Molecular Therapy
Nucleic Acids
Original Article

Prediction of m5C Modifications in RNA Sequences by Combining Multiple Sequence Features

Lijun Dou,[1,2,6] Xiaoling Li,[3,6] Hui Ding,[4] Lei Xu,[5] and Huaikun Xiang[1]

[1]School of Automotive and Transportation Engineering, Shenzhen Polytechnic, Shenzhen, China; [2]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China; [3]Department of Oncology, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China; [4]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China; [5]School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China

5-Methylcytosine (m5C) is a well-known post-transcriptional modification that plays significant roles in biological processes, such as RNA metabolism, tRNA recognition, and stress responses. Traditional high-throughput techniques on identification of m5C sites are usually time consuming and expensive. In addition, the number of RNA sequences shows explosive growth in the post-genomic era. Thus, machine-learning-based methods are urgently requested to quickly predict RNA m5C modifications with high accuracy. Here, we propose a noval support-vector-machine (SVM)-based tool, called iRNA-m5C_SVM, by combining multiple sequence features to identify m5C sites in *Arabidopsis thaliana*. Eight kinds of popular feature-extraction methods were first investigated systematically. Then, four well-performing features were incorporated to construct a comprehensive model, including position-specific propensity (PSP) (PSNP, PSDP, and PSTP, associated with frequencies of nucleotides, dinucleotides, and trinucleotides, respectively), nucleotide composition (nucleic acid, di-nucleotide, and tri-nucleotide compositions; NAC, DNC, and TNC, respectively), electron-ion interaction pseudopotentials of trinucleotide (PseEIIPs), and general parallel correlation pseudo-dinucleotide composition (PC-PseDNC-general). Evaluated accuracies over 10-fold cross-validation and independent tests achieved 73.06% and 80.15%, respectively, which showed the best predictive performances in *A. thaliana* among existing models. It is believed that the proposed model in this work can be a promising alternative for further research on m5C modification sites in plant.

INTRODUCTION

To date, more than 150 types of RNA post-transcriptional modifications have been found in all kingdoms of life.[1–7] As one of most prevalent modifications, 5-methylcytosine (m5C) is catalyzed by RNA methyltransferase, in which a methyl group is attached to the fifth position of the cytosine ring. It has been reported that m5C sites are involved in many kinds of biological processes, including RNA structural stability and metabolism, tRNA recognition and stress responses,[8–14] and so forth. Additionally, it has also been proved that m5c modifications are associated with many diseases, such as breast cancer,[15] autosomal recessive intellectual disability,[16] amyotrophic lateral sclerosis,[17] and Parkinson's disease.[18] Thus, the accurate identification of m5C is the primary and crucial task for carrying out the research on corresponding diseases and biological functions.[8,9,11–13,15–21] In experiments, several traditional high-throughput sequencing techniques, such as bisulfite conversion,[22] miCLIP,[23] and Aza-IP,[24] have been developed to detect m5C sites. More details about m5C biological mechanisms and related diseases can be found in Chen et al.[25] and literature therein. However, considering the time-consuming and labor-intensive nature of these techniques, it is challenging to keep pace with the dramatic increase of the number of RNA sequences in the post-genome era. Therefore, the identification of m5C and non-m5C sequences using computational methods is of great significance and necessity.

Eight computational predictors have been proposed to detect m5C sites in RNA sequences, including m5C-PseDNC,[26] iRNAm5C-PseDNC,[27] M5C-HPCR,[28] pM^{5}CS-Comp-mRMR,[29] RNAm5C-finder,[30] PEA-m5C,[31] iRNA-m5C,[32] and RNAm5CPred.[33] Related species, feature-extraction techniques, and classifiers are listed in Table 1. It can be seen that there were a total of four species investigated: *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana*. In specific, Feng et al.[26] first provided the m5C-PseDNC tool based on the support vector machine (SVM) in *H. sapiens*. By applying pseudo-dinucleotide composition (PseDNC) features with three physiochemical properties, the accuracy over the jackknife test achieved 90.42%. Qiu et al.[27] also used PseDNC features with 10 properties to construct the random forest (RF) model called

Received 18 February 2020; accepted 4 June 2020;
https://doi.org/10.1016/j.omtn.2020.06.004.

[6]These authors contributed equally to this work.

Correspondence: Lei Xu, School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China.
E-mail: csleixu@szpt.edu.cn
Correspondence: Huaikun Xiang, School of Automotive and Transportation Engineering, Shenzhen Polytechnic, Shenzhen, China.
E-mail: xianghuaikun@szpt.edu.cn

Table 1. Eight Proposed Methods to Identify m5C Sites in RNA Sequences

Method	Species	Feature Extraction/Selection	Classifiers
m5C-PseDNC[26]	*H. sapiens*	PseDNC (3 properties)	SVM
iRNAm5C-PseDNC[27]	*H. sapiens*	PseDNC (10 properties)	RF
M5C-HPCR[28]	*H. sapiens*	HPCR	SVM
pM^{5}CS-Comp-mRMR[29]	*H. sapiens*	Kmer (k = 2, 3, and 4) /mRMR	SVM
RNAm5Cfinder[30]	*H. sapiens*, *M. musculus*	BE	RF
PEA-m5C[31]	*A. thaliana*	BE + Kmer + PseDNC	RF
iRNA-m5C[32]	*H. sapiens*, *S. cerevisiae*, *M. musculus*, *A. thaliana*	Kmer + BE + NV + PseKNC	RF
RNAm5CPred[33]	*H. sapiens*	Kmer + KSNPF + PseDNC	SVM

iRNAm5C-PseDNC, where the jackknife test gave an accuracy of 92.37%. Later, Zhang et al.[28] introduced the m5c-HPCR model, with a higher Matthew's correlation coefficient (MCC) of 0.859 and area under the receiver operating characteristic (ROC) curve (AUC) of 0.962, where a novel heuristic nucleotide physicochemical property reduction (HPCR) algorithm was applied. Then, Sabooh et al.[29] presented the pM^{5}CS-Comp-mRMR method, with an accuracy of 93.33%, where the minimum redundancy and maximum relevance (mRMR) method was used to select effective features from Kmer features with *k*s = 2, 3, and 4 (corresponding to di-nucleotide composition, tri-nucleotide composition, and tetra-nucleotide composition; DNC, TNC, and TetraNC, respectively). For the m5C sites in *A. thaliana*, Song et al.[31] first developed the predictor PEA-M5C, where an independent test showed an overall accuracy of 83.5% with the MCC of 0.688. In this method, three kinds of feature-encoding techniques—binary encoding (BE), Kmer, and PseDNC—were incorporated to give combined performances. Li et al.[30] designed the RNAm5Cfinder using BE features to analyze m5C sites in *H. sapiens* and *M. musculus*, where comprehensive and cell-specific predictors gave AUC values of 0.77 and 0.87, respectively. Recently, Lv et al.[32] established a novel approach, iRNA-m5C, to systematically diagnose m5C sites in four species, where Kmer, BE, pseudo-k-tuple nucleotide composition (PseKNC), and natural vector (NV) were incorporated to obtain overall results. Optimal models of four species gave evaluated accuracies of 92.90%, 100.00%, 100.00%, and 70.70% on training datasets and 74.00% on testing datasets in *A. thaliana*. Also recently, Fang et al.[33] constructed an accurate RNAm5CPred tool in *H. sapiens*, where Kmer (described as K-nucleotide frequencies [KNFs] in their paper), K-spaced nucleotide pair frequencies (KSNPFs), and PseDNC were combined to represent RNA samples.

Generally, except for the PEA-M5C[31] model, which was focused on *A. thaliana*, seven other tools[26–30,32,33] all gave better performances in *H. sapiens*, where the average accuracy was higher than 90%. As for *S. cerevisiae* and *M. musculus*, it was noted that only 97 and 211 positive samples were experimentally validated, where the remaining sequences, by removing sequence similarity, were too few to construct computational predictors (i.e., lacking of statistical significance; details can be found in Sun et al.[5] and Lv et al.[32]). In addition, reported accuracies using the original data were adequately equal to 100.00%. It is hoped that more ideal/reliable models will be built in the future, with more experiment-proven sequences. As for the only plant, *A. thaliana*, there were only two predictors developed: PEA-m5C[31] and iRNA-m5C.[32] Especially, the latest iRNAm5C method presented accuracies of 70.7% and 74% over 10-fold cross-validation (CV) and independent tests using combined features "KNFC + MNBE + NV," respectively. On the other hand, only a few feature-extraction techniques have been used in two published methods. Therefore, there is still a big hope for improving predictive performances by applying other new feature-encoding techniques. In summary, we were mainly focused on improving the performances of the identification of m5C sites in *A. thaliana* in this article (Table 1).

We first investigated eight kinds of sequence-representing methods; namely, position-specific propensity (PSP), Kmer, enhanced nucleic acid composition (ENAC), xxKGap, electron-ion interaction pseudopotentials (EIIPs) and EIIPs of trinucleotides (PseEIIPs), general parallel correlation PseDNC (PC-PseDNC-general), nucleotide chemical property and nucleotide density (NCP + ND), and BE. Then, four well-performing features, "PSP + Kmer + PseEIIP + PseDNC," chosen by preliminary results, were incorporated to build the prediction model. Four different classifiers (SVM, RF, AdaBoost, and Naive Bayes [NB]) were separately applied for comparison, where the best performing model was optimized using the SVM method. The schematic flowchart of this work is shown in Figure 1.

RESULTS AND DISCUSSION

Predictive Performances Using One Kind of Feature

First, we plotted enriched and depleted nucleotides of the training datasets in Figure 2, which directly reflected the differences of position-specific nucleotide frequencies between positive and negative samples by $Z_{i,j} = Z_{i,j}^{+} - Z_{i,j}^{-}$ (i.e., the position-specific nucleotide propensity (PSNP) matrix described in Materials and Methods). Obvious differences can be observed between m5C and non-m5C sequences as well as upstream and downstream regions. Generally, the C and U bases are almost enriched in positive samples, whereas the A and G bases are almost enriched in negative sequences. However, nucleotides near the center (C, labeled as 0) show a completely different distribution, where C and U are more likely located in negative samples at

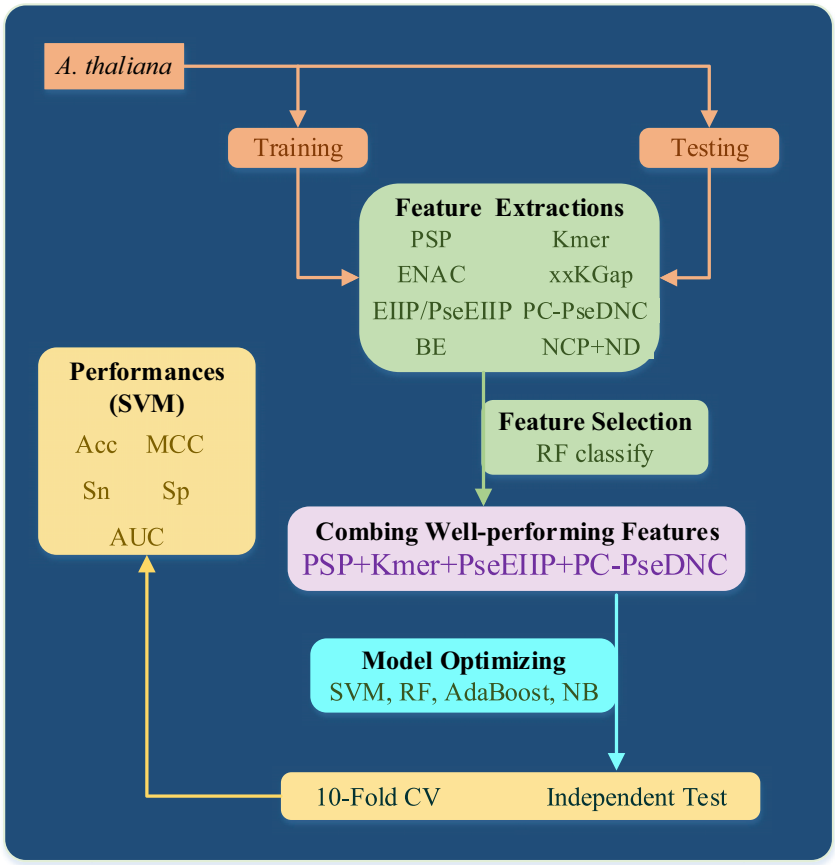

Figure 1. The Flowchart of the Proposed Predictor for m5C Identification by Combining Multiple Sequence Features

positions 1, 2, and 4 and −6, −3, −2 and −1, respectively. At the same time, A and G refer to distribution in positive samples at positions 1, 2, 4, and 10. On the other hand, occupied distinction downstream is obviously weaker than upstream. Specifically, C is, on average, 5% enriched in positive samples, and A is enriched 3% in negative samples upstream. However, the average difference of enriched and depleted nucleotides is approximately 1.4% downstream. It can be generally concluded that the characteristics of nucleotide location between m5C and non-m5C instances can be obviously found; i.e., m5C sites could be identified using the sequence information. Furthermore, the position-specific property is hoped to be an effective feature-extraction method to directly represent RNA sequence.

Many kinds of feature-extraction approaches have been developed to effectively encode RNA sequences, which can be conveniently obtained using several state-of-the-art toolkits, such as Pse-in-One2.0,[34] BioSeq-Analysis2.0,[35] iLearn,[36] PyFeat,[37] and so forth. Here, four kinds of feature-representing techniques associated with nucleotide frequencies were first investigated, including PSP, Kmer, ENAC, and xxKGap. Corresponding experimental results using the RF classifier are listed in Table 2, where 10-fold CV, and independent tests were used for training (left) and testing datasets (right), respectively. For three kinds of PSP features (i.e., PSNP, PSDP, and PSTP, associated with frequencies of nucleotides, dinucleotides, and trinucleotides, respectively), performances were gradually increased. It can be seen that accuracies over 10-fold CV and independent tests were only 65.48% and 65.05% for PSNP features; however, accuracies of 67.29% and 74.98%, respectively, were quickly achieved for PSTP. Compared to the latest tool, iRNAm5C,[32] the accuracy over the independent test using only 39-dimensional PSTP features has achieved 74.00%, although it was 3.41% lower over 10-fold CV. Thus, the distribution of trinucleotides is exactly an effective description to represent m5C sequences. As for three Kmer features (i.e., nucleic acid composition [NAC], DNC, and TNC, associated with ks = 1, 2, and 3, respectively), predictive accuracies increased with k, where TNC features showed better accuracies of 69.26% and 72.55% for training and testing datasets. As a variation of the NAC technique, ENAC also showed good performances, with accuracies of 69.11% and 71.9% on two datasets. Additionally, xxKGap results were also listed with different conditions, including monoMonoKGap (mMKGap), monoDiKGap (mDGap), and diMonoKGap (dMGap), with ks = 1, 2, and 3, corresponding to dinucleotide and trinucleotide frequencies within kgaps. It can be observed that there were not obvious improvements for those listed nine features with k increasing, and mM2Gap showed relatively best performances with 10-fold and independent accuracies of 68.80% and 77.20%.

Additionally, other five kinds of feature vectors, including EIIP, PseEIIP, PC-PseDNC-general (λ=3, ω=0.2), BE, and NCP + ND were also applied for model constructing; the evaluated results are listed in Table 3. It can be found that PseEIIP and PseDNC features performed well among those five approaches, where corresponding training accuracies achieved 69.24% and 68.63% with testing accuracies of 72.60% and 72.65%, respectively. It was also noted that predictive performances of BE were actually unsatisfied, where training accuracy is only 66.55%. For the PC-PseDNC method implemented in Pse-in-One 2.0,[34] two important parameters, λ and w, were optimized using the grid search ($1 \leq \lambda \leq 10$ with $\Delta\lambda = 1$; $0.1 \leq w \leq 1$ with $\Delta w = 0.1$). Combining predictive accuracies and number of features, PC-PseDNC-general (3,0.2) (i.e., $\lambda = 3$, $w = 0.2$; abbreviated as PC-PseDNC hereinafter) was finally chosen.

In general, evaluated accuracies were approximately 68%–69% (10-fold CV) and 72%–73% (independent test) for several well-performing features, including PSTP (independent test: 74.98%), DNC, TNC, ENAC, xxKGAP (mM2Gap: independent test, 77.20%), PseEIIP, and

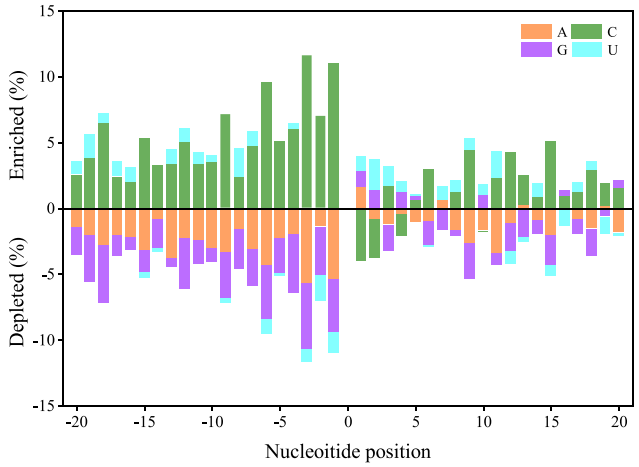

Figure 2. Differences of Position-Specific Nucleotide Frequencies between Positive and Negative Samples by $Z_{ij} = Z_{ij}^{+} - Z_{ij}^{-}$

Enriched nucleotides correspond to the condition $Z_{ij} > 0$ while depleted to $Z_{ij} < 0$.

PseDNC. It is known that PSP features reflect characteristics of statistical frequencies for positive and negative samples. Thus, the PSP-based model cannot convince researchers if the number of training instances does not reach a certain level. Additionally, compared with the reported tools, evaluated accuracies were not exactly satisfactory. At the same time, a single kind of feature can only indicate one aspect of sequence information. Therefore, we further incorporated multiple kinds of sequence-encoding methods to obtain comprehensive predictors, which can well reflect sequence information of nucleotide frequencies, physiochemical properties, electron-ion interaction, and so forth.

Predictive Performances Using Combined Features

Based on the discussion earlier, comprehensive predictive performances of multiple features proceeded further and are summarized in Table 4, where the second column "Fea_num" indicates the number of combined features. For the integration of three PSPs "PSNP + PSDP + PSTP," predictive accuracies were 67.39% and 73.30% over 10-fold CV and independent tests, respectively. Also, 84-dimensional Kmer features "NAC + DNA + TNC" displayed better results (for the 10-fold CV test: accuracy, 69.13%; MCC = 0.39; for the independent test: accuracy, 73.85%, MCC = 0.48). When the two features were integrated as "PSP + Kmer," training and testing accuracies were rapidly increased to 71.47% and 77.60%, respectively. Besides, when we incorporated all four kinds of frequency-associated features as "PSP + Kmer + ENAC + mM2Gap," better training and testing accuracies of 71.72% and 78.15%, respectively, were obtained. As for the combination of "PseEIIP + PC-PseDNC," no better results were obtained. It is also noted that the feature combination of four kinds of feature-extraction methods, "PSP + Kmer + PseEIIP + PC-PseDNC," showed the best performances (in total, 287 features), where overall accuracies reached 71.77% and 78.30% over 10-fold CV and independent tests, respectively. In addition, ENAC features were also combined with the 287 features mentioned earlier, written as "PSP + Kmer + PseEIIP + PC-PseDNC + ENAC," where the accuracy of training datasets was only improved 0.59% but −1.55% for testing

Table 2. Evaluated Performances of Frequency-Associated Feature-Extraction Techniques Using the RF Classifier, Where 10-fold CV, Left, and Independent Tests, Right, Were Separately Used for Training and Testing Datasets

	Training Datasets				Testing Datasets			
Feature Subset	Acc (%)	MCC	Sn (%)	Sp (%)	Acc (%)	MCC	Sn (%)	Sp (%)
PSNP	65.48	0.31	57.78	73.19	65.05	0.32	49.60	80.50
PSDP	65.07	0.31	56.78	73.36	67.52	0.36	57.54	77.50
PSTP	67.29	0.35	61.30	73.28	74.98	0.51	65.87	84.10
NAC	64.96	0.30	61.32	68.60	68.75	0.38	69.70	67.80
DNC	68.74	0.38	64.17	73.30	72.60	0.45	70.40	74.80
TNC	69.26	0.39	61.92	76.59	72.55	0.45	68.90	76.20
ENAC	69.11	0.38	64.53	73.68	71.90	0.44	71.90	71.90
mM1GAP	68.11	0.36	62.94	73.28	71.45	0.43	69.50	73.40
mM2GAP	68.80	0.38	63.32	74.29	77.20	0.55	80.60	73.80
mM3GAP	69.09	0.38	63.75	74.42	73.50	0.47	71.40	75.60
mD1GAP	67.57	0.36	60.33	74.82	72.15	0.44	68.80	75.50
mD2GAP	68.33	0.37	60.92	75.74	72.10	0.44	68.00	76.20
mD3GAP	68.38	0.37	60.41	76.35	72.70	0.46	68.60	76.80
dM1GAP	68.05	0.37	60.52	75.57	72.95	0.46	69.00	76.90
dM2GAP	68.39	0.37	60.37	76.40	72.10	0.44	68.10	76.10
dM3GAP	68.43	0.37	60.35	76.52	73.10	0.46	68.10	78.10

Acc, accuracy.

Table 3. Same as Table 2 but for Other Five Feature-Representing Methods

	Training Datasets				Testing Datasets			
Feature Subset	Acc (%)	MCC	Sn (%)	Sp (%)	Acc (%)	MCC	Sn (%)	Sp (%)
EIIP	66.65	0.34	59.27	74.02	70.85	0.42	68.40	73.30
PseEIIP	69.24	0.39	62.03	76.44	72.60	0.45	68.80	76.40
PC-PseDNC	68.63	0.37	63.47	73.79	72.65	0.45	70.00	75.30
BE	64.37	0.29	57.48	71.26	66.55	0.33	63.60	69.50
NCP + ND	66.67	0.34	60.92	72.41	70.25	0.41	69.30	71.20

Acc, accuracy.

datasets. If we considered all kinds of features listed in Tables 2 and 3 (for xxKGap, only mM2Gap was included), there were 1,571 features in total, with evaluated accuracies of 71.93% and 75.71% for training and testing datasets, respectively.

Considering the number of features and corresponding performances, the integration of four types of features, "PSP + Kmer + PseEIIP + PC-PseDNC," was finally used to optimize prediction model. Here, four different classifiers, including RF, SVM, AdaBoost, and NB implemented in the scikit-learn package (*sklearn*),[38] were separately applied to construct predictive models; the results are given in Table 5. It was found that three algorithms—RF, SVM, and AdaBoost—all showed better results, where average accuracies were up to 71.89% and 79.55% for the training and testing datasets. Here, default parameters were used in preliminary experiments, where n_esti = 100 was set as the number of decision trees in the RF method, and C = 1 and gamma = "scale," (i.e., gamma = 1/(num_fea · X.var()) were chosen in the SVM method. Among the four listed methods, the SVM classifier gave the overall best performance (10-fold CV: accuracy = 72.72%, MCC = 0.46; independent test: accuracy = 79.90%, MCC = 0.60), where the related AUC values achieved were 0.70 and 0.88, respectively.

Parameter Optimization and Comparison with Published Predictors

Parameter optimization is also a critical process for improving the performances of constructed models. Here, two important parameters of the SVM method, C and gamma, were simply selected using the dimension-reduction method.[38] The best performing model was finally obtained with C = 1.5 and default gamma, corresponding to predictive performances (for training datasets: accuracy = 73.06, MCC = 0.47, and AUC = 0.80; for testing datasets: accuracy = 80.15%, MCC = 0.60, and AUC = 0.88).

Table 6 gave a comparison of our introduced tool iRNA-m5C_SVM and the only two existing predictors, PEA-m5C[31] and iRNA-m5C,[32] in *A. thaliana*. For a fair comparison, the same independent datasets in this article were used to obtain performances of the PEA-m5C tool (see details in Lv et al.[32]). It can be seen that only 44.30% accuracy was obtained for the PEA-m5C model.[31] Compared with the latest iRNA-m5C method,[32] accuracies were improved from initially 70.70% to finally 73.06% and from 74.0% to 80.15% for training and testing datasets, respectively. Although predictive performances of 10-fold CV only improved 2.36%, the accuracy of the independent test was improved 6.15%. It has been mentioned earlier that the feature combination "KNFC + MNBE + NV" showed the best performance in the iRNA-m5C[32] predictor. However, besides the basic Kmer technique, the sequence information on PSP, electron-ion interaction potential, and physicochemical properties was considered in this method. At the same time, we also optimized the parameters of the SVM classifier to obtain the best results. Figure 3 visually demonstrated ROC curves of this method (left) and comparison between the latest iRNA-m5C tool[32] and our method (right). The AUC values for training and testing datasets achieved were 0.80 and 0.88, respectively, where the iRNA-m5C tool[32] reported AUC values of 0.77 over 10-fold CV. It is believed that our methods can obtain higher accuracies for m5C identification than two existing tools in *A. thaliana*. It is hoped that new benchmark datasets will be collected further with larger amounts of experiment-proved m5C sequences. Then, a more accurate machine-learning-based predictor can be established to predict m5C sites. On the other hand, although, in total, seven kinds of features have been investigated, there are still other powerful feature-extraction techniques worth exploring. Efficient machine learning classifiers and even deep learning methods also should be considered to improve performances.

Conclusions

As an important post-transcriptional modification, m5C plays crucial roles in the biological process. In this work, multiple sequence features were combined to construct a comprehensive SVM-based model to predict RNA m5C sites in *A. thaliana*. Specifically, four better performing feature-extraction techniques were incorporated, including PSP (PSNP, PSDP, and PSTP), nucleotide composition (NAC, DNC, and TNC), electron-ion interaction pseudopotentials of trinucleotide (PseEIIP), and physicochemical-property-incorporated dinucleotide composition (PC-PseDNC-general). Finally, the optimal model showed a prediction accuracy of 73.06%, with an AUC of 0.80 over 10-fold CV. As for the independent test, the accuracy achieved 80.15%, with an AUC of 0.88. Compared with the latest iRNA-m5C predictor, the evaluated accuracy was improved 4.25% on average. Although there is still some

Table 4. Performances of Combined Features Over 10-fold CV, in Training Datasets, and Independent Tests, in Testing Datasets

		Training Datasets				Testing Datasets			
Feature Combination	Fea_num[a]	Acc (%)	MCC	Sn (%)	Sp (%)	Acc (%)	MCC	Sn (%)	Sp (%)
PSP (PSNP + PSDP + PSTP)	120	67.39	0.35	60.88	73.89	73.30	0.48	63.30	83.30
Kmer (NAC + DNC + TNC)	84	69.13	0.39	63.41	74.85	73.85	0.48	71.80	75.90
PSP + Kmer	204	71.47	0.43	67.01	75.93	77.60	0.56	71.60	83.60
PSP + Kmer + ENAC	352	71.27	0.43	66.50	76.04	76.80	0.54	72.20	81.40
PSP + Kmer + ENAC + MM2Gap	384	71.72	0.44	67.86	75.59	78.15	0.56	74.10	82.20
PseEIIP + PseDNC	83	69.38	0.39	63.26	75.50	72.45	0.45	70.10	74.80
PSP + Kmer + PseEIIP + PseDNC[b]	287	71.77	0.44	67.56	75.99	78.30	0.57	73.90	82.70
PSP + Kmer + PseEIIP + PseDNC + MM2Gap	319	71.73	0.44	67.86	75.60	78.18	0.57	74.10	82.25
PSP + Kmer + PseEIIP + PC-PseDNC + ENAC	435	72.06	0.44	68.05	76.08	76.75	0.54	73.40	80.10
PSP + Kmer + PseEIIP + PC-PseDNC + ENAC + MM2Gap	476	71.74	0.44	67.44	76.04	77.00	0.54	74.50	79.48
All	1,571	71.93	0.44	68.18	75.69	75.71	0.51	74.50	76.92

Acc, accuracy.
[a]The "Fea_num" column indicates the number of combined features.
[b]Performances with maximum accuracies.

room for further improvement, we believe that the proposed model can be a useful choice to predict m5C sites in RNA sequences.

MATERIALS AND METHODS

Datasets

In this study, benchmark datasets constructed by Lv et al.[32] were applied, including 6,289 positive and 6,289 negative sequences. Specifically, positive samples were selected from Gene Expression Omnibus (GEO) datasets (https://www.ncbi.nlm.nih.gov/geo/) using the accession number GEO: gse94065,[39] where the CD-HIT package[40] was adapted to remove redundant sequences with a threshold of 80%. Then, 6289 negative samples were randomly chosen from their genomes to construct balanced benchmark datasets. Finally, 1,000 positive and 1,000 negative samples were randomly selected as independent datasets, and the rest were treated as training datasets, including 5,289 positive and 5,289 negative sequences (see details in Lv et al.[32]).

Feature-Extraction Methods

In the process of constructing a machine-learning-based predictor, feature extraction plays an extremely crucial role. In this paper, seven kinds of feature-encoding methods were chosen to represent the sequence information described as follows.

PSP

PSP is an effective nucleotide-encoding approach that has been successfully applied to the identification of many functional sites in biological sequences.[41–44] In this method, the position-specific information is well represented using occurrence frequencies in positive and negative samples. Considering an RNA sequence $R = R_1R_2R_3..R_{2\xi+1}$, the PSNP matrix can be written as a $[4 \times (2\xi + 1)]$-dimensional vector

$$Z_{PSNP} = [Z_1 Z_2 \ldots Z_{2\xi+1}] = \begin{bmatrix} Z_{1,1} & Z_{1,2} & \ldots & Z_{1,2\xi+1} \\ Z_{2,1} & Z_{2,2} & \ldots & Z_{2,2\xi+1} \\ Z_{3,1} & Z_{3,2} & \ldots & Z_{3,2\xi+1} \\ Z_{4,1} & Z_{4,2} & \ldots & Z_{4,2\xi+1} \end{bmatrix}, \quad \text{(Equation 1)}$$

where $Z_{i,j} = Z_{i,j}^{+} - Z_{i,j}^{-}$ gives the difference of frequencies of the ith nucleotide at the jth position between positive $(Z_{i,j}^{+})$ and negative $(Z_{i,j}^{-})$ samples. Finally, the $(2\xi + 1)$-length RNA sequence can be encoded as

$$V_{PSNP} = \left[f_1\ f_2 \ldots f_j \ldots f_{2\xi+1}\right]^T \quad \text{(Equation 2)}$$

Here, f_j is the element from the Z_{PSNP} matrix

$$f_i = \begin{cases} Z_{1,j} & when\ N_j = A, \\ Z_{2,j} & when\ N_j = C, \\ Z_{3,j} & when\ N_j = G, \\ Z_{4,j} & when\ N_j = U, \end{cases} j = 1, 2, \ldots, 2\xi + 1. \quad \text{(Equation 3)}$$

Similarly, PSDP-associated dinucleotides can be written as a $[16 \times (2\xi)]$-dimensional vector

$$Z_{PSDP} = \begin{bmatrix} Z_{1,1} & Z_{1,2} & \ldots & Z_{1,2\xi} \\ Z_{2,1} & Z_{2,2} & \ldots & Z_{2,2\xi} \\ \ldots & \ldots & \ldots & \ldots \\ Z_{16,1} & Z_{16,2} & \ldots & Z_{16,2\xi} \end{bmatrix} \quad \text{(Equation 4)}$$

The corresponding feature can be expressed as

$$V_{PSDP} = \left[f_1\ f_2\ ..\ f_{2\xi}\right]^T, \quad \text{(Equation 5)}$$

Table 5. Comparison of Different Classifiers Using the Feature Combination "PSP + Kmer + PseEIIP + PC-PseDNC"

	Training Datasets					Testing Datasets				
Classifier	Acc (%)	MCC	Sn (%)	Sp (%)	AUC	Acc (%)	MCC	Sn (%)	Sp (%)	AUC
RF	71.77	0.44	75.99	67.56	0.79	78.30	0.57	73.90	82.70	0.85
SVM[a]	72.72	0.46	65.46	79.98	0.80	79.90	0.60	79.40	80.40	0.88
AdaBoost	71.19	0.42	68.33	74.04	0.78	80.45	0.61	77.10	83.80	0.88
NB	66.60	0.34	55.08	78.12	0.71	69.82	0.40	73.00	66.63	0.77

Acc, accuracy.
[a]Performances with maximum accuracies using the SVM algorithm.

and PSTP-associated trinucleotides are displayed as a $[64 \times (2\xi - 1)]$-dimensional vector,

$$Z_{PSTP} = \begin{bmatrix} Z_{1,1} & Z_{1,2} & \cdots & Z_{1,2\xi-1} \\ Z_{2,1} & Z_{2,2} & \cdots & Z_{2,2\xi-1} \\ \cdots & \cdots & \cdots & \cdots \\ Z_{64,1} & Z_{64,2} & \cdots & Z_{64,2\xi-1} \end{bmatrix}. \quad \text{(Equation 6)}$$

The RNA sequence can be represented as

$$V_{PSTP} = \left[f_1 \; f_2 \; .. \; f_{2\xi-1} \right]^T. \quad \text{(Equation 7)}$$

Kmer

Kmer is a common method to represent RNA sequences, which is simply expressed as the occurrence frequencies of k-neighboring nucleotides in bioinformatics.[31,32,35,45] Here, we considered three kinds of feature vectors with ks = 1, 2, and 3, corresponding to NAC, DNC, and TNC, respectively.

ENAC

The ENAC is a variant of the NAC method, which calculates nucleotide occurrence frequencies in a length-fixed sequence window.[46] The window can continuously loop through all nucleotides from 5′ to the 3′ terminus. Here, the default length 5 was used, forming a $[(2\xi + 1 - 4) \times 4]$-dimensional feature vector.

xxKGAP

xxKGAP composition is a major method implemented in PyFeat,[37] which considered kgaps in the nucleotide sub-sequences. Frequencies of these sub-sequences are treated as prediction features. Specifically, for mMKGap features, if kgap = 1, the sequence can be encoded as frequencies of X_X, i.e., $4 \times 1 \times 4 = 16$-dimensional features. If kgap = 2, the sequence can be expressed as $4 \times 2 \times 4 = 32$ features. As for dMKGap, there are, in total, $4^2 \times n \times 4$. The number of features are increased with the n. In this paper, in total, nine kinds of features, including mMKGap, mDGKap, dMKGap with ks = 1, 2, and 3, were studied.

EIIP and PseEIIP

The EIIP approach directly uses EIIP values of 4 nucleotides to represent corresponding nucleotides (expressed as $EIIP_A$, $EIIP_C$, $EIIP_G$, and $EIIP_U$), which induces $(2\xi + 1)$-dimensional features.

Additionally, the PseEIIP vector can be written as the mean EIIP value of related trinucleotides:

$$V = \left[EIIP_{AAA}.f_{AAA}, EIIP_{AAc}.f_{AAc}, \ldots, EIIP_{UUU}.f_{UUU} \right], \quad \text{(Equation 8)}$$

where f_{XYZ} and $EIIP_{XYZ}$ are the normalized frequency and associated EIIP value of the ith trinucleotide XYZ by $EIIP_{XYZ} = EIIP_X + EIIP_Y + EIIP_Z$. These two methods showed good results for prediction problems.[43,47] It is noted that only EIIP values (A, 0.1260; C, 0.1340; G, 0.0806; and T, 0.1335)[48] were applied in the iLearn package to represent the DNA sequence.[36] Here, we still use the EIIP value 0.1335 for the U nucleotide in RNA sequences. It is obviously found that PseEIIP methods produce a 64-dimensional feature vector.

PC-PseDNC-General

The PC-PseDNC-general method[49–51] incorporates short-range and long-range information by dinucleotide composition and related correlations of physicochemical properties. Here, we extracted

Table 6. Comparison of the Constructed Model with Two Published Methods

	Training Datasets					Testing Datasets				
Method	Acc (%)	MCC	Sn (%)	Sp (%)	AUC	Acc (%)	MCC	Sn (%)	Sp (%)	AUC
PEA-m5C[a]						44.30	−0.11	43.20	45.40	
iRNA-m5C	70.70	0.42	65.70	75.70	0.77	74.00	0.48	72.40	75.60	
This work	73.06	0.47	66.42	79.70	0.80	80.15	0.60	79.40	80.90	0.88

Acc, accuracy.
[a]Results of the PEA-m5C tool[31] were excerpted from Lv et al.[32] (i.e., obtained using independent data objectively).

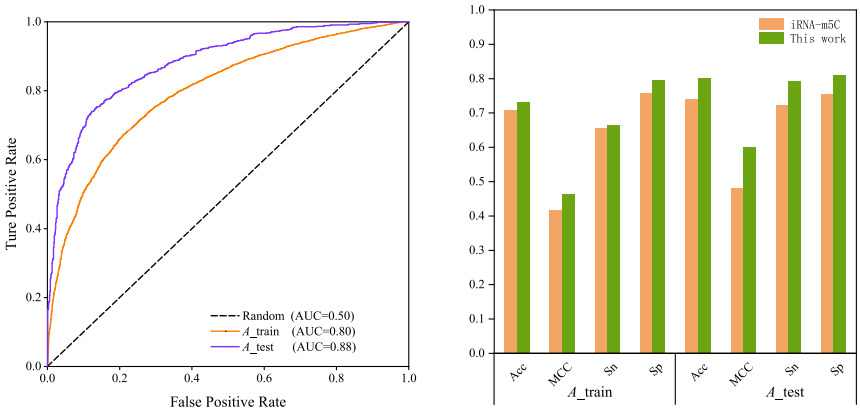

Figure 3. Evaluated Perfromances

Left: ROC curves for best performing feature combinations based on the SVM method. Right: comparison of our results (green) and the iRNA-m5C predictor (orange).

PC-PseDNC features by the Pse-in-One 2.0 package with 22 physicochemical properties included,[34] which can be written as a $(16 + \lambda)$-dimensional vector

$$V = (x_1 \cdots |x_{16}x_{16+1}| \cdots |x_{16+\lambda})^T, \quad \text{(Equation 9)}$$

where the parameter λ indicates the highest counted rank (or tier) in calculations. The detailed description can be found in Liu et al.[34]

BE

In the BE method, the sequence can be directly written as a $4\times(2\xi + 1)$-dimensional vector, in which A, C, G, and U are characterized as (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1), respectively.[52–54]

NCP + ND

Features NCP and ND are combined to encode RNA sequences with high performances.[55,56] The nucleotide N_i can be written as

$$N_i = (x_i, y_i, z_i, d_i), \quad \text{(Equation 10)}$$

where x_i, y_i, and z_i indicate the three properties of ring structure, functional group, and hydrogen bond, respectively. It is defined as:

$$x_i = \begin{cases} 1, & N_i \in [A, G] \\ 0, & N_i \in [C, U] \end{cases}; y_i = \begin{cases} 1, & N_i \in [A, C] \\ 0, & N_i \in [G, U] \end{cases}; z_i = \begin{cases} 1, & N_i \in [A, U] \\ 0, & N_i \in [C, G] \end{cases} \quad \text{(Equation 11)}$$

Additionally, d_i is the accumulated density

$$d_i = \frac{1}{\|S_i\|} \sum_{j=1}^{l} f(N_j), |f(N_j) = \begin{cases} 1, & |if| \ N_j \in [A, C, G, U] \\ 0, & other|cases \end{cases}, \quad \text{(Equation 12)}$$

here, $\|S_i|\|$ is the length of the subsequence ended in the relevant nucleotide.

Classifiers

Many kinds of machine-learning algorithms have been successfully applied in bioinformatics. Here, we used four classifiers implemented in the *sklearn* package[38,57] for comparison, including RF, SVM, AdaBoost, and NB.

RF

RF is a popular tree-based ensemble estimator, where the overall predictive accuracy is improved by combining a number of decision tree classifiers effectively.[58] It has been widely applied in fields of bioinformatics research.[30–32,35,59–61]

SVM

SVM is an efficient supervised machine-learning algorithm for classification, regression, and outlier detection.[62–64] It has been successfully applied in prediction subjects.[55,65–73] In this method, the original input vectors are transformed into a higher Hilbert space by kernel function. Here, the radial basis kernel function (RBF) was chosen to seek the best classification hyperplane.

In comparison, AdaBoost and NB were both used in this work. Specifically, the AdaBoost method is used to try to fit a sequence of weak learners (i.e., models that are only slightly better than random guessing, such as small decision trees) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction.[74,75] The NB method is from a set of supervised learning algorithms based on applying Bayes' theorem with the independent assumption.[76] Specifically, Gaussian NB algorithm was implemented for the classifier task.

CV Test

For a convenient and fair comparison with the newest predictor iRNA-m5C,[32] 10-fold CV and independent tests were separately used to evaluate constructed models for training and testing datasets. For the *k*-fold CV, benchmark datasets are equally divided into *k* subsets. Then, the k − 1 subsets are used to train the model, and the remaining one is used to test. This process is repeated *k* times until all subsets are used once for testing. The final performance is an average value of all *k* testing experiments.[77]

Performance Evaluation

For the two-label classification, four metrics are usually applied to evaluate performances of the proposed model, formulated as follows:[78–83]

$$\begin{cases} S_n = 1 - \dfrac{N_-^+}{N^+} & 0 \le S_n \le 1 \\ S_p = 1 - \dfrac{N_+^-}{N^-} & 0 \le S_p \le 1 \\ Acc = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le A_{CC} \le 1 \\ MCC = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le MCC \le 1 \end{cases}$$

(Equation 13)

Here, S_n, S_p, S_p, and MCC indicate sensitivity, specificity, accuracy, and Matthew's correlation coefficient, respectively. N^+ and N^- indicate the number of positive and negative sequences considered, in which incorrectly predicted samples are labeled as N_-^+ and N_+^-, respectively.

In addition, the graph of the ROC[84,85] is also widely used to intuitively display the performance. Specifically, vertical and horizontal coordinates are the true positive rate (TPR) and the false positive rate (FPR), respectively. Then, the AUC can be obtained to objectively evaluate performances of the proposed model.

AUTHOR CONTRIBUTIONS

L.X. and H.X. proposed the idea and designed the overall research. L.D. performed the experiments and wrote the manuscript. X.L. and H.D. helped to revise the paper. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (no. 61902259), the Natural Science Foundation of Guangdong Province (grant no. 2018A0303130084), and the Scientific Research Foundation in Shenzhen (JCYJ20170818100431895, JCYJ20180305163701198, and JCYJ20180306172207178).

REFERENCES

1. Machnicka, M.A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K.M., et al. (2013). MODOMICS: a database of RNA modification pathways–2013 update. Nucleic Acids Res. *41*, D262–D267.

2. Li, S., and Mason, C.E. (2014). The pivotal regulatory landscape of RNA modifications. Annu. Rev. Genomics Hum. Genet. *15*, 127–150.

3. Meyer, K.D., and Jaffrey, S.R. (2014). The dynamic epitranscriptome: N6-methyladenosine and gene expression control. Nat. Rev. Mol. Cell Biol. *15*, 313–326.

4. Kirchner, S., and Ignatova, Z. (2015). Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nat. Rev. Genet. *16*, 98–112.

5. Sun, W.J., Li, J.H., Liu, S., Wu, J., Zhou, H., Qu, L.H., and Yang, J.H. (2016). RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. Nucleic Acids Res. *44* (D1), D259–D265.

6. Roundtree, I.A., Evans, M.E., Pan, T., and He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation. Cell *169*, 1187–1200.

7. Boccaletto, P., Machnicka, M.A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T.K., de Crécy-Lagard, V., Ross, R., Limbach, P.A., Kotter, A., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res. *46* (D1), D303–D307.

8. Chen, Y., Sierzputowska-Gracz, H., Guenther, R., Everett, K., and Agris, P.F. (1993). 5-Methylcytidine is required for cooperative binding of Mg2+ and a conformational transition at the anticodon stem-loop of yeast phenylalanine tRNA. Biochemistry *32*, 10249–10253.

9. Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M., and Lyko, F. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. Genes Dev. *24*, 1590–1595.

10. Blanco, S., Kurowski, A., Nichols, J., Watt, F.M., Benitah, S.A., and Frye, M. (2011). The RNA-methyltransferase Misu (NSun2) poises epidermal stem cells to differentiate. PLoS Genet. *7*, e1002403.

11. Zhang, X., Liu, Z., Yi, J., Tang, H., Xing, J., Yu, M., Tong, T., Shang, Y., Gorospe, M., and Wang, W. (2012). The tRNA methyltransferase NSun2 stabilizes p16INK^4 mRNA by methylating the 3′-untranslated region of p16. Nat. Commun. *3*, 712.

12. Khoddami, V., and Cairns, B.R. (2013). Identification of direct targets and modified bases of RNA cytosine methyltransferases. Nat. Biotechnol. *31*, 458–464.

13. Hussain, S., Tuorto, F., Menon, S., Blanco, S., Cox, C., Flores, J.V., Watt, S., Kudo, N.R., Lyko, F., and Frye, M. (2013). The mouse cytosine-5 RNA methyltransferase NSun2 is a component of the chromatoid body and required for testis differentiation. Mol. Cell. Biol. *33*, 1561–1570.

14. Yang, X., Yang, Y., Sun, B.-F., Chen, Y.-S., Xu, J.-W., Lai, W.-Y., Li, A., Wang, X., Bhattarai, D.P., Xiao, W., et al. (2017). 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m^{5}C reader. Cell Res. *27*, 606–625.

15. Frye, M., Dragoni, I., Chin, S.-F., Spiteri, I., Kurowski, A., Provenzano, E., Green, A., Ellis, I.O., Grimmer, D., Teschendorff, A., et al. (2010). Genomic gain of 5p15 leads to over-expression of Misu (NSUN2) in breast cancer. Cancer Lett. *289*, 71–80.

16. Abbasi-Moheb, L., Mertel, S., Gonsior, M., Nouri-Vahid, L., Kahrizi, K., Cirak, S., Wieczorek, D., Motazacker, M.M., Esmaeeli-Nieh, S., Cremer, K., et al. (2012). Mutations in NSUN2 cause autosomal-recessive intellectual disability. Am. J. Hum. Genet. *90*, 847–855.

17. Ciccia, A., and Elledge, S.J. (2010). The DNA damage response: making it safe to play with knives. Mol. Cell *40*, 179–204.

18. Guy, M.P., Shaw, M., Weiner, C.L., Hobson, L., Stark, Z., Rose, K., Kalscheuer, V.M., Gecz, J., and Phizicky, E.M. (2015). Defects in tRNA Anticodon Loop 2′-O-Methylation Are Implicated in Nonsyndromic X-Linked Intellectual Disability due to Mutations in FTSJ1. Hum. Mutat. *36*, 1176–1187.

19. Hong, B., Brockenbrough, J.S., Wu, P., and Aris, J.P. (1997). Nop2p is required for pre-rRNA processing and 60S ribosome subunit synthesis in yeast. Mol. Cell. Biol. *17*, 378–388.

20. Alexandrov, A., Chernyakov, I., Gu, W., Hiley, S.L., Hughes, T.R., Grayhack, E.J., and Phizicky, E.M. (2006). Rapid tRNA decay can result from lack of nonessential modifications. Mol. Cell *21*, 87–96.

21. Gigova, A., Duggimpudi, S., Pollex, T., Schaefer, M., and Koš, M. (2014). A cluster of methylations in the domain IV of 25S rRNA is required for ribosome stability. RNA *20*, 1632–1644.

22. Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. USA *89*, 1827–1831.

23. Edelheit, S., Schwartz, S., Mumbach, M.R., Wurtzel, O., and Sorek, R. (2013). Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. PLoS Genet. *9*, e1003602.

24. Masiello, I., and Biggiogera, M. (2017). Ultrastructural localization of 5-methylcytosine on DNA and RNA. Cell. Mol. Life Sci. *74*, 3057–3064.

25. Chen, X., Sun, Y.Z., Liu, H., Zhang, L., Li, J.Q., and Meng, J. (2019). RNA methylation and diseases: experimental results, databases, Web servers and computational models. Brief. Bioinform. *20*, 896–917.

26. Feng, P., Ding, H., Chen, W., and Lin, H. (2016). Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. Mol. Biosyst. *12*, 3307–3311.

27. Qiu, W.R., Jiang, S.Y., Xu, Z.C., Xiao, X., and Chou, K.C. (2017). iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget *8*, 41178–41188.

28. Zhang, M., Xu, Y., Li, L., Liu, Z., Yang, X., and Yu, D.J. (2018). Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble. Anal. Biochem. *550*, 41–48.

29. Sabooh, M.F., Iqbal, N., Khan, M., Khan, M., and Maqbool, H.F. (2018). Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. J. Theor. Biol. *452*, 1–9.

30. Li, J., Huang, Y., Yang, X., Zhou, Y., and Zhou, Y. (2018). RNAm5Cfinder: A Web-server for Predicting RNA 5-methylcytosine (m5C) Sites Based on Random Forest. Sci. Rep. *8*, 17299.

31. Song, J., Zhai, J., Bian, E., Song, Y., Yu, J., and Ma, C. (2018). Transcriptome-Wide Annotation of m^5C RNA Modifications Using Machine Learning. Front. Plant Sci. *9*, 519.

32. Lv, H., Zhang, Z.M., Li, S.H., Tan, J.X., Chen, W., and Lin, H. (2020). Evaluation of different computational methods on 5-methylcytosine sites identification. Brief. Bioinform. *21*, 982–995.

33. Fang, T., Zhang, Z., Sun, R., Zhu, L., He, J., Huang, B., Xiong, Y., and Zhu, X. (2019). RNAm5CPred: Prediction of RNA 5-Methylcytosine Sites Based on Three Different Kinds of Nucleotide Composition. Mol. Ther. Nucleic Acids *18*, 739–747.

34. 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. Nat. Sci. *9*, 67–91.

35. 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. *47*, e127.

36. Chen, Z., Zhao, P., Li, F., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I., 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*, 1047–1057.

37. Muhammod, R., Ahmed, S., Md Farid, D., Shatabda, S., Sharma, A., and Dehzangi, A. (2019). PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences. Bioinformatics *35*, 3831–3833.

38. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

39. Cui, X., Liang, Z., Shen, L., Zhang, Q., Bao, S., Geng, Y., Zhang, B., Leo, V., Vardy, L.A., Lu, T., et al. (2017). 5-Methylcytosine RNA Methylation in Arabidopsis Thaliana. Mol. Plant *10*, 1387–1399.

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

41. Li, G.Q., Liu, Z., Shen, H.B., and Yu, D.J. (2016). TargetM6A: Identifying N^6-Methyladenosine Sites From RNA Sequences via Position-Specific Nucleotide Propensities and a Support Vector Machine. IEEE Trans. Nanobioscience *15*, 674–682.

42. He, W., Jia, C., Duan, Y., and Zou, Q. (2018). 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. BMC Syst. Biol. *12* (*Suppl 4*), 44.

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

44. 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. Brief. Funct. Genomics *18*, 367–376.

45. Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. IEEE/ACM Trans. Comput. Biol. Bioinformatics *11*, 192–201.

46. Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T.T., Wang, Y., Webb, G.I., Smith, A.I., Daly, R.J., Chou, K.C., and Song, J. (2018). iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics *34*, 2499–2502.

47. Jia, C., Yang, Q., and Zou, Q. (2018). NucPosPred: Predicting species-specific genomic nucleosome positioning via four different modes of general PseKNC. J. Theor. Biol. *450*, 15–21.

48. Nair, A.S., and Sreenadhan, S.P. (2006). A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation *1*, 197–202.

49. Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.C. (2015). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics *31*, 119–120.

50. Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: A Sequence-Based Predictor for Identifying 2′-O-Methylation Sites in Homo sapiens. J. Comput. Biol. *25*, 1266–1277.

51. Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Brief. Bioinform. *20*, 1280–1294.

52. Chen, Z., Zhou, Y., Song, J., and Zhang, Z. (2013). hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochim. Biophys. Acta *1834*, 1461–1467.

53. Wei, L., Luan, S., Nagai, L.A.E., Su, R., and Zou, Q. (2019). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics *35*, 1326–1333.

54. Chen, Z., Chen, Y.-Z., Wang, X.-F., Wang, C., Yan, R.-X., and Zhang, Z. (2011). Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. PLoS ONE *6*, e22930.

55. Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K.C. (2016). iRNA-PseU: Identifying RNA pseudouridine sites. Mol. Ther. Nucleic Acids *5*, e332.

56. Chen, W., Song, X., Lv, H., and Lin, H. (2019). iRNA-m2G: Identifying N^2-methylguanosine Sites Based on Sequence-Derived Information. Mol. Ther. Nucleic Acids *18*, 253–258.

57. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., et al. (2013). API design for machine learning software: Experiences from the scikit-learn project. arXiv, arXiv:1309.0238, http://arxiv.org/abs/1309.0238.

58. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

59. Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and Validation of Disease Genes Using HeteSim Scores. IEEE/ACM Trans. Comput. Biol. Bioinformatics *14*, 687–695.

60. 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. Genet. *10*, 33.

61. Ru, X., Li, L., and Zou, Q. (2019). Incorporating Distance-Based Top-n-gram and Random Forest To Identify Electron Transport Proteins. J. Proteome Res. *18*, 2931–2939.

62. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.

63. Cristianini, N., and Shawe-Taylor, J. (2000). An Introduction of Support Vector Machines and Other Kernel-based Learning Methods (Cambridge University Press).

64. Andrew, A.M. (2000). An Introduction to Support Vector Machines and Other Kernal-Based Learning Methods. Robotica *18*, 687–689.

65. Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. Inf. Sci. *418-419*, 546–560.

66. Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif. Intell. Med. *83*, 67–74.

67. Wei, L., Wan, S., Guo, J., and Wong, K.K. (2017). A novel hierarchical selective ensemble classifier with bioinformatics application. Artif. Intell. Med. *83*, 82–90.

68. Zhu, X.J., Feng, C.Q., Lai, H.Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowl. Base. Syst. *163*, 787–793.

69. Chen, W., Feng, P., Liu, T., and Jin, D. (2019). Recent Advances in Machine Learning Methods for Predicting Heat Shock Proteins. Curr. Drug Metab. *20*, 224–228.

70. Xiong, Y., Qiao, Y., Kihara, D., Zhang, H.Y., Zhu, X., and Wei, D.Q. (2019). Survey of Machine Learning Techniques for Prediction of the Isoform Specificity of Cytochrome P450 Substrates. Curr. Drug Metab. *20*, 229–235.

71. Li, Y.H., Li, X.X., Hong, J.J., Wang, Y.X., Fu, J.B., Yang, H., Yu, C.Y., Li, F.C., Hu, J., Xue, W.W., et al. (2020). Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. Brief. Bioinform. *21*, 649–662.

72. Liu, B., and Li, K. (2019). iPromoter-2L2.0: Identifying Promoters and Their Types by Combining Smoothing Cutting Window Algorithm and Sequence-Based Features. Mol. Ther. Nucleic Acids *18*, 80–87.

73. Liu, B., Li, C.-C., and Yan, K. (2019). DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. Brief. Bioinform. bbz098. Published October 28, 2019. https://doi.org/10.1093/bib/bbz098.

74. Freund, Y., and Schapire, R.E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J. Comput. Syst. Sci. *55*, 119–139.

75. Keogh, E., and Mueen, A. (2010). Curse of dimensionality. In Encyclopedia of Machine Learning, C. Sammut and G.I. Webb, eds. (Springer), pp. 257–258.

76. Zhang, H. (2004). The Optimality of Naive Bayes. In Proceedings of the 17th Florida Artificial Intelligence Research Society Conference, FLAIRS 2004, V. Barr and Z. Markov, eds. (AAAI Press), pp. 562–567.

77. Chou, K.C., and Zhang, C.T. (1995). Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. *30*, 275–349.

78. Chen, W., Feng, P.M., Lin, H., and Chou, K.C. (2013). iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. *41*, e68.

79. Xu, Y., Ding, J., Wu, L.-Y., and Chou, K.-C. (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*, e55844.

80. Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics *35*, 2796–2800.

81. 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.

82. Shen, Y., Tang, J., and Guo, F. (2019). Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. J. Theor. Biol. *462*, 230–239.

83. Ding, Y., Tang, J., and Guo, F. (2019). Identification of Drug-Side Effect Association via Semisupervised Model and Multiple Kernel Learning. IEEE J. Biomed. Health Inform. *23*, 2619–2632.

84. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognit. Lett. *27*, 861–874.

85. Davis, J., and Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning, W.W. Cohen and A. Moore, eds. (Association for Computing Machinery), pp. 233–240.