symmetry

MDPI

Article

iPVP-MCV: A Multi-Classifier Voting Model for the Accurate Identification of Phage Virion Proteins

Haitao Han, Wenhong Zhu, Chenchen Ding and Taigang Liu *

College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; m190711268@st.shou.edu.cn (H.H.); 1729223@st.shou.edu.cn (W.Z.); m190711300@st.shou.edu.cn (C.D.)
* Correspondence: tgliu@shou.edu.cn; Tel.: +86-21-61900624

Abstract: The classic structure of a bacteriophage is commonly characterized by complex symmetry. The head of the structure features icosahedral symmetry, whereas the tail features helical symmetry. The phage virion protein (PVP), a type of bacteriophage structural protein, is an essential material of the infectious viral particles and is responsible for multiple biological functions. Accurate identification of PVPs is of great significance for comprehending the interaction between phages and host bacteria and developing new antimicrobial drugs or antibiotics. However, traditional experimental approaches for identifying PVPs are often time-consuming and laborious. Therefore, the development of computational methods that can efficiently and accurately identify PVPs is desired. In this study, we proposed a multi-classifier voting model called iPVP-MCV to enhance the predictive performance of PVPs based on their amino acid sequences. First, three types of evolutionary features were extracted from the position-specific scoring matrix (PSSM) profiles to represent PVPs and non-PVPs. Then, a set of baseline models were trained based on the support vector machine (SVM) algorithm combined with each type of feature descriptors. Finally, the outputs of these baseline models were integrated to construct the proposed method iPVP-MCV by using the majority voting strategy. Our results demonstrated that the proposed iPVP-MCV model was superior to existing methods when performing the rigorous independent dataset test.

Keywords: phage virion protein; machine learning; support vector machine; position-specific scoring matrix

Citation: Han, H.; Zhu, W.; Ding, C.; Liu, T. iPVP-MCV: A Multi-Classifier Voting Model for the Accurate Identification of Phage Virion Proteins. *Symmetry* **2021**, *13*, 1506. https://doi.org/10.3390/sym13081506

Academic Editors: Jeffrey A. Thompson, Filip Jagodzinski and Ellen Palmer

Received: 8 July 2021
Accepted: 14 August 2021
Published: 17 August 2021

1. Introduction

The bacteriophages, also known informally as phages, are a form of viruses that infect and replicate within bacteria and archaea. Bacteriophages are among the most common and diverse entities on Earth, usually found wherever bacteria exist. Their interactions with microbial communities profoundly influence microbial ecology and biogeochemical cycling in various ecosystems [1]. The classic structure of a bacteriophage is commonly characterized by complex symmetry. The head of the structure features icosahedral symmetry, whereas the tail features helical symmetry. Recently, it has been shown that the abundant bacteriophages existing in the human gut microbiota heavily impact human metabolism and immunity [2], with evident therapeutic implications for some diseases [3]. Bacteriophages are composed of proteins that encapsulate a DNA or RNA genome [4]. They replicate within the bacterium following the injection of the genome into bacterial cytoplasm. Owing to their properties, no toxicity for human cells, harmless to normal flora, and their potential against antibiotic-resistant bacteria, phages are expected to become an alternative to antibiotics [5].

Bacteriophage proteins are fundamental materials of the infectious viral particles and are responsible for important biological functions in the interaction between bacteriophages and host cells. They are divided into two categories: structural proteins, also called phage virion proteins (PVPs), and nonstructural proteins, also called phage non-virion proteins

(non-PVPs). PVPs recognize the host, bind to its surface receptors, and deliver the phage's genome into the host's cell, while non-PVPs play crucial roles in the biological process of viral genome replication and expression [6–8]. The accurate identification of PVPs will not only contribute to further comprehending the molecular mechanisms of phage genetics but also be helpful in the development of antimicrobial drugs [9]. However, it is often quite expensive and time-consuming to annotate new PVPs based on the current experimental methods, such as mass spectrometry, protein arrays, and electron microscopy [10,11]. Therefore, with the explosive extension of protein sequence data, there is an urgent need to exploit computational methods to identify PVPs.

In a few years of recent, several sequence-based computational methods have been developed to predict PVPs, including PVPred [12], PVP-SVM [13], PhagePred [14], Pred-BVP-Unb [15], PVPred-SCM [16], Meta-iPVP [17], and so on [18–22]. For instance, Feng et al. [19] constructed a benchmark dataset with 99 PVPs and 208 non-PVPs and applied a naïve Bayes (NB) classifier to predict PVPs by using amino acid composition (AAC) and dipeptide composition (DPC). Based on jackknife cross-validation (CV), their method could distinguish the PVPs from the non-PVPs with an accuracy (ACC) of 79.15% [19]. In 2018, Manavalan et al. [13] developed a support vector machine (SVM)-based classifier, namely PVP-SVM, which extracted AAC, DPC, atomic composition (ATC), physicochemical properties (PCP), and chain–transition–distribution (CTD) as preliminary input features and then performed a feature selection protocol to identify the optimal features. PVP-SVM achieved an ACC of 79.80% when tested on the independent dataset [13]. Subsequently, Arif et al. [15] proposed an unbiased predictor called Pred-BVP-Unb for PVPs prediction, which utilized three protein encoding strategies, i.e., composition and translation (CT), split AAC, and bi-profile position-specific scoring matrix (PSSM). In addition, they applied the synthetic minority oversampling technique to reduce the bias of the unbalanced benchmark dataset and employed the recursive feature elimination algorithm to select the optimal feature subset [15]. Pred-BVP-Unb obtained the highest ACC value of 83.06% on the same independent dataset with PVP-SVM [15]. Recently, Charoenkwan et al. [16] developed an interpretable PVPs classifier named PVPred-SCM in a systematic manner based on the scoring card method combined with DPC, with an ACC of 77.66% on the independent dataset. Later, the same research group [17] first established a balanced benchmark dataset which contained 313 PVPs and 313 non-PVPs, and then explored a meta-predictor termed Meta-iPVP to distinguish whether a query protein is a PVP or not. Meta-iPVP adopted seven feature encodings and four machine learning algorithms to construct the final ensemble classifier, which performed well on the independent test. Many other reports about the computational identification of PVPs can be seen in the recent review articles [23–25].

As we all know, PSSM profiles, which represent the evolutionary information of protein sequence, have been widely applied and proved to have a remarkable contribution to protein function and attribute predictions [26]. However, it appears that evolutionary features derived from the PSSM profile have rarely been used to identify PVPs. Thus, there is still room for further improvement in PVPs prediction by extracting more effective evolutionary features from the PSSM profile. To this end, we present an ensemble model, called iPVP-MCV, to increase the annotation levels of PVPs by combining PSSM profiles and a multi-classifier voting scheme. The overall framework of the iPVP-MCV method is illustrated in Figure 1. First, three types of feature encodings were utilized to convert protein sequences into fixed-length feature vectors, including PSSM-AAC, PSSM-composition, and DP-PSSM. Second, these features were input into the baseline SVM classifiers to perform the prediction. Finally, iPVP-MCV integrated the outputs of these baseline models to predict whether the query protein was a PVP or not by using the majority voting strategy. Experimental results from the independent dataset test indicated that iPVP-MCV performed better than most of the current existing methods and could serve as a useful tool to help enhance the prediction performance of PVPs.

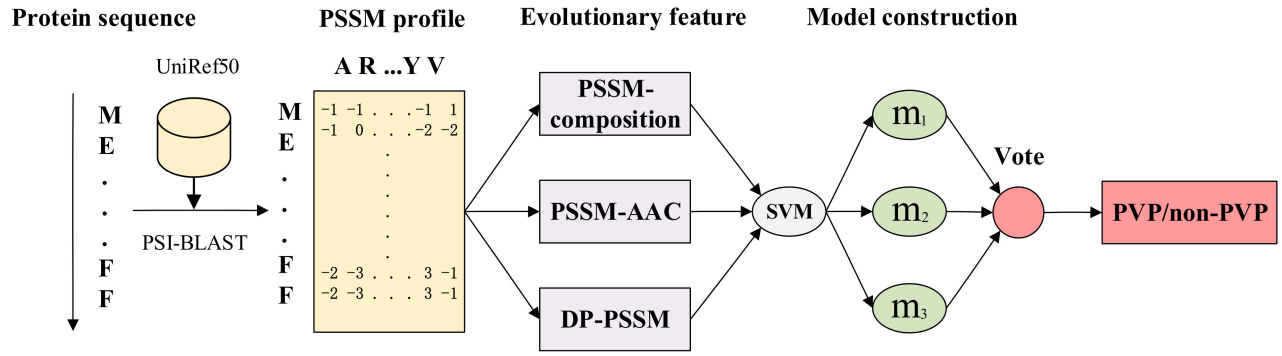


Figure 1. The overall framework of the proposed iPVP-MCV model. First, the query protein sequence was converted into its PSSM profile by running the PSI-BLAST program against the UniRef50 database. Second, three types of evolutionary features were extracted from the PSSM profile to characterize the query protein. Third, these features were input into the baseline SVM classifiers to perform the prediction. Finally, the outputs of three baseline models were integrated to predict whether the query protein was a PVP or not by using the majority voting strategy.

2. Materials and Methods

2.1. Datasets

A reliable and high-quality benchmark dataset is crucial for training and verifying the proposed prediction model. In this study, the benchmark dataset, which contained the training subset (termed D1_Train) and the independent subset (termed D1_Test), was directly collected from previous work to validate the proposed model [13]. Specifically, the training dataset contained 208 non-PVPs and 99 PVPs, while the independent dataset included 64 non-PVPs and 30 PVPs.

To objectively assess the robustness of the proposed method, another dataset constructed by Charoenkwan et al. [17] was further studied. This dataset first expanded the benchmark dataset and then randomly divided it into two parts: The first one included 250 PVPs and 250 non-PVPs (termed D2_Train), which was applied to train the model and perform the CV; and the other one consisted of 63 PVPs and 63 non-PVPs (termed D2_Test), which was used for the independent test.

The main reasons why we adopted the above two benchmark datasets were as follows: (i) they were extracted from the Universal Protein Resource (UniProt) [27]; (ii) they eliminated the sequences with nonstandard letters, such as "B", "X", or "Z", and (iii) they removed protein sequences with more than 40% similarity to avoid misleading results with the overestimated ACC.

2.2. Feature Extraction Algorithms

2.2.1. PSSM Profiles

A growing number of studies have found that evolutionary information embedded in the PSSM profiles highly represents the sequence information [28]. The PSSM profile of a query protein with the length of L can be represented as an $L \times 20$ matrix, which was generated by performing the PSI-BLAST search against the UniRef50 [29] database with three iterations and a threshold value of 0.001 [30]. The (i, j)th element in the matrix denotes the log odds of amino acid type j to appear at position i of the protein sequence. To facilitate the further analysis, the elements of the PSSM profile were normalized between 0 and 1 using the following formula:

$$f(x) = \frac{1}{1 + e^{-x}}. \quad (1)$$

The original and standardized PSSM profiles are respectively denoted as follows:

$$S = \left[s_{i,j}\right] \ (1 \le i \le L, 1 \le j \le 20), \quad (2)$$

$$M = \left[m_{i,j}\right] \ (1 \le i \le L, 1 \le j \le 20). \tag{3}$$

2.2.2. PSSM-AAC

AAC is a commonly used feature descriptor that computes the fractions of the 20 standard amino acid residues in each protein sequence [31]. The PSSM-AAC features derived from the PSSM profile can be defined as:

$$\left[\overline{M_1}, \overline{M_2}, \overline{M_3}, \ldots, \overline{M_{20}}\right], \tag{4}$$

where $\overline{M_j} = \frac{1}{L}\sum_{i=1}^{L} m_{i,j} (1 \le j \le 20)$.

Note that the AAC features based on the sequence (termed Seq-AAC) can be calculated by the following formula:

$$v_j = \frac{c_j}{L} (1 \le j \le 20), \tag{5}$$

where c_j is the number of amino acid of type j in the query protein sequence.

2.2.3. PSSM-Composition

PSSM-composition transforms the PSSM profile into a 20×20 matrix by summing up all rows for each naturally occurring amino acid type [32]:

$$R_i = \sum_{k=1}^{L} r_k \times \delta_k, \tag{6}$$

where

$$\begin{cases} \delta_k = 1, \text{ if } p_k = a_i \\ \delta_k = 0, \text{ if } p_k \ne a_i \end{cases} \quad (1 \le i \le 20), \tag{7}$$

R_i represents the ith row of the converted matrix, r_k denotes the kth row of the normalized PSSM profile, p_k denotes the kth amino acid in the original sequence, and a_i denotes the ith of 20 standard amino acids. Finally, a 400-dimensional numeric vector was generated by flattening the 20×20 matrix.

2.2.4. DP-PSSM

DP-PSSM is a similarity-based protein descriptor that reflects the hidden sequential order information in the real number and can avoid the cancellation of positive and negative terms in the average process [33]. The whole process is described below.

First, the original PSSM profile S was normalized to a matrix T by using the following three equations:

$$mean_i = \frac{1}{20}\sum_{k=1}^{20} s_{i,k}, \tag{8}$$

$$std_i = \sqrt{\frac{\sum_{k=1}^{20}(s_{i,k} - mean_i)^2}{20}}, \tag{9}$$

$$T_{i,j} = \frac{s_{i,j} - mean_i}{std_i}. \tag{10}$$

Second, a 40-dimensional feature vector T' was generated by computing the average of positive and negative terms in each column of the normalized matrix T, as shown in the Formulas (11) and (12):

$$T' = \left[\overline{T}_1^P, \overline{T}_1^N, \overline{T}_2^P, \overline{T}_2^N, \ldots, \overline{T}_{20}^P, \overline{T}_{20}^N\right], \tag{11}$$

where

$$\begin{cases} \overline{T}_j^P = \frac{1}{NP_j} \sum T_{i,j}, \text{if } T_{i,j} \geq 0 \\ \overline{T}_j^N = \frac{1}{NN_j} \sum T_{i,j}, \text{if } T_{i,j} < 0 \end{cases} \quad (1 \leq j \leq 20), \tag{12}$$

NP_j and NN_j represent the numbers of positive and negative terms in the jth column of the matrix T, respectively.

Third, a feature vector G' with the dimension of $20 \times (\alpha \times 2)$ was obtained by calculating the average of squared differences between entries corresponding to amino acids at the positions i and $i + k$ in each column of matrix T, as shown below:

$$G' = [G_1, G_2, \ldots, G_{20}], \tag{13}$$

where

$$G_j = \left[\overline{\Delta}_{1,j}^P, \overline{\Delta}_{1,j}^N, \overline{\Delta}_{2,j}^P, \overline{\Delta}_{2,j}^N, \ldots, \overline{\Delta}_{\alpha,j}^P, \overline{\Delta}_{\alpha,j}^N\right] \ (1 \leq j \leq 20), \tag{14}$$

$$\begin{cases} \overline{\Delta}_{k,j}^P = \frac{1}{NDP_j} \sum \left(T_{i,j} - T_{i+k,j}\right)^2, \text{if } T_{i,j} - T_{i+k,j} \geq 0 \\ \overline{\Delta}_{k,j}^N = \frac{-1}{NDN_j} \sum \left(T_{i,j} - T_{i+k,j}\right)^2, \text{if } T_{i,j} - T_{i+k,j} < 0 \end{cases} (1 \leq k \leq \alpha), \tag{15}$$

NDP_j and NDN_j are the numbers of positive and negative terms of $T_{i,j} - T_{i+k,j}$ $(1 \leq i \leq L)$, respectively.

Finally, the DP-PSSM descriptor was defined as a $40 + 40 \times \alpha$ dimensional feature vector by combining the generated T' and G'.

In this study, the value of the parameter α was set to 4, and we obtained a 200-dimensional numeric vector for each protein sequence. The experimental results of parameter selection are listed in Supplementary Table S1.

2.3. Support Vector Machine

Support vector machine (SVM), which is considered one of the most robust prediction algorithms [34], has been successfully used to solve the protein classification problems in bioinformatics [35,36]. Theoretically, SVM maps the training examples into a high-dimensional space and then finds an optimal hyperplane that maximized the margin between two classes. New test examples that are mapped into the same high-dimensional space will be predicted based on which side of the hyperplane they fall into. In addition to solving the linear classification problem, SVM can efficiently handle the non-linear problem by using the kernel trick. Due to its better performance, the radial basis kernel function (RBF) was adopted in this work. The effectiveness of the RBF-based SVM depends on the values of the kernel parameter γ and the regularization parameter C. In this study, to avoid overfitting to the testing set, we only performed a grid search method to select the best parameters $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}\}$ and $\gamma \in \{2^3, 2^5, \ldots, 2^{-13}, 2^{-15}\}$ of each SVM-based PVPs predictor on the training set by using the nested 10-fold CV.

2.4. Performance Evaluation

In this study, the proposed method was rigorously validated based on three widely used validation tests: the 10-fold CV, the jackknife CV, and the independent test. All results were reported by applying the four performance measurements: sensitivity (SN), specificity (SP), accuracy (ACC), and Matthew's correlation coefficient (MCC), formulated as follows:

$$\text{SN} = \frac{TP}{TP + FN}, \tag{16}$$

$$\text{SP} = \frac{TN}{TN + FP}, \tag{17}$$

$$\text{ACC} = \frac{TP + TN}{TP + FP + TN + FN}, \tag{18}$$

$$\mathrm{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}, \quad (19)$$

where TP, FP, TN, and FN represent the instances of true positive, false positive, true negative, and false negative, respectively. Moreover, we exploited a receiver operating characteristic (ROC) curve to illustrate the diagnostic ability of the proposed method. The ROC curve was created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Note that TPR and FPR were equal to SN and 1-SP, respectively. The area under the ROC curve (AUC) was also calculated for the model evaluation. The high value of AUC means that the model has a good performance.

3. Results and Discussion

3.1. Performance Comparison of Various Classifiers

In this section, five commonly used machine learning algorithms, including SVM, random forest (RF), extreme gradient boosting machine (XGB), extremely randomized trees (ERT), and NB, were adopted to compare their performance for the PVP's prediction by using three different feature encoding methods. All the experiments were performed on the D1_Train dataset by utilizing the jackknife CV, and the results are illustrated in Figures 2 and 3.

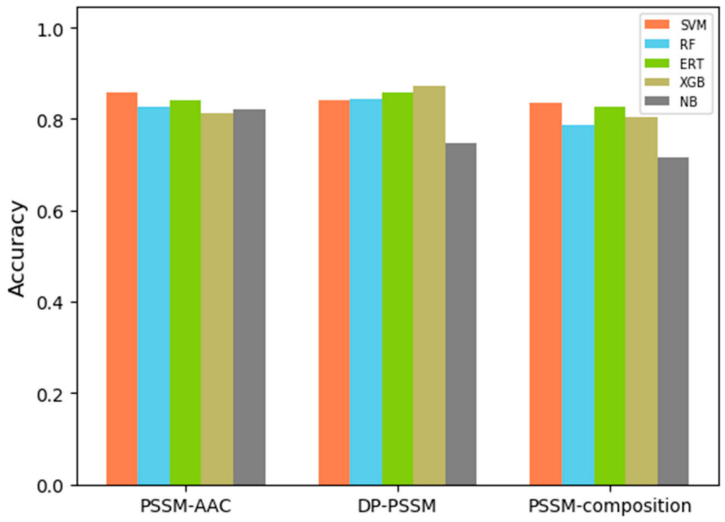
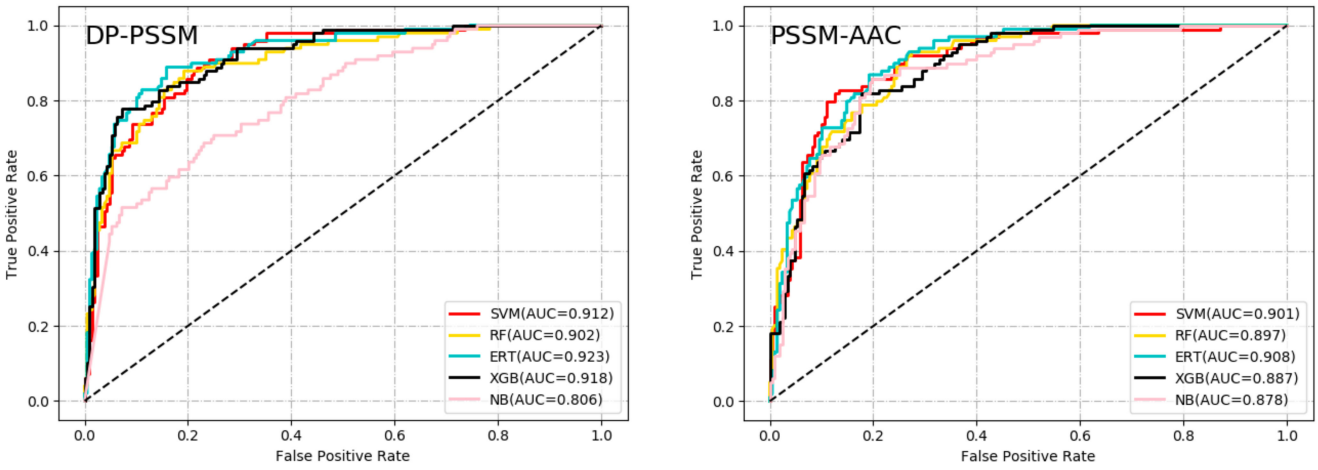
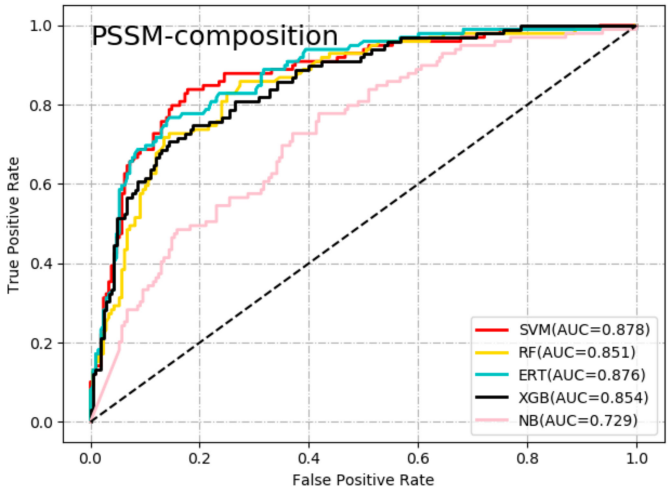


Figure 2. Prediction results of different classifiers with different features.



(**a**) (**b**)

Figure 3. *Cont.*



(**c**)

Figure 3. The ROC curves of different classifiers using different features. (**a**) ROC curves based on the DP-PSSM descriptor. (**b**) ROC curves based on the PSSM-AAC descriptor. (**c**) ROC curves based on the PSSM-composition descriptor.

As can be seen from Figure 2, the SVM, ERT, RF, and XGB classifiers obtained satisfactory ACC values with three different encoding schemes, while the NB algorithm performed worse in this task. Specifically, the SVM classifier reached the highest ACC values based on the PSSM-AAC and PSSM-composition descriptors and showed comparable performance with the XGB model when combined with the DP-PSSM encoding method. Accordingly, the SVM algorithm was selected as the final predictor in the subsequent analysis due to its robustness. In addition, the models based on the PSSM-AAC and DP-PSSM descriptors performed slightly better than the ones based on the PSSM-composition descriptor. This suggested that these three types of features may provide important clues for the identification of PVPs, and the predictive ability could be further improved by constructing the ensemble model. Furthermore, the ROC curves associated with these models were plotted in Figure 3, which led to similar conclusions as Figure 2.

3.2. Performance Comparison between the Base Models and the Ensemble Models

In this section, we assessed the performance of six SVM-based models on the D1_Train dataset by using the jackknife CV, including four base models with different feature encodings (i.e., Seq-AAC, PSSM-AAC, DP-PSSM, and PSSM-composition) and two ensemble models with different integration strategies (one was multi-feature fusion, and the other one, multi-classifier voting). The four metrics, including ACC, sensitivity (SN), specificity (SP), and Matthew's correlation coefficient (MCC), were applied to evaluate the performance of these models, and the corresponding results are reported in Table 1.

Table 1. Prediction results of different models on the D1_Train dataset.

Model	ACC	SN	SP	MCC
Seq-AAC	0.795	0.677	0.851	0.529
PSSM-AAC	0.857	0.798	0.885	0.676
DP-PSSM	0.840	0.687	0.913	0.625
PSSM-composition	0.834	0.737	0.880	0.619
All PSSM features	0.847	0.727	0.904	0.644
iPVP-MCV	0.879	0.778	0.928	0.720

As shown in Table 1, three PSSM-based models were superior to the one based on the Seq-AAC descriptor in terms of ACC, SN, SP, and MCC, which reconfirmed that the evolutionary information in the form of PSSM profiles may be more informative than the sequence information alone when tested on the PVPs identification. Among three PSSM-

based models, the PSSM-AAC model achieved the best ACC value (0.857), the DP-PSSM model obtained the highest SP value (0.913), and the PSSM-composition model could bear comparison with the DP-PSSM model in terms of SN. In addition, we attempted to further improve the prediction performance of PVPs by using two different ensemble methods. The first one constructed an SVM model by fusing three PSSM-based feature spaces, which outperformed the DP-PSSM model and the PSSM-composition model in terms of ACC and MCC but lagged behind the PSSM-AAC model. The second one (i.e., iPVP-MCV) combined the outputs of three individual SVM models into a final prediction by the majority vote, which performed better than the other predictors listed in Table 1 in terms of ACC (0.879), SP (0.928), and MCC (0.720). This suggested that the multi-classifier voting strategy was superior to the multi-feature fusion strategy for this task. The possible cause could be that either redundant or irrelevant features existed in the fused feature space and had an adverse effect on the prediction of PVPs.

3.3. Performance Comparison with Existing Methods

To objectively assess the performance of iPVP-MCV, we made fair comparisons with several existing methods on the two benchmark datasets by using the same validation schemes with previous works [13,17]. These methods included Feng et al.'s method [19], PVPred [12], PVP-SVM [13], Tan et al.'s method [21], PVP-SCM [16], and Meta-iPVP [17]. For this experiment, the jackknife CV was first performed on the D1_Train dataset to verify the performance of iPVP-MCV, and then the iPVP-MCV model trained with the D1_Train dataset was examined on the D1_Test dataset. Moreover, we carried out the 10-fold CV on the D2_Train dataset to validate the robustness of iPVP-MCV. The iPVP-MCV model trained with the D2_Train dataset was further evaluated on the D2_Test dataset. The prediction results of our method and several existing models were documented in Tables 2 and 3. Note that the prediction values were collected from the previous studies [15–17]. The results of two validation tests on the two training datasets were also listed in Supplementary Table S2 and the ROC curves of iPVP-MCV classifier on the independent testing datasets were illustrated in Supplementary Figure S1.

Table 2. Performance comparison on the D1_Train and D1_Test datasets.

Validation	Methods	ACC	SN	SP	MCC
Jackknife CV	Feng et al.'s method [19]	0.792	0.758	0.808	0.546
	PVPred	0.850	0.758	0.894	0.655
	PVP-SVM	0.870	0.737	0.933	0.695
	iPVP-MCV	0.879	0.778	0.928	0.720
Independent test	PVPred	0.713	0.600	0.765	0.357
	Tan et al.'s method [21]	0.755	0.700	0.781	0.464
	PVPred-SCM	0.777	0.767	0.781	0.523
	PVP-SVM	0.798	0.667	0.859	0.531
	iPVP-MCV	0.840	0.667	0.922	0.621

Table 3. Performance comparison on the D2_Train and D2_Test datasets.

Validation	Methods	ACC	SN	SP	MCC
10-fold CV	Meta-iPVP	0.846	0.860	0.832	0.698
	iPVP-MCV	0.864	0.852	0.876	0.728
Independent test	PVPred	0.730	0.892	0.663	0.505
	PVP-SVM	0.746	0.816	0.701	0.505
	PVPred-SCM	0.714	0.745	0.690	0.432
	Meta-iPVP	0.817	0.889	0.746	0.642
	iPVP-MCV	0.833	0.889	0.778	0.671

As shown in Table 2, the proposed iPVP-MCV model, together with the PVP-SVM model, exhibited the best performance with the ACC value higher than 0.87, the SP value better than 0.92, and the MCC value near or greater than 0.70 during the jackknife test on the D1_Train dataset. It is worth mentioning that the PVP-SVM predictor first extracted multiple features, including AAC, ATC, CTD, DPC, and PCP, to train an SVM model and then employed a feature selection protocol to remove the noisy features and enhance the predictive ability. Unlike the PVP-SVM predictor, our method utilized the majority voting strategy to improve the performance of base classifiers. The comparison results on the D1_Test dataset also validated the feasibility and effectiveness of two methods for the prediction of PVPs. In particular, the iPVP-MCV model achieved the highest ACC (0.840), SP (0.922), and MCC (0.621). Additionally, the best SN value was reached by the PVPred-SCM model, which adopted the scoring card method in conjunction with DPC to identify the PVPs.

To further evaluate the robustness of our method, two larger training and testing datasets were adopted to compare the iPVP-MCV method with the recent Meta-iPVP method and other existing tools. From Table 3, our model was slightly superior to the Meta-iPVP model in terms of ACC, SP, and MCC not only on the 10-fold CV but also on the independent test. The SN values of the two models were also level pegging. It should be noted that the Meta-iPVP model combined seven sequence-based feature encodings and four machine learning algorithms to construct a meta-predictor. These experiment results indicated that ensemble learning methods could indeed obtain better predictive performance than those based on the single feature encoding or the individual classifier. Additionally, our method showed substantial improvements for identifying PVPs, which may be attributed to the helpful evolutionary information extracted from the PSSM profile and the efficient multi-classifier voting technique. We anticipated that our method could become a useful tool or at least play a complementary role to existing methods for increasing the annotation levels of PVPs.

4. Conclusions

The medical and commercial value of PVPs has motivated the development of computational tools that facilitate the accurate annotation of PVPs based on their protein sequences. In this study, we proposed an ensemble model iPVP-MCV to further enhance the predictive accuracy of PVPs by using a multi-classifier voting strategy combined with evolutionary features extracted from the PSSM profile. First, three types of feature descriptors were designed to characterize all the PVPs and non-PVPs from the working datasets, including PSSM-AAC, PSSM-composition, and DP-PSSM. Second, a set of baseline models were trained based on each type of feature descriptor. Third, iPVP-MCV integrated the outputs of these baseline models to perform the prediction of PVPs by using the majority voting strategy. Finally, both the CV and the independent test were adopted to verify the performance of iPVP-MCV on the two benchmark datasets. The comparison results with the existing tools demonstrated that the proposed method was efficient, robust, and promising for the annotation of PVPs and could serve as an alternative tool to identify putative PVPs. The source code and all the datasets are freely available at https://github.com/taigangliu/iPVP-MCV, accessed on 14 August 2021. In the future, we will develop a user-friendly web server for the public use.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/sym13081506/s1, Table S1: The performance of DP-PSSM with different α on the D1_Train dataset using the 10-fold CV, Table S2: Prediction comparison on the D1_Train and D2_Train datasets by using the 10-fold CV and the jackknife CV, Figure S1: The ROC curves of iPVP-MCV on the independent testing datasets.

Author Contributions: Methodology, T.L.; software, H.H.; validation, C.D. and W.Z.; writing—original draft preparation, H.H.; writing—review and editing, T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number 11601324) and the National Key Research and Development Program of China (grant number 2018YFD0900600).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are freely available to the academic community at https://github.com/taigangliu/iPVP-MCV, accessed on 14 August 2021.

Acknowledgments: We thank the researchers for providing their datasets.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Auslander, N.; Gussow, A.B.; Benler, S.; Wolf, Y.I.; Koonin, E.V. Seeker: Alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Res.* **2020**, *48*, e121. [CrossRef] [PubMed]
2. Cani, P.D.; Possemiers, S.; Van de Wiele, T.; Guiot, Y.; Everard, A.; Rottier, O.; Geurts, L.; Naslain, D.; Neyrinck, A.; Lambert, D.M.; et al. Changes in gut microbiota control inflammation in obese mice through a mechanism involving GLP-2-driven improvement of gut permeability. *Gut* **2009**, *58*, 1091–1103. [CrossRef] [PubMed]
3. Tripathi, A.; Debelius, J.; Brenner, D.A.; Karin, M.; Loomba, R.; Schnabl, B.; Knight, R. The gut-liver axis and the intersection with the microbiome. *Nat. Rev. Gastroenterol. Hepatol.* **2018**, *15*, 397–411. [CrossRef]
4. Clark, J.R.; March, J.B. Bacteriophages and biotechnology: Vaccines, gene therapy and antibacterials. *Trends Biotechnol.* **2006**, *24*, 212–218. [CrossRef] [PubMed]
5. Lyon, J. Phage Therapy's Role in Combating Antibiotic-Resistant Pathogens. *JAMA* **2017**, *318*, 1746–1748. [CrossRef] [PubMed]
6. Aguilar, P.V.; Adams, A.P.; Wang, E.; Kang, W.; Carrara, A.-S.; Anishchenko, M.; Frolov, I.; Weaver, S.C. Structural and nonstructural protein genome regions of eastern equine encephalitis virus are determinants of interferon sensitivity and murine virulence. *J. Virol.* **2008**, *82*, 4920–4930. [CrossRef]
7. Moreland, N.J.; Tay, M.Y.F.; Lim, E.; Paradkar, P.N.; Doan, D.N.P.; Yau, Y.H.; Shochat, S.G.; Vasudevan, S.G. High Affinity Human Antibody Fragments to Dengue Virus Non-Structural Protein 3. *PLoS Negl. Trop. Dis.* **2010**, *4*, e881. [CrossRef] [PubMed]
8. Cantu, V.A.; Salamon, P.; Seguritan, V.; Redfield, J.; Salamon, D.; Edwards, R.A.; Segall, A.M. PhANNs, a fast and accurate tool and web server to classify phage structural proteins. *PLoS Comput. Biol.* **2020**, *16*, e1007845. [CrossRef]
9. Lekunberri, I.; Subirats, J.; Borrego, C.M.; Balcazar, J.L. Exploring the contribution of bacteriophages to antibiotic resistance. *Environ. Pollut.* **2017**, *220*, 981–984. [CrossRef]
10. Jara-Acevedo, R.; Diez, P.; Gonzalez-Gonzalez, M.; Degano, R.M.; Ibarrola, N.; Gongora, R.; Orfao, A.; Fuentes, M. Screening Phage-Display Antibody Libraries Using Protein Arrays. *Methods Mol. Biol.* **2018**, *1701*, 365–380. [CrossRef]
11. Lavigne, R.; Ceyssens, P.-J.; Robben, J. Phage proteomics: Applications of mass spectrometry. *Methods Mol. Biol.* **2009**, *502*, 239–251. [CrossRef]
12. 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]
13. 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]
14. 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]
15. Arif, M.; Ali, F.; Ahmad, S.; Kabir, M.; Ali, Z.; Hayat, M. Pred-BVP-Unb: Fast prediction of bacteriophage Virion proteins using un-biased multi-perspective properties with recursive feature elimination. *Genomics* **2020**, *112*, 1565–1574. [CrossRef] [PubMed]
16. Charoenkwan, P.; Kanthawong, S.; Schaduangrat, N.; Yana, J.; Shoombuatong, W. PVPred-SCM: Improved Prediction and Analysis of Phage Virion Proteins Using a Scoring Card Method. *Cells* **2020**, *9*, 353. [CrossRef]
17. Charoenkwan, P.; Nantasenamat, C.; Hasan, M.M.; Shoombuatong, W. Meta-iPVP: A sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. *J. Comput. Aided Mol. Des.* **2020**, *34*, 1105–1116. [CrossRef] [PubMed]
18. Seguritan, V.; Alves, N., Jr.; Arnoult, M.; Raymond, A.; Lorimer, D.; Burgin, A.B., Jr.; Salamon, P.; Segall, A.M. Artificial Neural Networks Trained to Detect Viral and Phage Structural Proteins. *PLoS Comput. Biol.* **2012**, *8*, e1002657. [CrossRef]
19. Feng, P.-M.; Ding, H.; Chen, W.; Lin, H. Naive Bayes Classifier with Feature Selection to Identify Phage Virion Proteins. *Comput. Math. Methods Med.* **2013**, *2013*, 530696. [CrossRef]
20. 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]
21. Tan, J.-X.; Dao, F.-Y.; Lv, H.; Feng, P.-M.; Ding, H. Identifying Phage Virion Proteins by Using Two-Step Feature Selection Methods. *Molecules* **2018**, *23*, 2000. [CrossRef]

22. Ru, X.Q.; Li, L.H.; Wang, C.Y. Identification of Phage Viral Proteins With Hybrid Sequence Features. *Front. Microbiol.* **2019**, *10*, 507. [CrossRef]
23. Yang, Y.; Fan, C.; Zhao, Q. Recent Advances on the Machine Learning Methods in Identifying Phage Virion Proteins. *Curr. Bioinform.* **2020**, *15*, 657–661. [CrossRef]
24. Meng, C.; Zhang, J.; Ye, X.; Guo, F.; Zou, Q. Review and comparative analysis of machine learning-based phage virion protein identification methods. *Biochim. Biophys. Acta-Proteins Proteom.* **2020**, *1868*, 140406. [CrossRef] [PubMed]
25. Chen, W.; Nie, F.; Ding, H. Recent Advances of Computational Methods for Identifying Bacteriophage Virion Proteins. *Protein Pept. Lett.* **2020**, *27*, 259–264. [CrossRef]
26. Wang, J.; Dai, W.; Li, J.; Xie, R.; Dunstan, R.A.; Stubenrauch, C.; Zhang, Y.; Lithgow, T. PaCRISPR: A server for predicting and visualizing anti-CRISPR proteins. *Nucleic Acids Res.* **2020**, *48*, W348–W357. [CrossRef]
27. Bateman, A.; Martin, M.J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bursteinas, B.; et al. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef]
28. Wang, J.; Yang, B.; Revote, J.; Leier, A.; Marquez-Lago, T.T.; Webb, G.; Song, J.; Chou, K.-C.; Lithgow, T. POSSUM: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics* **2017**, *33*, 2756–2758. [CrossRef] [PubMed]
29. Suzek, B.E.; Wang, Y.; Huang, H.; McGarvey, P.B.; Wu, C.H.; UniProt, C. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015**, *31*, 926–932. [CrossRef] [PubMed]
30. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.H.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
31. Liu, T.; Zheng, X.; Wang, J. Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. *Biochimie* **2010**, *92*, 1330–1334. [CrossRef] [PubMed]
32. Zou, L.; Nan, C.; Hu, F. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* **2013**, *29*, 3135–3142. [CrossRef]
33. Juan, E.Y.T.; Li, W.J.; Jhang, J.H.; Chiu, C.H. Predicting Protein Subcellular Localizations for Gram-Negative Bacteria using DP-PSSM and Support Vector Machines. In Proceedings of the 2009 International Conference on Complex, Intelligent and Software Intensive Systems, Fukuoka, Japan, 16–19 March 2009; pp. 836–841.
34. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef] [PubMed]
35. Garg, A.; Singhal, N.; Kumar, R.; Kumar, M. mRNALoc: A novel machine-learning based in-silico tool to predict mRNA subcellular localization. *Nucleic Acids Res.* **2020**, *48*, W239–W243. [CrossRef] [PubMed]
36. Bressin, A.; Schulte-Sasse, R.; Figini, D.; Urdaneta, E.C.; Beckmann, B.M.; Marsico, A. TriPepSVM: De novo prediction of RNA-binding proteins based on short amino acid motifs. *Nucleic Acids Res.* **2019**, *47*, 4406–4417. [CrossRef] [PubMed]