RESEARCH ARTICLE

A machine learning approach to identify *N*^{6}-methyladenine sites in the rice genome

Linghua Kong[1], Zhongwang Zhang[2], Xueda Zhao[1, *]

[1]School of Information Engineering, Dalian Ocean University, Dalian 116023, Liaoning, China. [2]School of Science, Dalian Maritime University, Dalian 116026, Liaoning, China

Received: January 16, 2023; **accepted:** February 20, 2023.

***N*^{6}-methyladenine (6mA) is a type of post-replication modification that exists across a wide range of DNA sequences. It has emerged as a key element in various biological processes, and due to this fact, has attracted great attention for its ability to be a target for disease treatment. Identifying 6mA sites has been the prelude to understand multiple biological functions including DNA replication, transcription, and repairing. Aside from a few numbers of experimental methods, a series of predictors were developed to distinguish 6mA sites in the whole genome. In this study, a 6mA site predictor named iDNA6mA-Pred was designed, which used a variety of valid feature descriptors to obtain informative characteristics. To improve computational efficiency and reduce the amount of redundant information, 34-dimensional features were selected with the aid of the *F*-score and were employed to learn support vector machine model. This study resulted an support vector machine (SVM)-based predictor to recognize 6mA sites of the rice genome, where nucleic acid composition (NAC), dinucleotide Composition (DNC), and Position-Specific Trinucleotide Propensity (PSTNP) were used to characterize the DNA sequences. Jackknife test results showed that the property of iDNA6mA-Pred was accurate with an accuracy rate of 92.16%. This predictor could be used for accurate identification of 6mA sites.**

Keywords: *N*^{6}-methyladenine; position-specific trinucleotide propensity; nucleic acid composition; Jackknife test; dinucleotide.

[*]**Corresponding authors:** Xueda Zhao, School of Information Engineering, Dalian Ocean University, Dalian 116023, Liaoning, China. Email: zxueda@dlou.edu.cn.

Introduction

In recent years, researchers identified DNA 6mA modification in the genomes of algae [1], drosophila [2], and nematode [3]. It showed that 6mA existed in the eukaryotic genome, which broke the traditional view that it was the specific epigenetic modification of prokaryotes and gave the study of 6mA a second life. Subsequently, more and more research teams joined in the study of 6mA, which showed that 6mA was not only spread in eukaryotic genomes, but also had a regulatory effect on gene expression. In 2018, 6mA modification maps of human cells were reported. 6mA is widely appeared in the human genome (including mitochondrial genome), and the abundance of 6mA in cancer tissues is significantly low. At the same time, N6AMT1 was identified as 6mA methylase [4]. In the same year, other researchers studied the function of 6mA in glioma in detail and determined that the content of 6mA in brain cancer stem cells was higher than that in normal nerve tissue, and targeting 6mA demethylase ALKBH1 was expected to be a new strategy for the treatment of this type of cancer [5]. Meanwhile, plant 6mA studies were also being carried out. The 6mA modification maps of rice and *Arabidopsis Thaliana* genomes were completed in 2018 [6-8].

As a methylation modification, DNA 6mA has potential effect on downstream biological functions and can determine expression levels of the genome. However, it is a very conservative DNA modification [2, 9–12]. This reversible modification allows for a more dynamic histone modification process, further affecting DNA structure, DNA transcription, and creating the ability for a limiting process of histone modification [13–15]. 6mA exists widely in prokaryotes, and its modification levels change dynamically over the course of an entire life cycle, influencing expression levels. Species diversity of plants and animals were closely related to the changes in 6mA, and DNA 6mA demethylase was found in drosophila species [2, 4, 9]. Recent work has reported that 6mA has been detected in rice, maize, and human cells by a series of experimental techniques [3, 4, 16-22] including single-molecule real-time sequencing [23], methylated DNA immunoprecipitation sequencing [24], and capillary electrophoresis, and laser-induced fluorescence [25]. Identifying 6mA methylation may bridge the gap for understanding the biological mechanisms. Apart from a group of biochemical experimental methods to identify presence of 6mA sites, a series of methods of calculation involving machine learning were developed to identify 6mA sites. In recent years, some researchers have developed several predictors to identify 6mA sites in biological sequences. The authors used support vector machine (SVM) to establish i6mA-Pred predictor with an accuracy of 83.13%, in which the characteristics were obtained by considering nucleotide frequency and chemical properties of nucleotides [16]. Tahir, *et al*. developed the iDNA6mA (5-step rule) predictor according to a deep learning approach and achieved an accuracy of 86.64% [17]. Chen, *et al*. also built the MethyRNA predictor and identified 6mA sites in *H. sapiens* and *M. musculus* with an accuracy of 90.38% and 88.39%, respectively [26, 27].

Although a few prediction models were developed for the classification and recognition of 6mA, there is still some rooms for prediction performances. To obtain higher prediction performances and examine the potential associated features from various aspects, this study developed a more comprehensive and balanced feature set by using three feature extraction methods including nucleic acid composition (NAC), di-nucleotide composition (DNC), and position-specific trinucleotide propensity (PSTNP). Further, the feature selection algorithm, Max-Relevance-Max-Distance (MRMD), was used to obtain the optimal feature set, which was used to train the SVM model and finally obtained a predictor with high performances. The results of this study will provide a potential method for identifying the site information of 6mA in other species of biology.

Materials and methods

Computational hardware and essential settings

All the experiments were conducted by using MATLAB 2020b (MathWorks, Natick, Massachusetts, USA) and LIBSVM package 3.22 (https://www.npackd.org/p/libsvm/3.22) on a personal computer with an intel i7 CPU, 16 GB memory, and 512 GB hard disk. Computational scheme with Chou's 5-step rule [28] provided some powerful and practical tools for identifying the recombination spots at high performance. Along with Chou's 5-step scheme, there were a series of publications [17, 27, 29–39] for analyzing biological sequences. Chou's 5-step scheme consisted of the following steps: (1) building a sound benchmark dataset for learning and testing predictors; (2) biological sequences were transformed into mathematical expressions to accurately express the intrinsic correlation of biological sequences; (3) developing an ideal predictor (or engine); (4) validating the identifying performance by using cross-validation tests, jackknife tests, or independent tests; (5) developing an accessible web-server to the public. The first four steps focused on designing a computing model, and the final step (setting up the webserver) was dedicated to achieving a simple and user friendly interface (Figure 1).

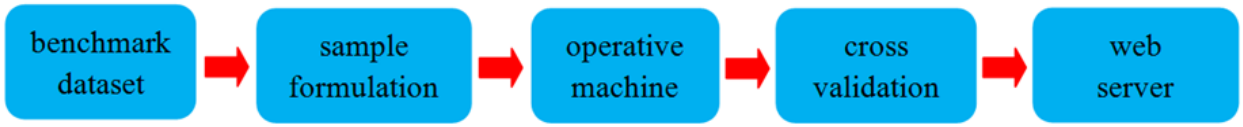

Figure 1. The Chou's five-step computational scheme.

Benchmark dataset

The experimental datasets for the 6mA editing sites in rice genome were collected from previous works conducted by Chen, *et al*. [16]. With the help of this dataset, the predicted value determined by the evaluation with other existing predicted values were compared. The adopted datasets met two sets of criteria including (1) the sites with a modified score less than 30 were filtered out; (2) the cutoff threshold was taken as 60%, which meant the sequences with values higher than the cutoff threshold were removed [17]. Notably, all sample lengths were 41 bp long. The benchmark dataset was described as follow:

$$S = S^{+} U S^{-} \quad (1)$$

where the positive subset S^{+} included 880 DNA sequences centered on 6mA sites, while the negative instances S^{-} contained 880 DNA sequences centered on non-6mA sites. The sign of U indicated the "union" of two sets [39, 40].

Sample sequence representation

Every sequence sample was denoted as follow:

$$S = N_1 N_2 N_3 \cdots N_L \quad (2)$$

where the length L = 41 [41, 42], and N_i denoted the nucleic acid at the *i*-th position, $N_i \in \{A, C, G, T\}, \quad (i = 1, 2, \ldots, L)$. Three feature extraction techniques including NAC, DNC, and PSTNP were adopted. These three methods could convert DNA sequences into numerical vectors for training classification algorithm.

(1) NAC:

As one of the most commonly used encoding methods, NAC has been used in various biological sequences [43-47]. Every sample was composed of four nucleotides, and the NAC described the frequency of all nucleic acid types in involved instances shown as follow:

$$f(t) = \frac{N(t)}{L}, \quad t \in \{A, C, G, T\} \quad (3)$$

where $N(t)$ represented the frequency and quantity of each nucleic acid type, respectively, and L represented the length of each sequence.

(2) DNC

DNC contains information about the occurrence of all nucleotide pairs [44, 47–51], which also has potential identification information for recognizing DNA N^6-methyladenine sites. Every sample consisted of four nucleotides, thus, the di-nucleotide composition had 16 descriptors. Dinucleotide composition could be denoted as:

$$D(r, s) = \frac{N_{rs}}{L - 1}, \quad r, s \in \{A, C, G, T\} \quad (4)$$

where $D(r, s)$ and N_{rs} represented the occurrence rate of different DNC types and the quantity of di-nucleotides in regard to nucleic acid types *r* and *s*. *L* was the length of each sequence. The DNC included 16 descriptors, and thus, 16-dimensional features were extracted by accounting for the occurrence rate information of DNC.

(3) PSTNP

PSTNP reflects the whole content of trinucleotides [47, 52] and the position information of each trinucleotide [53–55]. Position-specific trinucleotide propensity was used to characterize the differences in the position of trinucleotides in 6mA and non-6mA sequences. For a sequence with length *L* (*L* = 41),

a dimensional eigenvector was constructed as follows:

$$D = \left[\phi_1, \phi_2, \cdots, \phi_v, \cdots, \phi_{39}\right]^T \quad (5)$$

where $\emptyset_v$ was defined as:

$$\phi_v = \begin{cases} y_{1,v}, & \text{if } N_v N_{v+1} N_{v+2} = AAA \\ y_{2,v}, & \text{if } N_v N_{v+1} N_{v+2} = AAC \\ y_{3,v}, & \text{if } N_v N_{v+1} N_{v+2} = AAG \\ \vdots & \vdots \\ y_{64,v}, & \text{if } N_v N_{v+1} N_{v+2} = TTT \end{cases} \quad (1 \le v \le 39). \quad (6)$$

where $Y = \left(y_{i,j}\right)_{64*39}$ could be formation with equation (7).

$$y_{i,j} = F^+\left(\text{trinucleotide}_i \mid j\right) - F^-\left(\text{trinucleotide}_i \mid j\right) \quad (i = 1,2,\ldots,64;\ j = 1,2,\ldots 39) \quad (7)$$

$F^+\left(\text{trinucleotide}_i \mid j\right)$ and $F^-\left(\text{trinucleotide}_i \mid j\right)$ represented the occurrence rate of the *i*-th trinucleotide at the *j*-th position in positive and negative samples, respectively. The matrix *Y* could be expressed as follow:

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,39} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,39} \\ \vdots & \vdots & \vdots & \vdots \\ y_{64,1} & y_{64,2} & \cdots & y_{64,39} \end{bmatrix}. \quad (8)$$

In total, 59 dimensions of features including 4 dimensions for NAC, 16 dimensions for DNC, and 39 dimensions for PSTNP were obtained. Positive and negative samples were transformed into a numerical feature matrix with 880 × 59 dimensions, respectively. Numerical feature matrices were convenient for learning classifiers and also in designing computational models. The positive sample sequence representation was exhibited in Figure 2.

Feature selection

As a classical feature evaluation filtering method based on statistical measurement, *F*-score is widely used in various fields [56–58]. The features due to the statistical values derived from the method following the rule that the element with a high value had good discrimination ability were sorted in this study. The incremental feature selection (IFS) added features from high to low according to the *F*-score values. When a feature was added, a new feature set was formed and sorted according to the *F*-score values to find the optimal subset. Although 59 potential features were extracted from various methods, there might still exist some redundant features that would negatively impact the model learning [59, 60]. For the purpose to eliminate redundant features for enhancing the model's accuracy, a series of feature selection were conceived including new filter approaches, wrapper methods, and embedded strategies. In addition, filter methods, MRMD and *F*-score, were also employed [61–64].

Support vector machine and Performance evaluation

In this study, SVM was selected as the original classifier. Each sample was alternately selected as the test set and the rest as the training set in the process of test. By performing the above procedure, each sample played a role in the training and test sets. Further, jackknife test was employed in the study to evaluate the performance of designed model. Four measurements including sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthew's correlation coefficient (MCC) were included in this study [65, 66].

Results and discussion

To obtain higher performance of the computing model, potential feature information is extracted from all aspects, but at the cost of high-dimensional features and poor computing efficiency. Feature selection approaches are usually used to find optimize features subset to improve identification efficiency [61, 67–69]. The feature vectors can be evaluated by employing filter approaches, wrapper methods, and embedded approaches [70–73]. In the process of

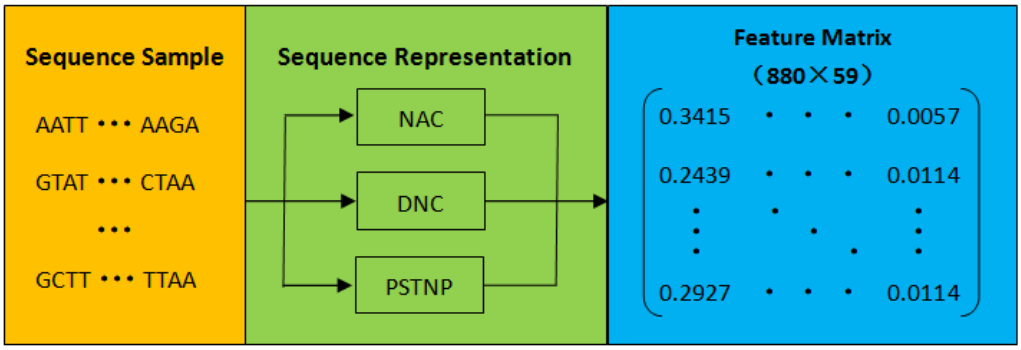

Figure 2. Positive sample sequence representation.

building the iDNA6mA-Pred model, feature extraction, feature selection, and classifier selection were the key components of target recognition. It is a common method to consider the synergetic effect of feature subset and classifier. In machine learning, SVM (also known as support vector network) is a supervised learning model and related learning algorithm for data classification and regression analysis [74]. In addition, SVM can effectively solve nonlinear classification problems by solving the linear classification problems in a multidimensional feature space derived from kernel methods. For classical SVM, the training data is equipped with classification labels, however, label instances are huge tasks. In industrial applications, clustering is often used as a preprocessing step for ranking when the instances are unlabeled, or some are labeled. In the evaluation of the computational model, three methods are usually used, namely jackknife test, cross-validation test, and the independence test, among which, the jackknife test is the most commonly used test method. It has known from the process of the jackknife test that the results are relatively objective, so the jackknife test has been usually used as the performance evaluation of the predictor.

Parameters of support vector machine

The radial basis function (RBF) serves as the nonlinear map (kernel function) of input data in the SVM model. For optimizing penalty parameter *C* (cost) and kernel parameter *g* (gamma), the grid search method with jackknife test was employed. The parameters *C* and *g* were in $[2^{-8}, 2^{8}]$ and $[2^{-8}, 2^{8}]$. The optimized parameters $C = 11.3137$ and $g = 32$ were assigned to predict the 6mA site in the rice genome.

Optimal feature analysis

The obtained 59 features were taken as the input vector for SVM for developing the predictor. The performance of the model was evaluated by the jackknife test. The Sn, Sp, Acc, and MCC of the prediction results were 0.8807, 0.8614, 0.8710, and 0.742, respectively. Although the performance was satisfied, there were still possibilities for improvement in reducing the number of redundant features and improving the performance of the predictor. Therefore, two additional methods were adopted to eliminate redundant features and select the best feature set. The first method was to use the MRMD and the IFS to find an ideal feature subset. A set of 57-dimension feature subset was obtained. The prediction result of SVM with the values of Sn, Sp, Acc, and MCC were 0.8920, 0.9193, 0.9057, and 0.8120, respectively, for the jackknife test. The second method was the combination of *F*-score ranking method with the IFS to eliminate redundant features and select for the best subset of features. The optimal feature set contained 34 feature dimensions for the input vector was used to build the predictor using the SVM model. The performance of the SVM model showed the values of Sn, Sp, Acc, and MCC as 0.9170, 0.9261, 0.9216, and 0.8430, respectively by applying jackknife test. The improvement in the performance of the predictors with the aid of non-redundant feature selection methods was

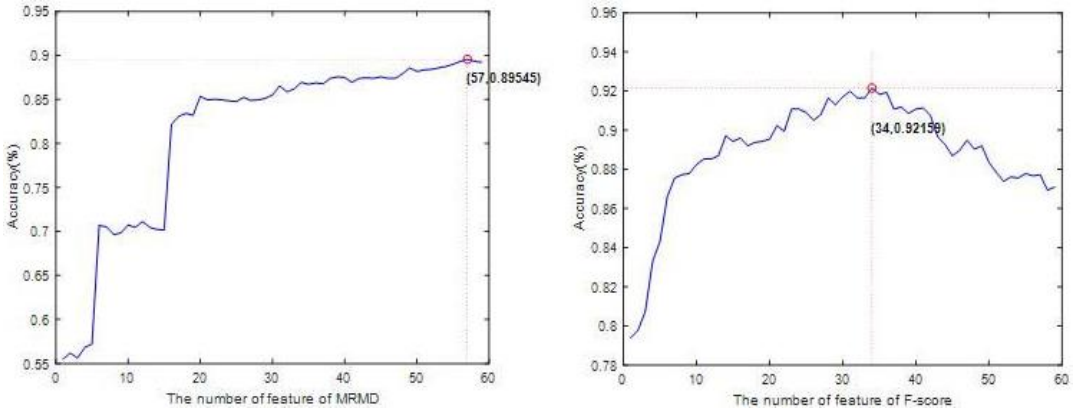

Figure 3. IFS curves of MRMD method and *F*-score ranking method.

observed. The results of predictor performance were listed in Table 1. The corresponding IFS curve was shown in Figure 3. By examining the classification performances, the *F*-score ranking method with the IFS was chosen to construct the predictor in this study, because it contributed to yield higher accuracy values than that in the MRMD method.

Table 1. Performance of predictor by jackknife test.

Features (dimension)	Sn (%)	Sp (%)	Acc (%)	MCC
All Feature (59)	88.07	86.14	87.10	0.742
MRMD (57)	89.20	91.93	90.57	0.812
F-score (34)	91.70	92.61	92.16	0.843

Comparison with other classifiers

Classifier selection is a very important step in designing the predictor. A range of classification learning algorithms such as K-nearest neighbor (KNN), logical regression, discriminatory analysis, and SVM have been well developed and successfully applied in bioinformatics [75]. In this study, 34-dimensional features obtained by *F*-score feature selection were obtained to train different classifiers for identifying 6mA sites across the rice genome. The prediction results of SVM were compared to the other classifiers. The relevant jackknife test data (based on the same sample) were shown in Table 2. In terms of Sn, Acc, and MCC, SVM achieved better results than that from the other classifiers. Therefore, SVM was chosen to build the computational model.

Table 2. Data of different classifiers.

Classifier	Sn (%)	Sp (%)	Acc (%)	MCC
SVM	91.70	92.61	92.16	0.842
KNN (K=5)	78.30	94.77	86.53	0.741
Logical Regression	89.55	91.36	90.45	0.809
Discriminatory Analysis	88.18	93.41	90.80	0.817

Table 3. Comparison of different methods.

Method	Sn (%)	Sp (%)	Acc (%)	MCC
i6mA-Pred	82.95	83.30	83.13	0.660
iDNA6mA	86.70	86.59	86.64	0.730
iDNA6mA-Pred	91.70	92.61	92.16	0.843

Comparisons with other methods

Many efforts have been undertaken to distinguish 6mA sites. To evaluate the performance, the designed iDNA6mA-Pred predictor from this study was compared to the previously developed i6mA-Pred [16] and iDNA6mA [17] predictors. The comparing data of Sn, Sp, Acc, and MCC were listed in Table 3. Comparing to the other two existing prediction systems, the newly designed method in this study was very accurate. Thus, this new method was effective and could be used as a powerful tool for predicting 6mA sites in the rice genome. In the experimental process, the *F*-score method was successfully used for feature selection. It was evident that this new model performed better

than the other existing models since the results indicated an elevation of 3.63% sensitivity, 6.47% specificity, 5.06% accuracy, and 10.10% MCC levels.

Conclusion

A new method of iDNA6mA-Pred was developed in this study to predict the rice genome 6mA sites. The method development process was as follows: (1) DNA sequence samples were converted into numerical vectors by three feature-extraction methods of NAC, DNC, and PSTNP; (2) 59 features were obtained to train newly developed method with the final jackknife test result of 87.10% in accuracy; (3) *F*-score and IFS strategies were applied in the study to select effective, non-redundant features from the initial set of 59 features, and to improve the accuracy of the model, which ended to 34 features to be used to train this newly developed method with the final jackknife test achieved an accuracy of 92.16%. By comparing this method to the different classifiers and other existing predictors, the results showed that the new method outperformed other existing methods. The results confirmed that this newly developed method could provide a powerful prediction of the 6mA sites in the rice genome. As the next step of this study, a web server with this new method will be built up for public access.

Acknowledgments

The authors acknowledge the General Research Project of Education Department of Liaoning Province, China (Grant: JL202014).

References

1. Fu Y, Luo GZ, Chen K, Deng X, Yu M, Han D, *et al*. 2015. N^6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. Cell. 161(4):879-892.
2. Zhang G, Huang H, Liu D, Cheng Y, Liu X, Zhang W, *et al*. 2015. N^6-methyladenine DNA modification in *Drosophila*. Cell. 161(4):893-906.
3. Greer EL, Blanco MA, Gu L, Sendinc E, Yang S. 2015. DNA methylation on N^6-adenine in *C. elegans*. Cell. 161(4):868-878.
4. Xiao CL, Zhu S, He M, Chen D, Yan GR. 2018. N^6-methyladenine DNA modification in the human genome. Mol Cell. 71(2):306-318, e307.
5. Xie Q, Wu TP, Gimple RC, Li Z, Prager BC, Wu QL, *et al*. 2018. N^6-methyladenine DNA modification in *Glioblastoma*. Cell. 175(5):1228-1243. e1220.
6. Zhou C, Wang CS, Liu HB, Zhou QW, Liu Q, Guo Y, *et al*. 2018. Identification and analysis of adenine N^6-methylation sites in the rice genome. Nat Plants. 4(8):554-563.
7. Zhang Q, Liang Z, Cui X, Ji C, Li Y, Zhang P, *et al*. 2018. N^6-methyladenine DNA methylation in Japonica and Indica rice genomes and its association with gene expression, plant development, and stress responses. Mol Plant. 11(12):1492-1508.
8. Liang Z, Shen L, Cui X, Bao S, Yu H. 2018. DNA N^6-adenine methylation in *Arabidopsis thaliana*. Dev Cell. 45(3):406-416, e403.
9. Luo GZ, Blanco MA, Greer EL. 2015. DNA N^6-methyladenine: a new epigenetic mark in eukaryotes. Nat Rev Mol Cell Biol. 16(12):705.
10. Wang Y, Sheng Y, Liu Y. 2017. N^6-methyladenine DNA modification in the unicellular eukaryotic organism *Tetrahymena thermophila*. Euro J Protistol. 58:94-102.
11. Gommers-Ampt JH, Borst P. 1995. Hypermodified bases in DNA. FASEB J. 9(11):1034-1042.
12. Korlach J, Turner SW. 2012. Going beyond five bases in DNA sequencing. Curr opin stru biol. 22(3):251-261.
13. Yao B, Li Y, Wang Z. 2018. Active N^6-methyladenine demethylation by DMAD regulates gene expression by coordinating with polycomb protein in neurons. Mol Cell. 71(5):848-857.
14. Shah K, Cao W, Ellison CE. 2019. Adenine methylation in *Drosophila* is associated with the tissue-specific expression of developmental and regulatory genes. G3: Genes, Genomes, Genetics. 9(6):1893-1900.
15. Summerer D. 2015. N^6-methyladenine: A potential epigenetic mark in eukaryotic genomes. Angewandte Chemie International Edition. 54(37):10714-10716.
16. Chen W, Lv H, Nie F. 2019. i6mA-Pred: Identifying DNA N N^6-methyladenine sites in the rice genome. Bioinformatics. 35(16):2796–2800.
17. Tahir M, Tayara H, Chong KT. 2019. iDNA6mA (5-step rule): identification of DNA N^6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. Chemometr Intell Lab Syst. 189:96-101.
18. Yao B, Cheng Y, Wang Z. 2017. DNA N^6-methyladenine is dynamically regulated in the mouse brain following environmental stress. Nat Commun. 8(1):1122.
19. Herna´ndez-Ledesma B, Da´valos A, Bartolome´ B. 2005. Preparation of antioxidant enzymatic hydrolysates from α-lactalbumin and β-lactoglobulin identification of active peptides by HPLC-MS/MS. J Agri Food Chem. 53(3):588-593.
20. Matuszewski BK, Constanzer ML, Chavez-Eng CM. 2003. Strategies for the assessment of matrix effect in quantitative

bioanalytical methods based on HPLC-MS/MS. Anal Chem. 75(13):3019-3030.
21. Churchwell MI, Twaddle NC, Meeker LR. 2005. Improving LC-MS sensitivity through increases in chromatographic performance: Comparisons of UPLC-ES/MS/MS to HPLC-ES/MS/MS. J Chromat B. 825(2):134-143.
22. Nayebare SR, Karthikraj R, Kannan K. 2018. Analysis of terephthalate metabolites in human urine by high-performance liquid chromatography-tandem mass spectrometry (HPLC-MS/MS). J Chromat B. 1092:473-479.
23. Flusberg BA, Webster DR, Lee JH. 2010. Direct detection of DNA methylation during single- molecule, real-time sequencing. Nat meth. 7(6):461-465.
24. Pomraning KR, Smith KM, Freitag M. 2009. Genome-wide high throughput analysis of DNA methylation in eukaryotes. Methods. 47(3):142-150.
25. Krais AM, Cornelius MG, Schmeiser HH. 2010. Genomic N6-methyladenine determination by MEKC with LIF. Electrophoresis. 31(21):3548-3551.
26. Chen W, Tang H, Lin H. 2017. MethyRNA: a web server for identification of N6 methyladenosine sites. J Biomol Stru Dyn. 35(3):683-687.
27. Chen W, Feng P, Yang H. 2018. iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. Mol Ther Nucleic Acids. 11:468-474.
28. Chou KC. 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J theo biol. 273(1):236-247.
29. Cheng X, Lin WZ, Xiao X. 2018. pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. Bioinform. 35(3):398-406.
30. Liu B, Li K, Huang DS. 2018. iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinform. 34(22):3835-3842.
31. Liu B, Weng F, Huang DS. 2018. iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC. Bioinform. 34(18):3086-3093.
32. Su ZD, Huang Y, Zhang ZY. 2018. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinform. 34(24):4196-4204.
33. Cheng X, Xiao X, Chou KC. 2017. pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. Mol BioSys. 13(9):1722-1727.
34. Feng P, Ding H, Yang H. 2017. iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Mol Ther Nucleic Acids. 7:155-163.
35. Jia C, Zuo Y. 2017. S-SulfPred: A sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling- synthetic minority oversampling technique. J theo biol. 422:84-89.
36. Jia J, Liu Z, Xiao X. 2015. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J theo biol. 377:47-56.
37. Jia J, Liu Z, Xiao X. 2016. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J theo biol. 394:223-230.
38. Song J, Wang Y, Li F. 2018. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief Bioinform. 20(2):638-658.
39. Hussain W, Khan YD, Rasool N. 2019. SPalmitoylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. Anal Biochem. 568:14-23.
40. Li T, Song R, Yin Q. 2019. Identification of S-nitrosylation sites based on multiple features combination. Sci Repo. 9(1):3098.
41. Chen W, Feng P, Yang H. 2017. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget. 8(3):4208.
42. Chou KC. 1995. A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. Protein Sci. 4(7):1365-1383.
43. 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. 43(W1): W65-W71.
44. Liu B, Wu H, Chou KC. 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. Nat Sci. 9(04):67.
45. Chen W, Feng P, Ding H, Lin H, Chou KC. 2015. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal Biochem. 490:26-33.
46. Sabooh MF, Iqbal N, Khan M, Khan M, Maqbool HF. 2018. Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. J theo biol. 452:1-9.
47. Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, *et al*. 2020. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief bioinform. 21(3):1047-1057.
48. Chen W, Feng PM, Lin H, Chou KC. 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 41(6):e68-e68.
49. Chen W, Feng PM, Lin H, Chou KC. 2014. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. BioMed Res Inte. 2014:623149.
50. Iqbal M, Hayat M. 2016. "iSS-Hyb-mRMR": Identification of splicing sites using hybrid space of pseudo trinucleotide and pseudo tetranucleotide composition. Comput Methods Programs Biomed. 128:1-11.
51. Kabir M, Hayat M. 2016. iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. Mol Genet Genomics. 291(1):285-296.
52. Chen W, Feng PM, Deng EZ, Lin H, Chou, KC. 2014. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem. 462:76-83.

53. He W, Jia C. 2017. EnhancerPred2.0: predicting enhancers and their strength based on position-specific trinucleotide propensity and electron–ion interaction potential feature selection. Mol BioSys. 13(4):767-774.
54. He W, Jia C, Zou Q. 2018. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinform. 35(4):593-601.
55. Jia C, Yang Q, Zou Q. 2018. NucPosPred: Predicting species-specific genomic nucleosome positioning via four different modes of general PseKNC. J theo biol. 450:15-21.
56. He W, Jia C, Duan Y, Zou Q. 2018. 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. BMC Sys Biol. 12:99-107.
57. Kumar N, Chandra MSS, Mohapatro S. 2021. F-score feature selection-based support vector regression for solar power forecasting. in proceedings of symposium on power electronic and renewable energy systems control: PERESC 2020. Springer Singapore. 2021:249-259.
58. Huang W, Yan H, Liu R, Zhu L, Zhang H, Chen H. 2018. F-score feature selection based Bayesian reconstruction of visual image from human brain activity. Neurocomput. 316:202–209.
59. Hu L, Gao W, Zhao K. 2018. Feature selection considering two types of feature relevancy and feature interdependency. Expert Syst Appl. 93:423-434.
60. Senawi A, Wei HL, Billings SA. 2017. A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. Pat Recog. 67:47-61.
61. Zou Q, Zeng J, Cao L, Ji R. 2017. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomput. 173:346–354.
62. Jia C, He W. 2016. EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features. Sci reports. 6(1):38741.
63. Chen W, Ding H, Feng P. 2016. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 7(13):16895.
64. Chen W, Ding H, Lin H. 2018. Classifying included and excluded exons in exon skipping event using histone modifications. Front Genetics. 9:433.
65. Wei L, Su R, Wang B, Li X, Zou Q, Gao X. 2019. Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites. Neurocomput. 324:3-9.
66. Xu Y, Ding J, Wu LY. 2013. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PloS One. 8(2):e55844.
67. Peng H, Long F, Ding C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 27(8):1226-1238.
68. Wei L, Xing P, Shi G, Ji ZL, Zou Q. 2017. Fast prediction of protein methylation sites using a sequence-based feature selection technique. IEEE/ACM Trans Comput Biol Bioinform. 16(4):1264-1273.
69. Zou Q, Wan S, Ju Y, Tang J, Zeng X. 2016. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. BMC Sys Biol. 10(4):401-412.
70. Kumar V, Minz S. 2014. Feature selection: a literature review. SmartCR. 4(3):211-229.
71. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. 2013. A review of feature selection methods on synthetic data. Knowl Infor Sys. 34:483–519.
72. Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Part of the book series: The Springer International Series in Engineering and Computer Science. 1998. Vol. 454.
73. Guo W, Liu X, Ma Y, Zhang R. 2021. iRspotDCC: Recombination hot/cold spots identification based on dinucleotidebased correlation coefficient and convolutional neural network. J Intell Fuzzy Syst. 41(1):1309-1317.
74. Chang CC, Lin CJ. 2011. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2(3):1-27.
75. Sun Y, Wong AK, Kamel MS. 2009. Classification of imbalanced data: A review. Int J Pat Recog Artif Intel. 23(4):687-719.