Pan et al. adopted a Multinomial Naïve Bayes classifier based on the discrete features obtained from the $g$-gap feature tree. They achieved a superb overall accuracy of 98.37% in the 10-fold cross-validation [8]. In this work, we achieved an overall accuracy of 87.95%, which is the second-best prediction result until now. Compared to the results with reference [4,5,7] except for the value of $S_p$, which is slightly lower than that of Manavalan's [7], the values of $S_n$, $Acc$, $Mcc$ and auROC are significantly higher than published results. These comparisons demonstrate a better performance of our proposed methods.

**Table 2.** Comparing the proposed method with other published methods.

| Ref. | $S_n$ (%) | $S_p$ (%) | $Acc$ (%) | $Mcc$ | auROC |
|---|---|---|---|---|---|
| [4] | 75.76 | 80.77 | 79.15 | - | 0.855 |
| [5] | 75.76 | 89.42 | 85.02 | - | 0.899 |
| [7] | 73.70 | 93.30 | 87.00 | 0.695 | 0.900 |
| [8] | 96.97 | 98.56 | 98.05 | 0.963 | 0.990 |
| This work | 83.83 | 89.90 | 87.95 | 0.761 | 0.915 |

### 3.4. Performance Evaluation Using an Independent Dataset

In general, the best model for the training dataset is not the optimal model for the independent dataset. Therefore, we repeated the feature selection procession, retrained the model, and validated the model on an independent dataset. The results for different classifiers were listed in Table 3. As indicated in Table 3, among the three compared methods, our method obtained the highest $S_n$ while the $S_p$, $Acc$, and $MCC$ is similar to the PVP-SVM, which obtained the best predictive results on the independent. This indicates that our method can play complementary roles to existing methods for identifying phage virion proteins.

**Table 3.** Comparing the proposed method with other published methods on the independent dataset.

| Ref. | $S_n$ (%) | $S_p$ (%) | $Acc$ (%) | $Mcc$ | auROC |
|---|---|---|---|---|---|
| [5] | 60.00 | 76.50 | 71.30 | 0.357 | 0.742 |
| [7] | 66.70 | 85.90 | 79.80 | 0.531 | 0.844 |
| This work | 70.00 | 78.13 | 75.53 | 0.464 | 0.651 |

## 4. Conclusions

In this work, we investigated the accuracies of different features for identifying phage virion proteins. The maximum overall accuracy (87.95%) was obtained by fusing 10 optimal $g$-gap (0 to 9) dipeptide compositions, which was obtained by fusing ANOVA and mRMR feature selection methods. Compared with the existing methods, the proposed model improved the overall accuracy. Therefore, the method can be used as a reliable tool for accurately predicting phage virion proteins. In our study, there are many problems worth investigating for phage virion protein prediction. For example, in order to build a high quality dataset, greater attention should be paid to the dynamic changes of the database. Second, the biological meanings of the selected optimal features also need to be clarified.

Third, considering the promising performance of the ensemble classification methods [55] and the deep learning technique [56–58] in bioinformatics, we will integrate multiple classification algorithms to build the model for identifying phage virion proteins. User-friendly and publicly accessible web tools including predictors [25,27,50,59,60] or databases [61–64] represent the future direction for developing a more useful bioinformatics method. In the future, we will establish a powerful tool for phage virion protein prediction. The feature selection strategy can be extended to other fields.

**Author Contributions:** P.-M.F. and H.D. conceived and designed the experiments. J.-X.T. performed the experiments. J.-X.T., F.Y.D., and H.L. analyzed the data. J.-X.T., F.Y.D., and H.L. wrote the paper.

**Funding:** This work was supported by the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015Z006, ZYGX2016J223).

**Conflicts of Interest:** The authors declare that there is no conflict of interest.

## References

1. Stella, E.J.; Franceschelli, J.J.; Tasselli, S.E.; Morbidoni, H.R. Analysis of novel mycobacteriophages indicates the existence of different strategies for phage inheritance in mycobacteria. *PLoS ONE* **2013**, *8*, e56384. [CrossRef] [PubMed]
2. Gibson, W. Structure and assembly of the virion. *Intervirology* **1996**, *39*, 389–400. [CrossRef] [PubMed]
3. Lavigne, R.; Ceyssens, P.J.; Robben, J. Phage proteomics: Applications of mass spectrometry. *Methods Mol. Biol.* **2009**, *502*, 239–251. [PubMed]
4. Feng, P.M.; Ding, H.; Chen, W.; Lin, H. Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Method Med.* **2013**, *2013*, 530696. [CrossRef] [PubMed]
5. Ding, H.; Feng, P.M.; Chen, W.; Lin, H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol. Biosyst.* **2014**, *10*, 2229–2235. [CrossRef] [PubMed]
6. Zhang, L.; Zhang, C.; Gao, R.; Yang, R. An Ensemble Method to Distinguish Bacteriophage Virion from Non-Virion Proteins Based on Protein Sequence Characteristics. *Int. J. Mol. Sci.* **2015**, *16*, 21734–21758. [CrossRef] [PubMed]
7. Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. *Front. Microbiol.* **2018**, *9*, 476. [CrossRef] [PubMed]
8. Pan, Y.; Gao, H.; Lin, H.; Liu, Z.; Tang, L.; Li, S. Identification of Bacteriophage Virion Proteins Using Multinomial Naive Bayes with g-Gap Feature Tree. *Int. J. Mol. Sci.* **2018**, *19*, 1779. [CrossRef] [PubMed]
9. UniProt, C. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **2013**, *41*, D43–D47.
10. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
11. Tang, H.; Chen, W.; Lin, H. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. *Mol. Biosyst.* **2016**, *12*, 1269–1275. [CrossRef] [PubMed]
12. Ding, H.; Yang, W.; Tang, H.; Feng, P.M.; Huang, J.; Chen, W.; Lin, H. PHYPred: A tool for identifying bacteriophage enzymes and hydrolases. *Virol. Sin.* **2016**, *31*, 350–352. [CrossRef] [PubMed]
13. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef] [PubMed]
14. Georgiou, D.N.; Karakasidis, T.E.; Megaritis, A.C.; Nieto, J.J.; Torres, A. An extension of fuzzy topological approach for comparison of genetic sequences. *J. Intell. Fuzzy Syst.* **2015**, *29*, 2259–2269. [CrossRef]
15. Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **2001**, *43*, 246–255. [CrossRef] [PubMed]
16. Yang, H.; Qiu, W.R.; Liu, G.Q.; Guo, F.B.; Chen, W.; Chou, K.C.; Lin, H. iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci.* **2018**, *14*, 883–891. [CrossRef] [PubMed]
17. Su, Z.D.; Huang, Y.; Zhang, Z.Y.; Zhao, Y.W.; Wang, D.; Chen, W.; Chou, K.C.; Lin, H. iLoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* **2018**. [CrossRef] [PubMed]

18. Zou, Q.; Wan, S.; Ju, Y.; Tang, J.; Zeng, X. Pretata: Predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* **2016**, *10*, 114. [CrossRef] [PubMed]
19. Tang, W.; Wan, S.; Yang, Z.; Teschendorff, A.E.; Zou, Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* **2018**, *34*, 398–406. [CrossRef] [PubMed]
20. Yang, H.; Tang, H.; Chen, X.X.; Zhang, C.J.; Zhu, P.P.; Ding, H.; Chen, W.; Lin, H. Identification of Secretory Proteins in *Mycobacterium tuberculosis* Using Pseudo Amino Acid Composition. *BioMed Res. Int.* **2016**, *2016*, 5413903. [CrossRef] [PubMed]
21. Chen, X.X.; Tang, H.; Li, W.C.; Wu, H.; Chen, W.; Ding, H.; Lin, H. Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. *BioMed Res. Int.* **2016**, *2016*, 1654623. [CrossRef] [PubMed]
22. Tang, H.; Zhao, Y.W.; Zou, P.; Zhang, C.M.; Chen, R.; Huang, P.; Lin, H. HBPred: A tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* **2018**, *14*, 957–964. [CrossRef] [PubMed]
23. Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinf. Comput. Biol.* **2005**, *3*, 185–205. [CrossRef]
24. Zou, Q.; Zeng, J.; Cao, L.; Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* **2016**, *173*, 346–354. [CrossRef]
25. Manavalan, B.; Shin, T.H.; Lee, G. DHSpred: Support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* **2018**, *9*, 1944–1956. [CrossRef] [PubMed]
26. Manavalan, B.; Basith, S.; Shin, T.H.; Choi, S.; Kim, M.O.; Lee, G. MLACP: Machine-learning-based prediction of anticancer peptides. *Oncotarget* **2017**, *8*, 77121–77136. [CrossRef] [PubMed]
27. Manavalan, B.; Lee, J. SVMQA: Support-vector-machine-based protein single-model quality assessment. *Bioinformatics* **2017**, *33*, 2496–2503. [CrossRef] [PubMed]
28. Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chen, W.; Chou, K.C. iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* **2018**. [CrossRef] [PubMed]
29. Lin, H.; Liang, Z.Y.; Tang, H.; Chen, W. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**. [CrossRef] [PubMed]
30. Chen, W.; Yang, H.; Feng, P.; Ding, H.; Lin, H. iDNA4mC: Identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* **2017**, *33*, 3518–3523. [CrossRef] [PubMed]
31. Tang, H.; Cao, R.Z.; Wang, W.; Liu, T.S.; Wang, L.M.; He, C.M. A two-step discriminated method to identify thermophilic proteins. *Int. J. Biomath.* **2017**, *10*, 1750050. [CrossRef]
32. Cao, R.; Wang, Z.; Wang, Y.; Cheng, J. SMOQ: A tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinform.* **2014**, *15*, 120. [CrossRef] [PubMed]
33. Ding, H.; Deng, E.Z.; Yuan, L.F.; Liu, L.; Lin, H.; Chen, W.; Chou, K.C. iCTX-type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Res. Int.* **2014**, *2014*, 286419. [CrossRef] [PubMed]
34. Lin, H.; Deng, E.Z.; Ding, H.; Chen, W.; Chou, K.C. iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* **2014**, *42*, 12961–12972. [CrossRef] [PubMed]
35. Guo, S.H.; Deng, E.Z.; Xu, L.Q.; Ding, H.; Lin, H.; Chen, W.; Chou, K.C. iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* **2014**, *30*, 1522–1529. [CrossRef] [PubMed]
36. Yang, H.; Lv, H.; Ding, H.; Chen, W.; Lin, H. iRNA-2OM: A sequence-based predictor for identifying 2′-*O*-methylation sites in Homo sapiens. *J. Comput. Biol.* **2018**. [CrossRef]
37. Zhao, Y.W.; Su, Z.D.; Yang, W.; Lin, H.; Chen, W.; Tang, H. IonchanPred 2.0: A Tool to Predict Ion Channels and Their Types. *Int. J. Mol. Sci.* **2017**, *18*, 1838. [CrossRef] [PubMed]
38. Li, D.; Ju, Y.; Zou, Q. Protein Folds Prediction with Hierarchical Structured SVM. *Curr. Proteom.* **2016**, *13*, 79–85. [CrossRef]
39. Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. AIPpred: Sequence-Based Prediction of Anti-inflammatory Peptides Using Random Forest. *Front. Pharmacol.* **2018**, *9*, 276. [CrossRef] [PubMed]
40. Manavalan, B.; Subramaniyam, S.; Shin, T.H.; Kim, M.O.; Lee, G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.* **2018**, *17*, 2715–2726. [CrossRef] [PubMed]

41. Zhang, C.J.; Tang, H.; Li, W.C.; Lin, H.; Chen, W.; Chou, K.C. iOri-Human: Identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* **2016**, *7*, 69783–69793. [CrossRef] [PubMed]
42. Chen, W.; Feng, P.M.; Lin, H.; Chou, K.C. iSS-PseDNC: Identifying splicing sites using pseudo dinucleotide composition. *BioMed Res. Int.* **2014**, *2014*, 623149. [CrossRef] [PubMed]
43. Chen, W.; Feng, P.M.; Deng, E.Z.; Lin, H.; Chou, K.C. iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.* **2014**, *462*, 76–83. [CrossRef] [PubMed]
44. Feng, P.M.; Chen, W.; Lin, H.; Chou, K.C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* **2013**, *442*, 118–125. [CrossRef] [PubMed]
45. Lai, H.Y.; Chen, X.X.; Chen, W.; Tang, H.; Lin, H. Sequence-based predictive modeling to identify cancerlectins. *Oncotarget* **2017**, *8*, 28169–28175. [CrossRef] [PubMed]
46. Feng, P.M.; Lin, H.; Chen, W. Identification of antioxidants from sequence information using naive Bayes. *Comput. Math. Method. Med.* **2013**, *2013*, 567529. [CrossRef] [PubMed]
47. Li, B.Q.; Zhang, Y.H.; Jin, M.L.; Huang, T.; Cai, Y.D. Prediction of Protein-Peptide Interactions with a Nearest Neighbor Algorithm. *Curr. Bioinform.* **2018**, *13*, 14–24. [CrossRef]
48. Naseem, I.; Khan, S.; Togneri, R.; Bennamoun, M. ECMSRC: A Sparse Learning Approach for the Prediction of Extracellular Matrix Proteins. *Curr. Bioinform.* **2017**, *12*, 361–368. [CrossRef]
49. Lin, Y.Q.; Min, X.P.; Li, L.L.; Yu, H.; Ge, S.X.; Zhang, J.; Xia, N.S. Using a Machine-Learning Approach to Predict Discontinuous Antibody-Specific B-Cell Epitopes. *Curr. Bioinform.* **2017**, *12*, 406–415. [CrossRef]
50. Kang, J.; Fang, Y.; Yao, P.; Li, N.; Tang, Q.; Huang, J. NeuroPP: A Tool for the Prediction of Neuropeptide Precursors Based on Optimal Sequence Composition. *Interdiscip. Sci.* **2018**. [CrossRef] [PubMed]
51. Li, N.; Kang, J.; Jiang, L.; He, B.; Lin, H.; Huang, J. PSBinder: A Web Service for Predicting Polystyrene Surface-Binding Peptides. *BioMed Res. Int.* **2017**, *2017*, 5761517. [CrossRef] [PubMed]
52. Zhu, P.P.; Li, W.C.; Zhong, Z.J.; Deng, E.Z.; Ding, H.; Chen, W.; Lin, H. Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. *Mol. Biosyst.* **2015**, *11*, 558–563. [CrossRef] [PubMed]
53. Li, W.C.; Deng, E.Z.; Ding, H.; Chen, W.; Lin, H. iORI-PseKNC: A predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. *Chemom. Intell. Lab. Syst.* **2015**, *141*, 100–106. [CrossRef]
54. Dao, F.Y.; Yang, H.; Su, Z.D.; Yang, W.; Wu, Y.; Hui, D.; Chen, W.; Tang, H.; Lin, H. Recent Advances in Conotoxin Classification by Using Machine Learning Methods. *Molecules* **2017**, *22*, 1057. [CrossRef] [PubMed]
55. Chen, W.; Xing, P.; Zou, Q. Detecting *N*(6)-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. *Sci. Rep.* **2017**, *7*, 40242. [CrossRef] [PubMed]
56. Peng, L.; Peng, M.M.; Liao, B.; Huang, G.H.; Li, W.B.; Xie, D.F. The Advances and Challenges of Deep Learning Application in Biological Big Data Processing. *Curr. Bioinform.* **2018**, *13*, 352–359. [CrossRef]
57. Patel, S.; Tripathi, R.; Kumari, V.; Varadwaj, P. DeepInteract: Deep Neural Network Based Protein-Protein Interaction Prediction Tool. *Curr. Bioinform.* **2017**, *12*, 551–557. [CrossRef]
58. Long, H.X.; Wang, M.; Fu, H.Y. Deep Convolutional Neural Networks for Predicting Hydroxyproline in Proteins. *Curr. Bioinform.* **2017**, *12*, 233–238. [CrossRef]
59. Cao, R.Z.; Adhikari, B.; Bhattacharya, D.; Sun, M.; Hou, J.; Cheng, J.L. QAcon: Single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics* **2017**, *33*, 586–588. [CrossRef] [PubMed]
60. Cao, R.; Freitas, C.; Chan, L.; Sun, M.; Jiang, H.; Chen, Z. ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules* **2017**, *22*, 1732. [CrossRef] [PubMed]
61. Zhang, T.; Tan, P.; Wang, L.; Jin, N.; Li, Y.; Zhang, L.; Yang, H.; Hu, Z.; Zhang, L.; Hu, C.; et al. RNALocate: A resource for RNA subcellular localizations. *Nucleic Acids Res.* **2017**, *45*, D135–D138. [PubMed]
62. Liang, Z.Y.; Lai, H.Y.; Yang, H.; Zhang, C.J.; Yang, H.; Wei, H.H.; Chen, X.X.; Zhao, Y.W.; Su, Z.D.; Li, W.C.; et al. Pro54DB: A database for experimentally verified sigma-54 promoters. *Bioinformatics* **2017**, *33*, 467–469. [CrossRef] [PubMed]

63. Cui, T.; Zhang, L.; Huang, Y.; Yi, Y.; Tan, P.; Zhao, Y.; Hu, Y.; Xu, L.; Li, E.; Wang, D. MNDR v2.0: An updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res.* **2018**, *46*, D371–D374. [CrossRef] [PubMed]
64. Yi, Y.; Zhao, Y.; Li, C.; Zhang, L.; Huang, H.; Li, Y.; Liu, L.; Hou, P.; Cui, T.; Tan, P.; et al. RAID v2.0: An updated resource of RNA-associated interactions across organisms. *Nucleic Acids Res.* **2017**, *45*, D115–D118. [CrossRef] [PubMed]