frontiers in Microbiology

ORIGINAL RESEARCH
published: 16 March 2018
doi: 10.3389/fmicb.2018.00476

PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine

Balachandran Manavalan[1], *Tae H. Shin*[1,2] *and Gwang Lee*[1,2]*

[1] Department of Physiology, Ajou University School of Medicine, Suwon, South Korea, *[2] Institute of Molecular Science and Technology, Ajou University, Suwon, South Korea*

Accurately identifying bacteriophage virion proteins from uncharacterized sequences is important to understand interactions between the phage and its host bacteria in order to develop new antibacterial drugs. However, identification of such proteins using experimental techniques is expensive and often time consuming; hence, development of an efficient computational algorithm for the prediction of phage virion proteins (PVPs) prior to *in vitro* experimentation is needed. Here, we describe a support vector machine (SVM)-based PVP predictor, called PVP-SVM, which was trained with 136 optimal features. A feature selection protocol was employed to identify the optimal features from a large set that included amino acid composition, dipeptide composition, atomic composition, physicochemical properties, and chain-transition-distribution. PVP-SVM achieved an accuracy of 0.870 during leave-one-out cross-validation, which was 6% higher than control SVM predictors trained with all features, indicating the efficiency of the feature selection method. Furthermore, PVP-SVM displayed superior performance compared to the currently available method, PVPred, and two other machine-learning methods developed in this study when objectively evaluated with an independent dataset. For the convenience of the scientific community, a user-friendly and publicly accessible web server has been established at www.thegleelab.org/PVP-SVM/PVP-SVM.html.

Keywords: bacteriophage virion proteins, feature selection, hybrid features, machine learning, support vector machine

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Wei Chen,
North China University of Science and Technology, China

****Correspondence:***
Gwang Lee
glee@ajou.ac.kr

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *07 December 2017*
Accepted: *28 February 2018*
Published: *16 March 2018*

Citation:
Manavalan B, Shin TH and Lee G (2018) PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. Front. Microbiol. 9:476. doi: 10.3389/fmicb.2018.00476

INTRODUCTION

Bacteriophages, also known as phages, are viruses that can infect and replicate in bacteria, and are found wherever bacteria survive. The phage virion is composed of proteins that encapsulate either DNA or RNA, which binds to bacterial surface and injects its genetic materials into the specific host bacteria. In lytic cycle, phage genes are expressed for proteins that poke hole in the cell membrane, which makes cell expand and burst. Subsequently, released phages from cell bursting spread and infects other host cells. Identification of phage virion proteins (PVPs) is important for understanding the relationship between phage and host bacteria and also development of novel antibacterial drugs or antibiotics (Lekunberri et al., 2017). For instance, phage encoded proteins including endolysins, exopolysaccharidases, and holins have been proven as promising antibacterial products (Drulis-Kawa et al., 2012). Experimental methods, including mass spectrometry, sodium dodecyl sulfate polyacrylamide gel electrophoresis, and protein arrays (Lavigne et al., 2009; Yuan and Gao, 2016; Jara-Acevedo et al., 2018) have been used to identify PVPs. However, these methods are expensive and often time-consuming. Therefore, computational methods to predict PVPs prior to *in vitro* experimentation are needed. It is difficult to predict the function of PVPs from sequence information because of relatively limited experimental data. However, machine-learning (ML) approaches have been successfully applied to several similar biological problems. Therefore, it may be possible to predict the functions of phage proteins using ML.

To this end, Seguritan et al., developed the first method to classify viral structure proteins using an artificial neural network, using amino acid composition (AAC) and protein isoelectric points as input features (Seguritan et al., 2012). Later, Feng et al., developed a naïve Bayesian method, with an algorithm utilizing AAC and dipeptide composition (DPC) as input features (Feng et al., 2013b). Subsequently, Ding et al., developed a support vector machine (SVM)-based prediction model called PVPred. In this method, analysis of variance was applied to select important features from g-gap DPC (Ding et al., 2014). Recently, Zhang et al., developed a random forest (RF)-based ensemble method to distinguish PVPs and non-PVPs (Zhang et al., 2015). PVPred is the only existing publicly available method that was developed using the same dataset as our method. Although the existing methods have specific advantages in PVPs prediction, it remains necessary to improve the accuracy and transferability of the prediction model.

It is worth mentioning that several sequence-based features including AAC, atomic composition (ATC), chain-transition-distribution (CTD), DPC, pseudo amino acid composition and amino acid pair, and several feature selection techniques including correlation-based feature selection, ANOVA feature selection, minimum-redundancy and maximum-relevance, RF-algorithm based feature selection have been successfully applied in other protein bioinformatics studies (Wang et al., 2012, 2016; Lin et al., 2015; Qiu et al., 2016; Tang et al., 2016; Gupta et al., 2017; Manavalan and Lee, 2017; Manavalan et al., 2017; Song et al., 2017). All these studies motivated us in the development of a new model in this study. Hence, we developed a SVM-based PVP predictor called PVP-SVM, in which the optimal features were selected using a feature selection protocol that has been successfully applied to various biological problems (Manavalan and Lee, 2017). We selected the optimal features from a large set, including AAC, DPC, CTD, ATC, and PCP. In addition to SVM (i.e., PVP-SVM), we also developed RF and extremely randomized tree (ERT)-based methods. The performance of PVP-SVM was consistent in both the training and independent datasets, and was superior to the current method and the RF and ERT methods developed in this study.

MATERIALS AND METHODS

Training Dataset

In this study, we utilized the dataset constructed by Ding et al., which was specifically used for studying PVPs (Ding et al., 2014). We decided to use this dataset for the following reasons: (i) it is a reliable dataset, constructed based on several filtering schemes; (ii) it is a non-redundant dataset and none of the sequences possesses pairwise sequence identity (>40%) with any other sequence. Hence, this dataset stringently excludes homologous sequences; and (iii) most importantly, it facilitates fair comparison between the current method and existing methods, which were developed using the same training dataset. Thus, the training dataset can be formulated as:

$$\boldsymbol{S} = \boldsymbol{S}^{+} \cup \boldsymbol{S}^{-} \tag{1}$$

where the positive subset $\boldsymbol{S}^{+}$ contained 99 PVPs, the negative subset $\boldsymbol{S}^{-}$ contained 208 non-PVPs, and the symbol ∪ denotes union in the set theory. Thus, *S* contained 307 samples.

Independent Dataset

We obtained PVP and non-PVP sequences from the Universal Protein Resource (UniProt) as previously described (Feng et al., 2013b; Ding et al., 2014; Zhang et al., 2015). To avoid overestimation in the prediction model, we excluded sequences that shared greater than 40% sequence identity with sequences in the training dataset. The final dataset contained 30 PVPs and 64 non-PVPs. We note that our independent dataset included Ding et al., independent dataset. The above two datasets can be downloaded from our web server.

Input Features

(i) AAC: The fractions of the 20 naturally occurring amino acid residues in a given protein sequence were calculated as follows:

$$\text{AAC}\,(i) = \frac{\textit{Frequency of amino acid}\,(i)}{\textit{Length of the protein sequence}} \tag{2}$$

where *i* can be any of the 20 natural amino acids.

(ii) ATC: The fraction of five atom types (C, H, N, O, and S) in a given protein sequence was calculated as previously reported (Kumar et al., 2015; Manavalan et al., 2017), with a fixed length of five features.

(iii) CTD: The global composition feature encoding method CTD comprises properties such as hydrophobicity, polarity, normalized van der Waals volume, polarizability, predicted secondary structure, and solvent accessibility. It was first proposed in protein folding class prediction (Dubchak et al., 1995). Composition (C) represents the composition percentage of each group in the peptide sequence. Transition (T) represents the transition probability between two neighboring amino acids belonging to two different groups. Distribution (D) represents the position of amino acids (the first 25, 50, 75, or 100%) in each group in the protein sequence. For each qualitative property of a given sequence, C, T, and D produce 3, 3, and 15-dimension features, respectively. As a result, $7 \times (3 + 3 + 15) = 147$ features can be generated for seven qualitative properties.

(iv) DPC: The fractions of the 400 possible dipeptides present in a given protein sequence were calculated as follows:

$$\text{DPC}(j) = \frac{\textit{Total number of dipeptide}\,(j)}{\textit{Total number of all possible dipeptides}} \tag{3}$$

where j can be any of the 400 possible dipeptides.

(v) PCP: We employed 11 representative PCP attributes of amino acids for feature extraction (polar, hydrophobic, charged, aliphatic, aromatic, positively charged, negatively charged, small, tiny, large, and peptide mass).

Note that all of the above features were in the range of [0, 1] as input for training and testing.

The Support Vector Machine

We employed a SVM as our classification algorithm, a well-known supervised ML method introduced in Boser et al. (1992) that has been applied to several biological problems (Wang et al., 2009; Eickholt et al., 2011; Deng et al., 2013; Cao et al., 2014; Manavalan et al., 2015). The objective of a SVM is to find the hyperplane with the largest margin to decrease the misclassification rate. Given a set of data points (input features) and an objective function associated with the data points (PVPs: 1 and non-PVPs: 0), SVM learn a function in the form of

$$y = \text{sign}\left(\sum\nolimits_{i=1}^{n} \alpha_i \, y_i \, K(x_i, x) + b\right) \qquad (4)$$

where y is the predicted class associated with an input feature vector of x; α_i is the adjustable weight assigned to the training data point x_i during training by minimizing a quadratic objective function; b is the bias term; and K is the Kernel function. Therefore, y can be viewed as a weighted linear combination of similarities between the training data points x_i and the target data point x. Data points with positive weights in the training dataset affect the final solution and are called support vectors. SVM is especially effective when the input data are not linearly separable. K is required to map the input data into a higher dimensional space to identify the optimal separating hyperplane (Scholkopf and Smola, 2001). Therefore, we experimented with several common Ks, including linear, Gaussian radial basis, and polynomial functions. The Gaussian radial basis K ($e^{(-\gamma \times \|x-y\|^2)}$; $\gamma = \frac{1}{\sigma^2}$) performed the best. Here, two critical parameters (γ and C) required optimization: γ controls how peaked Gaussians are centered on the support vectors, while C controls the trade-off between the training error and the margin size (Smola and Vapnik, 1997; Vapnik and Vapnik, 1998; Scholkopf and Smola, 2001). These two parameters were optimized using a grid search from 2^{-15}–2^{10} for C and 2^{-10}–2^{10} for γ, in $\log_2$ steps. In this study, we used a SVM implemented in the scikit-learn package (Pedregosa et al., 2011).

Cross-Validation and Independent Testing

As demonstrated in a series of studies (Feng et al., 2013a,c, 2018; Chen et al., 2014, 2017a,b), among three cross-validation methods, i.e., independent dataset test, K-fold cross-validation test and Leave-one-out cross-validation (LOOCV, also called jackknife cross validation), LOOCV is the most rigorous and objective evaluation methods. Accordingly, the jackknife test has been widely recognized and increasingly used to test the quality for various predictors. In LOOCV, each sample in the training dataset is in turn singled out as an independent test sample and all the rule parameters are calculated without including the one being identified. We performed LOOCV on the training dataset and the trained model was tested on the independent dataset to confirm the generality of the developed method.

Performance Evaluation Criteria

The following four metrics are commonly used in literature to measure the quality of binary classification (Xiong et al., 2012; Li et al., 2015): sensitivity, specificity, accuracy and Matthews' correlation coefficient (MCC), which are expressed as

$$\begin{cases} \mathit{Sensitivity} = \frac{TP}{TP + FN} \\ \mathit{Specificity} = \frac{TN}{TN + FP} \\ \mathit{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \end{cases} \qquad (5)$$

where TP is the number of PVPs predicted to be PVPs; TN is number of non-PVPs predicted to be non-PVP; FP is the number of non-PVPs predicted to be PVP; and FN is the number of PVPs predicted to be non-PVP.

To further evaluate the performance of the classifier, we employed a receiver operating characteristic (ROC) curve. The ROC curve was plotted with the false positive rate as the x-axis and true positive rate as the y-axis by varying the thresholds. The area under the curve (AUC) was used for model evaluation, with higher AUC values corresponding to better performance of the classifier.

RESULTS

Framework of the Proposed Predictor

Figure 1 illustrates the overall framework of the PVP-SVM method. It consisted of four steps: (i) construction of the training and independent datasets; (ii) extraction of various features from the primary sequences, including AAC, ATC, CTD, DPC, and PCP; (iii) generation of 25 different feature sets based on feature importance scores (FIS) computed using the RF algorithm. These different sets were inputted to the SVM to develop their respective prediction models; and (iv) the model producing the best performance in terms of MCC was considered the final model, and the corresponding feature set was considered the optimal feature set.

Feature Selection Protocol

Generally, high dimensional features can contain a higher degree of irrelevant and redundant information that may greatly degrade the performance of ML algorithms. Therefore, it is necessary to apply a feature selection protocol to filter the redundant features and increase prediction efficiency (Wang et al., 2012; Zheng et al., 2012; Manavalan et al., 2014; Manavalan and Lee, 2017; Song et al., 2017). Previously, Manavalan and Lee applied a systematic feature selection protocol and developed a novel quality assessment method called SVMQA (Manavalan and Lee, 2017), which was the best method in CASP12 blind prediction experiments (Elofsson et al., 2017; Kryshtafovych et al., 2017). We applied a similar protocol in our recent studies, including cell-penetrating peptide

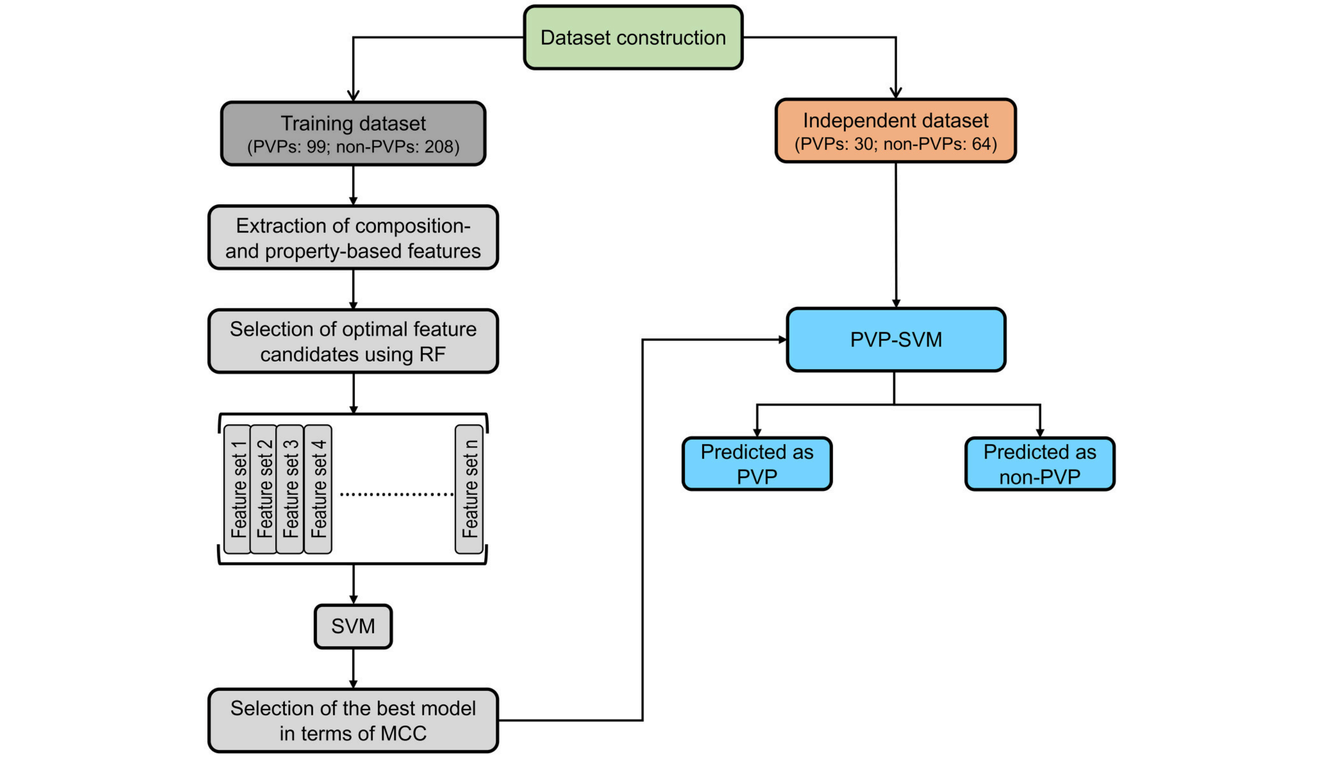
Dataset construction
Training dataset (PVPs: 99; non-PVPs: 208)
Independent dataset (PVPs: 30; non-PVPs: 64)
Extraction of composition- and property-based features
Selection of optimal feature candidates using RF
Feature set 1
Feature set 2
Feature set 3
Feature set 4
Feature set n
SVM
Selection of the best model in terms of MCC
PVP-SVM
Predicted as PVP
Predicted as non-PVP

FIGURE 1 | PVP-SVM development consisted of four steps: (i) dataset construction; (ii) feature extraction; (iii) development of the prediction model; and (iv) selection of the best model and construction of PVP-SVM.

and DNase I hypersensitivity predictions (Manavalan et al., 2018). Interestingly, this protocol significantly improved the performance of our method. Therefore, we extended this approach to the current problem. The current protocol differs slightly from the published protocol in terms of parameters (*ntree* and *mtry*) used in the RF algorithm, which is mainly due to the large number of features used in this study (i.e., 26-fold more features than were used in SVMQA).

In our study, each protein sequence was represented as 583 dimensional vectors, which was higher than the number of samples. In the first step, we applied the RF algorithm and estimated the FIS of 583 features (AAC: 20; DPC: 400; ATC: 5; PCP: 11; and CTD: 147) to distinguish PVPs and non-PVPs. A detailed description of how we computed the FIS scores of the input features has been reported previously (Manavalan et al., 2014; Manavalan and Lee, 2017). Briefly, we used all features as inputs in the RF algorithm and performed ten-fold cross-validation using the training dataset. For each round of cross-validation, we built 5,000 trees, and the number of variables at each node was chosen randomly from 1 to 100. The average FIS from all the trees are shown in **Figure 2A**, where most of the features had similar scores and only ∼5% (FIS $\geq$ 0.005) contributed significantly to PVP prediction. In the second step, we applied a FIS cutoff $\geq$ 0.001 and selected 477 features as optimal feature candidates (**Figure 2B**). Subsequently, we generated 25 different sets of features from the optimal feature candidates based on an FIS cut-off (0.001 $\leq$ FIS $\leq$ 0.004, with a step size of 0.0011). Basically, we considered each set of more important features in a step-wise manner. To identify the optimal feature set, we inputted each set into the SVM separately and performed LOOCV to evaluate their performance. The prediction model that produced the best performance (i.e., the highest MCC) was considered final, and the corresponding feature set was considered optimal.

Performance of Various Prediction Models on the Training Dataset

Figure 3A shows the performances of the SVM model using different sets of input features, in which the MCC gradually increased with respect to the different feature sets, peaked with the F136-based model, and then gradually declined. **Figure 3B** shows the classification accuracy vs. parameter variation (C and γ) of the final F136-based model. The maximal classification accuracy was 0.870, when the parameters $\log_2(C)$ and $\log_2(\gamma)$ were 6.72 and −2.18, respectively, with MCC, sensitivity, and specificity values of 0.695, 0.737, and 0.933, respectively. The feature type distribution of the optimal feature set and the total features employed in this study are shown in **Figure 3C**. Among 136 optimal features, there were eight AAC features, one ATC feature, 25 CTD features, 98 DPC features, and four PCP features, indicating that important properties from all five compositions contributed to PVP prediction.

To demonstrate the effect of our feature selection protocol, we compared the F136-based model with the

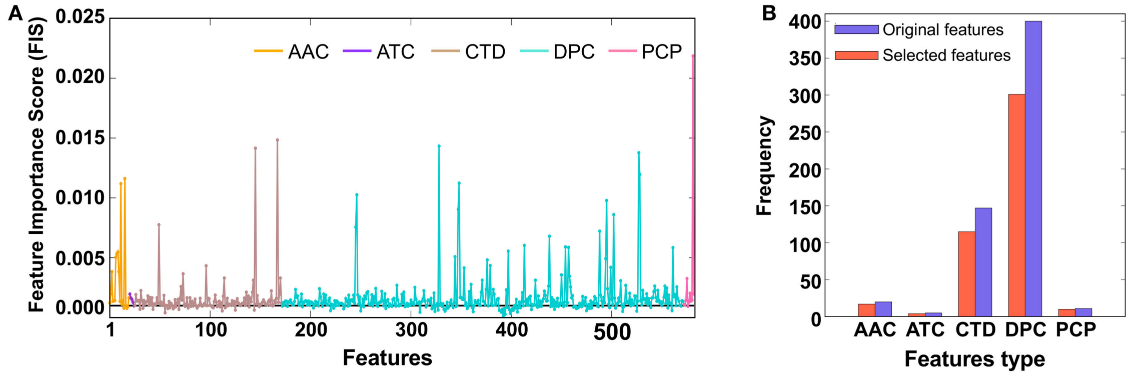

FIGURE 2 | (A) The x- and y-axes represent each feature and its feature importance scores (FIS), respectively. We applied a FIS cutoff ≥ 0.001 and selected 477 optimal feature candidates. **(B)** Distribution of each feature type in the optimal feature candidates and original feature set.

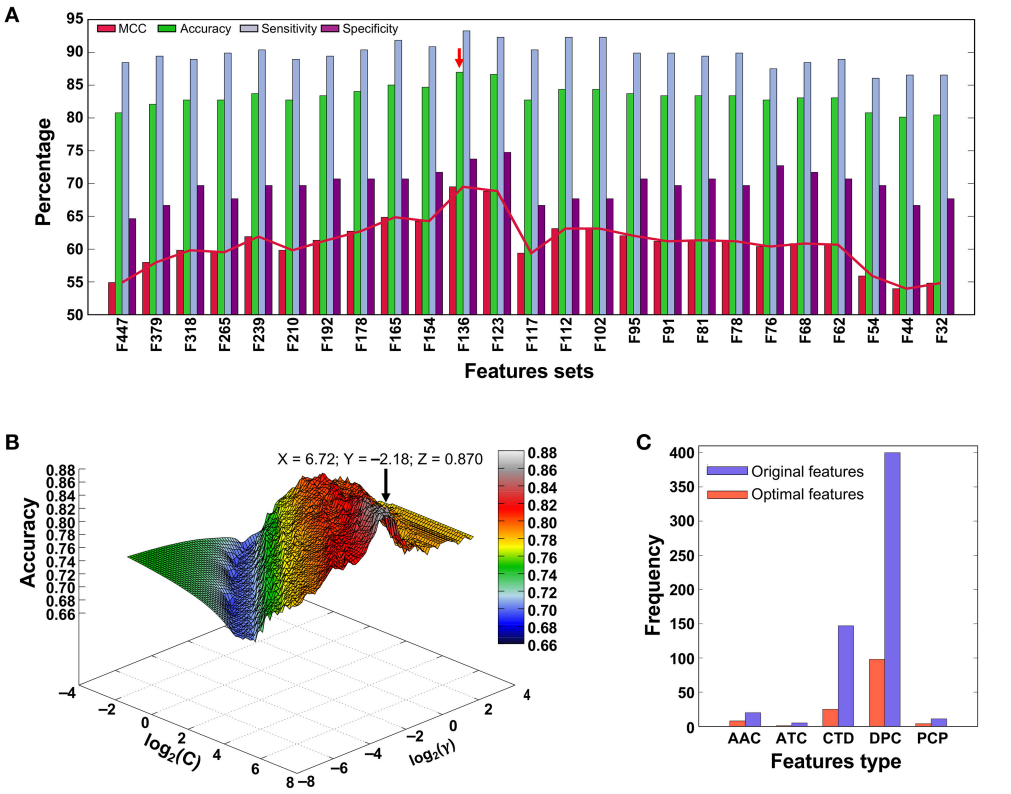

FIGURE 3 | (A) Performance of SVM-based classifiers in distinguishing between PVPs and non-PVPs. A total of 25 classifiers were evaluated using LOOCV, and their performances in terms of MCC, accuracy, sensitivity, and specificity are shown. The red arrow denotes the final selected model. **(B)** Classification accuracy of the F136-based (final selected) model with respect to variations in parameters *C* and γ. **(C)** Distribution of each feature type in the optimal feature set (136 features) and original feature set (583 features).

control SVM (using all features) and also an individual composition-based prediction model. As shown in **Table 1**, F136-based model accuracy, MCC, and area under curve (AUC) were 15–44, 6–17, and 6–11% higher, respectively, than the other models. These results demonstrate that the many redundant or uninformative features present in the original feature set were eliminated through our feature selection protocol, resulting in significant performance improvement.

Comparison of PVP-SVM With Other ML Algorithms

In addition to PVP-SVM, we also developed RF- and ERT-based models using the same feature selection protocol and training dataset (**Figures 4A,B**). These two methods have been described in detail in our previous study (Manavalan et al., 2017, 2018). The procedure for ML parameter optimization and final model selection was the same as for PVP-SVM. The performance of the final selected RF and ERT models was compared with PVP-SVM, as well as PVPred, which was constructed using the same training dataset. **Table 2** shows that the accuracy, AUC, and MCC of PVP-SVM were 2–4, 0.1–2, and 8–9% higher, respectively, than those achieved by other methods, indicating the superiority of PVP-SVM.

Method Performance Using an Independent Dataset

We evaluated the performance of our three ML methods and PVPred using an independent dataset. **Table 3** shows that PVP-SVM achieved the highest MCC and AUC values (0.531 and 0.844, respectively). Indeed, the corresponding metrics were 2.2–17.4% and 4.8–10.0% higher than those achieved by other methods, indicating the superiority of PVP-SVM. Specifically, PVP-SVM outperformed PVPred in all five metrics,

TABLE 1 | A comparison of the proposed predictor with the individual composition-based SVM model on training dataset.

Methods	MCC	Accuracy	Sensitivity	Specificity	AUC	*P*-value
PVP-SVM	0.695	0.870	0.737	0.933	0.900	
SVM control	0.554	0.811	0.636	0.894	0.837	0.068
AAC	0.525	0.792	0.841	0.687	0.841	0.086
DPC	0.395	0.743	0.837	0.546	0.760	***0.00023***
CTD	0.534	0.801	0.880	0.636	0.819	***0.022***
DPC	0.478	0.782	0.889	0.556	0.812	***0.014***
ATC	0.252	0.708	0.091	1.000	0.788	***0.002***

The first column represents the method name employed in this study. The second, the third, the fourth and the fifth respectively represent the MCC, accuracy, sensitivity, and specificity. The sixth column and the seventh represent the AUC and pairwise comparison of ROC area under curves (AUCs) between PVP-SVM and the other methods using a two-tailed t-test. A P ≤ 0.05 indicates a statistically meaningful difference between PVP-SVM and the selected method (shown in bold italic).

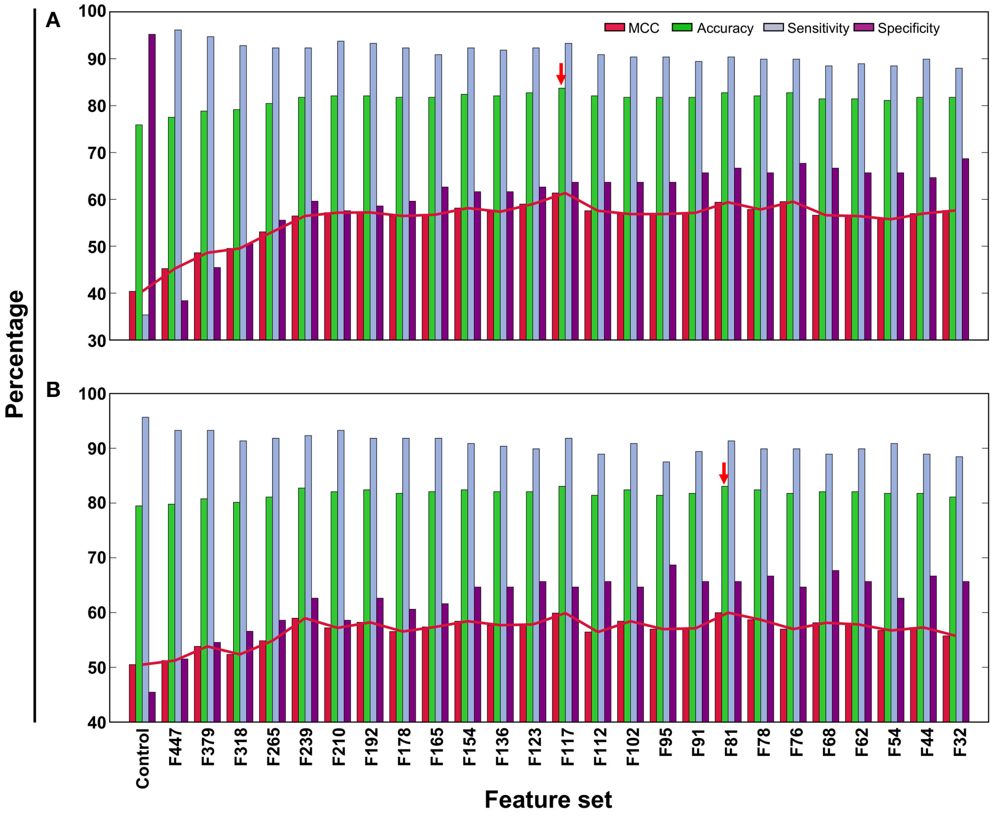

FIGURE 4 | Performance of ERT- and RF-based classifiers in distinguishing between PVPs and non-PVPs. A total of 26 classifiers were evaluated using LOOCV, whose performances in terms of MCC, accuracy, sensitivity, and specificity are shown. **(A)** ERT-based performance, **(B)** RF-based performance. Red arrow denotes the final selected models for each ML method.

TABLE 2 | A comparison of the proposed predictor with other ML-based methods on training dataset.

Methods	MCC	ACC	Sensitivity	Specificity	AUC	*P*-value
PVP-SVM	0.695	0.870	0.737	0.933	0.900	
PVPred	NA	0.850	0.758	0.894	0.899	0.974
RF	0.600	0.831	0.657	0.914	0.877	0.476
ERT	0.614	0.837	0.636	0.933	0.883	0.594

The first column represents the method name employed in this study. The second, the third, the fourth and the fifth respectively represent the MCC, accuracy, sensitivity, and specificity. The sixth column and the seventh represent the AUC and pairwise comparison of ROC area under curves (AUCs) between PVP-SVM and the other methods using a two-tailed t-test.

TABLE 3 | Performance of various methods on independent dataset.

Method	MCC	ACC	Sensitivity	Specificity	AUC	*P*-value
PVP-SVM	0.531	0.798	0.667	0.859	0.844	
ERT	0.509	0.798	0.533	0.922	0.778	0.367
RF	0.481	0.787	0.500	0.922	0.756	0.238
SVM control	0.414	0.755	0.533	0.859	0.796	0.505
PVPred	0.357	0.713	0.600	0.765	0.742	0.176

The first column represents the method name employed in this study. The second, the third, the fourth and the fifth respectively represent the MCC, accuracy, sensitivity, and specificity. The sixth column and the seventh represent the AUC and pairwise comparison of ROC area under curves (AUCs) between PVP-SVM and the other methods using a two-tailed t-test.

suggesting its usefulness as an improvement to existing tools for predicting PVPs.

In general, ML-based methods are problem-specific (Zhang and Tsai, 2005). Instead of selecting a ML method arbitrarily, it is necessary to explore different ML methods on the same dataset to select the best one. Hence, we explored three most commonly used ML methods (SVM, RF, and ERT), each having its own advantages and disadvantages. The PVP-SVM method performed consistently better than other two methods both with the training and independent datasets (**Figures 5A,B**). Although the differences in performance between these three methods were not significant ($P > 0.05$), SVM was superior to other ML methods in PVP prediction, consistent with a previous report (Ding et al., 2014). Hence, we selected PVP-SVM as the final prediction model.

Comparison of PVP-SVM and PVPred Methodology

A detailed comparison between our method and the existing method in terms of methodology is as follows: (i) the PVPred method utilizes only g-gap dipeptides as input features, and its optimal features were determined by an analysis of variance-based feature selection protocol. However, PVP-SVM utilizes AAC, ATC, CTD, and PCP in addition to DPC, with optimal features selected based on a RF algorithm; (ii) the number of optimal features used differs between the two methods; PVP-SVM uses 136 features, while PVPred uses 160; (iii) although the same ML method was used for the two methods, the parameter optimization procedure differed, as PVP-SVM used LOOCV, while PVPred used five-fold cross-validation.

Web Server Implementation

Several examples of bioinformatics tools/web servers utilized for protein function predictions have been reported in previous publications (Govindaraj et al., 2010, 2011; Manavalan et al., 2010a,b, 2011; Basith et al., 2011, 2013), and are of great practical use to researchers. To this end, an online prediction server for PVP-SVM was developed, which is freely accessible at the following link: www.thegleelab.org/PVP-SVM/PVP-SVM.html. Users can paste or upload query protein sequences in FASTA format. After submitting the input protein sequences, the results can be retrieved in a separate interface. All the curated datasets used in this study can be downloaded from the web server. PVP-SVM represents the second publicly available method for PVP prediction, and delivers a higher level of accuracy than PVPred.

DISCUSSION

PVPs play critical roles in adsorption between phages and their host bacteria, and are key in the development of new antibiotics. Phage-derived proteins are considered as safe and efficient antimicrobial agents due to its versatile properties, including bacteria-specific lytic mechanism, broad range of antibacterial spectrum, enhanced tissue penetration by small size, low immunogenicity, and reduced possibility for bacterial resistance (Drulis-Kawa et al., 2012). Thus, we have developed a novel computational method for predicting PVPs, called PVP-SVM. The molecular functions and biological activities of proteins can be predicted from their primary sequence (Lee et al., 2007); hence, we utilized the available PVPs sequences to develop the method.

A combination of AAC, ATC, DPC, CTD, and PCP features was used to map the protein sequences onto numeric feature vectors, which were inputted into the SVM to predict PVPs. Although AAC, CTD, and DPC features have been used previously (Feng et al., 2013b; Ding et al., 2014; Zhang et al., 2015), this is the first report including ATC and PCP. In ML-based predictions, feature selection is one of the most important steps because of redundant and non-informative features. Generally, high dimensional features contain numerous non-informative and redundant features, which affect prediction accuracy. Hence, the feature selection protocol is considered one of the most important steps in ML-based prediction (Wang et al., 2012; Manavalan et al., 2014; Manavalan and Lee, 2017; Song et al., 2017). To this end, we applied a feature selection protocol that has been proven effective in various biological applications (Manavalan and Lee, 2017; Manavalan et al., 2018), and identified the optimal features. Of those, the major contribution was from DPC (∼72%), followed by CTD, AAC, PCP, and ATC, indicating that information about the fraction of amino acids as well as their local order might play a major role in predicting PVPs. A previous study demonstrated that basic amino acids (Lys and Arg) usually occur in the flanking potential cleavage site in PVPs, as their side chain flexibility is required to accommodate the

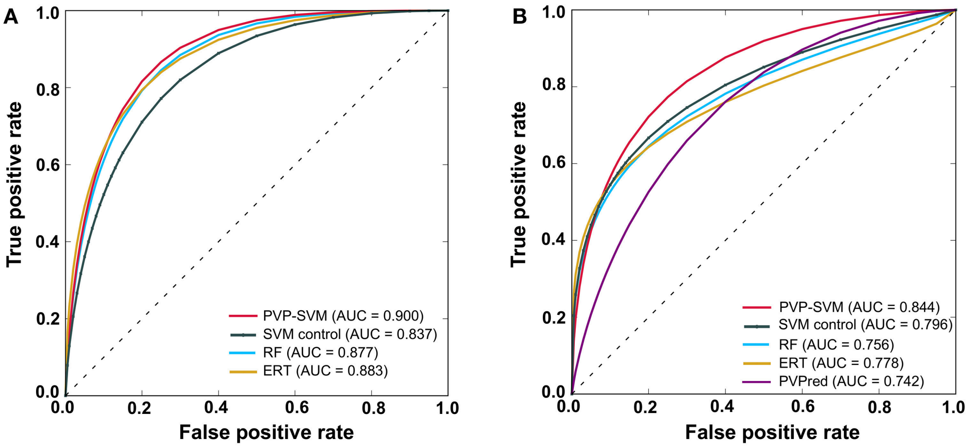

FIGURE 5 | Receiver operating characteristic curves of the prediction models. **(A)** LOOCV of the training dataset. **(B)** Evaluation with an independent dataset. Higher AUC values indicate better method performance.

change observed in the cleavage site (Coia et al., 1988; Speight et al., 1988). Interestingly, our optimal features contain these two important types of residues.

In general, if a prediction model is developed using a training dataset that contains highly homologous sequences, this method will overestimate the prediction accuracy. In this regard, Feng et al., and Ding et al., used a lower homology (<40% sequence identity) sequence dataset to develop their prediction models (Feng et al., 2013b; Ding et al., 2014). Zhang et al., developed their model using a highly homologous sequence dataset (<80% sequence identity); as a result, this method showed higher accuracy when evaluated with an independent dataset (Zhang et al., 2015). Furthermore, PVPred is the only publicly available method of the three, in the form of a web server, and was generated using the same dataset as our method. Therefore, we compared the performance of our method with PVPred only. Generally, a prediction model tends toward over-optimization in order to attain higher accuracy. Therefore, it is always necessary to evaluate the prediction model using an independent dataset, to measure the generalizability of the method (Chaudhary et al., 2016; Manavalan and Lee, 2017; Nagpal et al., 2017). Hence, we evaluated our three prediction models and PVPred on an independent dataset. Our study demonstrated that PVP-SVM consistently performed better than PVPred and the two other methods developed in this study on both datasets, indicating the greater transferability of the method.

The superior performance of PVP-SVM may be attributed to two important factors: (i) integration of previously reported features and inclusion of novel features that collectively make significant contributions to the performance; and (ii) a feature selection protocol that eliminates overlapping and redundant features. Furthermore, our approach is a general one, which is applicable to many other classification problems in structural bioinformatics. Although PVP-SVM displayed superior performance over the other methods, there is room for further improvements, including increasing the size of the training dataset based on the experimental data available in the future, incorporating novel features, and exploring different ML algorithms including stochastic gradient boosting (Xu et al., 2017) and deep learning (LeCun et al., 2015).

A user-friendly web interface has been made available, allowing researchers access to our prediction method. Indeed, this is the second method to be made publicly available, with higher accuracy than the existing method. Compared to experimental approaches, bioinformatics methods, such as PVP-SVM, represent a powerful and cost-effective approach for the proteome-wide prediction of PVPs. Therefore, PVP-SVM might be useful for large-scale PVP prediction, facilitating hypothesis-driven experimental design.

AUTHOR CONTRIBUTIONS

BM and GL conceived and designed the experiments; BM performed the experiments; BM and TS analyzed the data; BM and GL wrote paper. All authors reviewed the manuscript and agreed to this information prior to submission.

FUNDING

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education, Science, and Technology [2015R1D1A1A09060192 and 2009-0093826], and the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning [2016M3C7A1904392].

ACKNOWLEDGMENTS

The authors would like to thank Da Yeon Lee for assistance in the preparation of the manuscript.

REFERENCES

Basith, S., Manavalan, B., Gosu, V., and Choi, S. (2013). Evolutionary, structural and functional interplay of the IkappaB family members. *PLoS ONE* 8:e54178. doi: 10.1371/journal.pone.0054178

Basith, S., Manavalan, B., Govindaraj, R. G., and Choi, S. (2011). *In silico* approach to inhibition of signaling pathways of Toll-like receptors 2 and 4 by ST2L. *PLoS ONE* 6:e23989. doi: 10.1371/journal.pone.0023989

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers," in *Proceedings of the Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (Pittsburgh, PA: ACM).

Cao, R., Wang, Z., Wang, Y., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinformatics* 15:120. doi: 10.1186/1471-2105-15-120

Chaudhary, K., Nagpal, G., Dhanda, S. K., and Raghava, G. P. (2016). Prediction of immunomodulatory potential of an RNA sequence for designing non-toxic siRNAs and RNA-based vaccine adjuvants. *Sci Rep*.6:20678. doi: 10.1038/srep20678

Chen, W., Feng, P. M., Lin, H., and Chou, K. C. (2014). iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed. Res. Int*. 2014:623149. doi: 10.1155/2014/623149

Chen, W., Tang, H., and Lin, H. (2017a). MethyRNA: a web server for identification of N(6)-methyladenosine sites. *J. Biomol. Struct. Dyn*. 35, 683–687. doi: 10.1080/07391102.2016.1157761

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017b). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Coia, G., Parker, M. D., Speight, G., Byrne, M. E., and Westaway, E. G. (1988). Nucleotide and complete amino acid sequences of Kunjin virus: definitive gene order and characteristics of the virus-specified proteins. *J. Gen. Virol*. 69(Pt 1), 1–21.

Deng, X., Li, J., and Cheng, J. (2013). Predicting protein model quality from sequence alignments by support vector machines. *J. Proteomics Bioinform.* S9:001. doi: 10.4172/jpb.S9-001

Ding, H., Feng, P. M., Chen, W., and Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol. Biosyst.* 10, 2229–2235. doi: 10.1039/c4mb00316k.

Drulis-Kawa, Z., Majkowska-Skrobek, G., Maciejewska, B., Delattre, A. S., and Lavigne, R. (2012). Learning from bacteriophages - advantages and limitations of phage and phage-encoded protein applications. *Curr. Protein Pept. Sci.* 13, 699–722. doi: 10.2174/138920312804871193

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

Eickholt, J., Deng, X., and Cheng, J. (2011). DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. *BMC Bioinformatics* 12:43. doi: 10.1186/1471-2105-12-43

Elofsson, A., Joo, K., Keasar, C., Lee, J., Maghrabi, A. H. A., Manavalan, B., et al. (2017). Methods for estimation of model accuracy in CASP12. *Proteins* 86(Suppl. 1), 361–373. doi: 10.1101/143925

Feng, P. M., Chen, W., Lin, H., and Chou, K. C. (2013a). iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal Biochem.* 442, 118–125. doi: 10.1016/j.ab.2013.05.024

Feng, P. M., Ding, H., Chen, W., and Lin, H. (2013b). Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.* 2013:530696. doi: 10.1155/2013/530696

Feng, P. M., Lin, H., and Chen, W. (2013c). Identification of antioxidants from sequence information using naive Bayes. *Comput. Math. Methods Med*. 2013:567529. doi: 10.1155/2013/567529

Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K. C. (2018). iDNA6mA-PseKNC: identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics*. doi: 10.1016/j.ygeno.2018.01.005. [Epub ahead of print].

Govindaraj, R. G., Manavalan, B., Basith, S., and Choi, S. (2011). Comparative analysis of species-specific ligand recognition in Toll-like receptor 8 signaling: a hypothesis. *PLoS ONE* 6:e25118. doi: 10.1371/journal.pone.0025118

Govindaraj, R. G., Manavalan, B., Lee, G., and Choi, S. (2010). Molecular modeling-based evaluation of hTLR10 and identification of potential ligands in Toll-like receptor signaling. *PLoS ONE* 5:e12713. doi: 10.1371/journal.pone.0012713

Gupta, S., Mittal, P., Madhu, M. K., and Sharma, V. K. (2017). IL17eScan: a tool for the identification of peptides inducing IL-17 response. *Front. Immunol*. 8:1430. doi: 10.3389/fimmu.2017.01430

Jara-Acevedo, R., Diez, P., Gonzalez-Gonzalez, M., Degano, R. M., Ibarrola, N., Gongora, R., et al. (2018). Screening phage-display antibody libraries using protein arrays. *Methods Mol. Biol*. 1701, 365–380. doi: 10.1007/978-1-4939-7447-4_20

Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. (2017). Assessment of model accuracy estimations in CASP12. *Proteins* 86(Suppl. 1), 345–360. doi: 10.1002/prot.25371

Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., et al. (2015). An *in silico* platform for predicting, screening and designing of antihypertensive peptides. *Sci. Rep.* 5:12512. doi: 10.1038/srep12512

Lavigne, R., Ceyssens, P. J., and Robben, J. (2009). Phage proteomics: applications of mass spectrometry. *Methods Mol. Biol*. 502, 239–251. doi: 10.1007/978-1-60327-565-1_14

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* 8, 995–1005. doi: 10.1038/nrm2281

Lekunberri, I., Subirats, J., Borrego, C. M., and Balcazar, J. L. (2017). Exploring the contribution of bacteriophages to antibiotic resistance. *Environ. Pollut*. 220(Pt B), 981–984. doi: 10.1016/j.envpol.2016.11.059

Li, L., Xiong, Y., Zhang, Z.-Y., Guo, Q., Xu, Q., Liow, H.-H., et al. (2015). Improved feature-based prediction of SNPs in human cytochrome P450 enzymes. *Interdiscipl. Sci.* 7, 65–77. doi: 10.1007/s12539-014-0257-2

Lin, H., Liu, W. X., He, J., Liu, X. H., Ding, H., and Chen, W. (2015). Predicting cancerlectins by the optimal g-gap dipeptides. *Sci Rep*. 5:16964. doi: 10.1038/srep16964

Manavalan, B., Basith, S., Choi, Y. M., Lee, G., and Choi, S. (2010a). Structure-function relationship of cytoplasmic and nuclear IkappaB proteins: an *in silico* analysis. *PLoS ONE* 5:e15782. doi: 10.1371/journal.pone.0015782

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., and Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget.* 8, 77121–77136. doi: 10.18632/oncotarget.20365

Manavalan, B., Govindaraj, R., Lee, G., and Choi, S. (2011). Molecular modeling-based evaluation of dual function of IkappaBzeta ankyrin repeat domain in toll-like receptor signaling. *J. Mol. Recognit*. 24, 597–607. doi: 10.1002/jmr.1085

Manavalan, B., Kuwajima, K., Joung, I., and Lee, J. (2015). "Structure-based protein folding type classification and folding rate prediction," in *Proceedings of the Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on* (Washington, DC: IEEE).

Manavalan, B., and Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* 33, 2496–2503. doi: 10.1093/bioinformatics/btx222.

Manavalan, B., Lee, J., and Lee, J. (2014). Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS ONE* 9:e106542. doi: 10.1371/journal.pone.0106542

Manavalan, B., Murugapiran, S. K., Lee, G., and Choi, S. (2010b). Molecular modeling of the reductase domain to elucidate the reaction mechanism of reduction of peptidyl thioester into its corresponding alcohol in non-ribosomal peptide synthetases. *BMC Struct. Biol.* 10:1. doi: 10.1186/1472-6807-10-1

Manavalan, B., Shin, T. H., and Lee, G. (2018). DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 9, 1944–1956. doi: 10.18632/oncotarget.23099

Nagpal, G., Chaudhary, K., Dhanda, S. K., and Raghava, G. P. S. (2017). Computational prediction of the immunomodulatory potential of RNA sequences. *Methods Mol. Biol.* 1632, 75–90. doi: 10.1007/978-1-4939-7138-1_5

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Qiu, W. R., Sun, B. Q., Xiao, X., Xu, Z. C., and Chou, K. C. (2016). iHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget* 7, 44310–44321. doi: 10.18632/oncotarget.10027.

Scholkopf, B., and Smola, A. J. (2001). *Learning with Kernels*: *Support Vector Machines*, *Regularization, Optimization*, and *Beyond.* London: MIT Press.

Seguritan, V., Alves, N. Jr., Arnoult, M., Raymond, A., Lorimer, D., Burgin, A. B. Jr., et al. (2012). Artificial neural networks trained to detect viral and phage structural proteins. *PLoS Comput. Biol.* 8:e1002657. doi: 10.1371/journal.pcbi.1002657

Smola, A., and Vapnik, V. (1997). Support vector regression machines. *Adv. Neural Inf. Process. Syst.* 9, 155–161.

Song, J., Wang, H., Wang, J., Leier, A., Marquez-Lago, T., Yang, B., et al. (2017). PhosphoPredict: a bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Sci. Rep.* 7:6862. doi: 10.1038/s41598-017-07199-4

Speight, G., Coia, G., Parker, M. D., and Westaway, E. G. (1988). Gene mapping and positive identification of the non-structural proteins NS2A, NS2B, NS3, NS4B and NS5 of the flavivirus Kunjin and their cleavage sites. *J. Gen. Virol.* 69(Pt 1), 23–34.

Tang, H., Su, Z. D., Wei, H. H., Chen, W., and Lin, H. (2016). Prediction of cell-penetrating peptides with feature selection techniques. *Biochem. Biophys. Res. Commun*. 477, 150–154. doi: 10.1016/j.bbrc.2016.06.035

Vapnik, V. N., and Vapnik, V. (1998). *Statistical Learning Theory*. New York, NY: Wiley.

Wang, H., Feng, L., Zhang, Z., Webb, G. I., Lin, D., and Song, J. (2016). Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci. Rep*. 6:21383. doi: 10.1038/srep21383

Wang, M., Zhao, X. M., Takemoto, K., Xu, H., Li, Y., Akutsu, T., et al. (2012). FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PLoS ONE* 7:e43847. doi: 10.1371/journal.pone.0043847

Wang, Z., Tegge, A. N., and Cheng, J. (2009). Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins* 75, 638–647. doi: 10.1002/prot.22275.

Xiong, Y., Liu, J., Zhang, W., and Zeng, T. (2012). Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome Sci.* 10(Suppl. 1), S20. doi: 10.1186/1477-5956-10-S1-S20

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H.-Y., et al. (2017). PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol.* 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

Yuan, Y., and Gao, M. (2016). Proteomic analysis of a novel Bacillus jumbo phage revealing glycoside hydrolase as structural component. *Front. Microbiol.* 7:745. doi: 10.3389/fmicb.2016.00745

Zhang, D., and Tsai, J. J. P. (2005). *Machine Learning Applications in Software Engineering*. River Edge, NJ: World Scientific.

Zhang, L., Zhang, C., Gao, R., and Yang, R. (2015). An ensemble method to distinguish bacteriophage virion from non-virion proteins based on protein sequence characteristics. *Int. J. Mol. Sci.* 16, 21734–21758. doi: 10.3390/ijms160921734

Zheng, C., Wang, M., Takemoto, K., Akutsu, T., Zhang, Z., and Song, J. (2012). An integrative computational framework based on a two-step random forest algorithm improves prediction of zinc-binding sites in proteins. *PLoS ONE* 7:e49716. doi: 10.1371/journal.pone.0049716

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.