Interdisciplinary Sciences: Computational Life Sciences (2021) 13:413–425
https://doi.org/10.1007/s12539-021-00429-4

ORIGINAL RESEARCH ARTICLE

i6mA-VC: A Multi-Classifier Voting Method for the Computational Identification of DNA N6-methyladenine Sites

Tian Xue[1] · **Shengli Zhang**[1] · **Huijuan Qiao**[1]

Received: 19 December 2020 / Revised: 26 March 2021 / Accepted: 29 March 2021 / Published online: 8 April 2021

Abstract

DNA N6-methyladenine (6 mA), as an essential component of epigenetic modification, cannot be neglected in genetic regulation mechanism. The efficient and accurate prediction of 6 mA sites is beneficial to the development of biological genetics. Biochemical experimental methods are considered to be time-consuming and laborious. Most of the established machine learning methods have a single dataset. Although some of them have achieved cross-species prediction, their results are not satisfactory. Therefore, we designed a novel statistical model called i6mA-VC to improve the accuracy for 6 mA sites. On the one hand, kmer and binary encoding are applied to extract features, and then gradient boosting decision tree (GBDT) embedded method is applied as the feature selection strategy. On the other hand, DNA sequences are represented by vectors through the feature extraction method of ring-function-hydrogen-chemical properties (RFHCP) and the feature selection strategy of ExtraTree. After fusing the two optimal features, a voting classifier based on gradient boosting decision tree (GBDT), light gradient boosting machine (LightGBM) and multilayer perceptron classifier (MLPC) is constructed for final classification and prediction. The accuracy of Rice dataset and M.musculus dataset with five-fold cross-validation are 0.888 and 0.967, respectively. The cross-species dataset is selected as independent testing dataset, and the accuracy reaches 0.848. Through rigorous experiments, it is demonstrated that the proposed predictor is convincing and applicable. The development of i6mA-VC predictor will become an effective way for the recognition of N6-methyladenine sites, and it will also be beneficial for biological geneticists to further study gene expression and DNA modification. In addition, an accessible web-server for i6mA-VC is available from http://www.zhanglab.site/.

Keywords DNA N6-methyladenine sites · Ring-function-hydrogen-chemical properties · Light gradient boosting machine · Multilayer perceptron classifier · Voting

1 Introduction

As a significant part of epigenetics, DNA methylation can change gene expression without changing gene sequence. DNA methylation can cause variations in chromatin structure that controls gene expression. For eukaryotes, 5-methylcytosine (5mC) is regarded as the most widely distributed [1, 2], while N6-methyladenine (6 mA) is the easiest to be found in prokaryotes and some unicellular eukaryotes [3]. In previous reports, it has been claimed that 6 mA also exists in eukaryotes [1, 4]. The 6 mA modification is tightly associated to the processes of replication [5], transcription [6] and repair [7] of genetic information and cellular defense [8–10].

In the past few decades, the experimental technology of 6 mA modification has been explored and identified by a large number of researchers. Dunn et al. [3] developed a technique combining ultraviolet absorption spectra and electrophoretic mobility to identify purines in 1955, while this technique is extremely insensitive to the prediction of 6 mA sites in animals. In 1978, a restriction enzyme technique was further developed to predict 6 mA modification [11]. Unfortunately, this restriction enzyme method can only be used to identify adenosines occurring in restriction enzyme target sequences. Subsequently, methylated DNA immunoprecipitation sequencing [12] was combined with high-throughput sequencing technology to analyze DNA methylation in the whole genome in 2009. Single-molecule real-time

✉ Shengli Zhang
shengli0201@163.com

[1] School of Mathematics and Statistics, Xidian University, Xi'an 710071, People's Republic of China

sequencing [13] the following year detected DNA modified nucleotides by catalyzing nucleotides with DNA polymerase. Later, capillary electrophoresis and laser-induced fluorescence [14] were studied to quantify the global adenine methylation of DNA. In succession, a series of biological research methods have appeared to identify 6 mA modification sites [15]. Biochemical exploration technology can provide some information related to 6 mA sites, but they rely on higher economic cost and longer operation time. Therefore, it is indispensable to design an efficient and powerful computing model to make up for the shortcomings of traditional experimental and identify 6 mA sites for pre-existing experimental data.

The rapid increase of biological sequences in recent years has effectively promoted the application of machine learning in biological research. Some biologists analyzed the 6 mA map of Rice genome using traditional biochemical experiment method [16]. At the same time, the smooth progress of this research has led many researchers to gradually turn to the development of calculational models. Chen et al. [17] predicted the 6 mA sites using the feature extraction technology of nucleotide and chemical properties combined with support vector machine(SVM). Le et al. [18] treated DNA sequences through continuous bags of nucleobases, and then input them into SVM classifier for recognition. Feng et al. [19] proposed a predictor by combining the physical and chemical properties into Pseudo K-tuple Nucleotide Composition. A model named MM-6mAPred was developed by Pian et al. [20], which used Markov to identify 6 mA sites. After that, Huang et al. [21] designed a 6 mA-RicePred prediction model based on MM-6mAPred and nucleotide chemical properties. I6mA-DNCP proposed by Kong et al. [22] combined dinucleotide composition with properties, and then employed Bagging classifier for prediction. CsDMA model introduced by Liu et al. [23] was the first model applied to cross-species 6 mA identification using combination of motif, kmer and binary feature coding schemes. Subsequently, Wahab et al. [24] adopted convolution neural network to predict cross-species datasets. In addition, there are still many excellent models for identification of 6 mA sites, such as iDNA6mA [25], i6mA-DNC [26], iDNA6mA-Rice [27], SDM6A [28], SICD6ma [29], SNNRice6mA [30] and so on. However, some of them only predicted in a single species, and others predicted in multi-species datasets with unsatisfactory results. To realize 6 mA prediction of multi-species and obtain good results, we designed a novel and powerful model called i6mA-VC: (1) Kmer and binary are applied for feature extraction, and then GBDT feature selection is performed; (2) feature extraction based on RFHCP and feature selection based on ExtraTree are developed; (3) the two part features obtained from (1) and (2) are simply combined together; (4) the voting classifier based on GBDT, LightGBM and MLPC is carried out for the predictor; (5) cross-validation is adopted for Rice dataset and M.musculus dataset, and independent dataset testing is employed for cross-species dataset. Ultimately, our predictor has got better prediction accuracy for cross-species 6 mA sites prediction. Here, Fig. 1 shows the operating flow of i6mA-VC.

2 Materials and Methods

2.1 Benchmark Dataset

Reasonable construction of dataset is the beginning of developing a favorable predictor. To select a more suitable method for 6 mA recognition, we used the same benchmark datasets as the published papers, which are rice genome, M.musculus genome and cross-species dataset constructed from the above two genomes. Rice dataset was constructed by Chen et al. [17] in 2019, consisting of 880 6 mA samples and 880 non-6 mA samples. Meanwhile, M.musculus dataset was created by Feng et al. [19] in 2018, including 1934 6 mA samples and 1934 non-6 mA samples. In addition, cross-species dataset was from csDMA [23], which sets its threshold to 0.8 through CD-HIT-EST [31] to reduce the redundancy of sequences. It includes 2768 6 mA samples and 2716 non-6 mA samples, which is an independent dataset. To show the superiority of the method we chose, we employed the same training set and test set as the previous research [24], in which the training set contains 2214

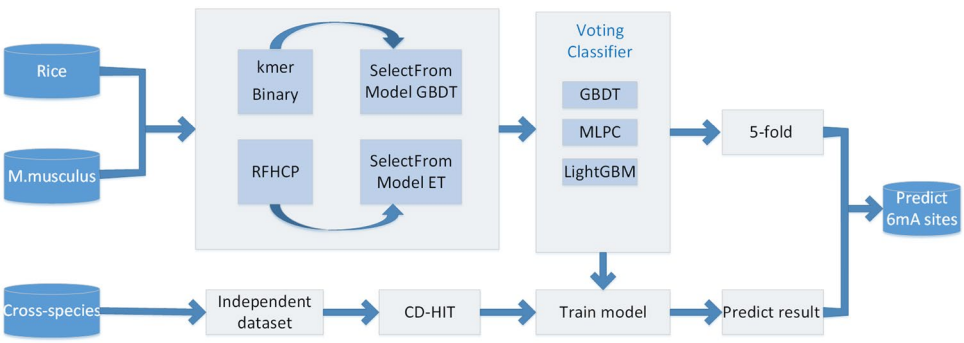

Fig. 1 The operating flow of i6mA-VC

6 mA samples and 2214 non-6 mA samples, and the test set includes 554 6 mA samples and 502 non-6 mA samples. The length of all samples is 41nt. The composition of the dataset is described in Table 1.

2.2 Feature Extraction

With the increasing number of gene sequences, fundamental biological knowledge provides effective ways for representing DNA sequences into vectors. At present, a large number of feature extraction techniques have been employed to process DNA samples, and the integration of various feature extraction methods has become the mainstream. In addition, there are many well-built servers that can output digital feature information after inputting DNA sequences, such as Pse-in-one [32, 33], ilearn [34], pyfeat [35], ifeature [36] and so on. In this paper, we extract the features by three methods, namely kmer, binary and RFHCP for a given DNA sequence among many different feature representing ways to obtain the best performance through extensive experiments.

2.2.1 Kmer

Kmer feature descriptor is one of the most common feature extraction method [37–39], which counts the frequency of nucleotide sequences with length k. When the value of k in a sequence is given, the corresponding feature descriptor can be obtained. When $k = 1$, the kmer descriptor represents four nucleotides. When $k = 2$, it means that any two nucleotides are combined, which are AA, AT, AC, AG, TA, TT, … GC, GG [40]. The feature dimension gradually increases with the increase of k. However, the excessively high characteristic dimension may cause the phenomenon of "dimension disaster". Here, we set the values of k as 1,2,3,4 and 5. Ultimately, it can be implemented to transform each sequence into a 1364-dimensional feature vector.

2.2.2 Binary Encoding

For DNA sequences, it is a simple and popular method to extract features using nucleotide binary encoding descriptor. It represents the sequence features by binary encoding system, in other words, adenine, cytosine, guanine and thymine can be converted to the following quaternion: (1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1). Eventually, the sequence of m length can be represented by $4 * m$-length vector [27]. In this study, each DNA sequence can be represented by a $41 * 4 = 164$ dimensional vector by binary encoding.

Table 1 Composition of datasets of different species

	6 mA sample	Non-6 mA sample	Total
Rice	880	880	1760
M. musculus	1934	1934	3868
Cross-species	2214 (training)	2214 (training)	4428 (training)
	554 (testing)	502(testing)	1056 (testing)

2.2.3 Ring-Function-Hydrogen-Chemical Properties (RFHCP)

Different types of nucleotides in each DNA sequence contain different ring numbers, chemical functions and hydrogen bond strengths according to their different chemical structures [41–45]. In terms of the number of rings, adenine and guanine have two rings, unlike cytosine and thymine, which have only one ring. As far as chemical functions are concerned, adenine and cytosine are distributed in the same amino group, which is different from guanine and thymine in ketone group; therefore, they can be divided into two groups. As for hydrogen bonds, nucleotides can be divided into two groups: adenine and thymine, cytosine and guanine due to the strength of hydrogen bonds. According to their chemical properties, we can use a four-dimensional vector (i, j, k, d_i) to express every nucleotide, in which i, j and k are calculated by the following formula:

$$\begin{aligned} i &= \begin{cases} 1 & \text{if } x \in \{\text{adenine,guanine}\} \\ 0 & \text{if } x \in \{\text{cytosine,thymine}\} \end{cases}; \\ j &= \begin{cases} 1 & \text{if } x \in \{\text{adenine,cytosine}\} \\ 0 & \text{if } x \in \{\text{guanine,thymine}\} \end{cases}; \\ k &= \begin{cases} 1 & \text{if } x \in \{\text{adenine,thymine}\} \\ 0 & \text{if } x \in \{\text{cytosine,guanine}\} \end{cases}; \end{aligned} \tag{1}$$

Therefore, the first three-dimensional vectors of adenine, cytosine, guanine and thymine can be expressed as (1,1,1), (0,1,0), (1,0,0) and (0,0,1), respectively. The fourth dimension d_i in the vector represents the density of nucleotide, which can be calculated as follows:

$$d_i = \frac{1}{|N_i|} \sum_{j=1}^{L} f(n_i), \tag{2}$$

where $|N_i|$ is the total length from the starting position of the sequence to the i-th position, and $f(n_i)$ is obtained by the following formula:

$$f(n_i) = \begin{cases} 1 \ if \ n_i = q \\ 0 \ \text{others} \end{cases}, q \in \{A, C, G, T\}. \tag{3}$$

At last, a sequence of m length can be converted into a $4 * m$ length feature vector through RFHCP feature descriptor, and the converted feature dimension is 41*4 = 164.

2.3 Feature Selection

Feature selection has a vital impact on the classification results in the classification process [46–48]. If the features are classified directly after feature extraction, it will often lead to too much calculation cost and even over-fitting, so it is particularly indispensable to select features in the process of building the model. Its main purpose is to eliminate redundant and unnecessary information through importance sorting, and at the same time achieve the effect of dimension reduction. The selectfrommodel in sklearn and GBDT classifier are employed to deal with kmer and binary encoding features, and the selectfrommodel module and ExtraTree (ET) classifier are adopted to optimize RFHCP. As the feature dimension is relatively high, it will cause too much computation cost and time consuming to fuse them directly. As a result, we consider selecting the features part by part and then fusing them all together, which can reduce the amount of calculation, improve the speed, and produce satisfactory results.

2.4 Voting Classifier

There are various classification algorithms in machine learning, and these different classifiers can affect the classification results strongly. In this study, we adopt the ensemble learning, concentrating on various algorithms and letting multiple algorithms predict the same model. Here, we use voting classifier for classification, and set the voting category to hard voting which means that the minority is subordinate to the majority. GBDT, LightGBM, and MLPC are selected as the three basic classifiers of voting classifier for the prediction of 6 mA sites.

2.4.1 Gradient Boosting Decision Tree (GBDT)

With the increase of genetic data year by year, ensemble learning has become the mainstream of classification algorithms. Boosting algorithm is an important branch of ensemble learning, which is a kind of common and effective statistical algorithm. It boosts the residue of the previous weaker classifier round by round to elevate the classification ability and achieve the goal which the objective function value is small enough. Among boosting algorithm, GBDT is the most common one, which was proposed by Freidman [49].

The core idea of GBDT is that the negative gradient value of loss function in the previous model is regarded as the approximate value of residual error to fit a regression tree. The GBDT algorithm is described as follows:

1. Assuming that the predicted value is constant for minimizing the loss function, the initial model is expressed by the following formula:

$$g_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c). \tag{4}$$

2. The negative gradient value is solved and recorded as r_{ni}, which can also be expressed as the approximate value of the residual error:

$$r_{ni} = -\left[\frac{\partial L(y_i, g(x_i))}{\partial g(x_i)}\right]_{g(x)=g_{n-1}(x)}, \tag{5}$$

where n represents the number of iterations and i represents the i-th sample. A regression tree is fitted for each r_{ni} to obtain the leaf node region R_{nj} of the n-th tree, and then the output value of each region is calculated as the following formula:

$$c_{nj} = \arg\min_c \sum_{x_i \in R_{nj}} L(y_i, g_{n-1}(x_i) + c). \tag{6}$$

The regression tree model is updated to

$$g_n(x) = g_{n-1}(x) + \sum_{j=1}^{J} c_{nj} I(x \in R_{nj}). \tag{7}$$

3. The final model is obtained as follows:

$$\hat{g}(x) = g_N(x) = \sum_{n=1}^{N} \sum_{j=1}^{J} c_{nj} I(x \in R_{nj}). \tag{8}$$

2.4.2 Light Gradient Boosting Machine (LightGBM)

LightGBM [50] is a distributed gradient lifting algorithm based on decision tree theory, which is improved from GBDT algorithm and has higher performance and stronger robustness. Similar to the principle of GBDT, it fits the new decision tree based on the residual of the previous round. It adds grandient-based one-side sampling (GOSS) algorithm and exclusive feature bunding (EFB) algorithm on the basis of GBDT. The core of GOSS is to optimize the way of sampling for training samples, and the core of EFB is to bind mutually exclusive features together, thereby reducing feature dimensions.

Compared with the traditional GBDT, LightGBM has higher training efficiency and accuracy, and supports parallel learning and large-scale data processing.

2.4.3 Multilayer Perceptron Classifier (MLPC)

MLPC is a kind of classifier with feedforward neural network as classification algorithm, which is a neural network model with forward propagation mechanism. It includes input layer, output layer, and hidden layer between them. There is no activate function in the input layer, and its nodes directly receive data for matrix operation. There are usually one or more hidden layers, and each layer of neurons has a sigmoid function to process the operation results. The neurons in the output layer also have an activation function, which is usually a SoftMax function, and the dimension of the output layer indicates the number of categories in the classifier. MLPC adopts BP algorithm in the learning process, and ultimately abstracts the problem into optimizing logistic loss function using L-BFGS.

2.5 Performance Metrics

To demonstrate the applicability and robustness of the model, cross-validation is considered to be the most persuasive method among all methods, including K-fold cross-validation, independent dataset test, and jackknife test [51]. To verify the validity of the proposed predictor, we use the same evaluation methods as published papers, and employ fivefold cross-validation for Rice dataset and M.musculus dataset. At the same time, the independent dataset testing method is employed to test cross-species dataset, and the same training set and test set as the previous articles are selected to illustrate the feasibility of the model.

Similarly, the four effective evaluation indexes are applied to measure the applicability and reliability of the predictor, which are ACC or accuracy, Sn or sensitivity, Sp or specificity and MCC or Matthew's correlation coefficient [52–59]. The related calculation of the four indicators is expressed as follows:

$$\begin{cases} \mathrm{ACC} = 1 - \frac{S_-^+ + S_+^-}{S^+ + S^-}, \\ \mathrm{Sn} = 1 - \frac{S_-^+}{S^+}, \\ \mathrm{Sp} = 1 - \frac{S_+^-}{S^-}, \\ \mathrm{MCC} = \frac{1 - \frac{S_-^+ + S_+^-}{S^+ + S^-}}{\sqrt{(1 - \frac{S_+^- + S^+}{S^+})(1 - \frac{S_-^+ + S_+^-}{S^-})}}, \end{cases} \tag{9}$$

where S^+ represents the total number of all 6 mA samples, S^- represents the total number of all non-6 mA samples, S_-^+ represents the total number of samples that were originally 6 mA samples but predicted to be non-6 mA samples and S_+^- represents the total number of samples that were originally non-6 mA samples but predicted to be 6 mA samples.

3 Results

3.1 Nucleotide Composition Analysis

The two sample logo [60] is used to detect the difference of nucleotide composition in DNA sequences, which is gradually used in the detection of various gene sample sequences [61–63]. As illustrated in Fig. 2, it shows the enrichment degree of nucleotides by T test ($p < 0.05$) in positive and negative sequences at 6 mA sites. The adenine is abundantly distributed in the positive sample sequence, while the guanine is more widely distributed in the negative sample sequence. The result indicates that the sequence of 6 mA sites can be determined by the enrichment degree of sequence information.

3.2 Comparison of Feature Extraction Methods

Feature extraction determines the quality of the model to a great extent, and the determination of feature extraction method is of great significance in building the model. In this research, kmer, binary encoding and RFHCP are adopted to extract features, and the simultaneous use of three methods makes our model obtain better results. We compared our experiment with the kmer, binary combining with GBDT feature selection method, and then with the RFHCP combining with ET feature selection method. As shown in Fig. 3, the four evaluation metrics are higher than the combination of kmer and binary, and higher than that of RFHCP alone. Therefore, it is sufficient to show that the combined use of multiple feature extraction methods is superior to the single feature use.

3.3 Comparison of Feature Selection Methods

For different feature extraction methods, several feature selection methods are compared. Meanwhile, the multiple feature selection in this paper is compared with the single method to reflect the advantages of multi-feature selection. For all feature extraction methods, the same feature selection method is adopted. As shown in Fig. 4, it is obviously observed that using different feature selection methods is indeed better than using only one feature selection method. In addition, we swapped the two feature selection methods, that is to say, ET is used to select features for kmer and binary, and GBDT is used to select features for RFHCP. Table 2 lists the comparison of the results after swapping the feature selection method with our method, which illustrates that the method before swapping feature selection is superior

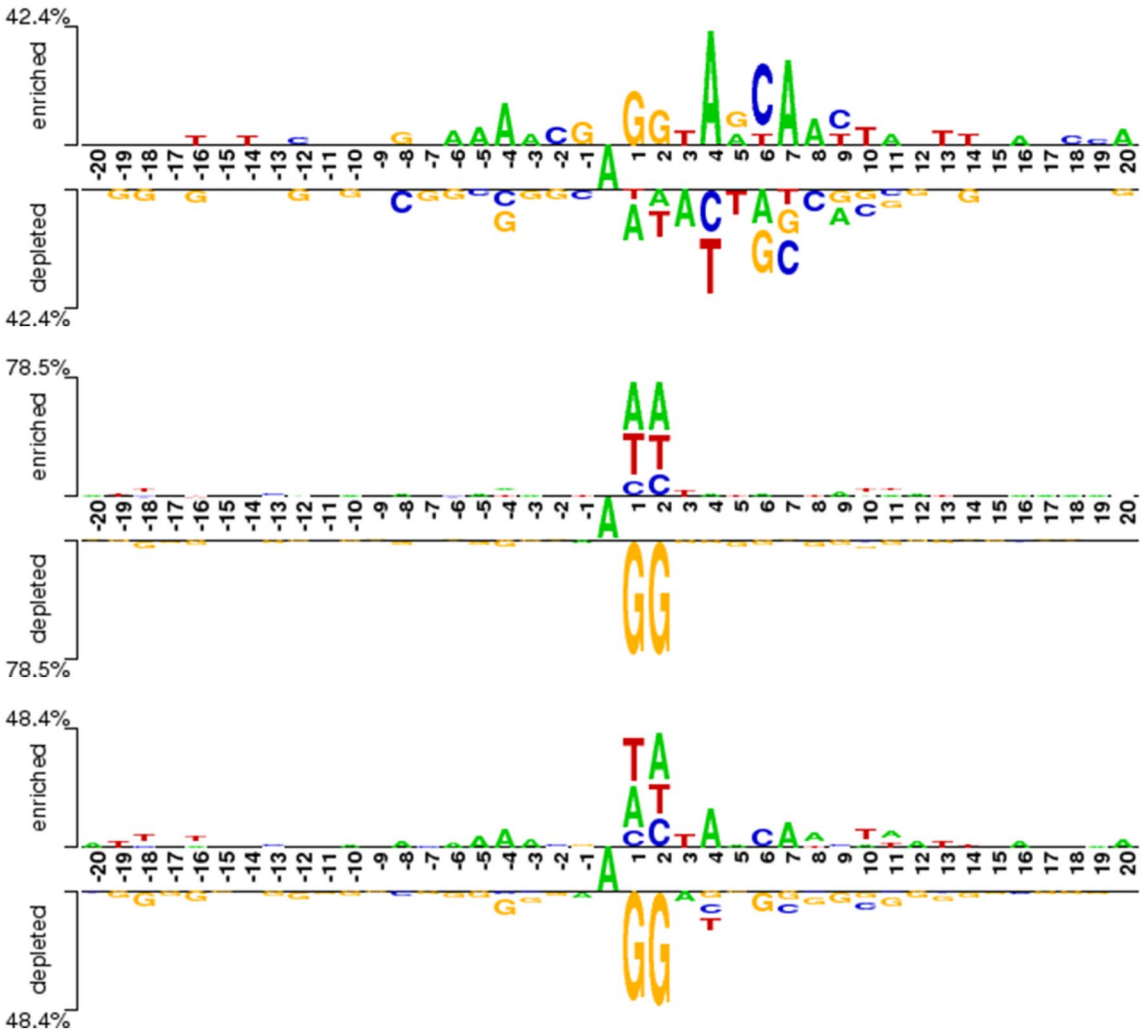

Fig. 2 Sequence analysis diagrams of three datasets. The upper part of the graph represents 6 mA sites, and the lower part represents non-6 mA sites

to the method after swapping feature selection. At the same time, we also compared other feature selection methods, such as MRMD2.0 [64, 65], F-score [61], Elastic Net [66], Recursive feature elimination(RFE), selectfrommodel LightGBM. Figure 5 displays the differences between the various methods. Although there is little difference among various feature selection methods on M. musculus dataset, on the whole, the feature selection method in our article is better than other methods on three datasets.

3.4 Comparison with Multiple Classifiers

The appropriate classifier directly determines the result of the model, so the selection of classifier is a key step for the assessment of the predictor. We carried out a series of experiments and operations in the process of selecting classifiers. Ultimately, we chose GBDT, LightGBM and MLPC as the base model, and then determined the final result by voting classifier. In the experiment, we compared the voting classifier with its three basic classifiers, the most common RF [67] and SVM [68], and XGboost [69] and ExtraTree [70], two excellent integrated classifiers. Figure 6 displays the comparison among the accuracy of various classifiers. It is enough to observe that the classification results of the three classifiers we selected are better than those of other classifiers. The classification results of the voting classifier are better than those of a single classifier. At last, ACC, Sp, Sn, and MCC of our model reach 0.888, 0.896, 0.881 and 0.776 on Rice dataset, 0.967,1, 0.933 and 0.935 on M.musculus dataset, and the four evaluation indexes on

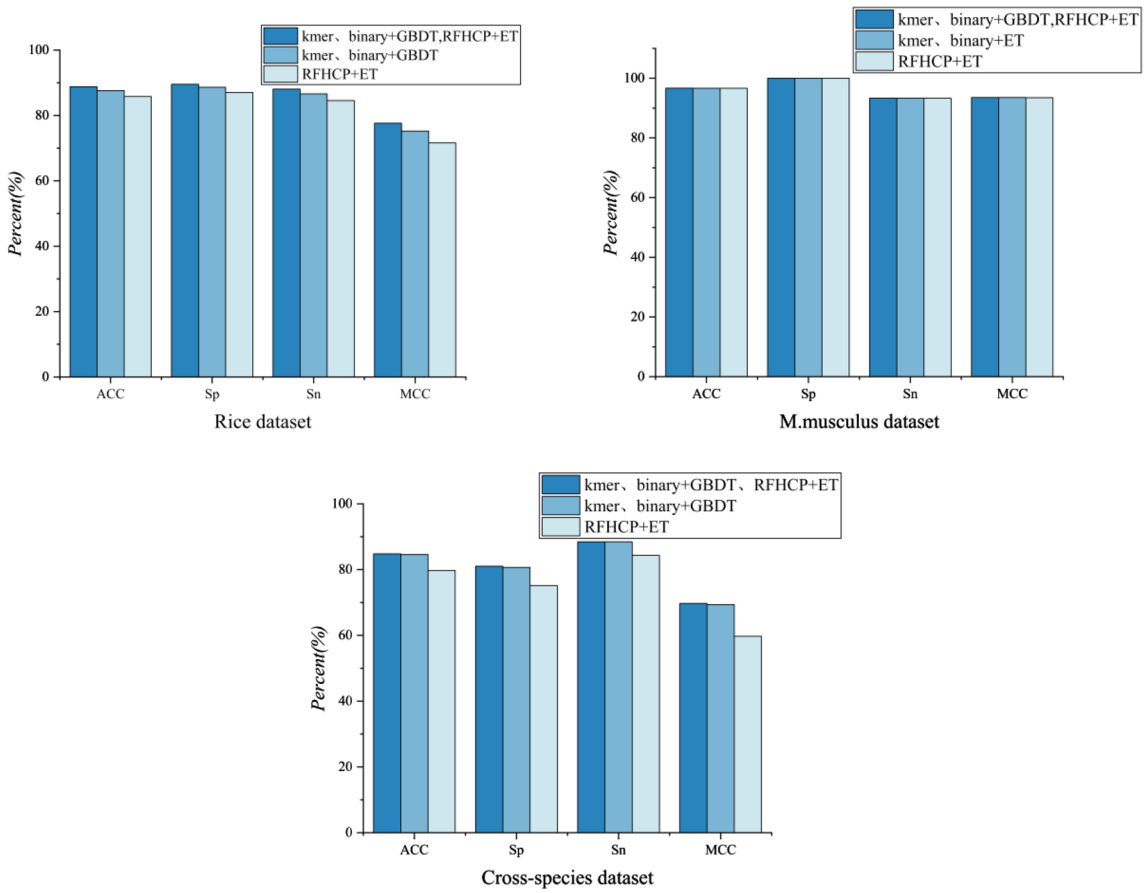

Fig. 3 Comparison of four evaluation indexes among multiple feature extraction methods and single feature extraction method

cross-species dataset are 0.848,0.810, 0.885, and 0.697, respectively.

3.5 Comparison with Existing Methods

In previous studies, many related articles about 6 mA sites identification have been published, but their results are not satisfactory. The results of i6mA-VC have been improved on the basis of the previous research. Table 3 lists the comparison of i6mA-VC with previous state-of-the-art predictors on three datasets. The four evaluation indexes of our model are 0.888, 0.896, 0.881 and 0.776, respectively, on Rice dataset, which exceed most published articles. And ACC, Sp, Sn, and MCC of our model reach 0.967,1, 0.933 and 0.935, respectively, for the M.musculus dataset. Although the accuracy is lower than iIM-CNN [24], our method performs better on other datasets. The independent dataset test is used to test cross-species dataset, and the results indicate that the index of our model is higher than that of existing models. Through experiments on three datasets, the feasibility of our predictor is demonstrated.

4 Web-server

For the i6mA-VC predictor, we developed a publicly available web-server to improve the practicality and meet the needs of biologists, and provided the operating guide of this server to meet the needs of users. In this process, users only need to select the species and input the sequences to get the desired results. Below, we will elaborate on how it is used.

The web-server is available from http://www.zhanglab.site/, and the interface is shown in Fig. 7. The source of our

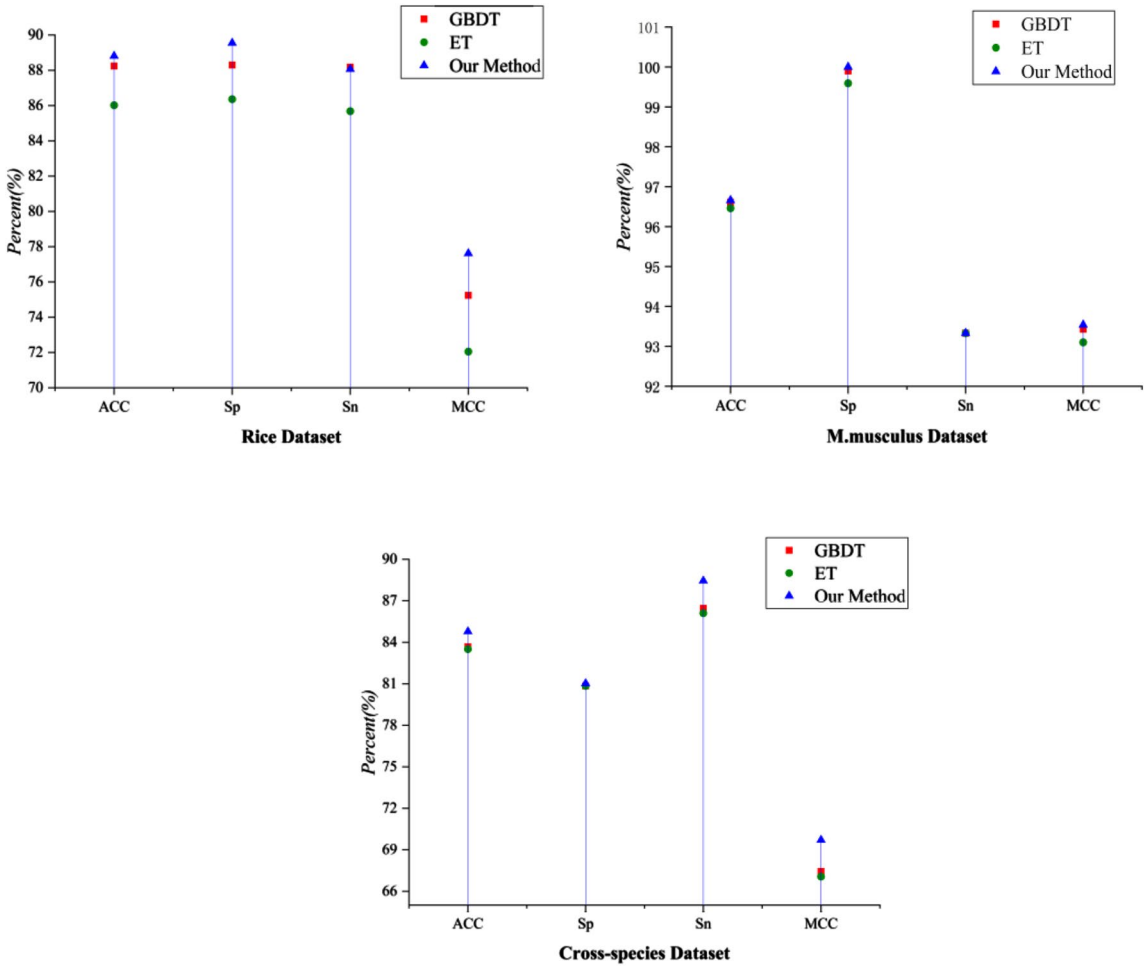

Fig. 4 Comparison of using multiple feature selection methods and single feature selection method

Table 2 Comparison between our experimental results and the results after swapping two feature selection methods

Methods	Species	ACC	Sn	Sp	MCC
Our method	Rice	0.888	0.881	0.896	0.776
	M. musculus	0.967	0.933	1	0.935
	Cross-species	0.848	0.885	0.810	0.697
After swapping feature selection methods	Rice	0.864	0.858	0.870	0.728
	M. musculus	0.967	0.933	1	0.935
	Cross-species	0.830	0.872	0.784	0.660

dataset is described in *Data*. A detailed description of the web-server is given in *Read Me*. The paper that made significant contributions to the i6mA-VC predictor is displayed in *Citation*.

In the interface, you just need to enter the sequence in the text box below and click *Submit* to get the predicted results. Users can click *Example* to view the standard DNA sequence format. If you need to enter the sequence for the second time, you can click *Clear* to clear the contents and re-enter. Here, it should be reminded that the input format of the sequence must be FASTA format and the data length must be 41nt.

5 Discussion

DNA N6-methyladenine is considered to be an essential component of epigenetic modification in biological evolution, and the prediction of its sites has become a crucial basis for study of corresponding biological functions. At present, many related experimental methods and computational

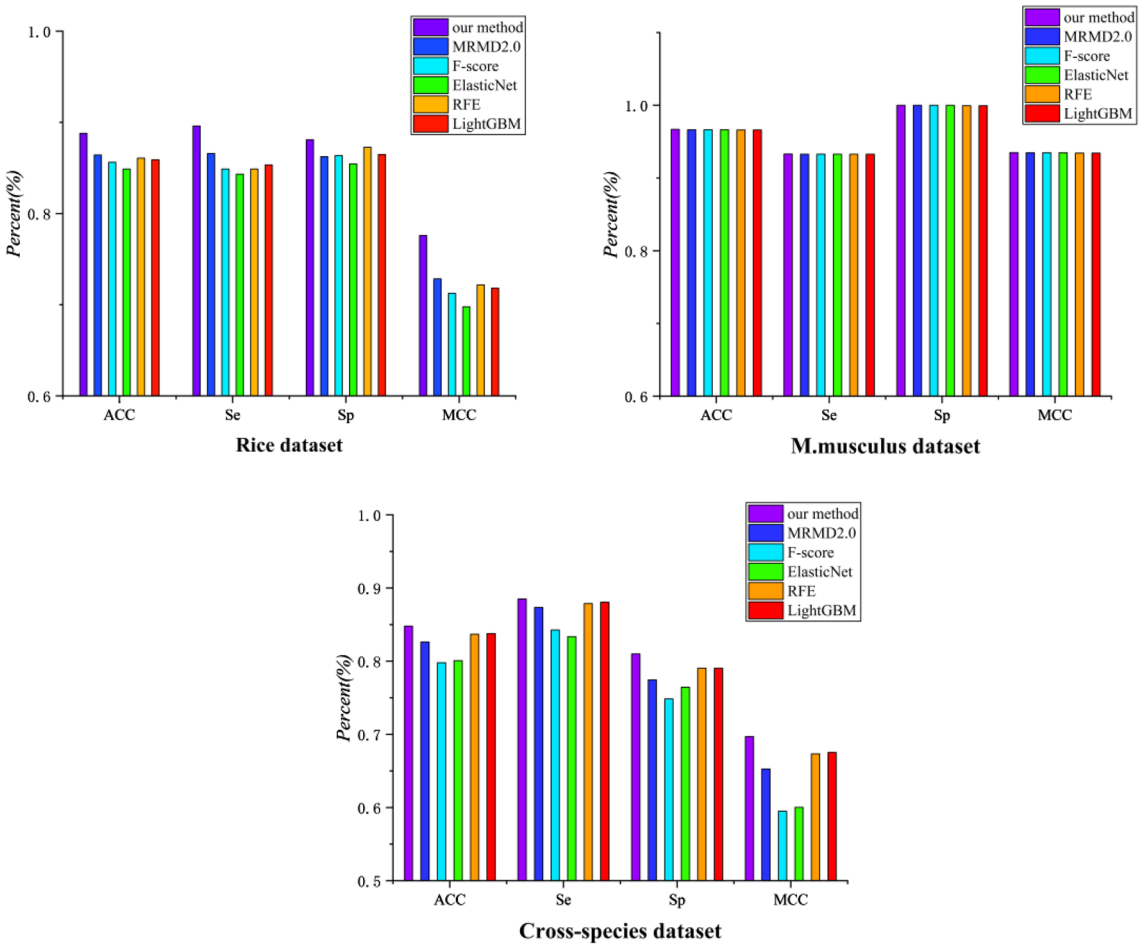

Fig. 5 Comparison of different feature selection methods

Fig. 6 Comparison for multiple classifiers

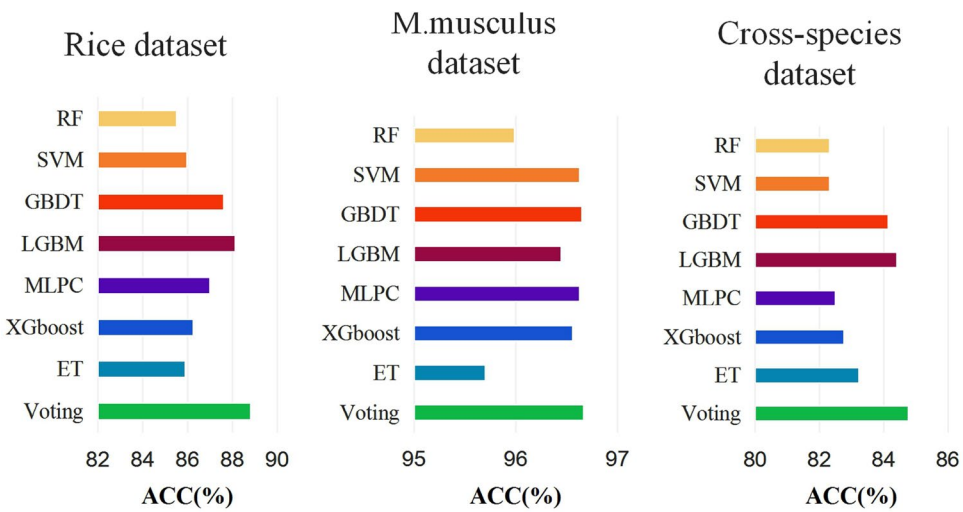

Table 3 Comparison with previous predictors

Species	Algorithm	ACC	Sp	Sn	MCC
Rice	iDNA6mA-PseKNC	0.641	0.721	0.569	0.394
	csDMA	0.861	0.880	0.842	0.723
	iIM-CNN	0.875	0.914	0.841	0.752
	i6mA-VC	0.888	0.896	0.881	0.776
M. musculus	iDNA6mA-PseKNC	0.935	1	0.869	0.877
	csDMA	0.966	1	0.932	0.935
	iIM-CNN	0.969	1	0.938	0.941
	i6mA-VC	0.967	1	0.933	0.935
Cross-species	iDNA6mA-PseKNC	0.765	0.769	0.762	0.531
	csDMA	0.799	0.735	0.863	0.603
	iIM-CNN	0.824	0.780	0.869	0.651
	i6mA-VC	0.848	0.810	0.885	0.697

models have been applied to identify 6 mA sites, but their results are not satisfactory and the experimental methods are not abundant enough. Therefore, we consider improving on the basis of previous research and introducing an effective statistical model called i6mA-VC. We adopt different feature selection methods for different feature extraction processes, and employ an ensemble voting classifier combined with a variety of effective classifiers, which obtain better results in the study of cross-species 6 mA sites, with the accuracy of 0.888, 0.967, and 0.848 on the rice, M.musculus and cross-species datasets, respectively. The establishment of the model lays a foundation for the discovery of potential effective drugs and the research of bioinformatics.

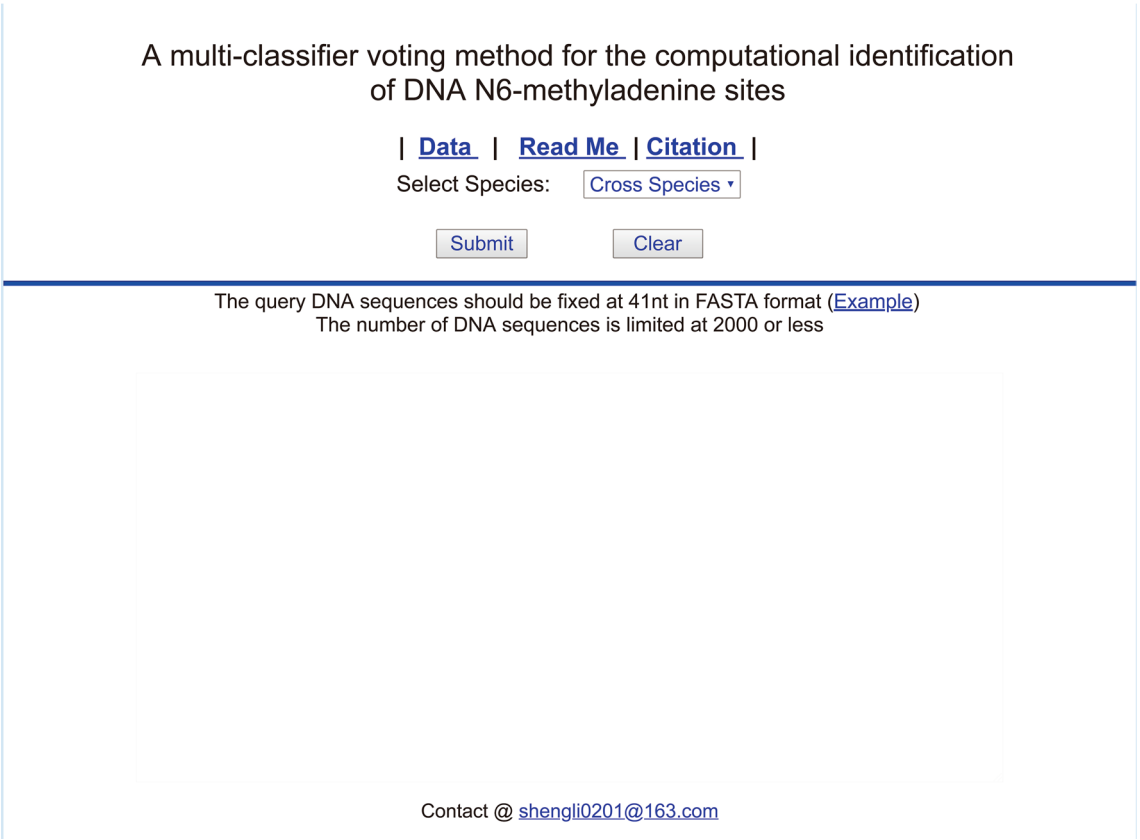

Fig. 7 The prediction interface of i6mA-VC web server

Acknowledgments This work was supported by the National Natural Science Foundation of China (No.11601407), the Natural Science Basic Research Program of Shaanxi (No. 2021JM-115), and the Fundamental Research Funds for the Central Universities (No. JB210715).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Vanyushin BF, Tkacheva SG, Belozersky AN (1970) Rare bases in animal DNA. Nature 225:948–949. https://doi.org/10.1038/225948a0
2. Vanyushin BF, Belozersky AN, Kokurina NA, Kadirova DX (1968) 5-Methylcytosine and 6-Methylaminopurine in bacterial DNA. Nature 218:1066–1067. https://doi.org/10.1038/2181066a0
3. Dunn DB, Smith JD (1955) Occurrence of a new base in the deoxyribonucleic acid of a strain of bacterium coli. Nature 175:336–337. https://doi.org/10.1038/175336a0
4. Unger G, Venner H (1966) Remarks on minor bases in spermatic desoxyribonucleic acid. Hoppe Seyler Z physiol Chem 344:280–283
5. Campbell JL, Kleckner N (1990) *E. coli* oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. Cell 62:967–979. https://doi.org/10.1016/0092-8674(90)90271-F
6. Robbins-Manke JL, Zdraveski ZZ, Marinus M, Essigmann JM (2005) Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase- and mismatch repair-deficient *Escherichia coli*. J Bacteriol 187:7027–7037. https://doi.org/10.1128/JB.187.20.7027-7037.2005
7. Pukkila PJ, Peterson J, Herman G, Modrich P, Meselson M (1983) Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in *Escherichia coli*. Genetics 104:571–582. https://doi.org/10.1093/genetics/104.4.571
8. Luria SE, Human ML (1952) A nonhereditary, host-induced variation of bacterial viruses. J Bacteriol 64:557–569. https://doi.org/10.1007/BF00410835
9. Meselson M, Yuan R (1968) DNA restriction enzyme from *E. coli*. Nature 217:1110–1114. https://doi.org/10.1038/2171110a0
10. Arber W, Dussoix D (1962) Host specificity of DNA produced by *Escherichia coli*. J Mol Biol 5:18–36. https://doi.org/10.1016/S0022-2836(62)80058-8
11. Bird AP (1978) Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. J. Mol. Biol. 118:49–60. https://doi.org/10.1016/0022-2836(78)90242-5
12. Pomraning KR, Smith KM, Freitag M (2009) Genome-wide high throughput analysis of DNA methylation in eukaryotes. Methods 47:142–150. https://doi.org/10.1016/j.ymeth.2008.09.022
13. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7:461–465. https://doi.org/10.1038/nmeth.1459
14. Krais AM, Cornelius MG, Schmeiser HH (2010) Genomic N6-methyladenine determination by MEKC with LIF. Electrophoresis 31:3548–3551. https://doi.org/10.1002/elps.201000357
15. Greer E, Blanco M, Gu L, Sendinc E, Liu J, Aristizabal-Corrales D, Hsu CH, Aravind L, He C, Shi Y (2015) DNA Methylation on N6-Adenine in *C. elegans*. Cell 161:868–878. https://doi.org/10.1016/j.cell.2015.04.005
16. Zhou C, Wang C, Liu H, Zhou Q, Liu Q, Guo Y, Peng T, Song J, Zhang J, Chen L, Zhao Y, Zeng Z, Zhou D-X (2018) Identification and analysis of adenine N6-methylation sites in the rice genome. Nat Plants 4:554–563. https://doi.org/10.1038/s41477-018-0214-x
17. Chen W, Lv H, Nie F, Lin H (2019) i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics 35:2796–2800. https://doi.org/10.1093/bioinformatics/btz015
18. Le NQK (2019) iN6-methylat (5-step): identifying DNA N6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule. Mol Genet Genomics 294:1173–1182. https://doi.org/10.1007/s00438-019-01570-y
19. Feng P, Yang H, Ding H, Lin H, Chen W, Chou KC (2018) iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics. https://doi.org/10.1016/j.ygeno.2018.01.005
20. Pian C, Zhang G, Li F, Fan X (2019) MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov Model. Bioinformatics 36:388–392. https://doi.org/10.1093/bioinformatics/btz556
21. Huang Q, Zhang J, Wei L, Guo F, Zou Q (2020) 6mA-RicePred: a method for identifying DNA N6-Methyladenine sites in the rice genome based on feature fusion. Front Plant Sci 11:4. https://doi.org/10.3389/fpls.2020.00004
22. Kong L, Zhang L (2019) i6mA-DNCP: computational identification of DNA N6-Methyladenine sites in the rice genome using optimized dinucleotide-based features. Genes 10:828. https://doi.org/10.3390/genes10100828
23. Liu Z, Dong W, Jiang W, He Z (2019) csDMA: an improved bioinformatics tool for identifying DNA 6 mA modifications via Chou's 5-step rule. Sci Rep-Uk 9:13109–13118. https://doi.org/10.1038/s41598-019-49430-4
24. Wahab A, Ali SD, Tayara H, Chong KT (2019) iIM-CNN: intelligent identifier of 6mA sites on different species by using convolution neural network. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2958618
25. Tahir M, Tayara H, Chong KT (2019) iDNA6mA (5-step rule): Identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via chou's 5-step rule. Chemometr Intell Lab 189:96–101. https://doi.org/10.1016/j.chemolab.2019.04.007
26. Park S, Wahab A, Nazari I, Ryu JH, Chong KT (2020) i6mA-DNC: Prediction of DNA N6-Methyladenosine sites in rice genome based on dinucleotide representation using deep learning. Chemometr Intell Lab 204:104102. https://doi.org/10.1016/j.chemolab.2020.104102
27. Hao L, Dao FY, Guan ZX, Zhang D, Lin H (2019) iDNA6mA-Rice: a computational tool for detecting n6-methyladenine sites in rice. Front Genet 10:793. https://doi.org/10.3389/fgene.2019.00793
28. Basith S, Manavalan B, Shin TH, Lee G (2019) SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. Mol Ther-Nucl Acids. https://doi.org/10.1016/j.omtn.2019.08.011
29. Liu W, Li H (2020) SICD6mA: identifying 6ma sites using deep memory network. BioRxiv. https://doi.org/10.1101/2020.02.02.930776
30. Yu H, Dai Z (2019) SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. Front Genet 10:1071–1077. https://doi.org/10.3389/fgene.2019.01071
31. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts565
32. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. https://doi.org/10.1093/nar/gkv458

33. Liu B, Wu H, Chou KC (2017) Pse-in-One 20: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nat Sci 9:67–91. https://doi.org/10.4236/ns.2017.94007
34. Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, Zhu Y, Powell DR, Akutsu T, Webb GI, Chou KC, Smith AI, Daly RJ, Li J, Song J (2019) iLearn: an integrated platform and meta-learner for feature engineering, machine learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform. https://doi.org/10.1093/bib/bbz041
35. Rafsanjani M, Sajid A, Dewan MF, Swakkhar S, Alok S, Abdollah D (2019) PyFeat: a Python-based effective feature generation tool for DNA RNA and protein sequences. Bioinformatics 35:3831–3833. https://doi.org/10.1093/bioinformatics/btz165
36. Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, Webb GI, Smith AI, Daly RJ, Chou KC, Song J (2018) iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 34:2499–2502. https://doi.org/10.1093/bioinformatics/bty140
37. He J, Fang T, Zhang Z, Huang B, Zhu X, Xiong Y (2018) PseUI: Pseudouridine sites identification based on RNA sequence information. BMC Bioinformatics 19:306. https://doi.org/10.1186/s12859-018-2321-0
38. Su ZD, Huang Y, Zhang ZY, Zhao YW, Wang D, Chen W, Chou KC, Lin H (2018) iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty508
39. Wang H, Ding Y, Tang J, Zou Q, Guo F (2021) Identify RNA-associated subcellular localizations based on multi-label learning using Chou's 5-steps rule. BMC Genomics 22:1–14. https://doi.org/10.1186/s12864-020-07347-7
40. Zhen C, Pan X, Yang Y, Huang Y, Shen HB (2018) The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. Bioinformatics 34:2185–2194. https://doi.org/10.1093/bioinformatics/bty085
41. Bari ATMG, Reaz MR, Choi HJ, Jeong BS (2013) DNA encoding for splice site prediction in large DNA sequence. Database Syst Adv Appl. https://doi.org/10.1007/978-3-642-40270-8_4
42. Chen W, Feng P, Tang H, Ding H, Lin H (2016) Identifying 2'-O-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions. Genomics 107:255–258. https://doi.org/10.1016/j.ygeno.2016.05.003
43. Chen W, Yang H, Feng P, Ding H, Lin H (2017) iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 33:3518–3523. https://doi.org/10.1093/bioinformatics/btx479
44. Wei L, Chen H, Su R (2018) M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. Mol Ther Nucleic Acids 12:635–644. https://doi.org/10.1016/j.omtn.2018.07.004
45. Wei L, Su R, Luan S, Liao Z, Manavalan B, Zou Q, Shi X (2019) Iterative feature representations improve N4-methylcytosine site prediction. Bioinformatics 35:4930–4937. https://doi.org/10.1093/bioinformatics/btz408
46. Lv Z, Jin S, Ding H, Zou Q (2019) A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. Front Bioeng Biotech 7(2019):215. https://doi.org/10.3389/fbioe.2019.00215.eCollection
47. Fu X, Cai L, Zeng X, Zou Q (2020) StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. Bioinformatics 36:3028–3034. https://doi.org/10.1093/bioinformatics/btaa131
48. Zhang S, Qiao H (2020) KD-KLNMF: identification of lncRNAs subcellular localization with multiple features and nonnegative matrix factorization. Anal Biochem. https://doi.org/10.1016/j.ab.2020.113995
49. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232. https://doi.org/10.2307/2699986
50. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: a highly efficient gradient boosting decision tree, In: 31st Conference Neural Information Processing Systems 30, pp 3149–3157. doi: https://doi.org/10.5555/3294996. 3295074.
51. Chou KC, Zhang CT (2008) Prediction of protein structural classes. Crit Rev Biochem Mol 30:275–349. https://doi.org/10.3109/10409239509083488
52. Su R, Hu J, Zou Q, Manavalan B, Wei L (2020) Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Brief Bioinform 21:408–420. https://doi.org/10.1093/bib/bby124
53. Manavalan B, Basith S, Shin TH, Wei L, Lee G (2019) mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 35:2757–2765. https://doi.org/10.1093/bioinformatics/bty1047
54. Jia J, Liu Z, Xiao X, Liu B, Chou KC (2015) iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol 377:47–56. https://doi.org/10.1016/j.jtbi.2015.04.011
55. Basith S, Manavalan B, Shin TH, Lee G (2018) iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. Comput Struct Biotec 16:412–420. https://doi.org/10.1016/j.csbj.2018.10.007
56. Manavalan B, Govindaraj RG, Shin TH, Kim MO, Lee G (2018) iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction. Front Immunol 9:1695. https://doi.org/10.3389/fimmu.2018.01695
57. Wei L, Luan S, Nagai LAE, Su R, Zou Q (2019) Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics 35:1326–1333. https://doi.org/10.1093/bioinformatics/bty824
58. Meng C, Guo F, Zou Q (2020) CWLy-SVM: a support vector Machine-based tool for identifying cell wall lytic enzymes. Comput Biol Chem 87:107304. https://doi.org/10.1016/j.compbiolchem.2020.107304
59. Zhang S, Zhu F, Yu Q, Zhu X (2021) Identifying DNA-binding proteins based on multi-features and LASSO feature selection. Biopolymers. https://doi.org/10.1002/bip.23419
60. Crooks GE (2004) WebLogo: a sequence logo generator. Genome Res 14:1188–1190. https://doi.org/10.1101/gr.849004
61. He W, Jia C, Zou Q (2018) 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics 35:593–601. https://doi.org/10.1093/bioinformatics/bty668
62. Wang J, Zhang S (2021) PA-PseU: an incremental passive-aggressive based method for identifying RNA pseudouridine sites via Chou's 5-steps rule. Chemometr Intell Lab. https://doi.org/10.1016/j.chemolab.2021.104250
63. Li J, Pu Y, Tang J, Zou Q, Guo F (2020) DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences. Brief Bioinform. https://doi.org/10.1093/bib/bbaa159
64. He S, Guo F, Zou Q, Ding H (2020) MRMD2.0: a python tool for machine learning with feature ranking and reduction. Curr. Bioinform. 15:1213–1221. https://doi.org/10.2174/1574893615999200503030350
65. Zhang YP, Zou Q (2020) PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and

adaptive feature representation learning. Bioinformatics 36:3982–3987. https://doi.org/10.1093/bioinformatics/btaa275
66. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc 67:768–768. https://doi.org/10.1111/j.1467-9868.2005.00527.x
67. Breiman L (2001) Random forest. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
68. Vapnik VN (1998) Statistical learning theory. In: New York: Wiley, p 1–768. doi: https://doi.org/10.1007/978-1-4419-1428-6_5864.
69. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. Acm sigkdd international conference on knowledge discovery and data mining, p 785–794 doi: https://doi.org/10.1145/2939672.2939785.
70. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42. https://doi.org/10.1007/s10994-006-6226-1