Hindawi
BioMed Research International
Volume 2022, Article ID 4035462, 11 pages
https://doi.org/10.1155/2022/4035462

Research Article

Predicting RNA 5-Methylcytosine Sites by Using Essential Sequence Features and Distributions

Lei Chen,[1,2] **ZhanDong Li**,[3] **ShiQi Zhang**,[4] **Yu-Hang Zhang**,[5] **Tao Huang**,[6,7] **and Yu-Dong Cai**[1]

[1]*School of Life Sciences, Shanghai University, Shanghai 200444, China*
[2]*College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[3]*College of Food Engineering, Jilin Engineering Normal University, Changchun, China*
[4]*Department of Biostatistics, University of Copenhagen, Copenhagen 2099, Denmark*
[5]*Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
[6]*Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[7]*CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Tao Huang; tohuangtao@126.com and Yu-Dong Cai; cai_yud@126.com

Received 15 September 2021; Revised 7 December 2021; Accepted 22 December 2021; Published 13 January 2022

Academic Editor: Hesham H. Ali

Methylation is one of the most common and considerable modifications in biological systems mediated by multiple enzymes. Recent studies have shown that methylation has been widely identified in different RNA molecules. RNA methylation modifications have various kinds, such as 5-methylcytosine (m^5C). However, for individual methylation sites, their functions still remain to be elucidated. Testing of all methylation sites relies heavily on high-throughput sequencing technology, which is expensive and labor consuming. Thus, computational prediction approaches could serve as a substitute. In this study, multiple machine learning models were used to predict possible RNA m^5C sites on the basis of mRNA sequences in human and mouse. Each site was represented by several features derived from *k*-mers of an RNA subsequence containing such site as center. The powerful max-relevance and min-redundancy (mRMR) feature selection method was employed to analyse these features. The outcome feature list was fed into incremental feature selection method, incorporating four classification algorithms, to build efficient models. Furthermore, the sites related to features used in the models were also investigated.

1. Introduction

Methylation is one of the most common and considerable modifications in biological systems mediated by multiple enzymes. The substrates of biological methylation are diverse, with DNA as the most common one. Previous studies on methylation mostly focused on DNA methylation, revealing its specific role in transcriptional activity regulation during development, aging, and pathogenesis [1]. However, recent studies have widely identified methylation among different RNA molecules, including mRNA, snoRNA, miRNA, and rRNA (not restricted to functional mRNAs) [2]. RNA methylation enables the posttranscriptional control of gene expression by changing how RNA interacts with other components of the cell as an important part of epitranscriptome [3]. RNA methylation is actively involved in posttranscriptional regulatory bioprocesses, like RNA splicing, transport, stability, and translatability, and it has strong relationships with mammalian development and diseases [4–6].

Among the various kinds of RNA methylation modifications, N^6-methyladenosine (m^6A), the methylation modification on the nitrogen at the sixth position of the adenosine base, is the most prevalent internal mRNA modification, accounting for 50% of the total methylated ribonucleotides

[2, 7]. M^6A broadly affects mRNA metabolism, and it is widely distributed in all kinds of RNA transcripts, including coding and noncoding regions. The deposition of m^6A modification in the transcriptome has its unique pattern: the m^6A modification sites have a typical consensus sequence DRACH (D = G, A, or U; R = G or A; H = A, C, or U), which is widely dispersed over coding sequence and untranslated region (UTR) and highly enriched near the stop codon area [8]. Recent evidence has proven that m^6A RNA methylation plays a vital role in pre-mRNA splicing, mRNA stability regulation, mRNA export, mRNA degradation, translation regulation, and miRNA processing [9–11]. M^6A modification is dynamic, it could be reversible, and it may vary between different genes and different tissues [12, 13]. With the increase in the number of m^6A mapping studies, the list of specific genes containing a disproportionately high level of m^6A was revealed. For example, Han et al. found a series of m^6A methylated genes related to the presynaptic membrane, the postsynaptic membrane, and the synaptic growth in Alzheimer's disease (AD) mouse models, suggesting that m^6A may be involved in the occurrence of AD [14]. While the function of m^6A modification is context-dependent and dynamic, many m^6A sites are evolutionally conserved among species. One-third of mammalian mRNAs share the same m^6A modifications, and many of them are conserved with single-nucleotide specificity [15].

Another kind of RNA methylation modification, namely, 5-methylcytosine (m^5C), which is the methylation of carbon 5 in cytosine, also acts as an important regulator in gene expression, including RNA localization, ribosome assembly, translation regulation, and mRNA stabilization. Among all the mRNA methylation sites, the proportion of m^5C could be up to 20% in human cells [16]. The distribution of m^5C sites in mRNA is not random; in HeLa and mouse cells, m^5C methylation were found to be enriched in $5'$ and $3'$ UTRs rather than coding regions [16]. Like m^6A, m^5C acts its function in dynamic ways. M^5C methylation occurs dynamically during testis development and helps maintain the stability of maternal mRNA in embryonic development [17].

Though RNA methylation plays a pivotal role in bioprocess and is of great importance to posttranscriptional regulation, their functions in individual methylation sites still remain to be elucidated. Testing of all the methylation sites relies heavily on high-throughput sequencing technology, which is expensive and labor consuming; thus, computational prediction approaches could serve as a substitute [18]. As mentioned above, the distribution of m^5C in mRNA has its own enrichment pattern and is not random. With adequate datasets and statistic method, predicting accurate m^5C RNA methylation sites and gaining an enhanced understanding of their functions are doable.

In this study, multiple machine learning models were applied to predict the possible m^5C RNA methylation sites in mRNA sequences of human and mouse. For each m^5C, a subsequence containing such site as center was extracted from the RNA sequence. The features of k-mers yielded by RNA2Vec [19] were refined to represent the subsequence. The powerful max-relevance and min-redundancy (mRMR) feature selection method [20] was employed to analyse all features. Obtained feature list was fed into incremental feature selection (IFS) [21] method, incorporating four classification algorithms, to build efficient models. In addition to prediction models, we also investigated the sites related to features used in the models, trying to discover special patterns around mouse and human m^5C sites. Comparison of those prediction results may help obtain a dynamic RNA methylation profile and build relationships between the RNA methylation sites and human diseases.

2. Materials and Methods

2.1. Data. M^5C is a common RNA modification in mammals. Human and mouse m^5C data were downloaded from one previous study (iRNA-m^5C, http://lin-group.cn/server/iRNA-m5C/download.html) [22]. In fact, the human m^5C data was first used in [23], which was extracted from the original data retrieved from RMBase database [24]. The original data was processed by CD-HIT program [25] so that the sequence similarity of any remaining sequences was less than 0.7. As a result, 120 positive and 120 negative m^5C sites were obtained. As for mouse m^5C data, it was constructed in [22]. It was directly retrieved from RMBase database [24] and was not processed by CD-HIT program [25] because its size was so small. The mouse data consisted of 97 positive and 97 negative m^5C sites. As the sites around the m^5C sites have some special patterns, which can help to identify m^5C sites in RNA sequence, 20 upstream sites and 20 downstream sites were picked up. These sites together with the m^5C site at the center constructed a subsequence with 41 bp. Some features would be extracted from this subsequence to represent the m^5C site.

2.2. Problem Description and Study Design. For a given RNA sequence, it is essential to identify m^5C sites in it. The machine learning models can give a deep investigation on current known m^5C sites and learn a special pattern to make prediction. The prediction procedure can be deemed as a function f, formulated by

$$f : \Psi \longrightarrow \{+,-\}, \tag{1}$$

where Ψ denoted the site set for human or mouse RNA sequences and $+(-)$ represented whether the input site was an m^5C site or not.

Generally, we want to discover an optimal function such that its loss was smallest. Because machine learning algorithms were employed to design such function, we adopted the following steps:

(1) For any site in the human or mouse m^5C data, sites around it were picked up to comprise a subsequence, which can indicate the surrounding information of the investigated site. This step was described in section "Feature Engineering"

(2) Each subsequence was represented by a number of features, which can reflect its essential information. This step was described in section "Feature Engineering"

(3) A feature selection method was adopted to analyse all features and produce a feature list. This step was described in section "Max-Relevance and Min-Redundancy (mRMR) Feature Selection"

(4) The IFS method was applied on such feature list to find out which classification algorithm and which features can yield the best performance (smallest loss). This step was described in section "Incremental Feature Selection (IFS)." The descriptions of four classification algorithms used in IFS method can be found in section "Classification Algorithm." The loss was determined by one measurement listed in section "Performance Measurement"

2.3. Feature Engineering. To build efficient models for identifying m^5C site in RNA sequence, it is very important to extract essential features from the subsequence consisting of this site, 20 upstream sites and 20 downstream sites. This study adopted a natural language processing approach to extract features, which were further used to represent the subsequence containing m^5C site.

RNA2Vec [19] was adopted to extract sequence features for each k-mers (subsequences of length k). In detail, this method employed the whole human genome as corpus. A sliding window technique was used to split the RNA sequence into several fix-length words. If an RNA sequence with length L was formulated by

$$S = R_1 R_2 \cdots R_i \cdots R_{L-1} R_L, \tag{2}$$

it was split into $L - k + 1$ words, say $R_1 R_2 \cdots R_k, R_2 R_3 \cdots R_{k+1}, \cdots, R_{L-k+1} R_{L-k+2} \cdots R_L$. All obtained words were fed into GloVe algorithm [26], a type of Word2vec method, to extract features of words, i.e., features of k-mers. Here, we selected $k = 4$. Features of 4-mers were directly retrieved from https://github.com/HsiaoYetGun/MiRLocator/blob/master/RNA2Vec/RNAVectors.txt. Each 4-mers was represented by 30 features.

Given a 41 bp long RNA subsequence SS, formulated by

$$SS = R_1 R_2 \cdots R_{20} R_{21} R_{22} \cdots R_{40} R_{41}, \tag{3}$$

where R_{21} was the m^5C site, we extracted all 4-mers from this subsequence. Because the R_{21} was always same for all investigated subsequences, the 4-mers containing this site were discarded. 34 4-mers can be obtained from each RNA subsequence. Their 30 features obtained by RNA2Vec were collected together to represent the subsequence SS. Accordingly, 1020 (34×30) features were adopted to encode each subsequence with 41 bp.

2.4. Max-Relevance and Min-Redundancy (mRMR) Feature Selection. The mRMR is a powerful feature selection method [20, 27–30], which evaluates the importance of features from two aspects: (1) relevance to class labels and (2) redundancies to other features. The mutual information (MI) is used to quantify the relevance and redundancy. For two variables x and y, their MI is computed by

$$MI(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy, \tag{4}$$

where $p(x)$ and $p(y)$ stand for the marginal probabilistic densities of x and y, respectively, and $p(x, y)$ stands for the joint probabilistic density of x and y. Generally, a high MI indicates the strong relevance or high redundancy of two variables. The mRMR method tries to keep features with high relevance to class labels and low redundancies to other features. However, this is a NP-hard problem. The mRMR method employed a heuristic way to evaluate features, which sorts all investigated features in a list, namely, mRMR feature list. At the beginning, this list is empty. For each feature f that is not in this list, compute its relevance to class labels, measured by $MI(f, c)$, where c is a variable representing class labels, and redundancies to features that are already in the list, measured by the average MI between f and features in the current list. The difference of these two values is computed. The feature with highest difference is selected and appended to the list. When all features have been in the list, the procedures stop. Feature ranks in this list indicate the importance of features. Generally, features with high ranks are more important than those with low ranks.

The mRMR program used in this study was downloaded from http://penglab.janelia.org/proj/mRMR/. For convenience, it was executed using its default parameters.

2.5. Incremental Feature Selection (IFS). Although mRMR method produced a feature list, it is still a problem that which features should be selected to construct the model. In view of this, this study employed the IFS method [21], which can aid to choose proper features for any given classification algorithm. In detail, on the basis of the mRMR feature list, IFS produces several feature subsets with a step interval as one. For instance, the first feature subset has the top feature in the mRMR list, and the second feature subset has the first two features, and so on. Then, a model based on a certain classification algorithm can be constructed on the training data, where samples are represented by feature in each feature subset. All constructed models are assessed by one cross-validation method [31]. The model yielding the best performance is picked up and called the optimum model. The feature subset used in this model is termed as the optimum feature subset.

2.6. Classification Algorithm. As mentioned above, IFS method needs one classification algorithm. Here, four classification algorithms were used, including (1) random forest (RF) [32], (2) support vector machine (SVM) [33], (3) K-nearest neighbor (kNN) [34], and (4) decision tree (DT) [35]. These algorithms have been widely used to tackle various medical problems [36–48]. Their brief descriptions are as follows.

2.6.1. Random Forest. RF is a powerful and classic classification algorithm. In fact, it is an ensemble algorithm that contains several DTs. Each DT is built using two random selection procedures. The first procedure is to select samples, whereas the second procedure is for the selection of features.

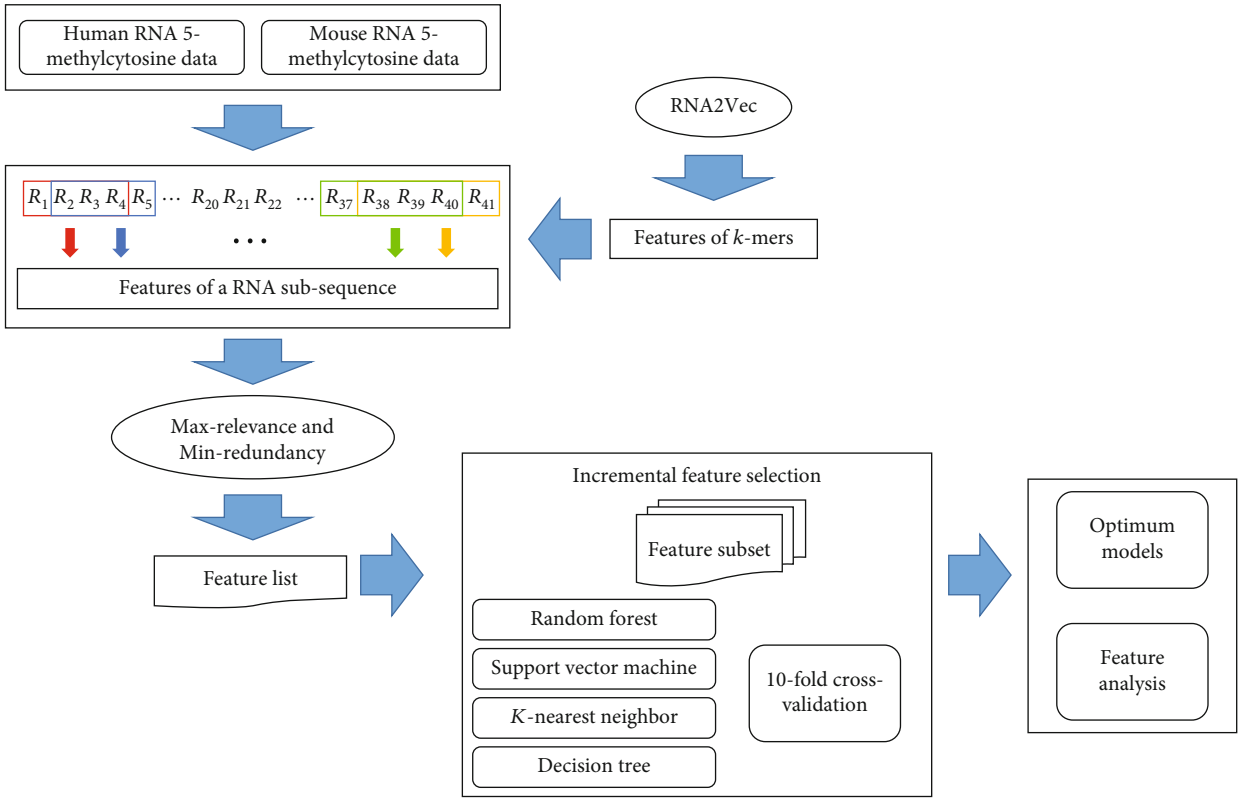

FIGURE 1: Flow chart to construct models for the prediction of m^5C sites. A subsequence with 41 bp is used to represent each m^5C site. Features of *k*-mers obtained by RNA2Vec are adopted to constitute features of the subsequence. All features are analysed by max-relevance and min-redundancy method. The outcome feature list is fed into incremental feature selection, incorporating four classification algorithms and 10-fold cross-validation, to construct optimum models.

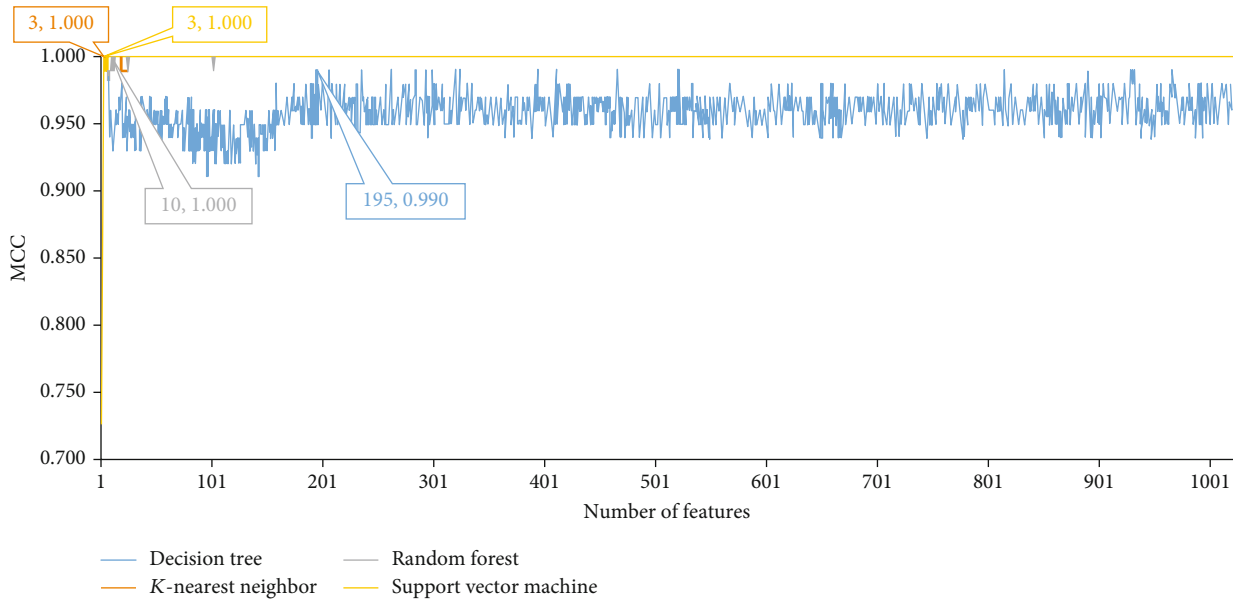

FIGURE 2: IFS curves with different classifiers on different numbers of sequence features on mouse m^5C data.

Table 1: Performance of models based on different classification algorithms for predicting mouse m^5C sites.

Classification algorithm	Number of features	SN	SP	ACC	MCC	Precision	$F1$-measure
Decision tree	195	1.000	0.990	0.995	0.990	0.990	0.995
K-nearest neighbor	3	1.000	1.000	1.000	1.000	1.000	1.000
Random forest	10	1.000	1.000	1.000	1.000	1.000	1.000
Support vector machine	3	1.000	1.000	1.000	1.000	1.000	1.000

Given a query sample, each DT yields the prediction. RF integrates these predictions with majority voting. Although DT is a quite weak classification algorithm, RF is much more robust. Thus, it is always an important candidate for constructing prediction models.

2.6.2. Support Vector Machine. SVM is another powerful and classic classification algorithm. Its main idea is to find out a hyperplane for separating samples in two classes. However, such hyperplane does not exist in many cases. SVM maps the original data with nonlinear pattern in low-dimensional space to a new data with linear pattern in high-dimensional space. Then, the hyperplane is constructed in such new space by maximizing interval between samples in two classes. Finally, it predicts the class label of a new sample according to which side of hyperplane this new data point belongs to.

2.6.3. K-Nearest Neighbor. kNN is a simple but also efficient classification algorithm. It is not a strict machine learning algorithm because there is no training procedures. Several computational steps are conducted to determine the class of a test sample, such as computing the distance between the test sample and all training samples, ranking all training samples by those distances, selecting the k high-ranked training samples (i.e., nearest k neighbors), estimating the class label distribution of such k samples, and predicting the class label of the test sample as the one with the highest distribution frequency.

2.6.4. Decision Tree. It aims to learn the human understanding classification and regression models. It generally uses IF–TEHN format to describe individual features' roles and weights in classification or regression models, thereby providing interpretative rules in a white box model. To date, several types of DT have been proposed. In this work, the CART algorithm with the Gini index was adopted to build DT model.

To quickly implement above-mentioned four classification algorithms, we employed corresponding packages collected in Scikit-learn (https://scikit-learn.org/stable/). They were executed using their default parameters.

2.7. Performance Measurement. In this study, the MCC [49] within 10-fold cross-validation [31] was used to evaluate each model's performance. A two-class classification model was obviously built here; thus, the MCC for binary problem was used as follows:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN\text{-}FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \quad (5)$$

where TP, TN, FP, and FN represent the sample numbers with true-positive, true-negative, false-positive, and false-negative predictions, respectively. The MCC value ranges from −1 to +1. When one classification model has the best performance, its MCC achieves +1.

Besides, we further computed other measurements to fully assess the performance of models, including sensitivity (SN) (same as recall), specificity (SP), accuracy (ACC), precision, and $F1$-measure. They can be calculated by

$$\begin{cases} \mathrm{SN} = \mathrm{Recall} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{SP} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{ACC} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{Precision} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\ F1\text{-measure} = \dfrac{2 \times \mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}}. \end{cases} \quad (6)$$

2.8. Feature Frequency Visualization. Each feature was related to four sites in the sequence to understand the biological meaning of the extracted sequence features. After the optimum features for one classification algorithm were obtained, the related sites of each feature were picked up, and the frequency of each site was counted and plotted as a bar illustration.

3. Results

In this study, we adopted the features of k-mers yielded by RNA2Vec to represent m^5C sites. Some machine learning algorithms were employed to analyse these features and further build efficient models for identifying m^5C site in RNA sequences. The whole procedures are shown in Figure 1. The detailed results were described in this section.

3.1. Selection of m^5C Methylation-Associated Features for Mouse. For mouse m^5C data, the mRMR method was first employed to analyse all 1020 features. An mRMR feature list was obtained. This list was fed into the IFS method that integrated one of four classification algorithms. On each feature subset, a model was built based on one classification algorithm and was further evaluated by 10-fold cross-validation. The performance of each model, including SN, SP, ACC, MCC, precision, and $F1$-measurem is provided in Supplementary file S1. MCC was selected as the key measurement. Accordingly, a curve is plotted in Figure 2 for each classification algorithm,

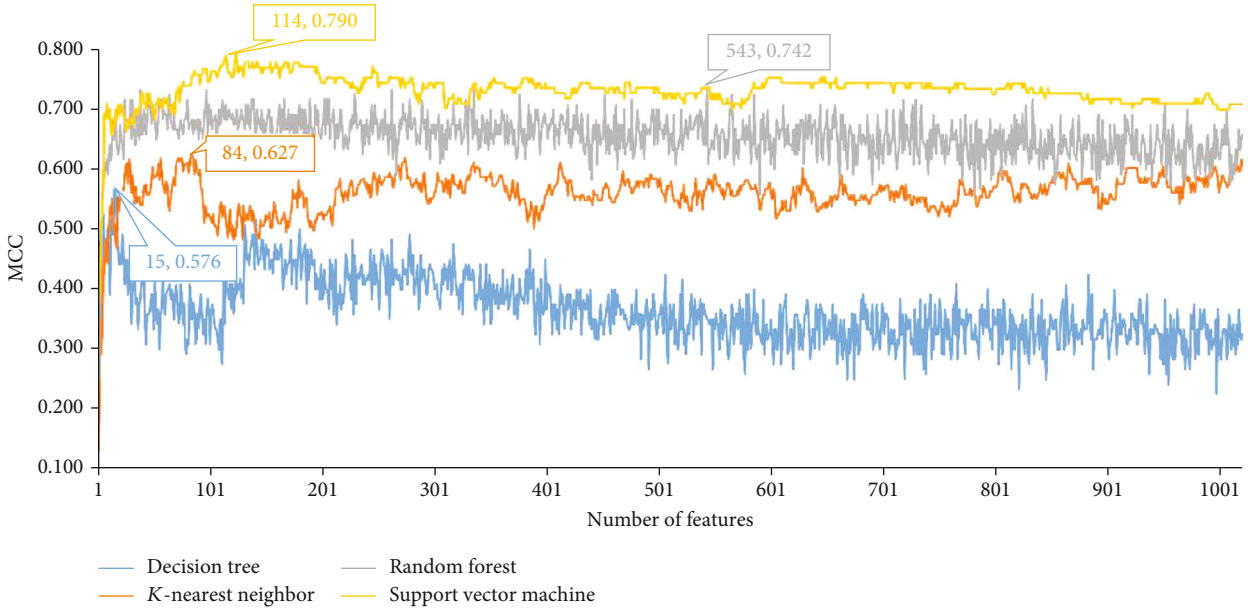

Figure 3: IFS curves with different classifiers on different numbers of sequence features on human m5C data.

Table 2: Performance of models based on different classification algorithms for predicting human m^5C sites.

Classification algorithm	Number of features	SN	SP	ACC	MCC	Precision	*F*1-measure
Decision tree	15	0.767	0.808	0.788	0.576	0.800	0.783
K-nearest neighbor	84	0.683	0.925	0.804	0.627	0.901	0.777
Random forest	543	0.875	0.867	0.871	0.742	0.868	0.871
Support vector machine	114	0.825	0.958	0.892	0.790	0.952	0.884

which defined MCC as *y*-axis and number of features as the *x*-axis. For kNN, RF, and SVM, they can provide perfect performance with MCC = 1 when top 3, 10, and 3 features were adopted. The corresponding optimum kNN/RF/SVM model can be built with these features. The detailed performance of these models is listed in Table 1. All measurements reached the maximum of 1.000. For DT, the highest MCC was 0.990, which can be obtained by using top 195 features. Accordingly, the optimum DT model was set up with these features. Its detailed performance is listed in Table 1. It can be observed that all measurements were very high. All these indicated that the models with features yielded by RNA2Vec were quite efficient for identification of mouse m^5C sites, also confirming the utility of these features to predict mouse m^5C sites.

3.2. Selection of m^5C Methylation-Associated Features for Human. For human m^5C data, the same procedures were conducted. The performance of four classification algorithms on all possible feature subsets is provided in Supplementary file S2. Similarly, one curve was plotted for each classification algorithm (as shown in Figure 3). It can be observed that four classification algorithms yielded the highest MCC values of 0.576, 0.627, 0.742, and 0.790, respectively. Such performance was obtained by using top 15, 84, 543, and 114 features. Accordingly, optimum DT/kNN/RF/SVM model can be set up with these features. The detailed performance of these models is listed in Table 2. Evidently, the performance of these models was much lower than that of models for mouse.

3.3. Feature Frequency Analysis. The purpose of this study was not only to set up efficient models for prediction of m^5C sites but also to discover novel patterns around the m^5C sites, thereby providing more biological insights. Thus, we conducted feature frequency analysis in this section.

For mouse m^5C data, four optimum models were built, which adopted some top features in the list. For each model, the number of selected features related to each site was counted. A bar chart was plotted to display such number of each site (as shown in Figure 4). Detailed discussion would be given in section "m5C Methylation-Associated Features in Mouse."

For human m^5C data, we conducted the same operations. For each optimum model, the number of selected features related to each site is shown in Figure 5. Evidently, Figures 4 and 5 displayed quite different patterns, indicating the difference between mouse and human m^5C sites. In section "m5C Methylation-Associated Features in Human," a discussion would be given.

3.4. Comparison with Previous Models. This study used the mouse and human m^5C data reported in [22]. In that study,

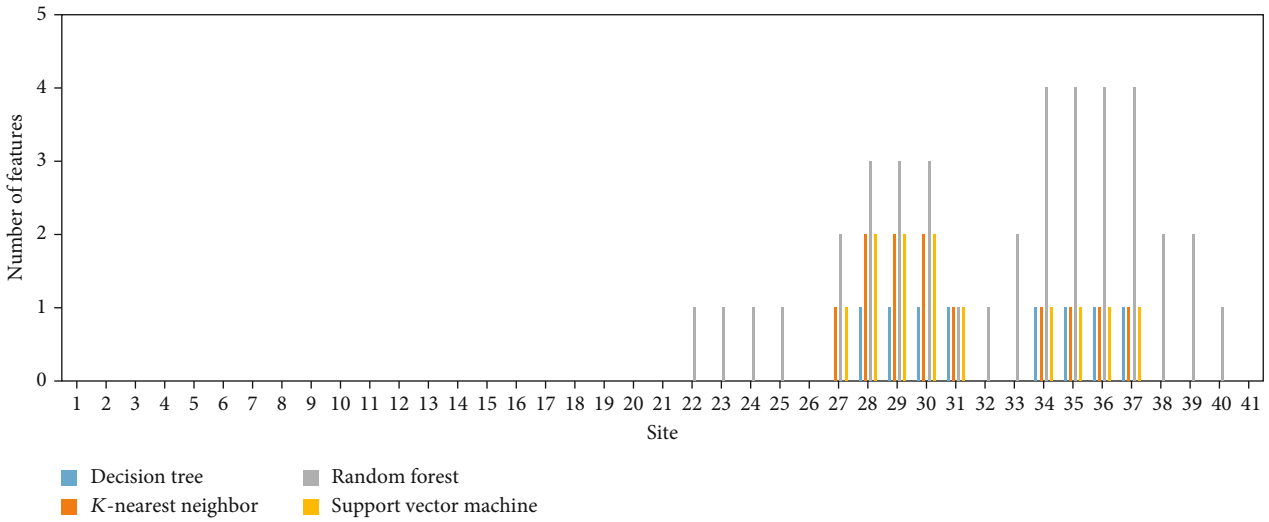

Figure 4: Frequency visualization for sequence features related to mouse m^5C.

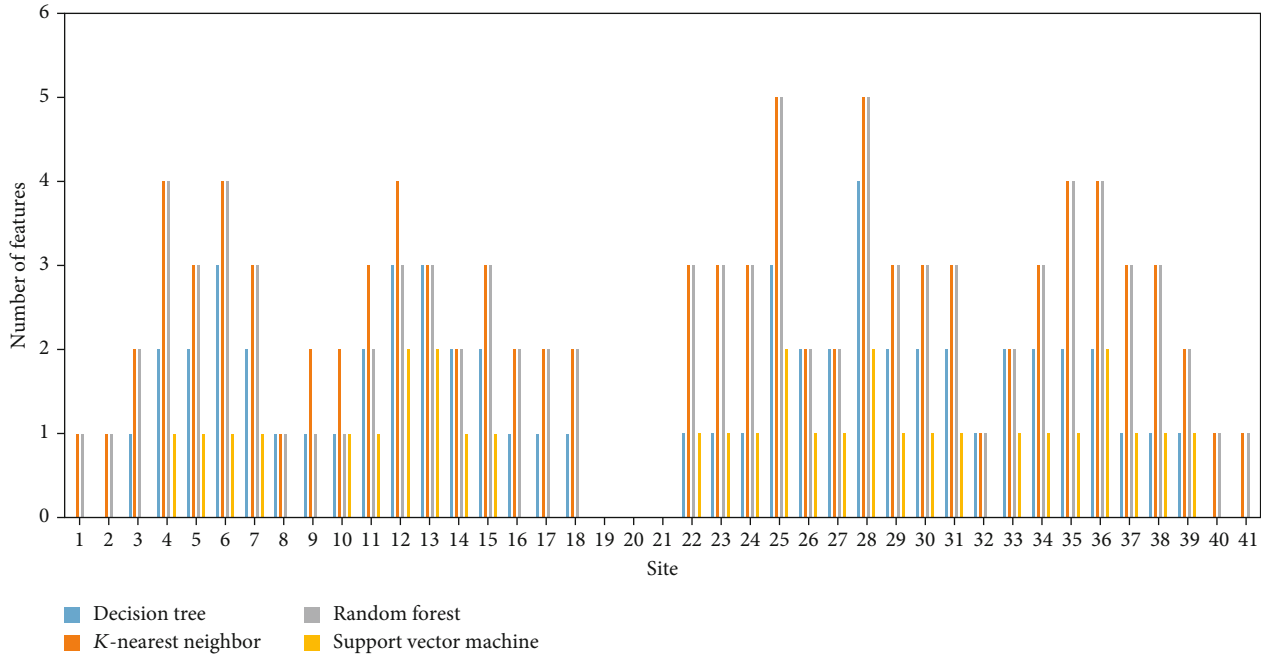

Figure 5: Frequency visualization for sequence features related to human m^5C.

several models with different classification algorithms were built and evaluated by 10-fold cross-validation, including DT, RF, SVM, Naïve Bayes, Bayes net, and logistic regression. The performance of models with DT, RF, and SVM is listed in Tables 3 and 4. For easy comparison, the performance of our models with same classification algorithms is also provided in these two tables. For mouse m^5C data, our model with DT was slightly superior to the model in [22] with the same classification algorithm. As for other two classification algorithms, all models with one of them gave perfect performance. For human m^5C data, DT provided better performance in our model than the model in [22], whereas other two classification algorithms generated lower performance in our model than the model in [22]. However, the gap was not very big. As a whole, our models and those in [22] were almost at the same level.

As mentioned in the above section, the purpose of this study further included the discovery of special patterns around m^5C sites. This was the exclusive contributions of this study compared with the previous study.

4. Discussion

Multiple machine learning models were used to distinguish samples/sites with or without a different kind of RNA methylation (human or mouse), focusing on the significant

Table 3: Comparison with previous models on mouse m^5C data.

Classification algorithm	Model	SN	SP	ACC	MCC
Decision tree	Our model	1.000	0.990	0.995	0.990
	Model in [22]	1.000	0.835	0.918	0.847
Random forest	Our model	1.000	1.000	1.000	1.000
	Model in [22]	1.000	1.000	1.000	1.000
Support vector machine	Our model	1.000	1.000	1.000	1.000
	Model in [22]	1.000	1.000	1.000	1.000

Table 4: Comparison with previous models on human m^5C data.

Classification algorithm	Model	SN	SP	ACC	MCC
Decision tree	Our model	0.767	0.808	0.788	0.576
	Model in [22]	0.783	0.783	0.783	0.567
Random forest	Our model	0.875	0.867	0.871	0.742
	Model in [22]	0.900	0.917	0.908	0.817
Support vector machine	Our model	0.825	0.958	0.892	0.790
	Model in [22]	0.842	0.967	0.904	0.815

pattern of RNA methylation as m^5C [50–52]. With the help of IFS, the optimal number of essential features was selected for RNA methylation prediction. The distribution of predicted features in the 41 nt sequence was summarized to evaluate the discriminative contributions of different RNA loci for RNA methylation [53]. The detailed analyses on the results of m^5C methylation in mouse or human tissues could be seen below, along with their respective distribution patterns.

4.1. m^5C Methylation-Associated Features in Mouse. Multiple physiochemical features were used to encode the 41 nt sequence [53] of RNA. For the evaluation of the differential contribution of RNA sites for m^5C methylation, four machine learning models were applied (DT, KNN, RF, and SVM) to identify the optimal combination of features for m^5C methylation prediction. The distribution of features' respective RNA loci is shown in Figure 4. As identified from the feature distribution, all the selected features belong to the back end of the selected sequence, from 23 nt to 41 nt, just behind the candidate m^5C methylation site (21 nt). In particular, two regions (27–31 nt and 34–37 nt) were predicted by at least three machine learning models to be associated with m^5C methylation. According to recent publications based on the biological functions of m^5C, the two kinds of m^5C sites in multiple subgroups of RNAs are (1) type I m^5C, which is followed by a G-rich triplet motif, and (2) type II m^5C, which is adjacent to a downstream UCCA motif; both have specific sequence characteristics in the following region of m^5C methylation loci [54], which corresponded with the prediction results in the present study. Further studies have also confirmed that specific regions in the downstream of m^5C loci may have different sequence contexts, indicating that the feature-enriched regions in the prediction list in the present study could definitely be associated with m^5C methylation efficiency. In 2019, a systematic analyses on mRNA 5-methylcytosine in mammals identified that the sequence context at the downstream of the captured m^5C loci was alternate with different m^5C locus methylation status, regulated by a specific 5-methylcytosine methyltransferase called NSUN2 [55, 56]. For comparison, the sequence before the m^5C loci in mouse did not considerably change with NSUN2 wild-type, knock-out, or rescue status, implying that the m^5C loci and their downstream sequence, especially for the following 10 nt sequence [55, 56], which corresponded with the prediction distribution in the present study. In addition, another similar 5-methylcytosine methyltransferase NSUN6 in mouse functioned as an mRNA m^5C methyltransferase [54]. As a methyltransferase of type II m^5C, the m^5C targets of such gene have a symbolic downstream UCCA tail located at the first ambiguous peak (only predicted via RF method) in the prediction result of the present study (1–4 nt following the methylation region) [54]. Furthermore, different from the biological regulatory effects of NSUN2, the flanking regions around 15 nt were found to have another low base-pairing regions, which include more variants, by using the same procedure that detects the sequences with methyltransferase knock-out, rescue, and wild-type statuses [54]. This finding indicated the importance of sequence around such region. All in all, the predicted distribution of m^5C methylation-associated loci has been validated by recent publications.

4.2. m^5C Methylation-Associated Features in Human. The m^5C-associated feature distribution among 40 flanking sequences (20 downstream and 20 upstream) from human

tissues was also identified. According to the same publications [54, 55], the following 1–4 nt (22–26 nt) and 13–15 nt (34–37 nt) were also associated with the efficacy of m^5C methylation, which corresponded with the prediction of the present study. As seen in Figures 4 and 5, the feature peaks in the downstream region (21–41 nt) were quite similar between the human and mouse data, reflecting the similarity of m^5C methylation-associated patterns among different species. However, obvious differences were also observed, implying the presence of biological differences in m^5C methylation among different species. In human beings, recent publications revealed that the distribution of RBP (RNA-binding protein) target density, which reflects the binding efficacy of the related region, was significant at the m^5C candidate site, and gradually, not suddenly going down in both directions [56, 57]. Therefore, the sequences around m^5C in each direction may also be not randomized but with specific sequence characteristics. Further, in 2015, an analysis on the regulatory homologous proteins of yeast and human from the same protein family (Nop2/NSUN/NOL family) showed that specific binding domains (e.g., SAM-binding domain) may be located behind the m^5C loci, and they may affect regulatory effects. Therefore, although they were not directly validated, some nucleotides located before the m^5C loci may be essential for the prediction of methylation status [58].

4.3. Biological Significance of Identified m^5C Methylation-Associated Features. As summarized above, we identified m^5C-associated features in mouse and human. The biological significance of identified m^5C methylation features can be clustered into two parts:

(i) The specific and diverse distribution of m^5C associated features in human or mouse. In this part, we identified that mouse m^5C methylations are generally only associated with 28-31 nt and 34-37 nt regions in the 41 nt subsequence, while in human tissues, apart from 19-21 nt regions, most positions of the 41 nt sequence are associated with m^5C methylation. These results identified key regulatory regions associated with m^5C methylation and the differences between regulatory effects on m^5C methylation in different species, reflecting the evolution conservation of m^5C methylation regulatory mechanisms

(ii) The downstream regulatory network associated with m^5C methylation is essential for gene transcription and translation. Generally, m^5C methylation can help bind hydrogen with guanine to stabilize the complete RNA structures and fold into unique spatial conformation [59]. According to recent publications, m^5C regulator *NSUN2* has been shown to alter m^5C capacity in certain RNA regions. Genes like *p27* (*KIPI*), *CDK1*, *p21*, and *ErbB2* have all been shown to be regulated by m^5C methylation and further related to tumorigenesis [59, 60]. The sequence loci of m^5C methylation have been shown to be specifically affects the downstream cell proliferation and inflammation associated pathway [61, 62], indicating the specific biological significance of m^5C methylation. Therefore, the identification of different contribution of nucleotide from different sequence location can help demonstrate the specific regulatory effects for abnormal m^5C methylation during different pathogenic conditions

Therefore, the identification of loci-related characters regulating m^5C methylation between different species can not only help us reveal the consistence and evolution conservation of m^5C methylation associated sequences but also connect specific sequence loci with significant m^5C methylation-associated phenotypes or diseases.

5. Conclusions

All in all, as discussed above, the top optimal methylation sites in the prediction list have been supported by recent publications. The RNA methylation patterns were validated to be different in multiple species by comparing the results of m^5C methylation-associated loci in human and mouse tissues. The discriminative feature distribution patterns for different methylation patterns were also detected by comparing the results of m^5C distribution patterns. Therefore, the results not only evaluated the discriminative contribution of different loci for important RNA methylation patterns but also revealed the site distribution differences of m^5C methylation types between species (human and mice).

Data Availability

The original data used to support the findings of this study are available at iRNA-m^5C (http://lin-group.cn/server/iRNA-m5C/download.html).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper."

Authors' Contributions

Lei Chen, ZhanDong Li, and ShiQi Zhang contributed equally to this work.

Acknowledgments

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA26040304 and XDB38050200), National Key R&D Program of China (2018YFC0910403), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

Supplementary Materials

Supplementary Material S1: performance of IFS for mouse m^5C sites. Supplementary Material S2: performance of IFS for human m^5C sites. *(Supplementary Materials)*

References

[1] M. V. Greenberg and D. Bourc'his, "The diverse roles of DNA methylation in mammalian development and disease," *Nature Reviews Molecular Cell Biology*, vol. 20, no. 10, pp. 590–607, 2019.

[2] H. Shi, J. Wei, and C. He, "Where, when, and how: context-dependent functions of RNA methylation writers, readers, and erasers," *Molecular Cell*, vol. 74, no. 4, pp. 640–650, 2019.

[3] X. Li, X. Xiong, and C. Yi, "Epitranscriptome sequencing technologies: decoding RNA modifications," *Nature Methods*, vol. 14, no. 1, pp. 23–31, 2017.

[4] K. du, L. Zhang, T. Lee, and T. Sun, "m6A RNA methylation controls neural development and is involved in human diseases," *Molecular Neurobiology*, vol. 56, no. 3, pp. 1596–1606, 2019.

[5] A. M. Heck and C. J. Wilusz, "Small changes, big implications: the impact of m^6A RNA methylation on gene expression in pluripotency and development," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1862, no. 9, article 194402, 2019.

[6] A. K. Chokkalla, S. L. Mehta, and R. Vemuganti, "Epitranscriptomic regulation by m6A RNA methylation in brain development and diseases," *Journal of Cerebral Blood Flow & Metabolism*, vol. 40, no. 12, pp. 2331–2349, 2020.

[7] C. Gu, X. Shi, C. Dai et al., "RNA m6A modification in cancers: molecular mechanisms and potential clinical applications," *The Innovation*, vol. 1, no. 3, article 100066, 2020.

[8] S. Zaccara, R. J. Ries, and S. R. Jaffrey, "Reading, writing and erasing mRNA methylation," *Nature Reviews Molecular Cell Biology*, vol. 20, no. 10, pp. 608–624, 2019.

[9] K. W. Min, R. W. Zealy, S. Davila et al., "Profiling of m6A RNA modifications identified an age-associated regulation of AGO 2 mRNA stability," *Aging Cell*, vol. 17, no. 3, article e12753, 2018.

[10] Y. Lee, J. Choe, O. H. Park, and Y. K. Kim, "Molecular mechanisms driving mRNA degradation by m6A modification," *Trends in Genetics*, vol. 36, no. 3, pp. 177–188, 2020.

[11] S. Lesbirel and S. A. Wilson, "The m^6A-methylase complex and mRNA export," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1862, no. 3, pp. 319–328, 2019.

[12] J.'. Liu, K. Li, J. Cai et al., "Landscape and regulation of m^6A and m^6Am methylome across human and mouse tissues," *Molecular Cell*, vol. 77, no. 2, pp. 426–440.e6, 2020.

[13] I. A. Roundtree, M. E. Evans, T. Pan, and C. He, "Dynamic RNA modifications in gene expression regulation," *Cell*, vol. 169, no. 7, pp. 1187–1200, 2017.

[14] M. Han, Z. Liu, Y. Xu et al., "Abnormality of m6A mRNA methylation is involved in Alzheimer's disease," *Frontiers in Neuroscience*, vol. 14, p. 98, 2020.

[15] Z. Zhang, L. Q. Chen, Y. L. Zhao et al., "Single-base mapping of m6A by an antibody-independent method," *Science Advances*, vol. 5, no. 7, article eaax0250, 2019.

[16] L. Trixl and A. Lusser, "The dynamic RNA modification 5-methylcytosine and its emerging role as an epitranscriptomic mark," *Wiley Interdisciplinary Reviews: RNA*, vol. 10, no. 1, article e1510, 2019.

[17] X. Yang, Y. Yang, B. F. Sun et al., "5-methylcytosine promotes mRNA export – NSUN2 as the methyltransferase and ALYREF as an m^5C reader," *Cell Research*, vol. 27, no. 5, pp. 606–625, 2017.

[18] X. Wu, Z. Wei, K. Chen et al., "m6Acomet: large-scale functional prediction of individual m6A RNA methylation sites from an RNA co-methylation network," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019.

[19] Y. Xiao, J. Cai, Y. Yang, H. Zhao, and H. Shen, "Prediction of MicroRNA Subcellular Localization by Using a Sequence-to-Sequence Model," in *2018 IEEE International Conference on Data Mining (ICDM)*, Singapore, 2018IEEE.

[20] H. Peng, L. Fulmi, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Ieee Transactions On Pattern Analysis And Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[21] H. A. Liu and R. Setiono, "Incremental feature selection," *Applied Intelligence*, vol. 9, no. 3, pp. 217–230, 1998.

[22] H. Lv, Z. M. Zhang, S. H. Li, J. X. Tan, W. Chen, and H. Lin, "Evaluation of different computational methods on 5-methylcytosine sites identification," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 982–995, 2020.

[23] P. Feng, H. Ding, W. Chen, and H. Lin, "Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions," *Molecular BioSystems*, vol. 12, no. 11, pp. 3307–3311, 2016.

[24] W. J. Sun, J. H. Li, S. Liu et al., "RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data," *Nucleic Acids Research*, vol. 44, no. D1, pp. D259–D265, 2016.

[25] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[26] J. Pennington, R. Socher, and C. D. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

[27] S. Zhang, T. Zeng, B. Hu et al., "Discriminating origin tissues of tumor cell lines by methylation signatures and Dys-methylated rules," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 507, 2020.

[28] S. Zhang, X. Y. Pan, T. Zeng et al., "Copy number variation pattern for discriminating MACROD2 states of colorectal cancer subtypes," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 407, 2019.

[29] L. Chen, T. Zeng, X. Pan, Y. H. Zhang, T. Huang, and Y. D. Cai, "Identifying methylation pattern and genes associated with breast cancer subtypes," *International Journal of Molecular Sciences*, vol. 20, no. 17, p. 4269, 2019.

[30] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.

[31] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *International Joint Conference On Artificial Intelligence*, vol. 14, no. 2, 1995.

[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[34] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[35] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[36] Y.-H. Zhang, Z. Li, T. Zeng et al., "Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles," *Frontiers in Genetics*, vol. 11, article 599970, 2021.

[37] X. Pan, H. Li, T. Zeng et al., "Identification of protein subcellular localization with network and functional embeddings," *Frontiers in Genetics*, vol. 11, article 626500, 2021.

[38] Y.-H. Zhang, T. Zeng, L. Chen, T. Huang, and Y. D. Cai, "Determining protein-protein functional associations by functional rules based on gene ontology and KEGG pathway," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1869, no. 6, article 140621, 2021.

[39] Y.-H. Zhang, H. Li, T. Zeng et al., "Identifying transcriptomic signatures and rules for SARS-CoV-2 infection," *Frontiers in Cell and Developmental Biology*, vol. 8, article 627302, 2021.

[40] M. Onesime, Z. Yang, and Q. Dai, "Genomic island prediction via chi-square test and random forest algorithm," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 9969751, 9 pages, 2021.

[41] Y. Wang, Y. Xu, Z. Yang, X. Liu, and Q. Dai, "Using recursive feature selection with random forest to improve protein structural class prediction for low-similarity sequences," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 5529389, 9 pages, 2021.

[42] Y. Yang and L. Chen, "Identification of drug–disease associations by using multiple drug and disease networks," *Current Bioinformatics*, vol. 16, 2021.

[43] J.-P. Zhou, L. Chen, T. Wang, and M. Liu, "iATC-FRAKEL: a simple multi-label web server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only," *Bioinformatics*, vol. 36, no. 11, pp. 3568-3569, 2020.

[44] Y. Zhu, B. Hu, L. Chen, and Q. Dai, "iMPTCE-Hnetwork: a multi-label classifier for identifying metabolic pathway types of chemicals and enzymes with a heterogeneous network," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 6683051, 2021.

[45] Y. Jia, R. Zhao, and L. Chen, "Similarity-based machine learning model for predicting the metabolic pathways of compounds," *IEEE Access*, vol. 8, pp. 130687–130696, 2020.

[46] L. Chen, S. Wang, Y. H. Zhang et al., "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[47] H. Liu, B. Hu, L. Chen, and L. Lu, "Identifying protein subcellular location with embedding features learned from networks," *Current Proteomics*, vol. 18, no. 5, pp. 646–660, 2021.

[48] W. Chen, L. Chen, and Q. Dai, "iMPT-FDNPL: identification of membrane protein types with functional domains and a natural language processing approach," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 7681497, 2021.

[49] B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[50] L. Jia, J. Chen, H. Liu et al., "Potential m6A and m5C methylations within the genome of a Chinese African swine fever virus strain," *Virologica Sinica*, vol. 36, no. 2, pp. 321–324, 2021.

[51] Z.-X. Liu, L. M. Li, H. L. Sun, and S. M. Liu, "Link between m6A modification and cancers," *Frontiers in Bioengineering and Biotechnology*, vol. 6, p. 89, 2018.

[52] S. Sommer, M. Salditt-Georgieff, S. Bachenheimer et al., "The methylation of adenovirus-specific nuclear and cytoplasmic RNA," *Nucleic Acids Research*, vol. 3, no. 3, pp. 749–766, 1976.

[53] F. Y. Dao, H. Lv, Y. H. Yang, H. Zulfiqar, H. Gao, and H. Lin, "Computational identification of N6-methyladenosine sites in multiple tissues of mammals," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1084–1091, 2020.

[54] J. Liu, T. Huang, Y. Zhang et al., "Sequence-and structure-selective mRNA m5C methylation by NSUN6 in animals," *National Science Review*, vol. 8, no. 6, 2021.

[55] T. Huang, W. Chen, J. Liu, N. Gu, and R. Zhang, "Genome-wide identification of mRNA 5-methylcytosine in mammals," *Nature Structural & Molecular Biology*, vol. 26, no. 5, pp. 380–388, 2019.

[56] Q. Li, X. Li, H. Tang et al., "NSUN2-mediated m5C methylation and METTL3/METTL14-mediated m6A methylation cooperatively enhance p21 translation," *Journal of Cellular Biochemistry*, vol. 118, no. 9, pp. 2587–2598, 2017.

[57] J. E. Squires, H. R. Patel, M. Nousch et al., "Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA," *Nucleic Acids Research*, vol. 40, no. 11, pp. 5023–5033, 2012.

[58] G. Bourgeois, M. Ney, I. Gaspar et al., "Eukaryotic rRNA modification by yeast 5-methylcytosine-methyltransferases and human proliferation-associated antigen p120," *PLoS One*, vol. 10, no. 7, article e0133321, 2015.

[59] K. E. Bohnsack, C. Höbartner, and M. T. Bohnsack, "Eukaryotic 5-methylcytosine (m5C) RNA methyltransferases: mechanisms, cellular functions, and links to disease," *Genes*, vol. 10, no. 2, p. 102, 2019.

[60] S. Xiang, Y. Ma, J. Shen et al., "m5C RNA methylation primarily affects the ErbB and PI3K–Akt signaling pathways in gastrointestinal cancer," *Frontiers in Molecular Biosciences*, vol. 7, 2020.

[61] W.-H. Lee, R. A. Morton, J. I. Epstein et al., "Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis," *Proceedings of the National Academy of Sciences*, vol. 91, no. 24, pp. 11733–11737, 1994.

[62] Y. Li, J. Li, M. Luo et al., "Novel long noncoding RNA NMR promotes tumor progression via NSUN2 and BPTF in esophageal squamous cell carcinoma," *Cancer Letters*, vol. 430, pp. 57–66, 2018.