frontiers in Plant Science

TECHNOLOGY AND CODE
published: 31 January 2020
doi: 10.3389/fpls.2020.00004

6mA-RicePred: A Method for Identifying DNA N^6-Methyladenine Sites in the Rice Genome Based on Feature Fusion

Qianfei Huang[1†], *Jun Zhang*[2†], *Leyi Wei*[1], *Fei Guo*[1*] *and Quan Zou*[3*]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China, [2] Rehabilitation Department, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China, [3] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China

Motivation: The biological function of N^6-methyladenine DNA (6mA) in plants is largely unknown. Rice is one of the most important crops worldwide and is a model species for molecular and genetic studies. There are few methods for 6mA site recognition in the rice genome, and an effective computational method is needed.

Results: In this paper, we propose a new computational method called 6mA-Pred to identify 6mA sites in the rice genome. 6mA-Pred employs a feature fusion method to combine advantageous features from other methods and thus obtain a new feature to identify 6mA sites. This method achieved an accuracy of 87.27% in the identification of 6mA sites with 10-fold cross-validation and achieved an accuracy of 85.6% in independent test sets.

Keywords: rice, model, fusion, DNA, 6mA

OPEN ACCESS

Edited by:
Fengfeng Zhou,
Jilin University, China

Reviewed by:
Balachandran Manavalan,
Ajou University, South Korea
Yushan Qiu,
Shenzhen University, China

****Correspondence:***
Fei Guo
fguo@tju.edu.cn
Quan Zou
zouquan@nclab.net

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Plant Science

Received: *04 December 2019*
Accepted: *06 January 2020*
Published: *31 January 2020*

Citation:
Huang Q, Zhang J, Wei L, Guo F and Zou Q (2020) 6mA-RicePred: A Method for Identifying DNA N^6-Methyladenine Sites in the Rice Genome Based on Feature Fusion. Front. Plant Sci. 11:4. doi: 10.3389/fpls.2020.00004

INTRODUCTION

DNA methylation plays crucial roles in many biological functions, and methylated DNA carries important epigenetic information. The modification of DNA methylation is a heavily researched topic in epigenetic research (Liu et al., 2016). Previously, DNA methylation was thought to comprise cytosine (5-methylcytosine, 5mC) methylation and N^4-methylcytosine (4mC) methylation (Chen et al., 2017; He et al., 2019; Tang et al., 2019a). However, with the rapid development of sequencing technology, a new type of DNA methylation modification, DNA-6mA methylation, has been identified and has become a heavily researched subject in the field of epigenetics (Xiao et al., 2018). N^6-methyladenine DNA (6mA) modification is the most prevalent type of DNA modification in prokaryotes. This modification plays important roles in DNA mismatch repair, chromosome replication, cell defense, cell cycle regulation, and transcription (Xu et al., 2017; He et al., 2019). 6mA shows similar properties in eukaryotes and prokaryotes (Hao et al., 2019).

Machine learning methods have overcome many problems in identifying 4mC (Chen et al., 2017) and 5mC modifications. The 6mA modification has become a heavily researched subject, and an increasing number of researchers are using machine learning to identify 6mA sites in the rice genome. The current machine learning algorithms perform notably well in recognizing 6mA sites in the rice genome. Many excellent features and algorithms have been applied to the recognition of 6mA sites. Regarding feature algorithms, nucleotide chemical property, nucleotide frequency (Pan et al., 2017;

Yin et al., 2019; Chen et al., 2019a), and mononucleotide binary encoding are often used in the recognition of 6mA sites in the rice genome. These methods all have some properties in common, including the unique representations of nucleotides. This property is also exhibited by our method. Regarding dimensionality reduction algorithms, MRMD (Zou et al., 2016a) performs well and is an excellent feature selection algorithm. Other highly efficient feature selection algorithms have been proposed in bioinformatics classification (Zou et al., 2015; Xu et al., 2017; Zhu et al., 2017; Pan et al., 2018; Wang et al., 2018; Cheng et al., 2018a; Zhu et al., 2018a; Cheng et al., 2018b; Zhu et al., 2018b; Lai et al., 2019; Dao et al., 2019; Yu et al., 2019; Yang et al., 2019; Ren Qi et al., 2019; Tang et al., 2019b). Regarding classification algorithms, an increasing number of classification methods are being used by researchers to identify 6mA sites, such as Random Forest, XGboost, support vector machine (SVM), and gradient boosted decision tree (GBDT). Research has proven that SVM and Random Forest perform better than the other classifier algorithms. In the present study, the performance of SVM was highly stable. A Markov model is used in MM-6mAPred (Pian et al., 2019) to identify 6ma sites and has achieved good results.

There are few computational methods to identify 6mA sites in the rice genome. Proposed methods include i6mA-Pred (Chen et al., 2019a), iDNA6mA-PseKNC (Feng et al., 2019a), MM-6mAPred (Pian et al., 2019), and iDNA6mA-Rice (Hao et al., 2019). i6mA-Pred uses nucleotide chemical property, nucleotide frequency, and SVM to identify 6ma sites. MM-6mAPred adopts a Markov model to identify 6mA sites. iDNA6mA-Rice uses mononucleotide binary encoding and Random Forest to identify 6mA sites. Feature fusion makes use of diverse features to build prediction models and has been successfully and widely applied in bioinformatics (Zhang et al., 2018a; Zhang et al., 2018b; Zhou et al., 2019; Zhang et al., 2019a; Zhang et al., 2019b). In this paper, we propose a feature fusion-based method to identify 6mA sites in the rice genome, in which nucleotide chemical properties, binary encoding, KMER, and Markov features are used to formulate DNA sequences. Our method combines these excellent features by using feature selection algorithms. The proposed model obtained an overall accuracy of 87.27% in identifying 6mA sites.

MATERIALS AND METHODS

Datasets

Two datasets were used in our study. One dataset comprised the same experimental benchmark data used by Chen et al. (Chen et al., 2019a) and has been used to train MM-6mAPred (Cheng et al., 2018b; Pian et al., 2019). This dataset contained 880 positive samples and 880 negative samples. The positive samples were obtained by setting the modification score and CD-HIT. Positive samples can improve the quality of the sequence and reduce redundancy. The second dataset comprised the same experimental benchmark data used to train iDNA6mA-Rice, and it contained 15,400 positive samples and 15,400 negative samples. All of the sequences in these two datasets measured 41 bp in length. All of the negative samples had non-methylated adenosine in the center, and all of the positive samples had a 6mA site in the center.

Table 1 shows the numbers of positive and negative samples for both datasets. Dataset 1 was mainly used for cross-validation. Dataset 2 was primarily used for independent testing. These two benchmark datasets are available at https://github.com/huangqianfei0916/6ma-rice.

Model Architecture

Feature extraction plays a crucial role in the construction of the model (Wang et al., 2019). Four feature extraction algorithms were adopted to formulate 6mA samples. Binary encoding, nucleotide chemical property (Xu et al., 2019), KMER, and Markov features were selected from among several feature algorithms, and **Table 1** shows the results of each algorithm. In order to reduce computation and optimize feature vectors, feature selection algorithms were used for each feature. The features after fusion were normalized, and the final feature vectors were the optimal representations of the sequence (Chen et al., 2019b). **Figure 1** illustrates the structure of the model.

Binary encoding and nucleotide chemical property are excellent feature algorithms extracted from iDNA6mA-Rice and i6mA-Pred. Kmer is a useful feature algorithm (Feng et al., 2019b) that we selected based on a large number of experiments. The Markov feature is a new feature extraction algorithm we introduced based on MM-6mAPred. Combining the best features does not necessarily produce the best results; for example, Kmer does not perform well when used alone, but does so when combined with other features. Feature selection solves this problem and reduces the amount of computation. Finally, the best features were obtained by normalization. Regarding the classifier, previous studies have shown that SVM and Random Forest perform better than other classifiers. In this study, the performance of SVM was significantly better than that of Random Forest.

Binary Encoding

Binary encoding is a simple and effective feature algorithm. This algorithm obtains sequence features by the binary representation of nucleotides (Zou et al., 2019). The binary encoding algorithm converts nucleotides into the following formats:

$$A \rightarrow [1\ ,0\ ,0\ ,0]$$
$$C \rightarrow [0\ ,1\ ,0\ ,0]$$
$$G \rightarrow [0\ ,0\ ,1\ ,0]$$
$$T \rightarrow [0\ ,0\ ,0\ ,1]$$

This algorithm can be understood as a unique representation of nucleotides and can be considered a one hot encoding algorithm. A random DNA sequence with m nucleotides can then be converted into a vector of $4 \times m$ features (Hao et al., 2019; Chen et al., 2019c). The representation of nucleotides is not unique, and the representations of A, T, G, and C are interchangeable.

TABLE 1 | All datasets.

Datasets	Positive	Negative	Total	Species
Dataset 1	880	880	1,760	Rice
Dataset 2	154,000	154,000	308,000	Rice

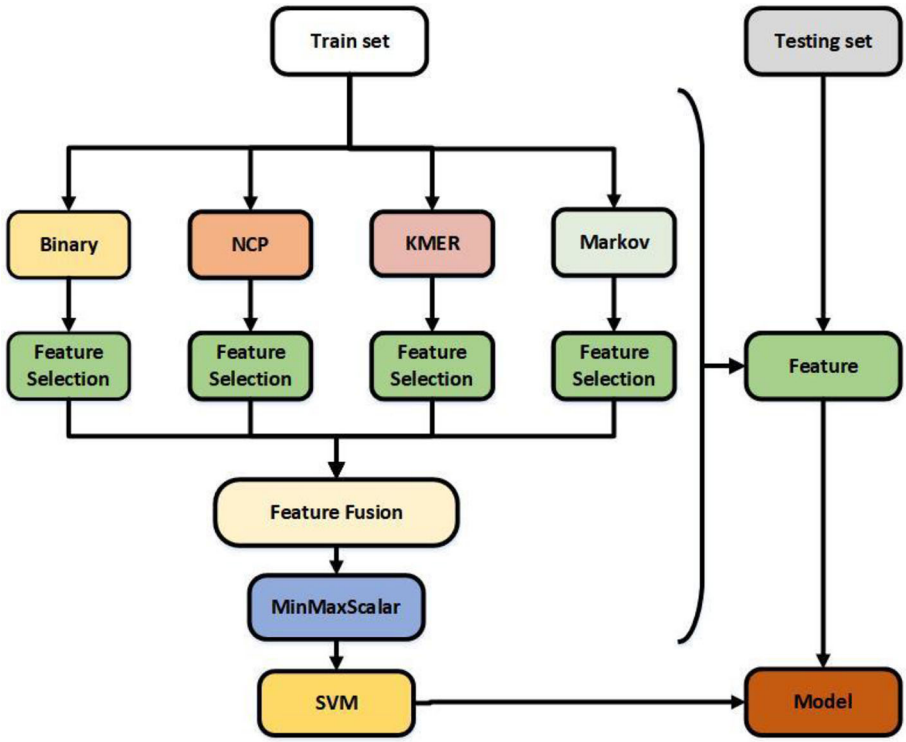

FIGURE 1 | Flowchart showing the construction of this model. The feature selection is selectfrommodel.

Nucleotide Chemical Property

DNA is composed of four types of nucleotides: adenine (A), cytosine (C), guanine (G), and cytosine (C). DNA has multiple properties, such as ring structures, functional groups, and hydrogen bonds (Fu et al., 2018; Wei et al., 2018; Xue et al., 2018; Tan et al., 2019a) (He et al., 2019). A and G each contain two rings, whereas C and T contain only one. Regarding secondary structures, A and T form weak hydrogen bonds, whereas C and G form strong hydrogen bonds. Regarding functional groups, A and C compose the amino group, whereas G and T compose the keto group. The feature extraction algorithm can be formulated as follows:

$$a = \begin{cases} 1 & n \in \{A, G\} \\ 0 & \text{others} \end{cases}$$

$$b = \begin{cases} 1 & n \in \{A, T\} \\ 0 & \text{others} \end{cases}$$

$$c = \begin{cases} 1 & n \in \{A, C\} \\ 0 & \text{others} \end{cases}$$

where n represents a nucleotide, which can be converted into the following format:

$$A \rightarrow [1, 1, 1] \; C \rightarrow [0, 0, 1]$$

$$G \rightarrow [1, 0, 0] \; T \rightarrow [0, 1, 0]$$

For instance, a DNA sequence "AATCGTA" can be transformed into a vector such as (1,1,1,1,1,1,0,1,0,0,0,1,1,0,0,0,1,0,1,1,1). Nucleotide chemical property has similar properties to binary encoding, both of which can be considered to yield unique representations of nucleotides.

KMER

Kmer is a highly common feature extraction algorithm and is easy to understand (Liu et al., 2015; He et al., 2018; Su et al., 2018; Zhu et al., 2019). When $k = 1$, Kmer denotes the frequency of the four nucleotides. When $k = 2$, the sequence can be represented by 16 features, i.e., AA, AT, AG, AC, TA, TT, TG, TC, …,CC (Cao et al., 2018). As the value of k increases, the dimension of the feature increases; thus, the difficulty of calculation increases. In this study, the k value that was employed was 3. Thus, a sequence could be represented as 64 features. We tested the results of k from 1 to 4 and chose 3. A k equal to 3 will not cause poor results because the features are too sparse. The Kmer ($k = 3$) descriptor can be calculated as follows:

$$p = \frac{t}{L-2} \qquad t \in \{AAA, AAT, \ldots\ldots, CCC\}$$

where L denotes the length of the sequence and t denotes the number of nucleotide occurrences. As the value of k increases, the results may improve, but the dimension will increase, causing the amount of calculation to increase. In this study, although Kmer yielded poor results when used alone, the information contained in Kmer was crucial in feature fusion.

Markov Feature

From MM-6mAPred, we can determine that the Markov chain achieves good performance in recognizing 6ma sites. Therefore, we introduced the Markov chain into DNA sequence analysis to improve the sequence representation. The algorithm constructed the first-order Markov chain (Kemeny and Snell, 1976) for each dataset. Before obtaining the sequence features, the algorithm must calculate the transition probability of the dataset. **Figure 2** shows a schematic of feature extraction from a DNA sequence with the first-order Markov chain. A, T, G, C are equivalent to four states, and P^{i}_{NN} is the transition probabilities between the ith nucleotide and the $(i + 1)$th nucleotide (Nigatu et al., 2017; Pian et al., 2019). Thus, a transition probability matrix is generated between every two nucleotides. A sequence of 41 bp can generate 40 transition probability matrices.

A sequence of 41 bp can be represented as a 40-dimensional vector. We did not use the initial probability because it did not improve the experimental results. The transition probability between two nucleotides is used to represent sequence information. Sequence information between the ith nucleotide and the $(i + 1)$th nucleotide is obtained from the ith transition probability matrix. Moreover, the features result can be optimized by adjusting the length of the sequence. The sequence contains the transition probability information, and a sequence can be represented by the transition probabilities.

Performance Evaluation of Different Algorithms

The type of feature algorithm has strong effects on experimental results (Liu et al., 2017; Cheng et al., 2018c; Zheng et al., 2019; Cheng et al., 2019a; Zhang et al., 2019c). After testing many features and classifications, three best-performing classifiers were selected to test the feature descriptors. **Table 2** reports the 10-fold cross-validation results for the classifiers identifying the 6mA sites in dataset 1. Binary encoding, NCP (Chen et al., 2019c), Markov features, and ENAC were selected for feature selection. Experimentation revealed that Kmer is a better choice than ENAC. **Table 2** shows that the results of ENAC are considerably better than those of Kmer, whereas the results from using Kmer fusion are better than those from using ENAC fusion. This finding shows that merging the best-performing features may not be the optimal option. The binary encoding feature algorithm was used in iDNA6mA-Rice, and the NCP feature algorithm was used in i6mA-Pred. Our experimental results were consistent with the results of previous studies. The Markov feature algorithm is a new feature algorithm that we created based on MM-6mAPred. To improve the experimental results and reduce the amount of calculation, the feature selection algorithm is applied for each feature. Feature selection after fusion can also reduce the amount of calculation, but does not achieve as good results. As an alternative approach to feature selection, feature selection can be performed before fusion and again after fusion; however, this approach will result in a few dimensions. Thus, feature selection before fusion is the best approach.

Support Vector Machine

SVM is a widely used machine learning algorithm (Ding and Li, 2015; Li et al., 2015; Zeng et al., 2017; Ding et al., 2017a; Zhang et al., 2019; Tan et al., 2019a) and was used in this study to identify 6mA sites in the rice genome. SVM is also widely used in bioinformatics fields (Zou et al., 2016b; Wang et al., 2018; Wei et al., 2018; Xiong et al., 2018; Zeng et al., 2018a; Xu et al., 2018a; Xu et al., 2018b; Xu et al., 2018c; Li et al., 2019). Our experiments showed that SVM was more suitable for the purposes of the present study than were the other algorithms. We used the libsvm package available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The radial basis kernel function (RBF) was used to obtain the classification hyperplane. The two main parameters of SVM, C and gamma, were optimized by grid search. The optimization ranges about C and gamma were $(2^{-5},2^{5})$ and $(2^{-5},2^{5})$, respectively, and the values of C and gamma were 1.0 and 0.125, respectively. In this study, SVM performed better than the other classifiers.

Feature Selection

Feature selection algorithms are widely used in machine learning (Liu X. et al., 2019; Zeng et al., 2019a; Zeng et al., 2019b), and

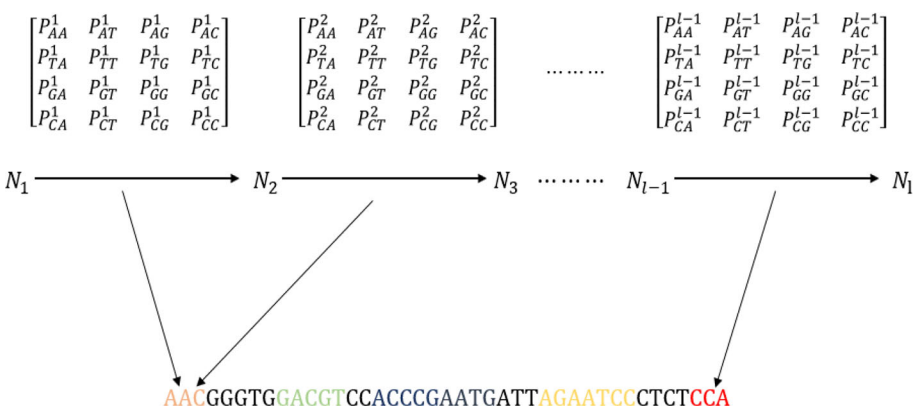

FIGURE 2 | Schematic showing the process of extracting features from the transition probability matrix of the DNA sequence. The sequence "AATACATGGGGTTATGTGCCACCGGTCATAATATCTAGGGT" is used as an example to explain the process.

TABLE 2 | Performance of different feature descriptors and classifiers.

Feature descriptors	SVM (Acc%)	XGboost (Acc%)	GBDT (Acc %)	Vote (Acc %)
EIIP	63.9	83.9	84.0	83.9
ANF	54.2	60.7	61.1	61.7
BINARY	82.8	84.4	83.6	84.7
DNC	58.4	61.0	59.7	61.2
NCP	82.8	83.3	83.9	84.3
PseEIIP	53.9	66.5	65.3	65.9
TNC	56.8	66.5	65.3	66.0
KMER	53.0	64.2	64.8	65.1
ENAC	73.5	79.4	78.8	79.0
NAC	56.3	55.5	54.6	55.5
CKSNAP	57.2	65.3	65.3	65.8
RCKMER	55.0	62.9	62.3	62.3
MAKOV	83.75	85.17	84.7	85.0

feature selection is necessary with our method. Feature selection removes redundant and uncorrelated information from the sequence and increases computational speed. In this study, we chose the selectfrommodel module of sklearn and the classifier XGboost (Chen and Guestrin, 2016; Zhou et al., 2019). Feature selection can optimize features and reduce the number of calculations. The results of our experiments proved that feature selection can improve results and reduce computation. Feature selection was able to identify the better features, and XGboost was the best-performing classifier. We investigated other feature selection methods, but did not obtain high-quality results. Feature selection can be performed in three ways: before fusion, after fusion, and both before and after fusion. The experimental results showed that before fusion is the best approach.

Performance Evaluation

It is important to evaluate the results of a new model, and several evaluation metrics are available. Sensitivity (Sn), specificity (Sp), accuracy (Acc), and Mathew's correlation coefficient (MCC) are often used to evaluate the quality of a model in machine learning (Liu B. et al., 2019; Cheng et al., 2012; Cheng et al., 2016; Ding et al., 2016b; Mariani et al., 2017; Ding et al., 2017; Xu et al., 2017; Wei et al., 2017a; Wei et al., 2017b; Hu et al., 2018; Zhang et al., 2018c; Ding et al., 2019; Shan et al., 2019; Xu et al., 2019; Tan et al., 2019b; Cheng et al., 2019b). These metrics are formulated as follows:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

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

These metrics are commonly used in machine learning. TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. The above mathematical expressions clearly describe the meanings of the four metrics. In model evaluation methods, independent dataset testing and cross-validation are often used to evaluate the prediction ability of the model. In this study, dataset 1 was mainly used for cross-validation and training, and dataset 2 was mainly used for independent testing. In the independent test experiment, dataset 1 was used for training and dataset 2 was used for testing.

In addition to the above metrics, area under the ROC curve (AUC) and receiver operating characteristic (ROC) were also used to evaluate model quality.

RESULTS AND DISCUSSION

Analysis of the Algorithms

After the feature fusion, we tested the feature using voting techniques and three different classifiers with 10-fold cross-validation and independent test experiments. The 10-fold cross-validation results of the different methods in identifying 6mA sites by using the benchmark dataset 1 are reported in **Figure 3A**. The independent test results of the different methods in identifying 6mA sites by using the benchmark dataset 2 are reported in **Figure 3B**. **Figure 3A** shows that the test results of the three classifiers were highly similar. However, **Figure 3B** shows that SVM performed significantly better than the other classifiers. Based on the experimental results, we chose the SVM classifier in this study.

Before feature extraction, we conducted a simple optimization of the sequence length. In addition, we tested the original sequence and the optimal sequence using our method. The experimental results are reported in **Table 3**. As depicted in the table, feature selection is an excellent choice; the results of the best sequence were considerably better than those of the original sequence. The results revealed no significant improvement; however, reducing the length of the sequence reduces the amount of calculation.

Many experiments have been conducted regarding the selection of feature algorithms and classifiers. Our experiments revealed that binary encoding, NCP, and the Markov feature were effective, and previous studies have shown that they yield good results when used alone. We visualize the features by reducing the dimensionality, and **Figure 4** reports the distribution of each feature method. Therefore, we combined these excellent features to improve representation. In the selection of feature selection methods, we tested several widely used methods, and the experimental results are shown in **Figure 5**. To further optimize the features, we applied MinMaxScaler to the features after fusion. The differences between **Figures 3A, B** indicate that the SVM was highly robust. Similarly, the model obtained by learning the optimized features with SVM was highly powerful. The method can be applied to computational intelligence techniques, such as neural networks (Chen et al., 2016; Song et al., 2018; Cabarle et al., 2019; Hong et al., 2019; Zhong et al., 2019; Zhou et al., 2019b; Wang et al., 2019b), evolutionary algorithms (Xu et al., 2019; Xu et al., 2019; Zeng et al., 2019b),

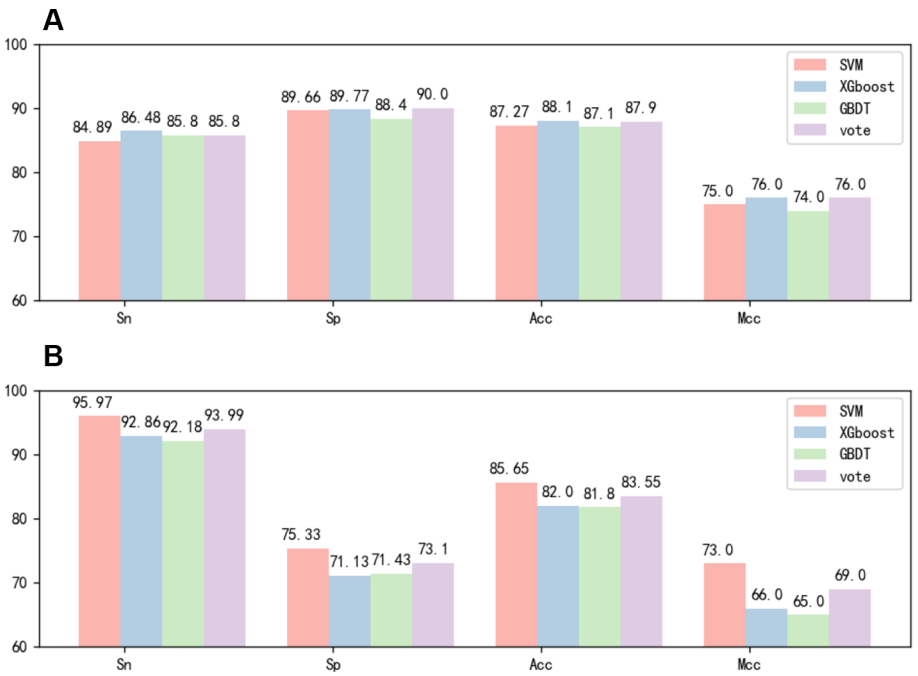

FIGURE 3 | (A) Tenfold cross-validation performance of different classifiers based on dataset 1. **(B)** Independent test performance of different classifiers based on dataset 1 and dataset 2.

TABLE 3 | Cross-validation performance of different methods based on dataset 1.

Method	Sn	Sp	Acc	Mcc
Best sequence—no fs	81.81	88.30	85.1	0.702
Original sequence—no fs	84.20	84.77	84.49	0.690
Best sequence—fs	84.89	89.66	87.27	0.746
Oraigin sequence—fs	85.0	89.20	87.10	0.742

and unsupervised learning (Zeng et al., 2018b; Zou et al., 2019), in future research.

Comparisons With Other Methods

To prove that our method is superior to other methods, we compared it with MM-6mAPred and i6mA-Pred, which are two excellent methods. i6mA-Pred uses nucleotide chemical property and nucleotide frequency as the features and reduces the dimensions by MRMD (Zou et al., 2016a). This approach then uses SVM to identify the 6mA sites in the rice genome. MM-6mAPred identifies 6mA sites based on the Markov model. The i6mA-Pred method is easy to understand, and its performance is good. The MM-6mA-Pred method is novel and has achieved good results. MM-6mA-Pred constructs multiple transition probability matrices for positive and negative samples. The new sample calculates the product of the transition probabilities in the two sets of transition probability matrices. The sample makes predictions based on the ratio of the two products. In addition, MM-6mA-Pred optimizes the length of the sequence to achieve optimal results, and similar operations are performed in our method. In general, the two methods yield effective models from different perspectives.

To improve experimental results, the main features of both methods are included in our method. In this study, we used feature selection and sequence length optimization, and we used 10-fold cross-validation and independent testing to evaluate the method. To conduct comparative tests, we reproduced the MM-6mAPred model with python3 and used the metrics we used previously for evaluation. In the cross-validation experiment, we performed 10-fold cross-validation based on dataset 1. The MM-6mAPred model that we reproduced obtained an accuracy of 84.7%, which is lower than the 89.7% reported in the paper in which the model is proposed. In the independent test experiment, dataset 1 was used for training and dataset 2 was used for testing. The model that we reproduced with python3 and the model implemented with MATLAB by the authors of the source paper yielded consistent results. The independent testing experiments revealed that the accuracy of MM-6mAPred was only 83.06%, whereas our method achieved 85.65% accuracy. Similar tests were performed with i6mA-Pred, and the results are reported in **Tables 4** and **5**. The experimental results show that our method is superior to other methods. In addition, the results obtained with MM-6mAPred were better than the results obtained with i6mA-Pred. Our reproduced MM-6mAPred code has been deposited on GitHub at https://github.com/huangqianfei0916/Markov.

Allowing further comparisons of these methods, ROC and AUC are shown in **Figure 6**. The area under the curve values (AUCs) of 6mA-ricePred, MM-6mAPred, and i6mA-Pred were

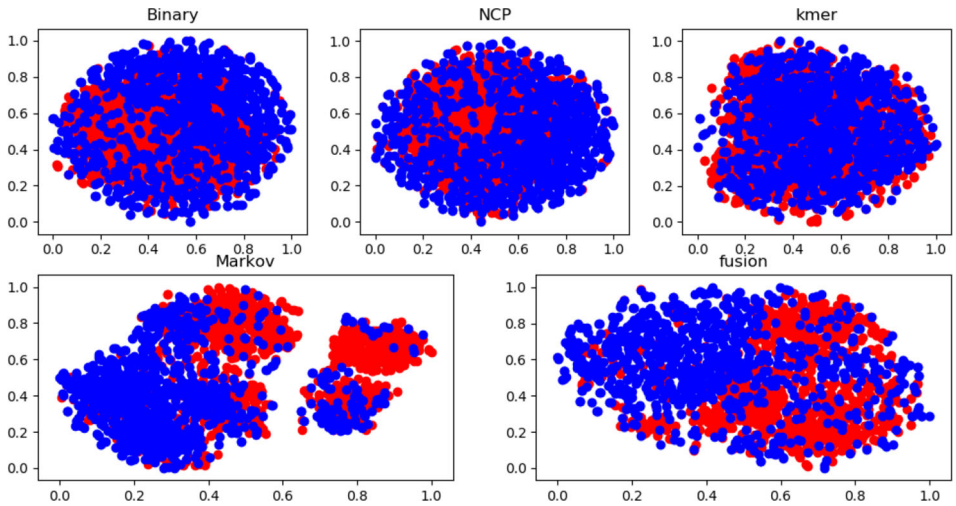

FIGURE 4 | Feature distribution of different feature methods based on dataset 1.

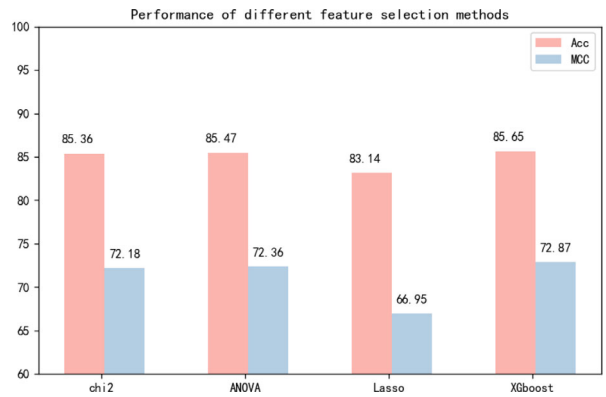

FIGURE 5 | Independent test performance of different feature selection methods based on dataset 1 and dataset 2.

0.945, 0.928, and 0.904, respectively. **Figure 6** shows that our method performed better than the other two methods.

CONCLUSION

Accuracy in identifying DNA N^6-methyladenine sites is highly important. The chemical properties of the nucleotides and the Markov model were used in i6mA-Pred and MM-6mAPred, respectively, and achieved good results. Our method, which is based on feature fusion, achieved better results than these previous methods in identifying 6mA sites in the rice genome. Our method obtains a more powerful model by combining multiple effective methods. These experiments proved that the proposed method is superior to other methods, and it integrates the main features of previous methods.

TABLE 4 | Cross-validation performance of different methods based on dataset 1.

Method	Sn	Sp	Acc	Mcc
Our method	84.89	89.66	87.27	0.746
MM-6mAPred	84.31	85.22	84.77	0.695
i6mA-Pred	82.95	83.30	83.13	0.662

TABLE 5 | Independent test performance of different methods based on dataset 1 and dataset 2.

Method	Sn	Sp	Acc	Mcc
Our method	95.97	75.33	85.65	0.73
MM-6mAPred	95.81	70.30	83.06	0.68
I6mA-Pred	94.24	66.59	80.42	0.63

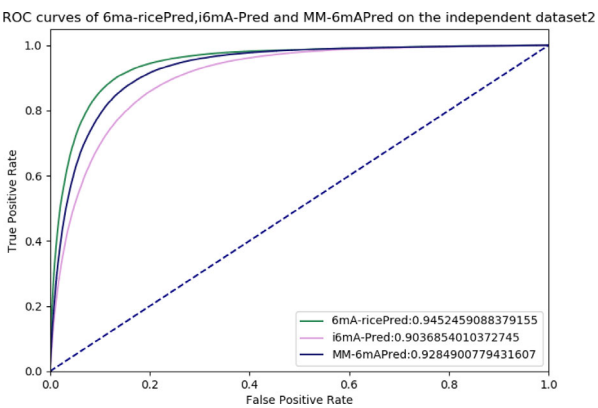

FIGURE 6 | Receiver operating characteristic (ROC) curves of 6ma-ricePred, MM-6mAPred, and i6mA-Pred.

We found that in addition to machine learning, the attention mechanism of deep learning can be applied to the recognition of 6mA sites in the rice genome. Amplifying the information of the 6mA sites by assigning attention can improve the recognition rate. The deep learning attention mechanism can be introduced into machine learning by, for example, multiplying different features by different weights and dynamically adjusting the weights according to the importance of the features.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/huangqianfei0916/6ma-rice/tree/master/dataset.

AUTHOR CONTRIBUTIONS

QH and JZ provided the data and did a lot of experiments. QZ and FG guided and modified the paper. LW provides a lot of good advice.

FUNDING

The work was supported by the National Natural Science Foundation of China (no. 91935302, no. 61922020, no. 61771331) and the Scientific Research Foundation in Shenzhen (JCYJ20180306172207178).

REFERENCES

Cabarle, F. G. C., de la Cruz, R. T. A., Cailipan, D. P. P., Zhang, D. F., Liu, X. R., and Zeng, X. X. (2019). On solutions and representations of spiking neural P systems with rules on synapses. *Inf. Sci.* 501, 30–49. doi: 10.1016/j.ins.2019.05.070

Cao, Z., Pan, X., Yang, Y., Huang, Y., and Shen, H.-B. (2018). The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics* 34 (13), 2185–2194. doi: 10.1093/bioinformatics/bty085

Chen, T., and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (ACM), 785–794. doi: 10.1145/2939672.2939785

Chen, X., Pérez-Jiménez, M. J., Valencia-Cabrera, L., Wang, B., and Zeng, X. (2016). Computing with viruses. *Theor. Comput. Sci.* 623, 146–159. doi: 10.1016/j.tcs.2015.12.006

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33 (22), 3518–3523. doi: 10.1093/bioinformatics/btx479

Chen, W., Lv, H., Nie, F., and Lin, H. (2019a). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics*. doi: 10.1093/bioinformatics/btz015

Chen, C., Zhang, Q., Ma, Q., and Yu, B. (2019b). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemom. Intell. Lab. Syst.* 191, 54–64. doi: 10.1016/j.chemolab.2019.06.003

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., et al. (2019c). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings In Bioinf.* doi: 10.1093/bib/bbz041

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., et al. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PloS Comput. Biol.* 8 (5), e1002503. doi: 10.1371/journal.pcbi.1002503

Cheng, L., Sun, J., Xu, W. Y., Dong, L. X., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep34820

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018a). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19 (Suppl 1), 919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018b). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinf.* 34 (11), 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018c). 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.2018.00657

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019a). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief Bioinform.* 20 (1), 203–209. doi: 10.1093/bib/bbx103

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2019b). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* doi: 10.1093/nar/gkz843

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2019). Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics* 35 (12), 2075–2083. doi: 10.1093/bioinformatics/bty943

Ding, H., and Li, D. (2015). Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids* 47 (2), 329–333. doi: 10.1007/s00726-014-1862-4

Ding, Y., Tang, J., and Guo, F. (2016b). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinf.* 17 (1), 398. doi: 10.1186/s12859-016-1253-9

Ding, Y., Tang, J., and Guo, F. (2017). Identification of protein–ligand binding sites by sequence information and ensemble classifier. *J. Chem. Inf. Model.* 57 (12), 3149–3161. doi: 10.1021/acs.jcim.7b00307

Ding, Y., Tang, J., and Guo, F. (2017a). Identification of drug-target interactions via multiple information integration. *Inf. Sci.* 418, 546–560. doi: 10.1016/j.ins.2017.08.045

Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.-C. (2019a). iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* 111 (1), 96–102. doi: 10.1016/j.ygeno.2018.01.005

Feng, C. Q., Zhang, Z. Y., Zhu, X. J., Lin, Y., Chen, W., Tang, H., et al. (2019b). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 35 (9), 1469–1477. doi: 10.1093/bioinformatics/bty827

Fu, J., Tang, J., Wang, Y., Cui, X., Yang, Q., Hong, J., et al. (2018). Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front. In Pharmacol.* 9, 681. doi: 10.3389/fphar.2018.00681

Hao, L., Dao, F.-Y., Guan, Z.-X., Zhang, D., Tan, J.-X., Zhang, Y., et al. (2019). iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Front. In Genet.* 10, 793. doi: 10.3389/fgene.2019.00793

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinf.* 19 (1), 306. doi: 10.1186/s12859-018-2321-0

He, W., Jia, C., and Zou, Q. (2019). 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* 35 (4), 593–601. doi: 10.1093/bioinformatics/bty668

He, S., Zhang, G., Wang, J., Gao, Y., Sun, R., Cao, Z., et al. (2019). 6mA-DNA-binding factor Jumu controls maternal-to-zygotic transition upstream of Zelda. *Nat. Commun.* 10 (1), 2219. doi: 10.1038/s41467-019-10202-3

Hong, Z., Zeng, X., Wei, L., and Liu, X. J. B. (2019). identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics*. doi: 10.1093/bioinformatics/btz694

Hu, Y., Zhao, T., Zhang, N., Zang, T., Zhang, J., and Cheng, L. (2018). Identifying diseases-related metabolites using random walk. *BMC Bioinf.* 19 (Suppl 5), 116. doi: 10.1186/s12859-018-2098-1

Kemeny, J. G., and Snell, J. L. (1976). *Markov Chains* (New York: Springer-Verlag).

Lai, H. Y., Zhang, Z. Y., Su, Z. D., Su, W., Ding, H., Chen, W., et al. (2019). iProEP: a computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids* 17, 337–346. doi: 10.1016/j.omtn.2019.05.028

Li, W., Yu, J., Lian, B., Sun, H., Li, J., Zhang, M., et al. (2015). Identifying prognostic features by bottom-up approach and correlating to drug repositioning. *PloS One* 10 (3), e0118672. doi: 10.1371/journal.pone.0118672

Li, J., Li, H., Zhang, L., Xu, Q., Ping, Y., Jing, X., et al. (2019). "Prediction of Human LncRNAs Based on Integrated Information Entropy Features," in *International Conference on Intelligent Computing* (Springer), 333–343. doi: 10.1007/978-3-030-26969-2_32

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43 (W1), W65–W71. doi: 10.1093/nar/gkv458

Liu, J., Zhu, Y., Luo, G.-Z., Wang, X., Yue, Y., Wang, X., et al. (2016). Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nat. Commun.* 7, 13052. doi: 10.1038/ncomms13052

Liu, B., Wu, H., and Chou, K. C. (2017). Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Natural Sci.* 09 (4), 67–91. doi: 10.4236/ns.2017.94007

Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* doi: 10.1093/nar/gkz740

Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., Zou, Q., et al. (2019). Bib: computational methods for identifying the critical nodes in biological networks. *Briefings In Bioinf.* doi: 10.1093/bib/bbz011

Mariani, L., Weinand, K., Vedenko, A., Barrera, L. A., and Bulyk, M. L. (2017). Identification of human lineage-specific transcriptional coregulators enabled by a glossary of binding modules and tunable genomic backgrounds. *Cell Syst.* 5 (3), 187–201. e187. doi: 10.1016/j.cels.2017.06.015

Nigatu, D., Sobetzko, P., Yousef, M., and Henkel, W. (2017). Sequence-based information-theoretic features for gene essentiality prediction. *BMC Bioinf.* 18 (1), 473. doi: 10.1186/s12859-017-1884-5

Pan, Y., Liu, D., and Deng, L. (2017). Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. *PloS One* 12 (6), e0179314. doi: 10.1371/journal.pone.0179314

Pan, Y., Wang, Z., Zhan, W., and Deng, L. (2018). Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* 34 (9), 1473–1480. doi: 10.1093/bioinformatics/btx822

Pian, C., Zhang, G., Li, F., and Fan, X. (2019). MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics*. doi: 10.1093/bioinformatics/btz556

Ren Qi, A. M., Qin, M., and Quan, Z. (2019). : Clustering and classification methods for single-cell RNA-sequencing data. *Brief Bioinf.* doi: 1093/bib/bbz062

Shan, X., Wang, X., Li, C.-D., Chu, Y., Zhang, Y., Xiong, Y. I., et al. (2019). Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method. *J. Chem. Inf Model.* doi: 10.1021/acs.jcim.9b00749

Song, T., Rodríguez-Patón, A., Zheng, P., and Zeng, XJIToC (2018). Systems D: spiking neural p systems with colored spikes. *IEEE Trans. Cogn. Dev. Syst.* 10 (4), 1106–1115. doi: 10.1109/TCDS.2017.2785332

Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics*. doi: 10.1093/bioinformatics/bty508

Tan, J. X., Lv, H., Wang, F., Dao, F. Y., Chen, W., and Ding, H. (2019a). A Survey for predicting enzyme family classes using machine learning methods. *Curr. Drug Targets* 20 (5), 540–550. doi: 10.2174/1389450119666181002143355

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019b). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16 (4), 2466–2480. doi: 10.3934/mbe.2019/23

Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., Yang, Q., et al. (2019a). ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform*. doi: 10.1093/bib/bby127

Tang, J., Fu, J., Wang, Y., Luo, Y., Yang, Q., Li, B., et al. (2019b). Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol. Cell. Proteomics : MCP* 18 (8), 1683–1699. doi: 10.1074/mcp.RA118.001169

Wang, H., Liu, C., and Deng, L. (2018). Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Sci. Rep.* 8 (1), 14285. doi: 10.1038/s41598-018-32511-1

Wang, Y., Yang, S., Zhao, J., Du, W., Liang, Y., Wang, C., et al. (2019). Using machine learning to measure relatedness between genes: a multi-features model. *Sci. Rep.* 9 (1), 4192. doi: 10.1038/s41598-019-40780-7

Wang, Y., Zhang, S., Yang, L., Yang, S., Tian, Y., and Ma, Q. (2019b). relatedness between genes using fully convolutional neural network. *Front. Genet.* 10, 1009. doi: 10.3389/fgene.2019.01009

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. In Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Tang, J., and Zou, Q. (2017b). PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Trans. nanobioscience* 16 (4), 240–247. doi: 10.1109/TNB.2017.2661756

Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Therapy-Nucleic Acids* 12, 635–644. doi: 10.1016/j.omtn.2018.07.004

Xiao, C.-L., Zhu, S., He, M., Chen, D., Zhang, Q., Chen, Y., et al. (2018). N6-Methyladenine DNA Modification in the Human Genome. *Mol. Cell* 71 (2), 306–318.e307. doi: 10.1016/j.molcel.2018.06.015

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D. Q. (2018). PredT4SE-Stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol* 9, 2571. doi: 10.3389/fmicb.2018.02571

Xu, Y., Liu, L., Guo, D., Jin, G., and Zhou, X. (2017). Alternative splicing links histone modifications to cell-cycle progression contributing to ESC differentiation. *Genome Biol.* Under revision. doi: 10.1186/s13059-018-1512-3

Xu, Y., Wang, Y., Luo, J., Zhao, W., and Zhou, X. (2017). Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. *Nucleic Acids Res.* 45 (21), 12100–12112. doi: 10.1093/nar/gkx870

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

Xu, L., Liang, G., Shi, S., and Liao, C. (2018a). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *Int. J. Mol. Sci.* 19 (6), 1773. doi: 10.3390/ijms19061773

Xu, L., Liang, G., Wang, L., and Liao, C. (2018b). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* 9 (3), 158. doi: 10.3390/genes9030158

Xu, L., Liang, G., Liao, C., Chen, G.-D., and Chang, C.-C. (2018c). An efficient classifier for Alzheimer's disease genes identification. *Molecules* 23 (12), 3140. doi: 10.3390/molecules23123140

Xu, Z. C., Feng, P. M., Yang, H., Qiu, W. R., Chen, W., and Lin, H. (2019). iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics*. doi: 10.1093/bioinformatics/btz358

Xu, L., Liang, G., Liao, C., Chen, G.-D., and Chang, C.-C. (2019). k-Skip-n-Gram-RF: a random forest based method for alzheimer's disease protein identification. *Front. In Genet.* 10 (33). doi: 10.3389/fgene.2019.00033

Xu, H., Zeng, W., Zeng, X., and Yen, G. G. (2019). An evolutionary algorithm based on minkowski distance for many-objective optimization. *IEEE Trans. Cybern.* 49 (11), 3968–3979. doi: 10.1109/TCYB.2018.2856208

Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., et al. (2018). What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem. Neurosci.* 9 (5), 1128–1140. doi: 10.1021/acschemneuro.7b00490

Yang, Q., Li, B., Tang, J., Cui, X., Wang, Y., Li, X., et al. (2019). Consistent gene signature of schizophrenia identified by a novel feature selection strategy from

comprehensive sets of transcriptomic data. *Brief Bioinform*. doi: 10.1093/bib/bbzo49

Yin, J., Sun, W., Li, F., Hong, J., Li, X., Zhou, Y., et al. (2019). VARIDT 1.0: variability of drug transporter database. *Nucleic Acids Res*. doi: 10.1093/nar/gkz779

Yu, L., Yao, S. Y., Gao, L., and Zha, Y. H. (2019). conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments. *Front. In Genet*. 9. doi: 10.3389/fgene.2018.00745

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Trans. Comput. Biol. Bioinf*. 14 (3), 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X. X., Liu, L., Lu, L. Y., and Zou, Q. (2018a). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 34 (14), 2425–2432. doi: 10.1093/bioinformatics/bty112

Zeng, X., Lin, J., Lin, C., Liu, X., and Rodríguez-Patón, A. J. I. A. (2018b). Structural hole spanner in humannet identifies disease gene and drug targets. *IEEE Access* 6, 35392–35401. doi: 10.1109/ACCESS.2018.2849216

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019a). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*. doi: 10.1093/bioinformatics/btz418

Zeng, X., Wang, W., Chen, C., and Yen, G. G. (2019b). JIToC: a consensus community-based particle swarm optimization for dynamic community detection. *IEEE Trans. Cybern*. doi: 10.1109/TCYB.2019.2938895

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018a). SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PloS Comput. Biol*. 14 (12), e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018b). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.neucom.2017.07.065

Zhang, W., Yue, X., Huang, F., Liu, R., Chen, Y., and Ruan, C. (2018c). Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. *Methods* 145, 51–59. doi: 10.1016/j.ymeth.2018.06.001

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. J. (2019). IAToCB, Bioinformatics: meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans. Comput. Biol. Bioinf*. 16 (1), 283–291. doi: 10.1109/TCBB.2017.2776280

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019a). SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions. *Inf. Sci*. 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019b). A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. bioinformatics/IEEE ACM*. doi: 10.1109/TCBB.2019.2931546

Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2019c). KATZLGO: large-scale prediction of LncRNA functions by using the KATZ measure based on multiple networks. *IEEE/ACM Trans. Comput. Biol. Bioinf*. 16 (2), 407–416. doi: 10.1109/TCBB.2017.2704587

Zheng, N., Wang, K., Zhan, W., and Deng, L. (2019). Targeting virus-host protein interactions: feature extraction and machine learning approaches. *Curr. Drug Metab*. 20 (3), 177–184. doi: 10.2174/1389200219666180829121038

Zhong, B. N., Bai, B., Li, J., Zhang, Y. L., and Fu, Y. (2019). Hierarchical tracking by reinforcement learning-based searching and coarse-to-fine verifying. *IEEE Trans. Image Process* 28 (5), 2331–2341. doi: 10.1109/TIP.2018.2885238

Zhou, H., Chen, C., Wang, M., Ma, Q., and Yu, B. (2019). Predicting golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion. *IEEE Access* 7, 144154–144164. doi: 10.1109/ACCESS.2019.2938081

Zhou, Q. Q., Zhong, B. N., Zhang, Y. L., Li, J., and Fu, Y. (2019b). Deep alignment network based multi-person tracking with occlusion and motion reasoning. *IEEE Trans. Multimedia* 21 (5), 1183–1194. doi: 10.1109/TMM.2018.2875360

Zhu, P. F., Zhu, W. C., Hu, Q. H., Zhang, C. Q., and Zuo, W. M. (2017). Subspace clustering guided unsupervised feature selection. *Pattern Recognition* 66, 364–374. doi: 10.1016/j.patcog.2017.01.016

Zhu, P. F., Xu, Q., Hu, Q. H., and Zhang, C. Q. (2018a). Co-regularized unsupervised feature selection. *Neurocomputing* 275, 2855–2863. doi: 10.1016/j.neucom.2017.11.061

Zhu, P. F., Xu, Q., Hu, Q. H., Zhang, C. Q., and Zhao, H. (2018b). Multi-label feature selection with missing labels. *Pattern Recognition* 74, 488–502. doi: 10.1016/j.patcog.2017.09.036

Zhu, X., He, J., Zhao, S., Tao, W., Xiong, Y., and Bi, S. (2019). A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae. *Briefings In Funct. Genomics*. doi: 10.1093/bfgp/elz018

Zou, Q., Li, J., Hong, Q., Lin, Z., Shi, H., Wu, Y., et al. (2015). Prediction of microRNA-disease associations based on social network analysis methods. *BioMed. Res. Int*. 2015, 810514. doi: 10.1155/2015/810514

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016a). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016b). Similarity computation strategies in the microRNA-disease network: a survey. *Briefings In Funct. Genomics* 15 (1), 55–64. doi: 10.1093/bfgp/elv024

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. *Nat. Genet*. 51 (1), 12–18. doi: 10.1038/s41588-018-0295-5

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. J. (2019). Bib: Sequence clustering in bioinformatics: an empirical study. *Briefings In Bioinf*. 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.