frontiers in Genetics

METHODS
published: 09 October 2020
doi: 10.3389/fgene.2020.569100

Identification of Proteins of Tobacco Mosaic Virus by Using a Method of Feature Extraction

*Yu-Miao Chen, Xin-Ping Zu and Dan Li**

Information and Computer Engineering College, Northeast Forestry University, Harbin, China

Tobacco mosaic virus, TMV for short, is widely distributed in the global tobacco industry and has a significant impact on tobacco production. It can reduce the amount of tobacco grown by 50–70%. In this research of study, we aimed to identify tobacco mosaic virus proteins and healthy tobacco leaf proteins by using machine learning approaches. The experiment's results showed that the support vector machine algorithm achieved high accuracy in different feature extraction methods. And 188-dimensions feature extraction method improved the classification accuracy. In that the support vector machine algorithm and 188-dimensions feature extraction method were finally selected as the final experimental methods. In the 10-fold cross-validation processes, the SVM combined with 188-dimensions achieved 93.5% accuracy on the training set and 92.7% accuracy on the independent validation set. Besides, the evaluation index of the results of experiments indicate that the method developed by us is valid and robust.

Keywords: feature extraction, physicochemical properties, identification, tobacco mosaic virus, machine learning

OPEN ACCESS

Edited by:
Xiaoming Song,
North China University of Science and Technology, China

Reviewed by:
Fenglong Wang,
Tobacco Research Institute (CAAS), China
Guang Song,
Johns Hopkins University, United States
Hao Lin,
University of Electronic Science and Technology of China, China

****Correspondence:***
Dan Li
ld@nefu.edu.cn

Specialty section:
This article was submitted to Plant Genomics, a section of the journal Frontiers in Genetics

Received: *03 June 2020*
Accepted: *09 September 2020*
Published: *09 October 2020*

Citation:
Chen Y-M, Zu X-P and Li D (2020) Identification of Proteins of Tobacco Mosaic Virus by Using a Method of Feature Extraction. Front. Genet. 11:569100. doi: 10.3389/fgene.2020.569100

INTRODUCTION

Tobacco mosaic virus is worldwide distribution and is the furthest invasive virus which is most harmful to crops. Tobacco is one of the important economic crops in our country, however, the existence of tobacco mosaic disease has greatly reduced the yield and quality of tobacco. Since plants do not have a complete immune system, once infected, the leaves can show mosaic symptoms or even deformities and the growth can also be chronically diseased, which makes tobacco mosaic virus is very difficult to control (Hu and Lee, 2015).

The study of viruses has attracted many scholars, and with the development of computer machine learning algorithms, many scholars have applied machine learning algorithms to the study of viruses. Metzler and Kalinina (2014) used one-class SVM method to detect atypical genes in viral families based on their statistical features, without the need for explicit knowledge of the source species. The simplicity of the statistical features used allows the method to be applied to a variety of viruses. Salama et al. (2016) predicted new drug-resistant strains that facilitate the design of antiviral therapies. In this study, neural network techniques were used to predict new strains, and using a rough set theory based on algorithm to extract these points mutation patterns. For phage virion proteins (PVPs) prior to *in vitro*, Manavalan et al. (2018) developed a SVM-based predictor that exhibited good performance and avoided the expensive costs required for experiments.

Using biochemical experiments to study all tobacco mosaic virus is a challenge because it is expensive and a waste of researchers' time, and there is no specific predictor to predict tobacco mosaic virus. So, in this research we evaluated the predictive performance of different classifiers in combination with different feature extraction methods. We have chosen classical machine

learning algorithms and classical feature extraction methods. The feature extraction methods AAC (Chou, 2001), 188-dimensions (Dubchak et al., 1995) and CKSAAGP (Chen et al., 2018) and their combination were chosen for the reasons. AAC is the first proposed feature extraction method that is widely used to predict the function of proteins. It based on their amino acid composition. CKSAAGP describes the spatial distribution information of amino acids, 188 dimensions in addition to the physicochemical properties of amino acids. Three feature extraction methods from different aspects, so these three feature extraction methods were chosen. The combined feature extraction method was attempted considering the expectation of better results. We finally chose the combination of support vector machine (SVM) with 188-dimensions as the final predictor because it has the best prediction effect.

MATERIALS AND METHODS

Our method was developed based on three steps (**Figure 1**). Step 1: we collected the data and preprocessed the dataset to obtain a non-redundant benchmark dataset that does not contain non-standard characters. Step 2: We used Amino Acid Composition (AAC), feature extraction method based on the composition of amino acid sequence and physicochemical properties (188-dimensions), composition of k-spaced amino acid pairs (CKSAAGP), and the combined methods AAC_CKSAAGP and 188_CKSAAGP which are proposed in this paper to extract features from protein sequences. Step 3: Five algorithms and 10-fold cross-validation are used to build and estimate the models, which are Random Forest (RF), Bagging, K-Nearest Neighbor (KNN), Naive Bayes (NB), and Support Vector Machine (SVM). Then, we validated experimental results using an independent validation set.

BENCHMARK DATASET

High-quality baseline data sets contribute to the accuracy of model predictions (Yang et al., 2019b; Cheng et al., 2020; Zhu et al., 2020). The dataset obtained for this experiment was derived from the Swiss-Prot database in The Uniprot (2018). Firstly, we used the keyword search method to collect data from the UniProt database. By entering the keywords "Tobacco mosaic virus" and "Tobacco leaf not virus" to obtain the positive and negative data needed for the experiment. For the sake of improving the reliability of the data, the following operations were performed: (1), Deleted the protein sequences containing non-standard letters, i.e. "B," "X," "Z," etc.; thus, we obtained 5,309 protein sequences of tobacco mosaic virus and 45,827 protein sequences of non-tobacco mosaic virus. (2), If the sample contains multiple similarity sequences, this sample is not statistically representative. We used the CD-HIT program (Fu et al., 2012) to delete sequences with similarity surpass 40% in positive and negative data sets (Zou et al., 2018). After removing the redundant sequences, we eventually obtained a dataset of 715 protein sequences of TMV proteins and 17,983 protein sequences of tobacco leaf proteins.

There are 715 sequences in the positive datasets and 17,893 sequences in the negative datasets. Much more negative data than positive data. For the purpose of balancing the datasets, we took a downsampling approach. We split the negative data by the size of the positive data. And randomly selected 10 of these copies as the negative dataset, so we obtained a negative dataset containing 7,150 sequences. The resulting positive and negative datasets were divided proportionally. The final training dataset consists of 500 positive data and 5,000 negative data. The test dataset consists of 215 positive data and 2,150 negative data. These data are available in our software package.

FEATURE EXTRACTION

Feature selection will affect the performance of machine learning methods for bioinformatics problems (Zhao et al., 2015). In the research of this paper, five feature extraction methods are selected, including amino acid composition (AAC), composition of k-spaced amino acid pairs (CKSAAGP), 188-dimensions feature extraction method, and the combined methods AAC_CKSAAGP and 188_CKSAAGP which are proposed in this paper.

Amino Acid Composition (AAC)

The coded amino acid composition coding scheme (Bhasin and Raghava, 2004) calculates the probability of occurrence of 20 natural amino acids (i.e., "ACDEFGHIKLMNPQRSTVWY") in protein sequences or peptide chains (Zhong et al., 2020). The calculation formula for each amino acid is as follows:

$$v_i = \frac{c_i}{len\left(seq\right)}, i \in (A, C, \cdots, Y) \tag{1}$$

Where c_i and $len(seq)$ represent the number of occurrences of amino acid i in the sequence or peptide chain and the length of the sequence or peptide chain, respectively (Lin et al., 2005; Lv et al., 2020).

Composition of k-Spaced Amino Acid Pairs (CKSAAGP)

The composition of K-spaced amino acid pairs (Chen et al., 2018) can be regarded as a variant of CKSAAP, which calculates the frequency of amino acid pairs separated by any k residues (the default maximum for k is set to 5) (Chen et al., 2020). Taking K = 0 as an e.g., a feature vector is defined as:

$$\gamma_0 = \left(\frac{C_{g1g1}}{C_n}, \frac{C_{g1g2}}{C_n}, \frac{C_{g1g3}}{C_n}, \cdots, \frac{C_{g5g5}}{C_n}\right)_{25} \tag{2}$$

wherein, $(g1g1, g1g2, g1g3, \cdots g5g5)$ represents 0-spacing amino acid group pairs. There are 25 groups in total, and each descriptor represents the composition of the corresponding residue pair in the protein sequence (Zhu et al., 2020). C_{g1g1} (Zhang et al., 2014) represents the number of times the residue pair $g1g1$ appears in the sequence and Cn represents the total number of residue pairs with a gap of 0 in the sequence. In a protein sequence of length N, for different values of K, the value of n can be defined as:

$$n = P - K - 1, K \in (0, 1, 2, 3, 4, 5) \tag{3}$$

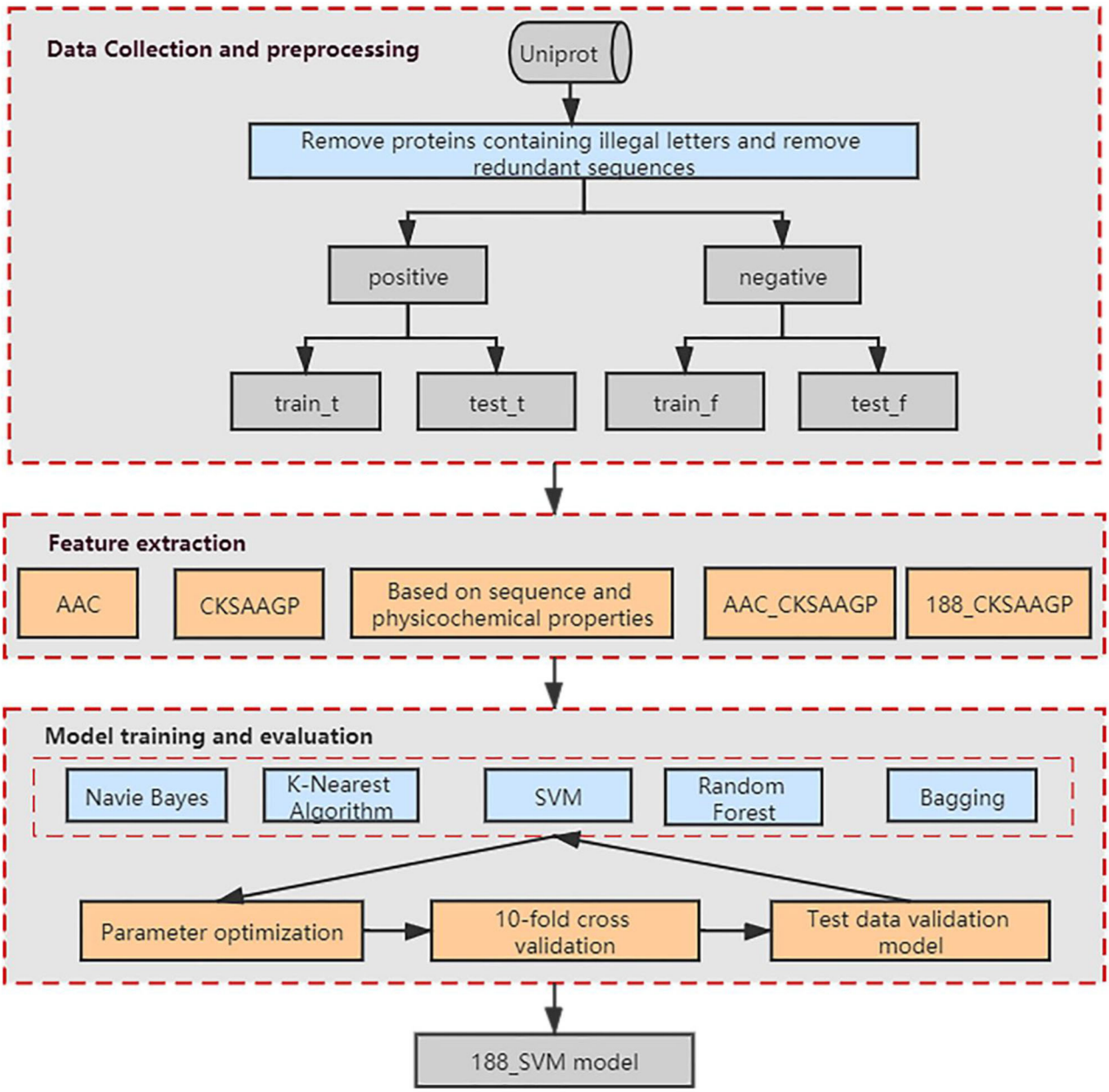

FIGURE 1 | Schematic showing the workflow used to develop our method. AAC, CKSAAGP, 188-dimensions, AAC_CKSAAGP, 188_ CKSAAGP are our five feature extraction methods. Naive Bayes, K-Nearest Algorithm, SVM, Random Forest and Bagging are our five classification algorithms.

When K takes each value, the CKSAAGP feature vector $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$ has a total size of 150 dimensions.

188-Dimensions

Each amino acid sequence has different physical and chemical properties including amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, and solvent accessibility for each residue in the sequence (Cai et al., 2003). The feature extraction method for protein P is formulated as follows:

$$P = \{C, Tr, D\} \tag{4}$$

Where C represents the frequency of a kind of specific attribute (such as polarity) amino acid appearing in the global sequence. Tr represents the global percentage of transitions between a specific amino acid and another amino acid of a specific property. D is used to describe the first, 25%, 50%, 75% and last position of each specific amino acid in the peptide chain.

Assume that the protein sequence "VNVVVNVNNVVVVVNVNNNVVNNVNNNVVNVN" has 17 valines ($n_1 = 17$) and 15 asparagines ($n_2 = 15$). The components of these two amino acids are $\frac{n_1}{n_1+n_2} \times 100.00 = 53.13$ and $\frac{n_2}{n_1+n_2} \times 100.00 = 46.87$, respectively. The number of transitions from V to N or N to V is 17, so the percentage of these transitions is $(17/32) \times 100.00 = 53.13$. The first, 25, 50, 75% and last positions of valine in the peptide chain are 1, 5, 13, 21, 31, respectively, so the D-attribute of valine is D_V = (3.13, 15.63, 40.63, 65.23, 96.88). In the same way, the position of asparagine in the peptide chain can be found. In summary, the amino acid composition descriptors are C = (53.13, 64.87), Tr = (53.13), and D = (3.13, 15.63, 40.63, 65.23, 96.88, 6.25, 28.13, 68.75, 78.13, 100.00). Descriptors of other attributes can be described through a similar process, and then all the descriptors are combined to form a 188-dimensions feature vector.

Sequence	V	N	V	V	V	N	V	N	N	V	V	V	V	V	N	V	N	N	N	V	V	N	N	V	N	N	N	V	V	N	V	N
Sequence index	1				5					10					15					20					25					30		
Index for V	1		2	3	4		5			6	7	8	9	10		11				12	13			14				15	16		17	
Index for N		1				2		3	4						5		6	7	8			9	10		11	12	13			14		15
V/N transitions	|	|			|	|	|		|					|	|			|		|	|		|		|	|	|		

Combined Method

A combination of AAC, CKSAAGP, and feature extraction based on a combination of sequence and physicochemical properties constitute a new feature extraction method. The number of characterization dimensions for the AAC_CKSAAGP combination is 170, and the number of characterization dimensions for the 188_CKSAAGP combination is 338. Since the information of amino acid content is included in the 188-dimensions feature extraction method, the combination of AAC and 188-dimensions feature extraction method is not used in this paper.

Classifier

In this paper, five classifiers were used for the experiments. these classifiers were implemented through Waikato Environment for Knowledge Analysis software (Azuaje et al., 2006).

Random Forest

Random Forest (RF) is an integrated learning method first proposed by Leo Breiman and Adele Cutler (Azuaje et al., 2006; Goldstein et al., 2011; Cheng et al., 2018b,d), and it is a combination of multiple decision trees. Nowadays, many bioinformatics' problems use Random Forest (Tastan et al., 2012; Jamshid et al., 2018; Lyu et al., 2019; Ru et al., 2019; Lv et al., 2020). For processing large amounts of data, Random Forest is characterized by high accuracy, high speed and good robustness. In RF, we need to input the prediction samples into each tree for prediction, and finally use the voting algorithm to determine the result of prediction. The voting algorithm is shown speed and good robustness. The anti-noise capability of RF is strong, and it often shows good robustness when processing high-dimensional data. The voting algorithm is shown below:

$$\text{result} = \text{sgn}\left(\sum_{\text{i}=1}^{\text{n}} \text{pre}_{\text{label}}\right) \tag{5}$$

Where result is the final prediction, pre_{label} represents the predicted result for each decision tree, which equals to 1 or −1, and n represents the number of Decision Trees in the model.

Support Vector Machine

Support Vector Machine (SVM) is often applied to classification problems and is a supervised learning approach (Huang et al., 2012; Jiang et al., 2013; Xing et al., 2014; Kumar et al., 2015; Zhao et al., 2015; Liao et al., 2018; Wang et al., 2019). There are already a number of software packages that support the SVM algorithms. In this experiment, we used Libsvm (Chang and Lin, 2007) in Weka (version 3-8-2) (Hall et al., 2008) to implement the SVM, where we chose RBF, a radial basis function, to classify the proteins of tobacco mosaic virus. Then we determined the regularization parameter C and the kernel parameter g through grid search and 10-fold cross-validation (Wang et al., 2011).

K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm (Zhang and Zhou, 2007; Lan et al., 2013; Deng et al., 2016) which is one of the simplest, most convenient, and highly effective algorithms. Now it has been frequently used in the functional classification of proteins problems. The key step of KNN prediction is to find the K neighbors closest to the test data from the training set, and then use the category with the most K neighbors as the final category of the test data. In this experiment, we adopt the KNN algorithm based on the Harmanton distance and the Harmanton distance formula is summarized as follows:

$$L_{1(x_i,x_j)} = \sum_{k=1}^{n} |x_i^{(k)} - x_j^{(k)}| \tag{6}$$

where: x_i^k(k=1, 2, 3, ……n) is characteristic of the training set and x_j^k(k=1, 2, 3, ……n) is characteristic of the test dataset.

Naive Bayes

The Naive Bayes (NB) is a easily understand classification algorithm (Xue et al., 2006; Wang et al., 2008; Feng et al., 2013), which is based on the Bayesian classifier and assumes that the feature attributes of the data are simple and independent. In the

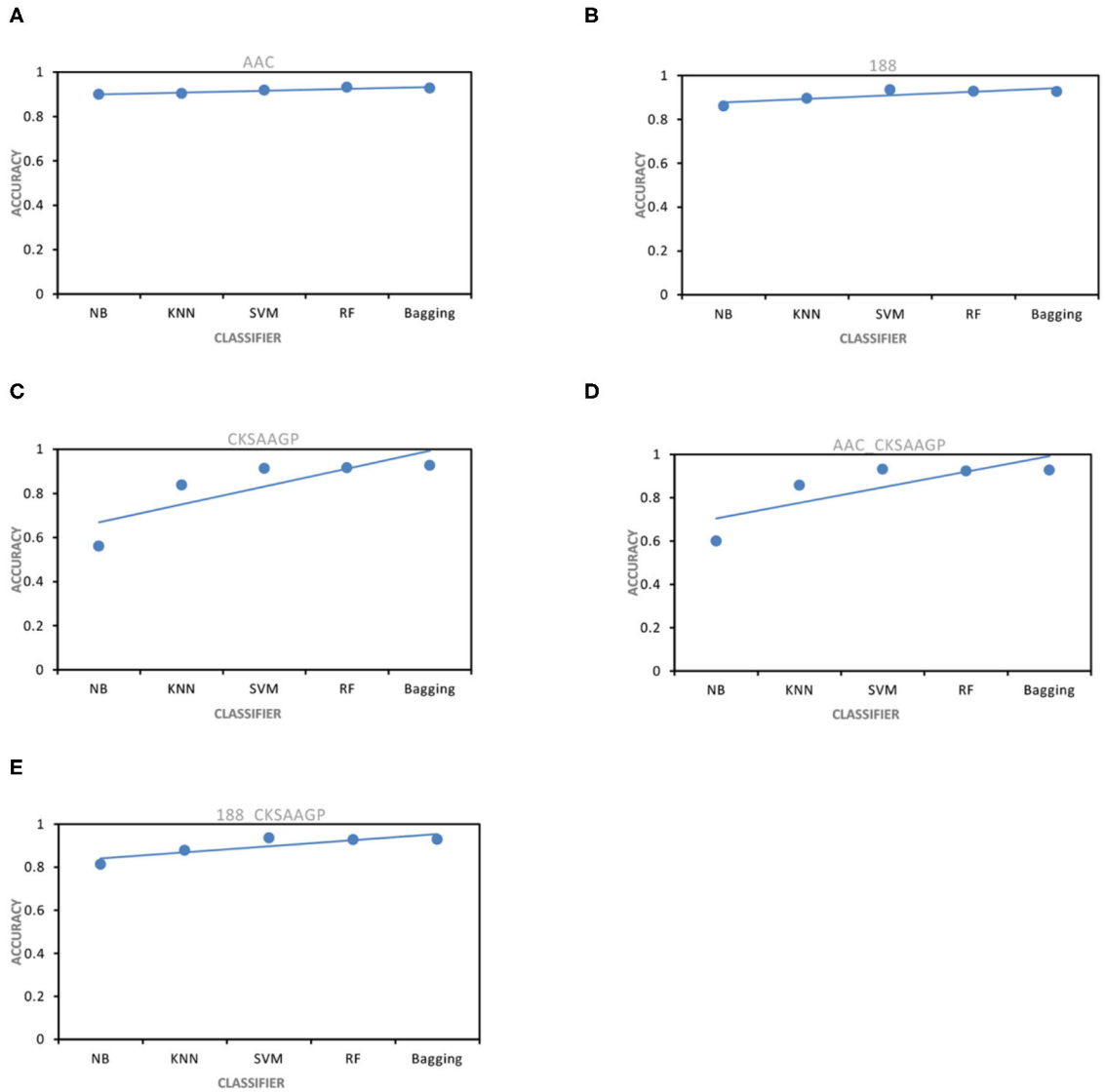

FIGURE 2 | ACC obtained by five feature extraction methods [**(A–E)** AAC, 188-dimensions, CKSAAGP, AAC_CKSAAGP, and 188-CKSAAGP] combined with five classifiers (NB, KNN, SVM, RF, Bagging) through 10-fold cross-validation.

classification scenario, it greatly reduces the complexity of the Bayesian classification algorithm. Suppose the sample data set is:

$$\left(x_1^{(1)}, x_2^{(1)}, \cdots, x_n^{(1)}, s_1\right), \left(x_1^{(2)}, x_2^{(2)}, \cdots, x_n^{(2)}, s_2\right), \cdots \left(x_1^{(m)}, x_2^{(m)}, \cdots, x_n^{(m)}, s_m\right) \quad (7)$$

there are m samples and each sample have n features. The data set has a total of class variable. Generally speaking, the m samples can be classified into s categories, where n features are independent of each other. The category of S is defined as follows:

$$S = \{s_1, s_2, s_3, \cdots, s_m\} \quad (8)$$

Among them, there are M class variables in the set S. Naive Bayes formula is defined as follows:

$$P\left(y_i|x_1, x_2, \cdots, x_n\right) = \frac{P_{y_i} \prod_{j=1}^{n} P\left(x_j|s_i\right)}{\prod_{j=1}^{n} P\left(x_j\right)} \quad (9)$$

Bagging

Bagging is a typical integrated learning algorithm (Abellán et al., 2017), which is directly based on autonomous sampling. For the input sample set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, a weak learner algorithm is used to classify each time, and a total of T iterations are made, and finally we will obtain a powerful classifier. Since it samples each training model, it has a strong generalization ability that can significantly reduce the variance of the training model.

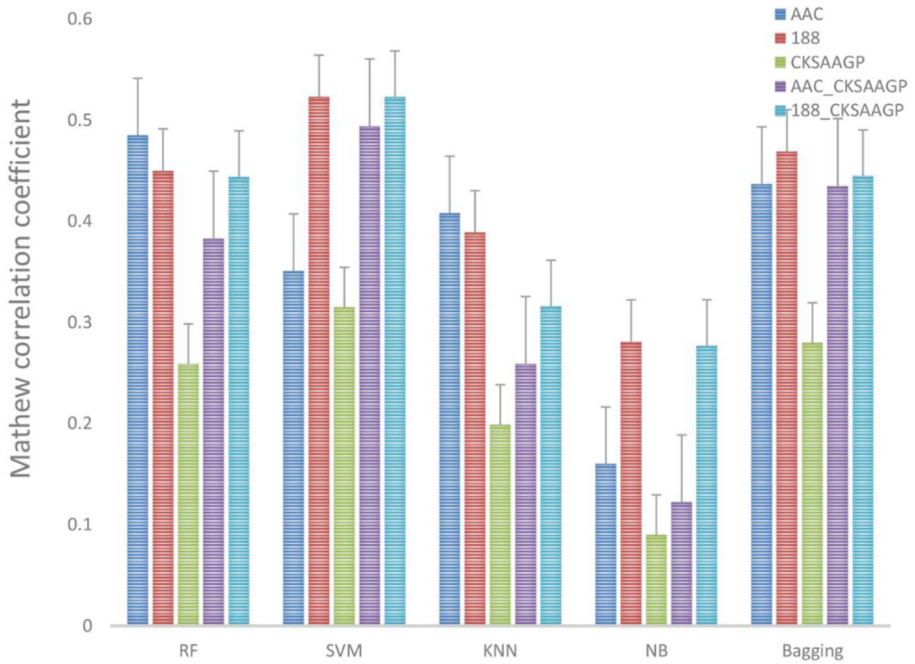

FIGURE 3 | Different feature extraction methods were used to predict the performance of different classification methods. The average MCC value of the classifier was tested by 10-fold cross-validation.

PERFORMANCE EVALUATION

There are five main parameters (Kou and Feng, 2015) to evaluate the predictive performance of this experiment, namely, sensitivity (Sn), specificity (Sp) (Ding et al., 2012; Tan et al., 2019), accuracy (ACC) (Thakur et al., 2016; Cheng et al., 2018a,b), Matthews correlation coefficient (MCC) (Yang et al., 2019a,b) and area of ROC curve (AUC) (Lobo et al., 2008; Li and Fine, 2010; Wang et al., 2010; Hajian-Tilaki, 2014; Baratloo et al., 2015). Defined as follows:

$$S_n = \frac{TP}{TP + FN} \tag{10}$$

$$S_P = \frac{TN}{TN + FP} \tag{11}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$MCC = \frac{(TP + TN) - (FP + FN)}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{13}$$

$$AUC = \frac{\frac{\sum_{i \in positiveClAss} rank_i - M \times (M+1)}{2}}{M \times N} \tag{14}$$

Where TP represents the amount of tobacco mosaic virus correctly predicted by the model (Dong et al., 2015); TN indicates the amount of non-tobacco mosaic virus correctly predicted by the model (Niu et al., 2018); FN indicates the amount of non-tobacco mosaic virus incorrectly predicted by the model (Kim et al., 2016); FP indicates the amount of non-tobacco mosaic virus predicted by the model; M and N indicate the amount of positive and negative data, respectively; and $rank_i$ is the score of the i-th positive sample was calculated by classification. The higher the value of the five evaluation indicators above, the better the model prediction.

RESULTS AND DISCUSSION

Performance Evaluation of Different Classifiers

The ACC and MCC of SVM and RF were mostly higher than the predictors of NB, KNN and Bagging under different feature extraction methods (**Figures 2**, **3**). When the feature extraction method selects 188_CKSAAGP or 188-dimensions, SVM reach the highest ACC. When the feature extraction method uses AAC, RF achieves the highest ACC. Through 10-fold cross-validation, the MCC of SVM is higher than that of RF (**Figure 3**).

However, when comparing predictor superiority, it is possible to use not only the predicted ACC and MCC comparison, but also the trade-off between Sn and Sp. The sensitivity (Sn) and specificity (Sp) of SVM, RF, and Bagging predictor variables are greater than those of NB and KNN (**Table 1** and **Figure 5**). This result shows that SVM, RF, and Bagging predict tobacco mosaic virus are better than NB and KNN due to the difference in the ability of these five common classification algorithms to handle multidimensional datasets. NB is a naive algorithm based on the assumption that the individual properties are independent of each other, and NB

TABLE 1 | Predictive effect of NB, KNN, SVM, RF, and Bagging on different trait extraction methods for tobacco mosaic virus.

Method	Feature	Sp	Sn	ACC (%)	MCC	AUC
NB	AAC	0.866	0.899	89.95	0.16	0.719
KNN		0.902	0.905	90.45	0.408	0.699
SVM (c = 9, g = 3)		0.901	0.918	91.80	0.351	0.61
RF		**0.932**	**0.932**	**93.22**	**0.485**	**0.833**
Bagging		0.922	0.928	92.76	0.437	0.799
NB	188-dimensions	0.882	0.862	86.16	0.281	0.732
KNN		0.899	0.897	89.65	0.389	0.699
SVM		**0.936**	**0.936**	**93.58**	**0.523**	0.657
RF		0.929	0.929	92.91	0.45	**0.838**
Bagging		0.924	0.93	93.04	0.469	0.807
NB	CKSAAGP	0.859	0.561	56.05	0.09	0.64
KNN		0.868	0.839	83.87	0.199	0.618
SVM (c = 5, g = 3)		0.894	0.914	91.38	**0.315**	0.601
RF		0.917	0.916	91.58	0.259	**0.779**
Bagging		**0.919**	**0.917**	**91.69**	0.28	0.711
NB	AAC_CKSAAGP	0.866	0.601	60.07	0.122	0.669
KNN		0.878	0.858	85.82	0.259	0.648
SVM (c = 3, g = 3)		0.923	**0.931**	**93.15**	**0.494**	0.675
RF		**0.927**	0.924	92.36	0.383	**0.831**
Bagging		0.921	0.927	92.75	0.435	0.795
NB	188_ CKSAAGP	0.885	0.814	81.40	0.277	0.726
KNN		0.887	0.878	87.80	0.316	0.669
SVM (c = 3, g = −11)		**0.936**	**0.936**	**93.58**	**0.523**	0.657
RF		0.929	0.929	92.85	0.444	**0.823**
Bagging		0.925	0.93	93.00	0.463	0.803

Sp, specificity; Sn, sensitivity; ACC, accuracy; MCC, Matthews correlation coefficient; AUC, area of ROC curve. AAC, amino acid composition; 188-dimensions, composition based on sequence and physicochemical properties; CKSAAGP, composition of k-spaced amino acid pairs; AAC_CKSAAGP, combination of AAC and CKSAAGP; 188_CKSAAGP, combination of 188-dimensions and CKSAAGP. The bold values represent the highest score of current feature extraction method in different classifiers.

TABLE 2 | Through the use of test data, the evaluation results of the prediction model of NB, KNN, SVM, RF, and Bagging combined with different types of extraction methods.

Method	Feature	Sp	Sn	ACC(%)	MCC	AUC
NB	AAC	0.845	0.89	0.8905	0.051	0.677
KNN		0.896	0.901	0.9011	0.368	0.673
SVM		0.893	0.914	0.9142	0.289	0.508
RF		**0.912**	**0.923**	**0.9226**	**0.372**	**0.781**
Bagging		**0.912**	0.922	0.9222	**0.372**	0.734
NB	188-dimensions	0.877	0.859	0.8588	0.255	0.709
KNN		0.886	0.888	0.8875	0.311	0.653
SVM		**0.925**	**0.928**	**0.9277**	**0.434**	0.615
RF		0.92	0.924	0.9243	0.393	**0.799**
Bagging		0.909	0.922	0.9218	0.377	0.774
NB	CKSAAGP	0.858	0.558	0.5577	0.086	0.634
KNN		0.86	0.832	0.8321	0.15	0.587
SVM		0.886	0.91	0.9099	**0.268**	0.582
RF		**0.922**	**0.915**	0.9146	0.235	**0.736**
Bagging		0.917	**0.915**	**0.9150**	0.243	0.692
NB	AAC_CKSAAGP	0.861	0.589	0.5886	0.104	0.647
KNN		0.87	0.859	0.8592	0.213	0.615
SVM		0.915	**0.926**	**0.9256**	**0.423**	0.63
RF		**0.922**	0.919	0.9188	0.312	**0.777**
Bagging		0.915	0.923	0.9226	0.373	0.726
NB	188_CKSAAGP	0.879	0.811	0.8106	0.244	0.705
KNN		0.877	0.87	0.8702	0.257	0.636
SVM		**0.925**	**0.928**	**0.9277**	**0.434**	0.615
RF		0.92	0.925	0.9247	0.399	**0.776**
Bagging		0.915	0.924	0.9239	0.393	0.75

The bold values represent the highest score of current feature extraction method in different classifiers.

is very friendly to low dimensional features. However, for multidimensional datasets, there is often some correlation between attribute features. The low ACC of KNN may be because the small size of the training datasets. The SVM, RF and Bagging classification algorithms do not require much in terms of dataset dimensionality, and they can handle high-dimensional, noisy and missing datasets with strong correlation between attributes.

In addition, we also used the test datasets to verify the model. The results are shown in **Table 2**. The results show that the model constructed by SVM combined with 188-dimensions or 188_CKSAAGP achieves a high AAC, which shows that this model is reliable. Although the model constructed by SVM combined with 188-dimensions or 188_CKSAAGP is lower than other algorithms in terms of AUC, evaluation indicators such as Sp, Sn, and MCC have all achieved the best results. Therefore, the SVM algorithm is very promising in TMV classification. Due to the above reasons, this experiment chose SVM as the final classifier to predict TMV.

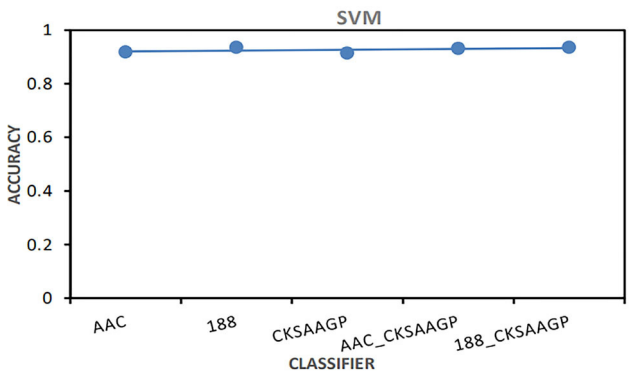

FIGURE 4 | The AAC value obtained by the predictor constructed by SVM combined with five feature selection methods (AAC, 188-dimensions, CKSAAGP, AAC_CKSAAGP, and 188-CKSAAGP).

Performance Evaluation of Different Feature Extraction Methods

Among different feature extraction, Bagging predictor combined with 188-dimensions feature extraction to obtain the best

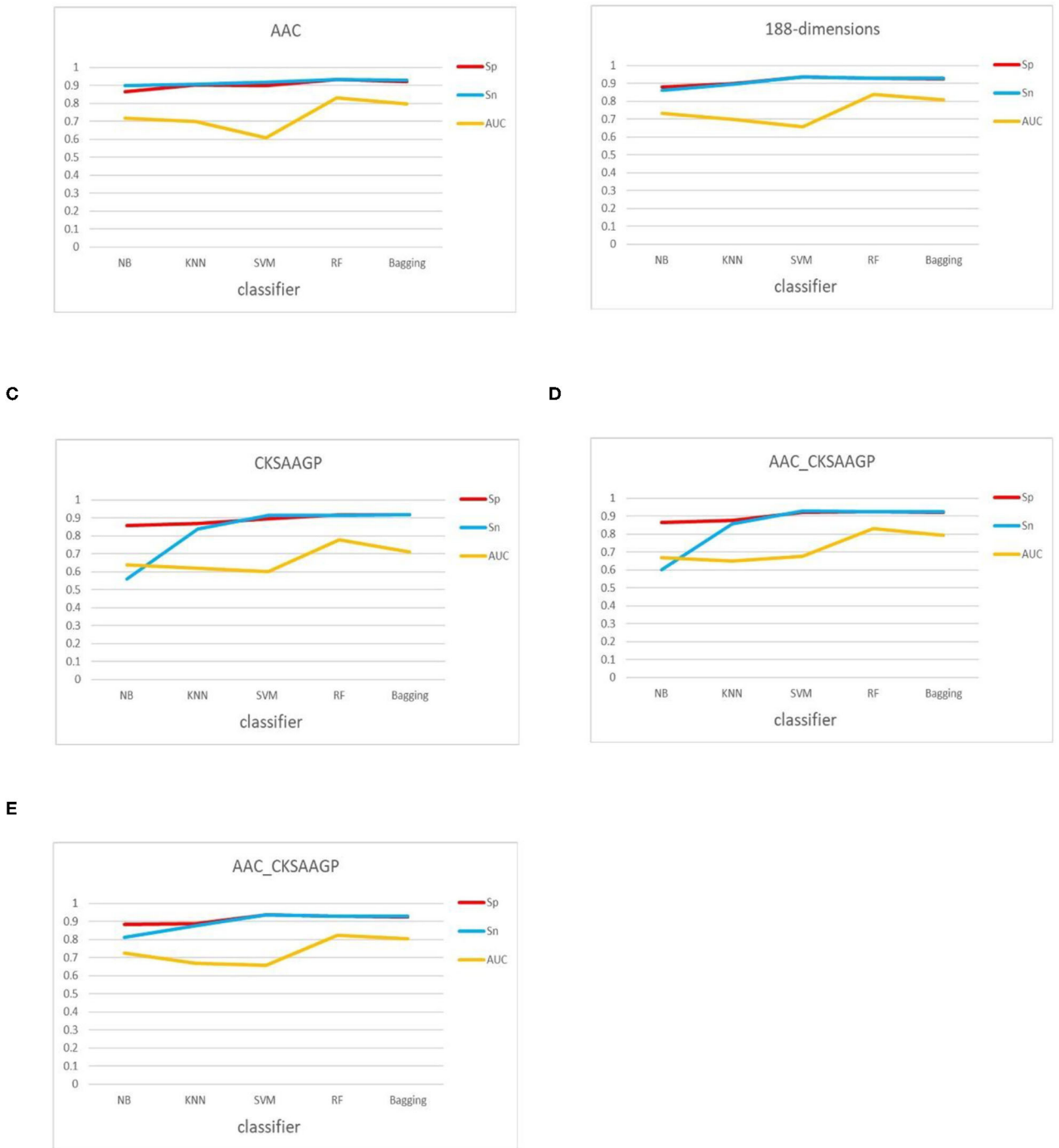

FIGURE 5 | Sn, Sp, and AUC values obtained by 10-fold cross-validation through five feature extraction methods [**(A–E)** AAC, 188-dimensions, CKSAAGP, AAC_CKSAAGP, and 188-CKSAAGP] combined with five classifiers (NB, KNN, SVM, RF, Bagging).

prediction performance. NB, KNN and RF predictors combined with AAC feature extraction to obtain the best prediction performance. The prediction models built by SVM combined with 188-dimensions or 188_CKSAAGP feature extraction have obtained the best prediction results (**Figure 4**).

In addition, the classification effect of the classifier constructed by 188-dimensions combined with different classification

algorithms in terms of Sn, Sp, MCC, AUC is higher than other feature extraction methods, which proves that the prediction model of the former is better than other models (**Figure 5**). In the test datasets, SVM combined with 188-dimensions obtained a prediction accuracy of 92.77%, which proves that the prediction model is reliable. Therefore, in this study, we use 188-dimensions as the final feature extraction method.

CONCLUSION

Rapid and accurate identification of tobacco mosaic virus is the key to successfully protecting tobacco from poison. Kumar and Prakash (2016) used direct antigen coating enzyme linked immunoassay (DAC-ELISA) technique to detect the TMV virus from pepper samples. However, this method is very complicated in sample preparation and detection processes, which is time-consuming and labor-intensive. Our goal was to distinguish between tobacco mosaic virus proteins and healthy tobacco leaf proteins in a large amount of data. The work in this paper provides an effective method to solve this problem.

In this experiment, first, we constructed a high-quality benchmark tobacco mosaic virus protein data set, which ensures the reliability of the classification tool. Secondly, we compared the performance of five feature extraction and five classifier constructs as predictors through 10-fold cross-validation, and then validated each model with the test datasets. The results show that SVM combined with 188-dimensions feature extraction method has the best prediction performance. It has obtained 93.58% accuracy on the train datasets and 92.77% accuracy on the test datasets, which proves that the prediction model has good robustness, so this paper chooses support vector machine as the prediction engine. We hope that these findings will help the development of identification of tobacco mosaic virus.

In future research, because feature selection technology has been successfully applied to some biological information experiments (Dong et al., 2015), feature selection on protein data can improve the prediction effect of the classifier. In addition, we will also try to use machine learning methods to solve analytical problems in genomics (Cheng et al., 2018c), epigenomics (Wang et al., 2017), and other proteomics fields.

DATA AVAILABILITY STATEMENT

Experimental data can be obtained from the corresponding author according to the reasonable request.

AUTHOR CONTRIBUTIONS

Y-MC collected the datasets. X-PZ processed the datasets. Y-MC and X-PZ designed the experiments. X-PZ did and analyzed the experiments' result. Y-MC contributed to the writing of this paper. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.569100/full#supplementary-material

REFERENCES

Abellán, J., Castellano, J. G., and Mantas, C. J. (2017). A new robust classifier on noise domains: bagging of credal C4.5 trees. *Complexity* 2017:9023970. doi: 10.1155/2017/9023970

Azuaje, F., Witten, I. H., Frank, E. (2006). Data mining: practical machine learning tools and techniques 2nd edition. *BioMed. Eng. Online* 5:51. doi: 10.1186/1475-925X-5-51

Baratloo, A., Hosseini, M., Negida, A., and El Ashal, G. (2015). Part 1: simple definition and calculation of accuracy, sensitivity and specificity. *Emergency* 3, 48–49. doi: 10.1111/j.1945-5100.2007.tb00551.x

Bhasin, M., and Raghava, G. P. S. (2004). Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* 279, 23262–23266. doi: 10.1074/jbc.M401932200

Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., and Chen, Y. Z. (2003). SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31, 3692–3697. doi: 10.1093/nar/gkg600

Chang, C.-C., and Lin, C.-J. (2007). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27. doi: 10.1145/1961189.1961199

Chen, Z., Zhao, P., Li, F., André, L., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

Chen, Z., Zhao, P., Li, F., Leier, A., Song, J., Marquez-Lago, T. T., et al. (2020). PROSPECT: A web server for predicting protein histidine phosphorylation sites. *J. Bioinform. Comput. Biol.* 18:2050018. doi: 10.1142/S0219720020500183

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018a). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 11, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018b). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19:919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48, D556–D560. doi: 10.1093/nar/gkaa511

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2018c). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144. doi: 10.1093/nar/gky1051

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018d). Exposing the causal effect of c-reactive protein on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Front. Genet.* 9:657. doi: 10.3389/fgene.2019.00085

Chou, K.-C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43, 246–255. doi: 10.1002/prot.1035

Deng, Z., Zhu, X., Cheng, D., Zong, M., and Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing* 195, 143–148. doi: 10.1016/j.neucom.2015.08.112

Ding, C., Yuan, L. F., Guo, S. H., Lin, H., and Chen, W. (2012). Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. *J. Proteomics* 77, 321–328. doi: 10.1016/j.jprot.2012.09.006

Dong, Q., Shanyi, W., Kai, W., Xuan, L., and Liu, B. (2015). "Identification of DNA-binding proteins by auto-cross covariance transformation," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Washington, DC), 470–5. doi: 10.1109/BIBM.2015.7359730

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. doi: 10.1073/pnas.92.19.8700

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

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Goldstein, B. A., Polley, E. C., and Briggs, F. B. S. (2011). Random forests for genetic association studies. *Stat. Appl. Genet. Mol. Biol.* 10, 32–32. doi: 10.2202/1544-6115.1691

Hajian-Tilaki, K. (2014). Sample size estimation in diagnostic test studies of biomedical informatics. *J. Biomed. Inform.* 48, 193–204. doi: 10.1016/j.jbi.2014.02.013

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2008). The WEKA data mining software: an update. *SIGKDD Explor. Newsl* 11, 10–18. doi: 10.1145/1656274.1656278

Hu, T.-W., and Lee, A. H. (2015). Commentary: Tobacco control and tobacco farming in African countries. *J. Public Health Policy* 36, 41–51. doi: 10.1057/jphp.2014.47

Huang, Y., Liu, N., Wang, J. P., Wang, Y. Q., Yu, X. L., Wang, Z. B., et al. (2012). Regulatory long non-coding RNA and its functions. *Biochemistry* 68, 611–618. doi: 10.1007/s13105-012-0166-y

Jamshid, P., Ali Reza, K., and Maryam, J. (2018). GENIRF: an algorithm for gene regulatory network inference using rotation forest. *Curr. Bioinform.* 13, 407–419. doi: 10.2174/1574893612666170731120830

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. J. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* 8, 282–293. doi: 10.1504/IJDMB.2013.056078

Kim, H.-J., Jo, N.-O., and Shin, K.-S. (2016). Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. *Expert Syst. Appl.* 59, 226–234. doi: 10.1016/j.eswa.2016.04.027

Kou, G., and Feng, Y. J. (2015). Identify five kinds of simple super-secondary structures with quadratic discriminant algorithm based on the chemical shifts. *J. Theor. Biol.* 380, 392–398. doi: 10.1016/j.jtbi.2015.06.006

Kumar, M., Gromiha, M. M., and Raghava, G. P. S. (2015). SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J. Mol. Recognit.* 24, 303–313. doi: 10.1002/jmr.1061

Kumar, S., and Prakash, H. S. (2016). Detection of tobacco mosaic virus and tomato mosaic virus in pepper seeds by enzyme linked immunosorbent assay (ELISA). *Arch. Phytopathol. Plant Protect.* 49, 59–63. doi: 10.1080/03235408.2012.658991

Lan, L., Djuric, N., Guo, Y., and Vucetic, S. (2013). MS-kNN: protein function prediction by integrating multiple data sources. *BMC Bioinform.* 14(Suppl. 3):S8. doi: 10.1186/1471-2105-14-S3-S8

Li, J., and Fine, J. P. (2010). Weighted area under the receiver operating characteristic curve and its application to gene selection. *J. Royal Stat. Soc.* 59, 673–692. doi: 10.1111/j.1467-9876.2010.00713.x

Liao, Z., Li, D., Wang, X., Li, L., and Zhou, Q. (2018). Cancer diagnosis through isomiR expression with machine learning method. *Curr. Bioinform.* 13, 57–63. doi: 10.2174/1574893611666160609081155

Lin, H., Han, L. Y., Cai, C. Z., Ji, Z., and Chen, Y. Z. (2005). Prediction of transporter family from protein sequence by support vector machine approach. *Proteins* 62, 218–231. doi: 10.1002/prot.20605

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* 17, 145–151. doi: 10.1111/j.1466-8238.2007.00358.x

Lv, H., Dao, F.-Y., Zhang, D., Guan, Z.-X., Yang, H., Su, W., et al. (2020). iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* 23:100991. doi: 10.1016/j.isci.2020.100991

Lyu, Z., Jin, S., Ding, H., and Zou, Q. (2019). A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.* 7:215. doi: 10.3389/fbioe.2019.00215

Manavalan, B., Shin, T. H., 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

Metzler, S., and Kalinina, O. V. (2014). Detection of atypical genes in virus families using a one-class SVM. *BMC Genomics* 15:913. doi: 10.1186/1471-2164-15-913

Niu, M., Li, Y., Wang, C., and Ke, H. (2018). RFAmyloid: a web server for predicting amyloid proteins. *Int. J. Mol. Sci.* 19:2071. doi: 10.3390/ijms19072071

Ru, X., Li, L., and Zou, Q. (2019). Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. *J. Proteome. Res.* 18, 2931–2939. doi: 10.1021/acs.jproteome.9b00250

Salama, M. A., Hassanien, A. E., and Mostafa, A. (2016). The prediction of virus mutation using neural networks and rough set techniques. *EURASIP J. Bioinform. Syst. Biol.* 2016:10. doi: 10.1186/s13637-016-0042-0

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Tastan, O., Qi, Y., Carbonell, J. G., and Kleinseetharaman, J. (2012). Prediction of interactions between HIV-1 and human proteins by information integration. *Pac. Symp. Biocomput.* 2009, 516–527. doi: 10.1142/9789812836939_0049

Thakur, A., Rajput, A., and Kumar, M. (2016). MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine. *Mol. Syst.* 12, 2572–2586 doi: 10.1039/C6MB00241B

The Uniprot, C. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2017). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151. doi: 10.1093/nar/gkx1096

Wang, G., Wang, Y., Feng, W., Wang, X., Yang, J. Y., Zhao, Y., et al. (2008). Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genomics* 9:S22. doi: 10.1186/1471-2164-9-S2-S22

Wang, G., Wang, Y., Teng, M., Zhang, D., Li, L., and Liu, Y. (2010). Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon γ-stimulated HeLa cells. *PLoS ONE* 5:e11794. doi: 10.1371/journal.pone.0011794

Wang, Y., Shi, F., Cao, L., Dey, N., Wu, Q., Ashour, A. S., et al. (2019). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinform.* 14, 282–294. doi: 10.2174/1574893614666190304125221

Wang, Y., Wang, Y., Yang, Z., and Deng, N. (2011). Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical *context*. *BMC Syst. Biol.* 5 (Suppl. 1):S6. doi: 10.1186/1752-0509-5-S1-S6

Xing, Y.-Q., Liu, G.-Q., Zhao, X.-J., Zhao, H.-Y., and Cai, L. (2014). Genome-wide characterization and prediction of Arabidopsis thaliana replication origins. *Biosystems* 124, 1–6. doi: 10.1016/j.biosystems.2014.07.001

Xue, Y., Chen, H., Jin, C., Sun, Z., and Yao, X. (2006). NBA-Palm: prediction of palmitoylation site implemented in Naïve Bayes algorithm. *BMC Bioinform.* 7:458. doi: 10.1186/1471-2105-7-458

Yang, H., Yang, W., Dao, F.-Y., Lv, H., Ding, H., Chen, W., et al. (2019a). A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. *Brief. Bioinform.* 21, 1568–1580. doi: 10.1093/bib/bbz123

Yang, W., Zhu, X.-J., Huang, J., Ding, H., and Lin, H. (2019b). A brief survey of machine learning methods in protein sub-golgi localization. *Curr. Bioinform.* 14, 234–240. doi: 10.2174/1574893613666181113131415

Zhang, M., and Zhou, Z. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern. Recogn.* 40, 2038–2048. doi: 10.1016/j.patcog.2006.12.019

Zhang, S., Hao, L., and Zhang, T. (2014). Prediction of protein–protein interaction with pairwise kernel support vector machine. *Int. J. Mol. Sci.* 15, 3220–3233 doi: 10.3390/ijms15023220

Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA promoter identification in arabidopsis using multiple histone markers. *BioMed. Res. Int.* 2015:861402. doi: 10.1155/2015/861402

Zhong, W., Zhong, B., Zhang, H., Chen, Z., Chen, Y. (2020). Identification of anti-cancer peptides based on multi-classifier system. *Comb. Chem. High Throughput Screen.* 22, 694–704 doi: 10.2174/1386207322666191203141102

Zhu, Y., Jia, C., Li, F., and Song, J. (2020). Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling. *Anal. Biochem.* 593:113592. doi: 10.1016/j.ab.2020.113592

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). *Sequence clustering in* bioinformatics: an empirical study. *Brief Bioinform.* 21, 1–10. doi: 10.1093/bib/bby090

Conflict of Interest: 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.