Journal of Computers Vol. 34 No. 5, October 2023, pp. 317-328
doi: 10.53106/199115992023103405024

A Machine Learning Based Computational Method for Prediction of Functional SNPs in Rice Genome

Rong Li, Zhi-e Lou*

Telecommunication and Networks National Laboratory, Nanjing University of Posts and Telecommunications,
Nanjing 210003, P.R. China

{lirong, louzhie}@njupt.edu.cn

Received 22 April 2022; Revised 5 May 2022; Accepted 11 May 2022

Abstract. Single nucleotide polymorphisms (SNPs) are the most prevalent and stable class of genetic diversity that exist in most organisms. Functional SNPs are the most commonly used genetic markers for diversity study and molecular breeding in plants, and their quick recognition is in urgent demand. In this work, a computational approach to identify functional SNPs in rice genome based on machine learning is presented. To characterize and prioritize variants, two different categories of features, the nucleotide-sequence based features and the allele-specific based features, are extracted. In particular, the weighted Euclidean distance is employed to measure the changes of the transcription factors (TFs) binding affinities caused by SNPs. To deal with the classification problem on unbalanced data, the support vector machine (SVM) together with an oversampling method is employed. We use mRMR to find the optimal feature set, and the result shows that our method can achieve accuracy with sensitivity of ~74.2% and specificity of ~72.3% after 10-fold cross-validation. Furthermore, the sources of data to build the proposed prediction model are mainly sequence context of SNP and TF profiles in JASPAR database, which are all easy to be acquired. So, the prediction method can be easily applied to other plant species.

Keywords: transcription factor binding affinity, position weight matrix, functional SNP, support vector machine

1 Introduction

Rice (Oryza sativa L.), one of the most important food crops, provides the daily dietary intake for approximately 50% of the worldwide human populations. The genome size of rice is limited and diploid, owing to this reason, rice has become one of the most excellent choices for initiating genomic studies among the cereal food species. So, rice serves as a model organism for agricultural research and plant biology. Since the first two subspecies of rice, i.e., japonica (cultivar Nipponbare) and indica (cultivar 93-11), have been sequenced in 2002 [1, 2], numerous rice accessions in the germplasm have been genotyped in the past decade [3]. At the same time, huge amounts of rice variation databases have been constructed in the wake of developments in sequencing technologies [4]. All these genomic data provide us with abundant resources for rice genomics research and breeding.

It's known that nucleotide variants can lead to different gene-phenotype or gene-trait associations and thus translate into phenotypic diversity of plants, so they are playing increasingly important roles in plant breeding [5]. Among all the nucleotide variants, Single Nucleotide Polymorphism (SNP), the DNA sequence polymorphism at the genomic level caused by variant of a single nucleotide, is the most common genetic variant, and SNP is also the most prevalent and stable type of genetic diversity that exists in most organisms [6-8]. As SNP genotyping is becoming faster and more cost-effective, it is widely used as the genetic marker for diversity study and molecular breeding in plants [9, 10]. However, the number of SNP in rice genome is very huge, with nearly 32M [11], thus bringing a very large obstacle to conducting genomic studies. The cost and time of genotyping is one of the most important factors to consider in plant molecular breeding, as molecular breeding usually requires the rapid genotyping of thousands of samples, often within days or even hours. So, low SNP density genotyping technologies are very necessary, because they can offer great flexibility through the rapid detection of a small quantity of candidate SNPs which have the ability to mark thousands of DNA samples [12]. In this context, the question facing us is how to quickly and accurately identify those functional trait-associated SNPs for complex phenotypes.

* Corresponding Author

Although the number of SNPs in genomes is very large, only a small fraction of them have been found to be associated with phenotype or trait. In the era of biological data outbreak, it is time- and money-consuming to identify functional SNPs only by experimental methods. Therefore, computational methods are very essential to be used to help identify functional SNPs. Previous studies of functional SNPs identification through computational methods have mainly focused on disease-associated SNPs in the human genome. GWAVA [13] trains random forest classifiers to differentiate functional SNPs based on numerous features extracted from different annotation sources, such as open chromatin, RNA polymerase binding and evolutionary conservation. CERENKOV2 [14] constructs a 248-dimensional feature matrix using a large amount of genome annotation data from the Encyclopedia of DNA Elements (ENCODE) project [15], the Ensembl project [16], the Genotype Tissue Expression (GETx) project [17], etc., and then trains gradient boosted decision trees to identify functional variants. However, few of these methods can be directly used to identify the functional SNPs in plants, for reason that these methods always need large amounts of human genome annotation data. But for plant genomes, such a large number and large scale of sequencing and annotation projects have not been launched yet, making the same type of annotation data difficult to be obtained in plant genomes. Kharabian et al. estimate the influence of SNPs in rice GBSSI gene through a series of functional elements prediction tools [18]. The limitation of this method is that the functional elements prediction tools they used are mainly designed for human genome, so the effectiveness of these tools remains to be tested and verified for plants, and it isn't suitable for the rapid prediction of large-scale functional SNPs.

Here, we are targeted at the prediction of functional SNPs in rice genome. A set of documented trait-associated SNPs found by Genome Wide Association Studies (GWAS) and functionally neutral SNPs were used to build the prediction model. To characterize and prioritize the variants, an extensive range of characteristics, such as the position weight matrix (PWM) scores of *k*-mers, significant motif scores and changes of transcription factor (TF) binding affinities are analyzed. After feature selection using mRMR, the SVM classifier together with a powerful over-sampling method – G-SMOTE is utilized to find the optimal feature sets and present the final prediction results of the unbalanced data. We hope such a prediction method will help geneticists to rapidly assess likely functional SNPs from massive background genetic polymorphisms for feature diversity study and molecular breeding research in plants.

2 Materials and Methods

2.1 Datasets

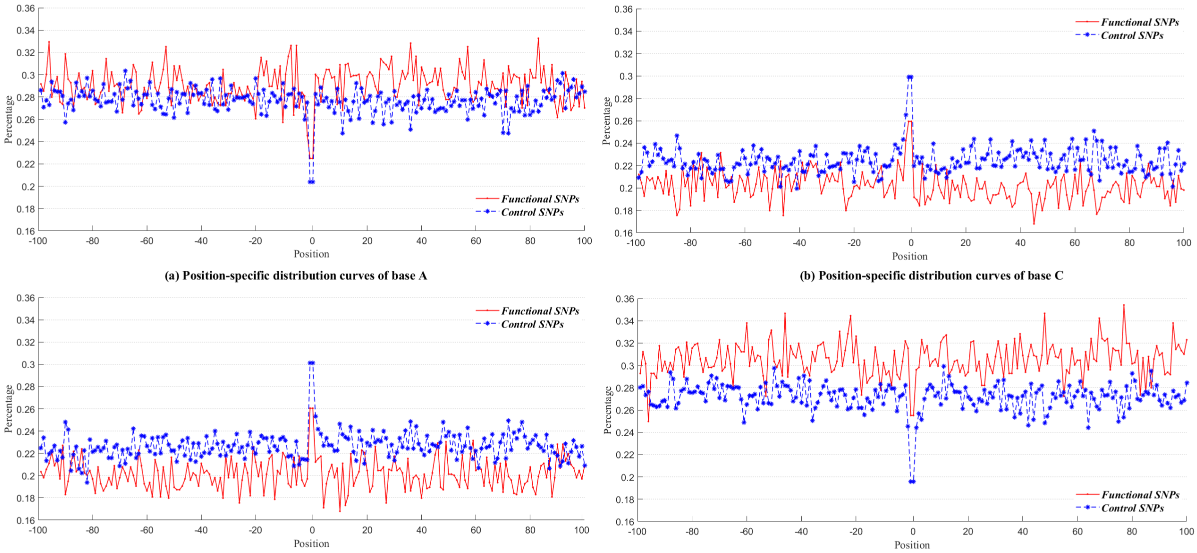
The functional SNPs were retrieved from the Rice SNP-Seek Database (http://snp-seek.irri.org/) [11]. To obtain a high-quality functional SNPs dataset, all SNPs with association *p*-values less than 10^{-5} to the recorded traits for Nipponbare (japonica) rice were downloaded first. 1106 SNPs were collected at this step, and they were SNPs with high association to all the 11 reported traits for Nipponbare rice, including grain weight, grain width, culm length, leaf length, seeding height, etc. Then these SNPs were mapped to dbSNP database (dbSNP build 151, Rice Genome Build IRGSP-1.0) [19] to ensure their validity. Finally, 929 SNPs were left as the positive samples. From the chromosome and position information of the SNP nucleotide provided by dbSNP, the SNP sequence of any length can be retrieved from the complete sequences of rice genome stored in NCBI [20].

For control set, neutral SNPs were retrieved from dbSNP. 350 SNPs were randomly chosen from each chromosome first and then those also exist in our collected functional SNPs dataset were removed. Then the redundant data with a sequence identity cut-off threshold of 0.8 were picked away using CD-HIT [21], a widely used program for clustering biological sequences thus reducing redundancy. The purpose of removing redundancy was to get a high quality control dataset and meanwhile reduce the difference in the number of positive and negative samples. 3747 SNPs were remained as the functionally neutral control samples finally.

2.2 Feature Extraction

Nucleotide-sequence Based Features. Nucleotide-sequence based features just take the characteristics of the reference sequence of an SNP into consideration while regardless of the variant nucleotide. This kind of feature is designed to find the genome region in which if a variant happens, the variant would more likely to cause trait changes of an organism. We counted the position-specific nucleotide distribution profiles of sequences that surrounding the SNP nucleotide with length 201bp (+/-100bp of the SNP site) and the result was shown in Fig. 1. The base A and base T of functional SNPs seemed to appear more frequently than that of control SNPs in positions both upstream and downstream of the SNP nucleotide, while the base C and base G revealed the opposite situation. The PWM scores of k-mers were used to describe the differences in position-specific nucleotide distributions between functional and control SNPs accordingly.

***k*-mers PWM Scores.** The term k-mers originally refers to all substrings of length k in a string. In computational genomics, k-mers refer to all subsequences of length k in the sequence fragments obtained by DNA sequencing.



(a) Position-specific distribution curves of base A

(b) Position-specific distribution curves of base C

(c) Position-specific distribution curves of base G

(d) Position-specific distribution curves of base T

Fig. 1. Position-specific distribution profiles of nucleotides for both functional and control SNPs, X-axis represents the nucleotide position with 0 being the SNP site

As there are only 4 bases in DNA, that are A, C, G, and T, the total number of k-mers is 4^k. For example, the all 16 forms for 2-mers are AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT. The k-mers strings are often used to find genomic regions of interest and can also be used to predict biological patterns of interest by calculating the probability distributions of a large number of k-mers [22]. PWM, also known as the position-specific scoring matrix (PSSM), was put forward for the first time by Stormo et al. as an alternative to consensus sequences [23]. PWM has been widely used to depict the conservative sequence patterns in computational biology [22]. The mathematical expression formula of the k-mers PWM is as follows:

$$W=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1i} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2i} & \cdots & w_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{j1} & w_{j2} & \cdots & w_{ji} & \cdots & w_{jm} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{ni} & \cdots & w_{nm} \end{bmatrix}. \quad (1)$$

where $n=4^k$ is the number of all k-mers; $m=l-k+1$ is the total number of position-specific k-mers in the sequences of length $l;$ the log-likelihood ratio of finding a pattern (the jth k-mer, $1\leq j\leq 4^k$) at a given position i was denoted as $w_{j,i}$, and it was calculated as follows:

$$w_{j,i} = \ln\frac{p_{j,i}}{p_{j,0}} = \ln\frac{(f_{j,i}+\sqrt{N_i}\,/\,4^k)\,/\,(N_i+\sqrt{N_i})}{(1/4)^k} \quad . \tag{2}$$

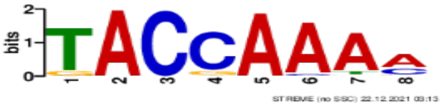
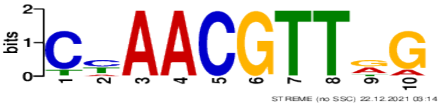
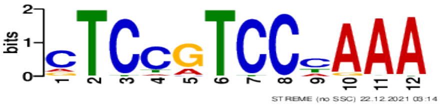
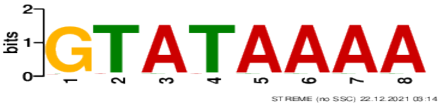
where $p_{j,i}$ represents the probability of finding a pattern (the jth k-mer) at the given position i, and $p_{j,0}$ is the priori probability of finding the pattern in genome with $p_{j,0}$=$(1/4)^k$ given here; $f_{j,i}$ refers to the occurrence frequency of the jth k-mer at given position i; N_i=$\sum_j f_{j,i}$ is the number of all samples at position i; $\sqrt{N_i}\,/\,4^k$ denotes a pseudo-count proportional to the standard deviation of the counted frequencies according to Kielbasa et al. [24]. Then, the final PWM score S_N was computed by accumulating all the weights corresponding to different patterns observed in a sequence within a window of length N:

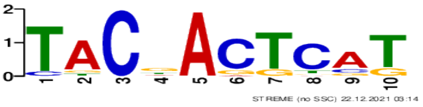
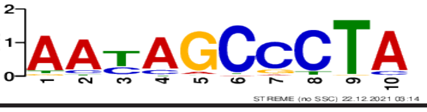
$$S_N = \sum_j \sum_{i=1}^{N-k+1} w_{j,i} \quad . \tag{3}$$

For an input sequence of the same length with the calculation window, a score can be generated corresponding to equation (3), and the larger the score is, the more likely it will be the functional site described by the matrix. The PWM score can effectively reflect the overall similarity between the testing sequence and the training sample sequences. In this work, we chose k=2 and a sliding window of length 10 and overlap 5 was used to catch S_N. Scores were calculated by $w_{j,i}$ generated from both the positive and negative datasets using equations (1) and (2). Therefore, a 4(L/5-1)-dimensional feature vector was extracted here to an SNP sequence with both flanking sequence lengths being L.

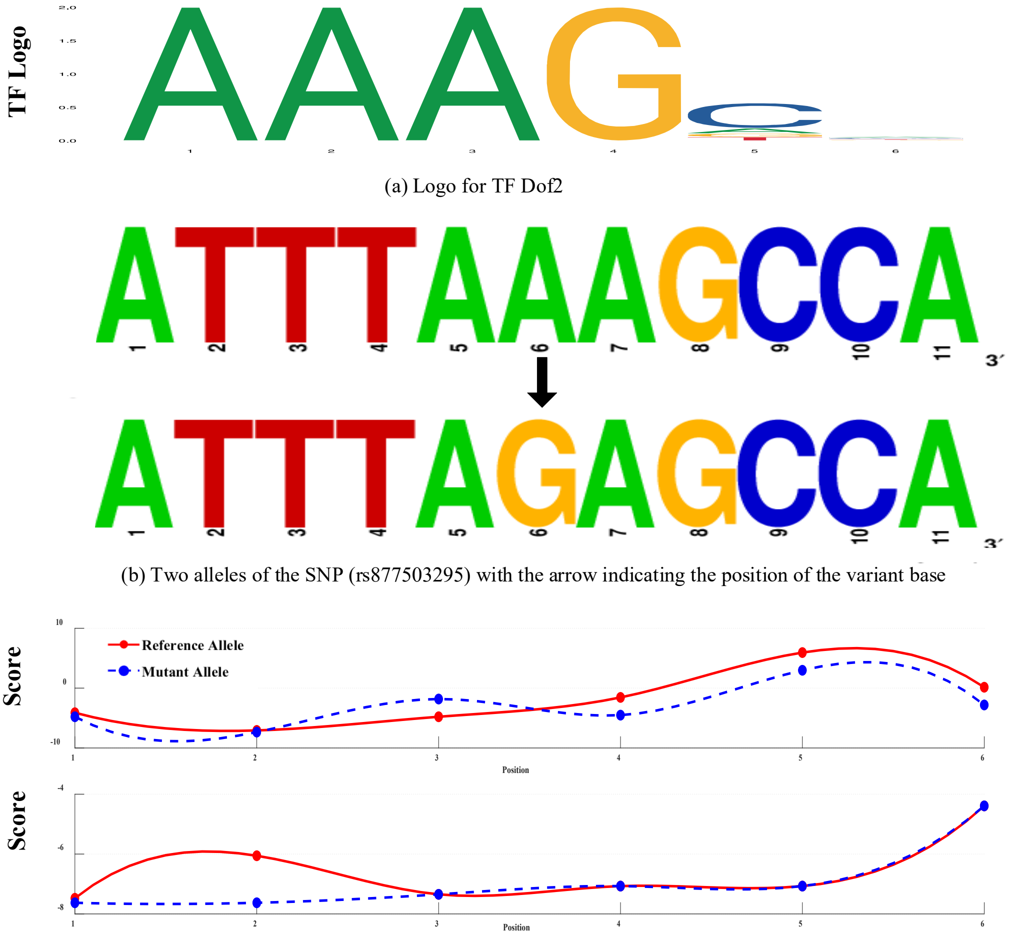
Significant Motif Scores. STREME [25] (version 5.4.1) was used to discover the most statistically significant motifs in our positive dataset. As the result shown in Table 1, 6 motifs were found to have statistically significance (p-value < 0.05). Subsequently, the position-specific frequency matrices (PSFMs), that were matrices comprising of the frequencies of nucleotides occurring at a specific location (namely $f_{j,i}$ in equation (2)), of these 6 motifs were obtained from STREME tool. Then the PSFMs were converted to PWM using equation (2) and the matching scores between a given DNA fragment and these frequent motifs were calculated by equation (3). When calculating the matching score for each SNP sequence, the window size N was equal to the length of the corresponding motif. In other words, a group of matching scores for a long sequence at each position were generated using a shifting window with the same length as the motif to be matched. So, for an SNP sequence with length L, L-N+1 scores will be generated totally for each motif. Finally, the maximum score was selected as the final feature to measure the matching degree of the SNP sequence and each motif. Thus, a 6-dimensional feature vector for each SNP was generated here.

Table 1. The most statistically significant motifs found by STREME and their corresponding p-values

No.	Motif	Logo	p-value
1	TACCAAAA		5.6×10^{-5}
2	CYAACGTTRG		3.2×10^{-4}
3	CTCCGTCCCAAA		7.6×10^{-4}
4	GTATAAAA		1.7×10^{-3}

5	TACKACTCAT		2.6×10^{-2}
6	AATAGCCCTA		5.0×10^{-2}

Allele-specific Based Features. The allele-specific based features were drawn to catch the effects caused by the variant base. Feature of this category was delta score of the TF binding affinities caused by SNPs. Kasowski et al. [26] analyzed the binding patterns of the TFs Pol-II and NF-kB on the genomes of different human individuals, including the gorilla. They found that the binding patterns of Pol-II and NF-kB were 75% and 92.5% similar among different individuals respectively, while the remaining ones were greatly affected by genetic variants. Further studies revealed that the differential binding patterns of TFs on noncoding DNA caused by variants can influence the expression of surrounding genes, thus relating to individual phenotypic differences and disease susceptibility. Variants in plant genomes have also been found to be phenotype- or trait-associated by changing TFs bindings. For example, the nucleotide variants in the promoter region of a TF (Ghd8) were found to be able to control grain number, plant height and heading date in rice [27]. So, the changes of TF binding affinities caused by SNP variants were used here to distinguish functional SNPs from the neutral ones.



(a) Logo for TF Dof2

(b) Two alleles of the SNP (rs877503295) with the arrow indicating the position of the variant base

(c) Scores generated by Dof2 PWM for both the forward (the top figure) and reverse (the bottom figure) sequences

Fig. 2. The changes of Dof2 TF binding scores caused by the SNP rs877503295

The data of TF binding profiles was got from JASPAR [28], an open access database containing manually curated, non-redundant TF binding profiles for TFs across six taxonomic groups. The profiles of all 656 TFs contained in the core collection for plants in JASPAR were downloaded. With the frequency matrix of these TFs, TFs binding affinities can be measured using equation (3), and then the changes on TFs binding affinities caused by SNP variant can be observed. To clarify further, we use an SNP (rs877503295, shown in Fig. 2(a)) as an example to demonstrate how to calculate the changes on TFs binding affinities caused by the variant. The Dof2 (shown in Fig. 2(b)) was chosen as the example TF, as the Dof family TFs had been reported for their involvement in numerous important developmental processes and responses to various environmental stresses in rice [29]. The influence of the TF binding score at each position caused by the variant was shown in Fig. 2(c) for both the forward and reverse sequences. The number of nucleotide positions where the binding scores will be changed by SNP variant is equal to the length of the corresponding TF. Then, the score change of TF binding affinity ($\Delta Score$) was measured by the weighted Euclidean distance as follows:

$$\Delta Score = \sqrt{\sum_i w_i \times \left(Sr_i - Sa_i\right)^2} = \sqrt{\sum_i \max_i \left(Sr_i, Sa_i\right) \times \left(Sr_i - Sa_i\right)^2} \ . \quad \textbf{(4)}$$

where Sr_i and Sa_i are binding scores at the position i of the reference and alternative allele respectively, and i (i= SNP_{pos}-TF_{len}+1: SNP_{pos}, SNP_{pos} is the position of SNP site and TF_{len} is the length of the potential TF) represents the position where the binding scores will be affected by SNP variant. We chose the maximum value among Sr_i and Sa_i as the weight, considering that a larger binding score means that the site is more likely to be a TF binding site, and so variant happened at this site is more likely to be functional. $\Delta Score$ for both the forward and reverse sequences were calculated, and then the larger one was chosen. Among the 656 TFs, 94 TFs show statistically significant differences ($p<0.01$) in binding scores between the collected functional and background SNPs using Mann-Whitney U-test. Here, a 94-dimensional feature vector for each SNP sample was generated.

2.3 Feature Selection

The mRMR (Minimum Redundancy Maximum Relevance) method [30], one of the most widely used feature selection algorithms, was used to select the relevant features in our work. mRMR can rank a set of features according to their importance based on their relevance to the target for a given classification task, and the redundancy of features will be penalized at the same time. In other words, mRMR reduces the feature dimensions by finding a set of features that have the most correlation with the final output, but the least correlation with each other. The maximum dependency between a set of features and the class was measured by mutual information. The mutual information (denoted as I) between feature pairs was defined as follows:

$$I\left(A;B\right) = \sum_{b \in B} \sum_{a \in A} p\left(a,b\right) \log\left(\frac{p\left(a,b\right)}{p\left(a\right)p\left(b\right)}\right) . \quad \textbf{(5)}$$

where p(a) and p(b) are the marginal probabilities of the two features vectors A and B, and $p(a, b)$ is the joint probability between them. Given a dataset $D=\{(x_i, y_i)\}, i=1,2,\ldots,n$, where $x_i=[f_1,f_2,\ldots,f_p]^T \in R^p$ is a p-dimensional feature vector and $y_i \in \{1,2,\ldots,k\}$ is the corresponding class label, the purpose of mRMR is to select a feature subset that has the maximum correlation between features and the label while has the minimum relevance among features. Due to the huge search space of feature subsets, the incremental feature selection strategy was adopted in mRMR, that is, the feature having the maximum mutual information with the class label was selected at first and then the m+1 feature was chosen based on the fore m ($m \geq 1$) features according to the criteria shown in the following equation:

$$\max_{f_j \in G - S_m} \left[I\left(f_j;y\right) - \frac{1}{m} \sum_{f_i \in S_m} I\left(f_i;f_j\right) \right] . \quad \textbf{(6)}$$

where, G is the complete feature set; S_m is the subset consisting of the selected fore m features; $I(f_j; y)$ is the relevance between feature f_j and class label y; and $I(f_i; f_j)$ is the mutual information between feature f_i and feature f_j.

2.4 G-SMOTE Sampling

To solve the problem of biased error caused by imbalanced data, a powerful oversampling method – G-SMOTE (Geometric Synthetic Minority Over Sampling Technique) was applied before classifying [31]. The main principle of G-SMOTE is increasing the number of the minority class by oversampling and new synthetic samples are created by data interpolation into a geometric region around each selected minority instance. Here, the number of neighbors was chosen as 5 and 2818 feature vectors were interpolated into the functional SNP feature space. So, the positive and negative datasets will have a sample size ratio from 1:4 to 1:1.

2.5 SVM Classifier

The support vector machine (SVM) is a very useful method for two-target classification problems [32, 33]. In our work, the freely available LIBSVM toolbox was employed to actualize the training and prediction procedures on our collected dataset [34]. When facing nonlinear samples, the SVM method transforms the input feature space into a high-dimensional space through a nonlinear transformation defined by the inner product function (namely the kernel function), and then SVM looks for the linear relationships between the input variables and the output in this high-dimensional space. The kernel function is one of the most important parts in SVM. However, there is still no theoretical basis for selecting the right kernel function, and it's mainly tried in experiments. The most commonly used kernel functions are Linear kernel, Polynomial kernel and Radial Basis Function (RBF) kernel. For reasons that RBF kernels are generally the most widely used and RBF kernels can not only classify more multidimensional functions compared to Linear kernel but also need less parameters compared to Polynomial kernel, the RBF kernel was chosen to apply LIBSVM in our work.

Parameter setting is also very important for conducting SVM method. The RBF kernel based SVM requires two important pre-determined parameters including the penalty coefficient *c* and the width coefficient of the kernel *g*. These two key tuning parameters were optimized by implementing a grid search of the parameter space using grid.py, a python script included inside LIBSVM. The grid search was repeated several times on different subsets generated by subset.py, which was also a python script in LIBSVM. Other insensitive parameters and parameters related to model settings were set with the default values. Since the SNP is only a single nucleotide, sequences of a certain length needed to be obtained by extending upstream and downstream nucleotides that centered on the SNP site when conducting parameter optimization and classification. In our work, the sequence length was chosen at 201bp (100bp for both flanks). After parameter optimization, finally, the two parameters were set at *c*=10 and *g*=0.02.

3 Results

3.1 Performance Evaluation

For unbalanced data, the accuracy (*ACC*) can't reflect the effect of the classification algorithm very well, as ACC will be more affected by the classification effect of the majority class. Sensitivity (*SN*) and specificity (*SP*) are the most commonly used indicators for evaluating unbalanced data classification problems. *SN* is the accurate identification rate of the model on positive samples, while *SP* measures the classification accuracy of the model on negative samples. The Matthews Correlation Coefficient (*MCC*) is also used to evaluate the comprehensive performance of our proposed method. Their calculation formulas are as follows:

$$\begin{gathered} SN=\frac{TP}{TP+FN};SP=\frac{TN}{TN+FP}; \\ ACC=\frac{TP+TN}{TP+FN+TN+FP}; \\ MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}\ . \end{gathered} \tag{7}$$

where *TP* is true positive; *TN* is true negative; *FN* is false negative and *FP* is false positive numbers.

3.2 Performance on Collected Functional SNPs

As an SNP is just a single nucleotide, the SNP sequence with a certain length can be obtained by expanding upstream and downstream. The sequence length we chose to build the prediction method was 201bp, that was a sequence with 100bp flanking sequences both upstream and downstream of the SNP site. For a sequence with length 201bp, there are 76 features of 2-mer PWM scores. So, after feature extraction, the total dimension of the feature vector was 176. In order to obtain a relatively objective result, 10-fold cross-validation was adopted. It means all data were divided into 10 groups randomly with one group selected as the test set while the remaining 9 as training sets, and then cycled for ten times until every last group was used as the test set.

To find the optimal feature combination, we examined the prediction effect of the proposed method with feature numbers from 1 to 176 selected by mRMR, and the result is shown in Fig. 3. From the figure, we can see the sensitivity increases with the feature dimension increases while specificity decreases. The *MCC* rises first and then decreases slightly with the feature dimension increasing. Taken together, when the feature dimension is 41, the method achieves the best comprehensive performance with *MCC* equal to 0.41. At this time, the sensitivity is 71.0% and the specificity is 77.1%. Of course, the aim we built the prediction model is to identify as many true functional SNPs as possible and meanwhile control the number of false positives. Under this criterion, a feature dimension of 42 may be a better choice, with the sensitivity being 74.2% and specificity being 72.3%.

The top 42 discriminating features selected by mRMR are listed in Table 2. Among these features, the kind of feature 2-mers PWM scores has the largest number, indicating that there are some differences between the base compositions of the functional and background SNP sequences. The features of significant motif scores also contribute much to discriminating functional SNPs in rice genome. The changes of 14 TFs binding affinities are also on the list, and previous studies have found that most of them are involved in the developmental and physiological processes of plants. For example, the ERF family are found to play an important role in signal transduction, plant growth, development, and response to various stresses [35], and the HSFs family are found to relate to the heat stress-responsive of plants thus significantly influencing plant growth and development [36].

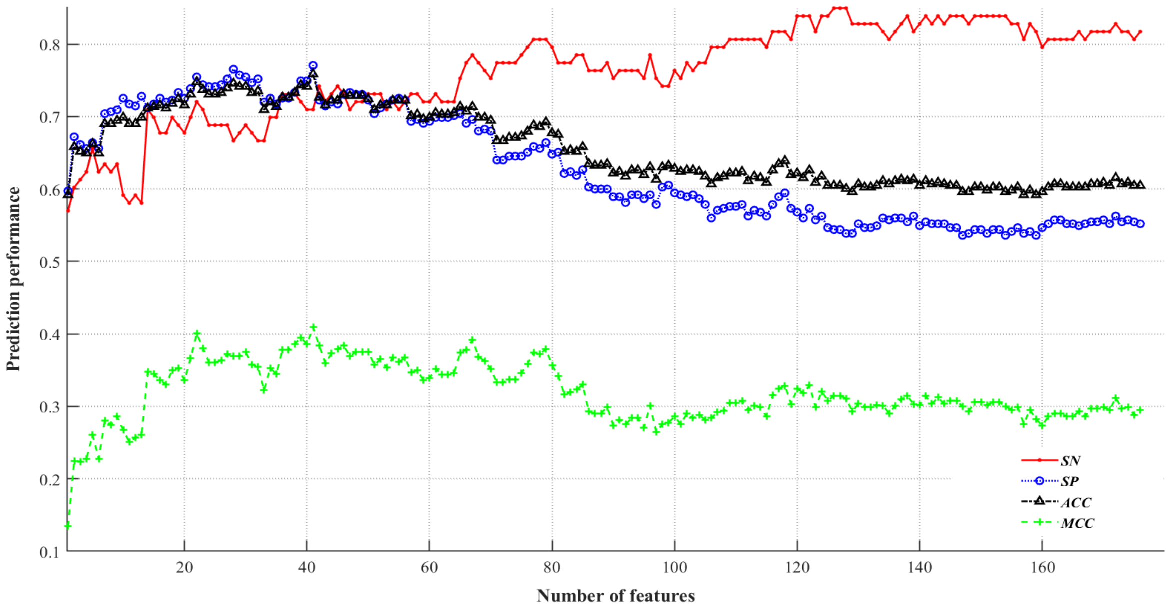


Fig. 3. The prediction performance of our method, X-axis represents the dimensions of different feature combinations using mRMR

3.3 Performance Comparison Against Other Methods

As there are few tools of functional SNPs prediction designed for plants, we compared the performance of our method to Blastn [37], a tool used for finding regions of similarity between biological sequences. Blastn can be used to infer functional and evolutionary relationships between nucleotide sequences. The principle of using Blastn to discover functional SNPs is that Blastn can find evolutionarily conserved regions between nucleotide sequences, and the conserved regions in genome are usually enriched with functional elements, so if variants happened in or near these regions, the variants are more likely to show function. After 10-fold cross-validation, 64.7% of testing positive samples are found to have some similarity with the training positive samples, under a relatively relaxed threshold (alignment length=30, %identity=70%). While 46.6% of testing negative samples are also found to have similarities with the training positive samples under the same threshold. Therefore, under the above-mentioned threshold, Blastn achieves performances of sensitivity 64.7% and specificity 53.4%. Of course, a more stringent threshold can significantly increase the specificity, but will also decrease the sensitivity. So, in general, the method we proposed performs better than Blastn.

Table 2. The top 42 features selected by mRMR

No.	Feature	No.	Feature
1	2-mers PWM score [1]	25	2-mers PWM score
2	Score of motif 6 [2]	26	Score of motif 2
3	ΔTF-MA1053.1_ERF109 [3]	27	ΔTF-MA1807.1_ZHD10
4-5	2-mers PWM score	28	2-mers PWM score
6	Score of motif 3	29	ΔTF-MA1666.2_HSFB2B
7	ΔTF-MA1188.1_At3g11280	30	2-mers PWM score
8	Score of motif 5	31	ΔTF-MA1060.1_SPL7
9	Score of motif 1	32-33	2-mers PWM score
10-11	2-mers PWM score	34	ΔTF-MA0549.1_BZR2
12	ΔTF-MA1761.1_HSFB3	35	ΔTF-MA0128.1_EmBP-1
13-16	2-mers PWM score	36-37	2-mers PWM score
17	ΔTF-MA0943.1_ARF5	38	ΔTF-MA1426.1_MYB124
18	2-mers PWM score	39	ΔTF-MA1794.1_NLP7
19	ΔTF-MA1083.2_WRKY30	40	2-mers PWM score
20-23	2-mers PWM score	41	ΔTF-MA1198.1_HAT2
24	ΔTF-MA0955.1_POPTR	42	2-mers PWM score

[1] This kind of features are *k*-mer PWM scores generated with a sliding window of length 10 and overlap 5.
[2] This kind of features are PWM scores of significant motifs, and the number here indicates the No. of motifs listed in Table 1.
[3] This kind of features are Δ score of TF binding affinities, and the symbol "-" follows by the potential TF matrix in JASPAR.

4 Conclusion

With the development of sequencing technologies, more and more SNPs are discovered in plants and other species. So, more efforts are in urgent demand in order to fully interpret their biological functions. In this paper, a new computational method for identifying functional SNPs in rice genome is proposed, according to sequence information and TF bindings. The original data we used to build the prediction model, which are sequence context of SNP and TF profiles in JASPAR, are all very easy to be obtained. So, the prediction method can be easily applied to other plant species. With a reliable recognition result, the proposed model might help experimental researchers to find potential SNPs for further diversity study and molecular breeding research in plants.

In addition, there is still much room for improvement in our proposed functional SNPs prediction model. Firstly, both the quantity and quality of documented trait-associated SNPs in rice should be greatly improved, thus we are able to further improve our model by parameter optimization. Secondly, we just discussed the effect of the SNP variant on TFs binding affinities, while more functional elements can be considered, such as promoter, miRNA, lncRNA, etc. Adding the analysis of these functional elements can not only help to further improve the prediction model but also help to understand the possible functional mechanism of functional SNPs in rice.

5 Acknowledgement

We are grateful to our colleagues in Telecommunication and Networks National Laboratory, Nanjing University of Posts and Telecommunications, for their help during the course of this work, in particular Prof. Xiao-yong Yan for his critical suggestions.

Funding. This research was sponsored by grants from the Scientific Research Foundation of Nanjing University of Posts and Telecommunications (NUPTSF Grant No. NY218143 and No. NY218141), the National Fund incubation project of Nanjing University of Posts and Telecommunications (NUPTSF Grant No. NY219116), and the Natural Science Fund for Colleges and Universities in Jiangsu province (No. 22KJB520024).

Data Availability Statement. The data of rice genome can be found in NCBI: [https://www.ncbi.nlm.nih.gov/]; The details of SNPs can be found in dbSNP: [https://www.ncbi.nlm.nih.gov/snp/]; the data of TF binding profiles of plants can be found in JASPAR: [https://jaspar.genereg.net/].

References

[1] J. Yu, S. Hu, J. Wang, G. K.-S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan, H. Yang, A draft sequence of the rice genome (Oryza sativa L. ssp indica), Science 296(5565)(2002) 79-92.

[2] S.A. Goff, D. Ricke, T.-H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B.M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W.-L. Sun, L. Chen, B. Cooper, S. Park, T.C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R.M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, S. Briggs, A draft sequence of the rice genome (Oryza sativa L. ssp japonica), Science 296(5565)(2002) 92-100.

[3] Z. Li, B.-Y. Fu, Y.-M. Gao, W.-S. Wang, J.-L. Xu, F. Zhang, X.-Q. Zhao, T.-Q. Zheng, Y.-L. Zhou, G. Zhang, S. Tai, J. Xu, W. Hu, M. Yang, Y. Niu, M. Wang, Y. Li, L. Bian, X. Han, J. Li, X. Liu, B. Wang, K.L. McNally, M.E.B. Naredo, S.M.Q. Mercado, M.C. Rellosa, R.A. Reaño, G.L.S. Capilit, F.C. de Guzman, J. Ali, N.R.S. Hamilton, R.P. Mauleon, N.N. Alexandrov, H. Leung, The 3,000 rice genomes project, Gigascience 3(2014) 7.

[4] S. Song, D. Tian, Z. Zhang, S. Hu, J. Yu, Rice Genomics: over the Past Two Decades and into the Future, Genomics Proteomics Bioinformatics 16(6)(2018) 397-404.

[5] Y.M. Liang, H.J. Liu, J.B. Yan, T. Feng, Natural Variation in Crops: Realized Understanding, Continuing Promise, Annual Review of Plant Biology 72(2021) 357-385.

[6] M. Thudi, R. Palakurthi, J.C. Schnable, A. Chitikineni, S. Dreisigacker, E. Mace, R.K. Srivastava, C.T. Satyavathi, D. Odeny, V.K. Tiwari, H.-M. Lam, Y.B. Hong, V.K. Singh, G. Li, Y. Xu, X. Chen, S. Kaila, H. Nguyen, S. Sivasankar, S. A Jackson, T.J. Close, W. Shubo, R.K. Varshney, Genomic resources in plant breeding for sustainable agriculture, Journal of Plant Physiology 257(2021) 153351.

[7] D. Robledo, C. Palaiokostas, L. Bargelloni, P. Martínez, R. Houston, Applications of genotyping by sequencing in aquaculture breeding and genetics, Reviews in Aquaculture 10(3)(2018) 670-682.

[8] G. Hemani, J. Bowden, G.D. Smith, Evaluating the potential role of pleiotropy in Mendelian randomization studies, Human Molecular Genetics 27(R2)(2018) R195-R208.

[9] S.B. Aglawe, A.K. Verma, A.K. Upadhyay, Bioinformatics Tools and Databases for Genomics-assisted Breeding and Population Genetics of Plants: A Review, Current Bioinformatics 16(6)(2021) 766-773.

[10] J.A. Garrido-Cardenas, C. Mesa-Valle, F. Manzano-Agugliaro, Trends in plant research using molecular markers, Planta 247(3)(2018) 543-557.

[11] L. Mansueto, R.R. Fuentes, F.N. Borja, J. Detras, J.M. Abriol-Santos, D. Chebotarov, M. Sanciangco, K. Palis, D. Copetti, A. Poliakov, I. Dubchak, V. Solovyev, R.A. Wing, R.S. Hamilton, R. Mauleon, K.L. McNally, N. Alexandrov, Rice SNP-seek database update: new SNPs, indels, and queries, Nucleic Acids Research 45(D1)(2017) D1075-D1081.

[12] J. Yan, D. Zou, C. Li, Z. Zhang, S. Song, X. Wang, SR4R: An Integrative SNP Resource for Genomic Breeding and Population Research in Rice, Genomics, Proteomics & Bioinformatics 18(2)(2020) 173-185.

[13] G.R.S. Ritchie, I. Dunham, E. Zeggini, P. Flicek, Functional annotation of noncoding sequence variants, Nature Methods 11(3)(2014) 294-296.

[14] Y. Yao, Z. Liu, Q. Wei, S.A. Ramsey, CERENKOV2: improved detection of functional noncoding SNPs using data-space geometric features, BMC Bioinformatics 20(1)(2019) 63.

[15] The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489(7414)(2012) 57-74.

[16] F. Cunningham, J.E. Allen, J. Allen, J. Alvarez-Jarreta, M.R. Amode, I.M. Armean, O. Austine-Orimoloye, A.G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L.D.R. Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El Houdaigui, T. El Naboulsi, R. Fatima, C.G. Giron, T. Genez, J. G. Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J.C. Marugán, S. Mohanan, A. Mushtaq, M. Naven, D.N. Ogeh, A. Parker, A. Parton, M. Perry, I. Piližota, I. Prosovetskaia, M.P. Sakthivel, A.I.A. Salam, B.M. Schmitt, H. Schuilenburg, D. Sheppard, J.G. Pérez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M.-M. Suner, M. Szpak, A. Thormann, F.F. Tricomi, D. Urbina-Gómez, A. Veidenberg, T.A.Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S.E. Hunt, G.R. IIsley, J.E. Loveland, F.J. Martin, B. Moore, J.M. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S.J. Trevanion, S. Dyer, P.W. Harrison, K.L. Howe, A.D. Yates, D.R. Zerbino, P. Flicek, Ensembl 2022, Nucleic Acids Research 50(D1)(2022) D988-D995.

[17] The Gtex Consortium, K.G. Ardlie, D.S. Deluca, A.V. Segrè, T.J. Sullivan, T.R. Young, E.T. Gelfand, C.A. Trowbridge, J.B. Maller, T. Tukiainen, M. Lek, L.D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C.D. Palmer, T. Esko, W. Winckler, J.N. Hirschhorn, M. Kellis, D.G. Macarthur, G. Getz, A.A. Shabalin, G. Li, Y.-H. Zhou, A.B. Nobel, I. Rusyn, F.A. Wright, T. Lappalainen, P.G. Ferreira, H. Ongen, M.A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J.M. Goldmann, D. Koller, R. Guigó, M.I. Mccarthy, E.T. Dermitzakis, E.R. Gamazon, H.K. Im, A. Konkashbaev, D.L. Nicolae, N.J. Cox, T. Flutre, X. Wen, M. Stephens, J.K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J.A. Thomas, J.T. Lonsdale, M.T. Moser, B.M. Gillard, E. Karasik, K. Ramsey, C. Choi, B.A. Foster, J. Syron, J. Fleming, Harold Magazine, R. Hasz, G.D. Walters, J.P. Bridge, M. Miklos, S. Sullivan, L.K. Barker, H.M. Traino, M. Mosavel, L.A. Siminoff, D.R. Valley, D.C. Rohrer, S.D. Jewell, P.A. Branton, L.H. Sobin, M. Barcus, L. Qi, J. Mclean, P. Hariharan, K.S. Um, S. Wu, D. Tabor, C. Shive, A.M. Smith, S.A. Buia, A.H. Undale, K.L. Robinson, N. Roche, K.M. Valentino, A. Britton, R. Burges, D. Bradbury, K.W. Hambright, J. Seleski, G.E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D.C. Mash, S. Volpi, J.P. Struewing, G.F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L.J. Carithers, H.M. Moore, P. Guan, C. Compton, S.J. Sawyer, J.P. Demchok, J.B. Vaught, C.A. Rabiner, N.C. Lockhart, K.G. Ardlie, G. Getz, F.A. Wright, M. Kellis, S. Volpi, E.T. Dermitzakis, The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans, Science 348(6235)(2015) 648-660.

[18] A. Kharabian, An efficient computational method for screening functional SNPs in plants, Journal of Theoretical Biology 265(1)(2010) 55-62.

[19] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Research 29(1)(2001) 308-311.

[20] E.W. Sayers, E.E. Bolton, J.R. Brister, K. Canese, J. Chan, D.C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B.W. Trawick, K.D. Pruitt, S.T. Sherry, Database resources of the national center for biotechnology information, Nucleic Acids Research 50(D1)(2022) D20-D26.

[21] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28(23)(2012) 3150-3152.

[22] M. Ghandi, M. Mohammad-Noori, M.A. Beer, Robust k-mer frequency estimation using gapped k-mers, Journal of Mathematical Biology 69(2)(2014) 469-500.

[23] G.D. Stormo, T.D. Schneider, L. Gold, A. Ehrenfeucht, Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in E. coli, Nucleic Acids Research 10(9)(1982) 2997-3011.

[24] S.M. Kielbasa, D. Gonze, H. Herzel, Measuring similarities between transcription factor binding sites, BMC Bioinformatics 6(2005) 237.

[25] T.L. Bailey, STREME: accurate and versatile sequence motif discovery, Bioinformatics 37(18)(2021) 2834-2840.

[26] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S.M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A.E. Urban, M.-Y. Hong, K.J. Karczewski, W. Huber, S.M. Weissman, M.B. Gerstein, J.O. Korbel, M. Snyder, Variation in Transcription Factor Binding Among Humans, Science 328(5975)(2010) 232-235.

[27] P. Wang, Y. Xiong, R. Gong, Y. Yang, K. Fan, S. Yu, A key variant in the cis-regulatory element of flowering gene Ghd8 associated with cold tolerance in rice, Scientific reports 9(2019) 9603.

[28] J.A.Castro-Mondragon, R. Riudavets-Puig, I. Rauluseviciute, R.B. Lemma, L. Turchi, R. Blanc-Mathieu, J. Lucas, P. Boddie, A. Khan, N.M. Pérez, O. Fornes, T.Y. Leung, A. Aguirre, F. Hammal, D. Schmelter, D. Baranasic, B. Ballester, A. Sandelin, B. Lenhard, K. Vandepoele, W.W. Wasserman, F. Parcy, A. Mathelier, JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles, Nucleic Acids Research 50(D1)(2022) D165-D173.

[29] I. Khan, S. Khan, Y. Zhang, J. Zhou, Genome-wide analysis and functional characterization of the Dof transcription factor family in rice (Oryza sativa L.), Planta 253(5)(2021) 101.

[30] S. Ramírez-Gallego, I. Lastra, D. Martínez-Rego, V. Bolón-Canedo, J.M. Benítez, F. Herrera, A. Alonso-Betanzos, Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data, International Journal of Intelligent Systems 32(2)(2017) 134-152.

[31] G. Douzas, F. Bacao, Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE, Information Sciences 501(2019) 118-135.

[32] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning 20(3)(1995) 273-297.

[33] W.S. Noble, What is a support vector machine?, Nature Biotechnology 24(12)(2006) 1565-1567.

[34] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, Acm Transactions on Intelligent Systems and Technology 2(3)(2011) 1-27.

[35] M.W. Riaz, J. Lu, L. Shah, L. Yang, C. Chen, X.D. Mei, L. Xue, M.A. Manzoor, M. Abdullah, S. Rehman, H. Si, C. Ma, Expansion and Molecular Characterization of AP2/ERF Gene Family in Wheat (Triticum aestivum L.), Frontiers in Genetics 12(2021) 632155.

[36] Q. Guan, C. Wen, H. Zeng, J. Zhu, A KH Domain-Containing Putative RNA-Binding Protein Is Critical for Heat Stress-Responsive Gene Regulation and Thermotolerance in Arabidopsis, Molecular Plant 6(2)(2013) 386-395.

[37] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T.L. Madden, BLAST plus: architecture and applications, BMC Bioinformatics 10(2009) 421.