Computational and Structural Biotechnology Journal 19 (2021) 5510–5524

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Machine learning applications in RNA modification sites prediction

A. El Allali [a,*], Zahra Elhamraoui [a], Rachid Daoud [a]

[a] *African Genome Center, University Mohamed VI Polytechnic, Morocco*

A R T I C L E I N F O

Article history:
Received 30 June 2021
Received in revised form 24 September 2021
Accepted 25 September 2021
Available online 29 September 2021

Keywords:
RNA modification
Machine learning
Deep learning
Feature extraction
Prediction

A B S T R A C T

Ribonucleic acid (RNA) modifications are post-transcriptional chemical composition changes that have a fundamental role in regulating the main aspect of RNA function. Recently, large datasets have become available thanks to the recent development in deep sequencing and large-scale profiling. This availability of transcriptomic datasets has led to increased use of machine learning based approaches in epitranscriptomics, particularly in identifying RNA modifications. In this review, we comprehensively explore machine learning based approaches used for the prediction of 11 RNA modification types, namely, m^1A, m^6A, m^5C, $5hmC$, ψ, $2\prime - O - Me$, $ac4C$, m^7G, $A - to - I$, m^2G, and D. This review covers the life cycle of machine learning methods to predict RNA modification sites including available benchmark datasets, feature extraction, and classification algorithms. We compare available methods in terms of datasets, target species, approach, and accuracy for each RNA modification type. Finally, we discuss the advantages and limitations of the reviewed approaches and suggest future perspectives.

Contents

[*] Corresponding author.
E-mail address: achraf.elallali@um6p.ma (A. El Allali).

https://doi.org/10.1016/j.csbj.2021.09.025
2001-0370/

1. Introduction

RNA modification refers to post-transcriptional modifications of RNA in eukaryotes and prokaryotes. Currently, more than 150 types of RNA modifications have been discovered. The most dominant is RNA methylation, which is often referred to as epitranscriptome. Common modification types include, N^6-methyladenosine (m^6A), pseudouridine (ψ), N^1-methyladenosine (m^1A), 5-hydroxymethylcytosine ($5hmC$), 5-methylcytosine (m^5C), and 2'-O-methylation of ribose ($2\prime - O - Me$). Other less common types, such as N4-acetylcytidine ($ac4C$), 7-methylguanosine (m^7G), N^2-methylguanosine (m^2G), dihydrouridine(D), and adenosine-to-inosine (A-to-I) also exist (Fig. 1) [1].

m^6A modification occurs at a different stage of mRNA metabolism, including ncRNA processing and CircRNA translation [2]. This modification occurs at the nitrogen-6 position of the adenine base (A). There is convincing evidence that m^6A methylation plays a significant role in several pathological and physiological immune responses, including the homeostasis and differentiation of T cells, inflammation, and the development of interferon type I [3]. m^1A is another methylation type that affects adenine. Most of what we know about m^1A originated from the analysis of tRNAs and rRNAs since its presence in mRNAs was only recently discovered. A methyl group is attached to the nitrogen atom at the first position of the adenine base in m^1A reaction. Research has shown that m^1A is related to respiratory chain malfunctioning and neurodevelopmental regression [4]. Another modification affecting adenine is A-to-I modification, the primary type of RNA modification in mammals. Depending on the environments in which RNA modification events occur, this deamination of adenine could influence gene regulation, expression, and functions concerning exchanges in the amino acid sequence and maturation [5].

Modifications that affect the cytosine base of RNA include $m^5C, ac4C$, and 5hmC. Recent research suggests that m^5C methylation influences the nuclear export efficiency of mRNA by impacting the function of the nuclear factor ALYREF/THOC4 [6]. It is also possible that m^5C has an impact on protein translation. It is established that m^5C is reversible, and TET-mediated oxidation of m^5C produces another form of RNA modification known as 5-hydroxymethylcytosine (5hmC) [7]. Recently, $ac4C$ modification was found in yeast and identified in human mRNA. It is the only acetylation modification found on cytidine in eukaryotic mRNA and of all the modifications that have regulatory potential. $ac4C$ is retained in all areas of life [8].

At the uridine (U) level, several modifications can occur, such as ψ. It was the first type of RNA modification to be discovered, and it remains the most common one. It is a C5-glycoside isomer of uridine and the only C–C bond connecting a nucleobase to a sugar known among nucleic acids. Although it was discovered in the 1950s, neither the enzymatic mechanism of its formation nor its function has been fully described. $2\prime - O - Me$ is a type of modification that alters ribose [9] and one of the most abundant types of RNA modification in various cellular RNAs. This modification occurs when the methyl group (-CH3) is attached to the 2'hydroxyl (-OH) of the ribose moiety, which can modify any ribonucleotide (Am, Cm, Gm, Um) [10]. The $2\prime - O - Me$ can protect RNAs from nuclease attacks, increase their hydrophobicity and affect their reactions with proteins or various RNAs [11].

m^7G is a positively charged RNA modification that occurs at the guanine (G) nucleotide. m^7G is among the most conserved modifications in tRNA, rRNA, and mRNA 5'cap [12]. It is an adjusted purine nucleoside that plays an essential role in regulating RNA function and cell viability. Even with its propagation within internal mRNA regions at the transcriptional level, its regulation remains largely unresolved. Dihydrouridine (D) is another typical RNA post-transcriptional modification in eukaryotes, bacteria, and some archaea. Because of this change, individual nucleotide bases can have more conformational versatility, and cancerous tissues have higher levels of this post-transcription modification [13].

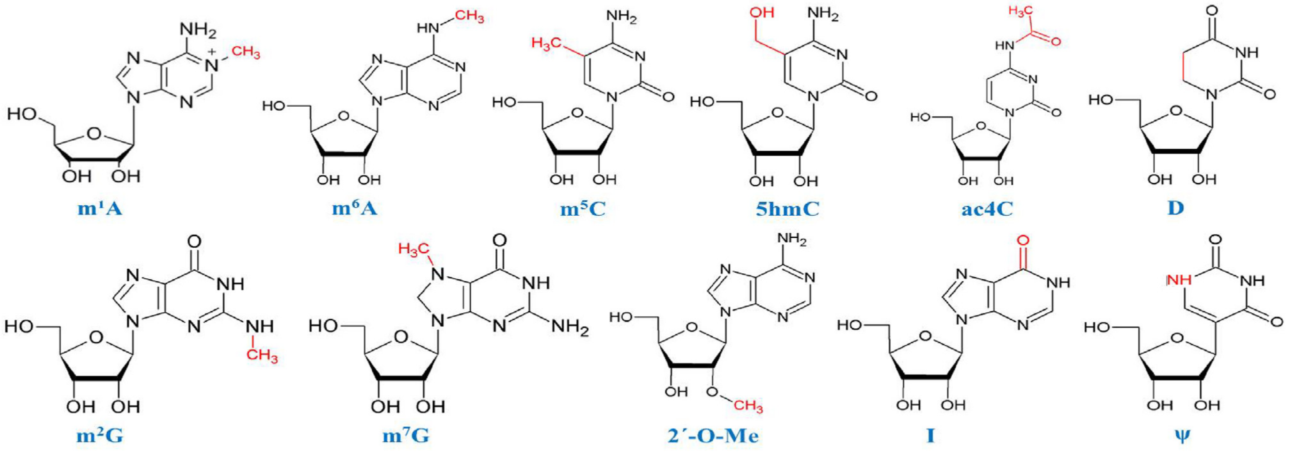

Fig. 1. Chemical structures of popular RNA modifications.

Table 1
Different databases used in post-transcriptional RNA modifications.

Database	Available RNA modification dataset	Link
RMBase V2 [1]	ψ (4128 *H. sapiens*, 3320 *M. musculus*, 2122 *S. cerevisiae*…) m^6A (477 452 *H. sapiens*, 490 704 *M. musculus*, 67 671 *S. cerevisiae*…) m^5C (680 *H. sapiens*, 97 *M. musculus*, 211 *S. cerevisiae*…), and other types	http://rna.sysu.edu.cn/rmbase/
MODOMICS [18]	ψ, m^6A, m^5C, $2\prime - O - Me$, m^1A, $m3C$, and other Types	http://modomics.genesilico.pl
DARNED [19]	A-to-I and other types	http://darned.ucc.ie/
RNAMDB [2]	M1Gm, m^1A, tm5U, tm5s2U, ac6A, and other types	Not available
m^6A-Atlas [20]	ψ, m^6A, m^5C, $2OMe$, m^1A, $m3C$ and other Types	http://www.xjtlu.edu.cn/biologicalsciences/atlas
REDIportal [21]	A-to-I	http://srv00.recas.ba.infn.it/atlas/
RADAR [22]	A-to-I (1 379 403 human, 8108 *M. musculus*, 2698 fly)	Not available
MeT-DB V2.0 [23]	m^6A	Not available
REPIC [24]	m^6A	https://repicmod.uchicago.edu/repic/
CVm6A [25]	m^6A	http://gb.whu.edu.cn:8080/CVm6A
ConsRM [26]	m^6A	http://180.208.58.19/conservation/
Gene-Expression Omnibus (GEO) [17]	Available as experimental data	https://www.ncbi.nlm.nih.gov/geo/
m7GHub [27]	m^7G	Not Available

RNA modification can be detected using several RNA biochemistry approaches, including different types of chromatography, such as TLC, HPLC, and LC-MS/MS [14]. It can also be detected based on the chemical properties of RNA modification, including RTL-P (Reverse Transcription at low [dNTP] with PCR) approach based on reverse transcription at low dNTP concentration, as well as a PCR approach [15]. There are also deep sequencing approaches for RNA modification identification. These approaches utilize next-generation sequencing (NGS) and computational techniques to identify RNA modification [16].

Even though they are time and labor-consuming, biochemical and deep sequencing approaches for identifying RNA modification sites yield effective results. Due to the increase in sequencing data generated in the post-genomic era, computational algorithms have surfaced to perform *in silico* prediction of RNA modification sites. In 2013, the first machine learning-based approach was introduced to detect m^6A sites. Later, multiple computational models were introduced to identify various RNA modifications. Machine learning approaches extract features from known benchmark datasets and build models that can predict modification sites. In this review, we explore machine learning approaches that predict the most common RNA modification sites. First, we introduce benchmark datasets used to predict post-transcription modifications. We then combine the most common statistical features used in machine learning approaches and summarize the feature selection methods used to identify modification sites. Finally, we review existing machine learning tools for each modification type in terms of methodology, performance, and accessibility.

2. Benchmark datasets

Perhaps the most important requirement when applying machine learning to modification sites prediction is the accessibility to curated data in order to achieve the best results. Most available benchmark datasets in the field are acquired from Gene Expression Omnibus (GEO) [17]. GEO is a database that contains experimental studies for gene expression and RNA modification profiling. Recently multiple databases were built for post-transcriptional modifications such as MODOMICS [18]. This database hosts a collection of modified RNA sequences, where modification records are highlighted. RMBase [2] is another popular RNA-modified database based on high throughput genome-wide data from 18 independent studies. RMBase V2.0 [1] is a 2017 update with more enhanced known RNA modification sites. These databases are used for constructing small benchmark datasets for machine learning based approaches for different modifications. Table 1 summarizes the most popular databases used in constructing RNA modification benchmark datasets. In order to build machine learning models, training datasets use experimentally identified RNA modifications as positive samples, while negative samples are randomly selected from non-modified samples. However, some of the negative samples used in training can be unidentified sites and introduce noise to the models.

3. Feature extraction

Feature extraction is an essential step in machine learning. In most cases, prediction models are as good as the features they use to distinguish between prediction classes. There are several features used to predict modification sites in RNA sequences. For example, K-nucleotide frequencies (KNF) [28] represent the frequencies of all possible k-nucleotides in each sequence. Position-specific dinucleotide sequence profile (PSDSP) [28] measures the difference of dinucleotide profiles between positive and negative samples. Other composition-based features such as the composition of k-spaced nucleic acid pairs (CKSNAP) and Ksnpf [29,24] rely on the frequencies of nucleotide pairs separated by k- residues. Dinucleotide-based auto-cross covariance (DACC) [30] incorporates information about the overall order of RNA sequences, while Enhanced nucleic acid composition (ENAC) [29] calculates the nucleic acid composition based on a fixed-length window.

Sequence and genome derived features are used by several algorithms, including binary variables that indicate whether modification sites are overlapped with the topological regions on the major RNA transcript or not, relative position on the region, the region length in bp, clustering information of modification sites from training data, scores related to evolutionary conservation, RNA secondary structure prediction score such as the minimum free energy (MFE), RNA binding protein annotation from MeTDB database, and gene characteristics such as annotation [12,31]. In addition to the features mentioned above, several more common features are further detailed in the following sections.

3.1. Nucleotide chemical property

Nucleic acids have common and distinctive properties in terms of the ring structure, hydrogen bond, and functional group. For example, adenine and guanine have a similar doubled ring structure, while cytosine and uracil contain a single ring. In addition, when forming a secondary structure, cytosine and guanine are strongly hydrogen-bonded, whereas adenine and uracil are weakly hydrogen-bonded. Chemical functionality allows us to divide nucleotides into two groups. Cytosine and adenine belong to the amino group, while guanine and uracil belong to the keto group. Given these different chemical properties, we can generally divide nucleotides into three distinct groups. Membership of each nucleotide in these groups allows the encoding of RNA sequences using a tuple (a,b,c) for each base. Therefore, A is represented as (1,1,1), C is represented as (0,1,0), G is represented as (1,0,0) and U is represented as (0,0,1) [32].

3.2. One-hot encoding

One-hot encoding techniques are often used to represent DNA and RNA sequences to transform categorical data to numerical form [33]. In particular, Adenine is encoded as (1,0,0,0), Uracil as (0,0,0,1), Cytosine as (0,1,0,0), and Guanine as (0,0,1,0). A matrix of 4 by n is then used to represent an RNA sequence of length n.

3.3. PseEIIP

Electron–Ion Interaction Pseudopotentials (EIIP) represent the energy of delocalized electrons in nucleotides. They have been used as a composition measure that has been effective in several bioinformatics algorithms [34,35]. Using the EIIP technique, every nucleotide in an RNA sequence is encoded using the distribution of free electron energies. The adenine, cytosine, guanine, and uracil values are 0.1260, 0.1340, 0.0806, and 0.1335, respectively. In order to apply pseudo-EIIP (PseEIIP) to trinucleotides, the mean EIIP value for each nucleotide is used as the composition measure.

$$\text{PseEIIP} = [EIIP_{AAA} \cdot f_{AAA}, EIIP_{AAC} \cdot f_{AAC}, \ldots, EIIP_{TTT} \cdot f_{TTT}]$$

3.4. K-mer content

K-mer content is a common feature extracted from nucleotide sequences and used in several prediction algorithms. The composition measure represents frequencies of all sub-sequences of length k (k = 1, 2, 3 ...) in a given sequence. For example, 1-mer contains single nucleotide frequencies, 2-mer represents dinucleotide frequencies, and 3-mer represents the frequency of triple nucleotides. K-mer content is used in conjunction with the position to enhance sensitivity to the position-specific sequence slope around RNA modification sites. This measure is often referred to as a position weight matrix (PWM) [33].

3.5. PseDNC and PseKNC measures

Another alternative to k-mer content is to formulate a dynamic coding scheme that simultaneously represents small local sequence patterns and global data arrangement. Due to the success of pseudo amino acid composition (PseAAC) in protein and peptide-related problems, another approach called pseudo dinucleotide composition (PseDNC) was introduced to encode modification in RNA sequences. The discrete measure describes the physical and chemical properties of oligonucleotides with the sequence. This type of pseudo-synthesis can retain much of the sequence-order information, especially the global or long-range sequence-order information. Compared to one-hot or k-mer encoding, both PseDNC and PseKNC associate short sequence order information with RNA physical–chemical properties. Type 2 PseKNC reflects both the local and the global sequence information of the nucleotide sequence.

4. Machine Learning

Machine Learning is a branch of artificial intelligence that involves computer self-learning to perform tasks automatically and is divided into three types: supervised learning, unsupervised learning, and reinforced learning. Supervised learning includes mapping an input to an output based on a set of known data. The output is either a class in the case of classification or a value in regression. Unsupervised learning involves studying the class itself by building models that are capable of describing the data and the relationships found in the data without the use of labels. Unsupervised learning includes dividing data into groups in the case of clustering and summarizing data distribution in density estimation. Reinforcement learning involves learning actions rather than class, and the input is mapped to actions based on feedback. The learning, in this case, is action-oriented, where the most rewarding actions are retained. Hybrid learning approaches combine supervised and unsupervised learning. When the training does not contain enough labeled data, semi-supervised learning builds models that take advantage of available unlabeled data. Self-learning is another hybrid approach that treats unsupervised problems as a supervised one. In this case, models are trained from related data and then applied to the original unlabeled data for prediction. Multi-instance learning is a supervised learning approach that uses unsupervised concepts in learning. This hybrid approach groups unlabeled data into labeled groups using unsupervised learning and uses the resulting labeled data in a supervised way. In our review, we are concerned with supervised learning and particularly classification. The prediction algorithms we studied use benchmark data with known RNA modification sites to predict whether novel RNA sequences have RNA modifications. Several pipelines exist to train a machine learning model and conduct a classification experiment. Fig. 2 shows an overview of the steps involved in performing supervised learning.

4.1. Feature Selection

Feature selection is a crucial step in machine learning as it has a direct role in the performance of prediction models. During feature selection, redundant and noisy features are illuminated, reducing overfitting and increasing the calculation speed. Feature selection is used to reduce the feature space by choosing the most relevant and discriminant features. For example, $BFS + LF$ [36] is an incremental feature selection method. Features that improve classification using a logistic function (LF) are kept while those that decrease the accuracy are removed. In addition, strategies such as principal component analysis (PCA) [37], n-Grams, minimal-redundance maximum-Relevance (mRMR) [38] are widely used in order to select a subset of the features. Studies show that applying feature selection algorithms produce better performance than using the extracted features directly or applying a multi-layer machine learning approach [39,40]. However, feature selection should be performed on a different dataset than the training to avoid biases in the performance analysis during testing. If the data is not enough, existing algorithms can help avoid feature selection bias [41,42].

4.2. Machine Learning Algorithms

Prediction algorithms learn a target function that best maps input variables to an output variable. The function is referred to

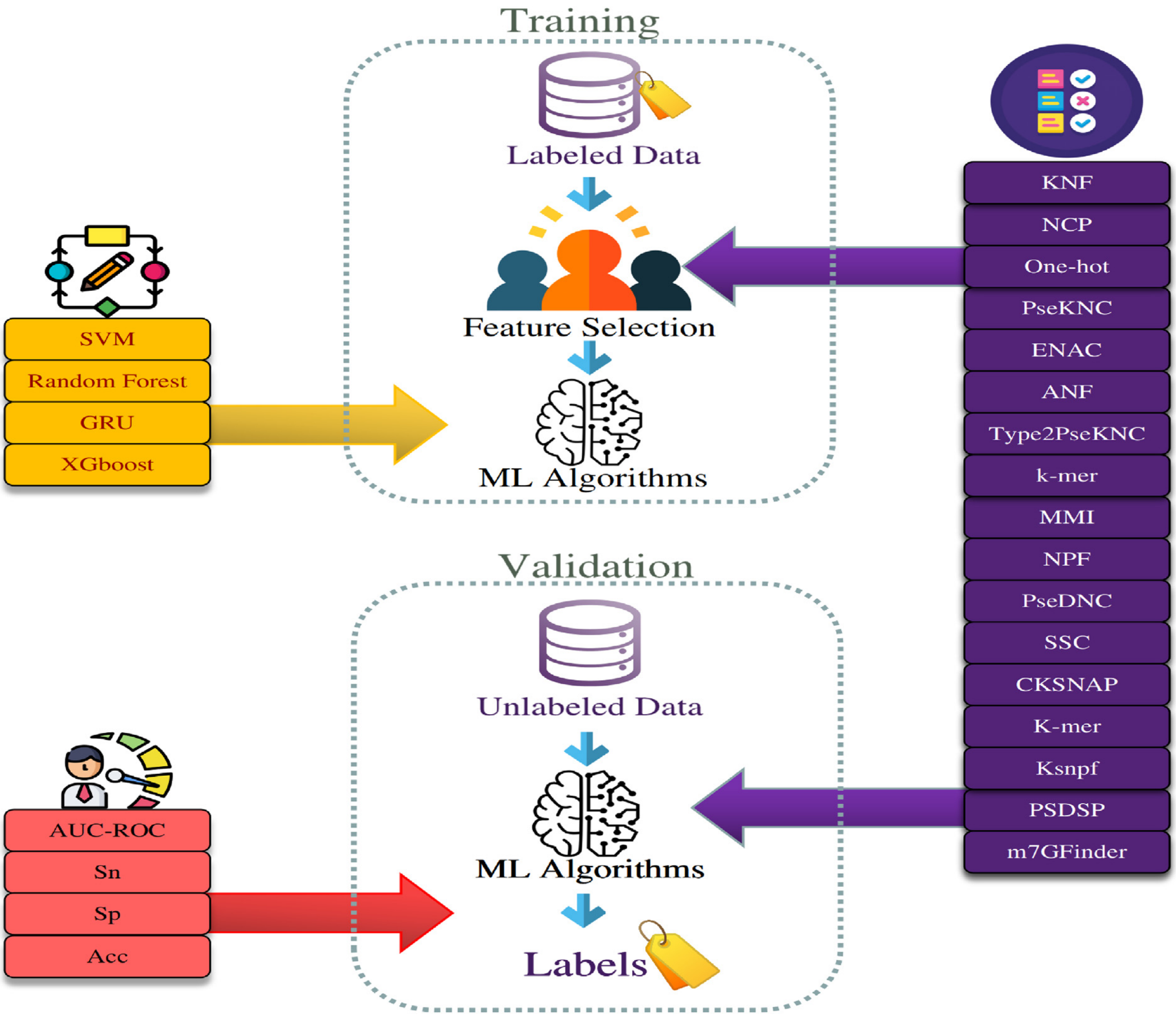

Fig. 2. Supervised learning workflow.

as the model, and the input variables are called features. If the output variable is a discrete value, then the prediction is called classification, while regression refers to prediction models that output continuous values. Support vector machines [43], XGboost [44], Random Forest [45], Convolutional neural network [46], and Recurrent Neural Network [47] are among the most commonly used machine learning algorithms for predicting RNA methylation sites.

Support vector machines (SVM) is a widespread algorithm for dealing with small sample sizes [43]. It is considered the most popular machine learning algorithm and the most commonly used in computational biology. The key idea behind SVM is to find the best separating hyperplane from data transformed into a high-dimensional feature space [48]. Another equally popular machine learning classifier is random forest (RF) [45] which is based on bagging. During training, features are randomly selected to create various decision trees. Each individual tree finds a sub-optimal solution. The average of predictions from all trees gives a better estimate of the true output. Likewise, eXtreme gradient boosting (XGboost) [10] is a boosting algorithm that uses tree models to classify data. XGboost's processing speed was increased by parallel computing. Furthermore, XGboost is extremely customizable, enabling users to identify their own optimization objectives and assessment criteria. Convolutional neural network (CNN) is a popular deep learning model initially introduced to classify images. Recent studies have demonstrated the utility of CNN in biological problems [46,33]. CNN automatically extracts the most suitable features from the input sequences and structures algorithms into layers to create an artificial neural network that can learn and make intelligent decisions on its own. Meanwhile, Recurrent Neural Network (RNN) [47] is a deep architecture that can recall context information, making it ideal for analyzing biological sequences. GRU, a simplified version of RNN, helps predict modification sites. CNN and RNN networks are often coupled with a strategy called transfer learning in order to overcome the problem of small data. In transfer learning, the training and testing datasets are not required to be independent and their distributions do not have to be identical. This allows the extrapolation of features learned from one domain to generate a pre-trained model that is further trained using a small training data from another closer domain.

4.3. Performance metrics

Several metrics are used to measure the performance of machine learning algorithms. In this review, we report the most common measures among the reviewed papers. The Area Under Curve score (AUROC) measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the false positive rate vs. the true positive rate. Similarly, the AUPRC score measures the

Table 2
Performance of *m^6A* modification sites prediction tools. NS refers to the number of samples and WS represents the word size.

Predictor	ML-Algorithm	Features	Testing	Species	NS	WS	Sn	Sp	ACC	ROC	PRC	Link
iN6-Methyl (5-step) [56]	CNN and Chou's 5-step rules	PseKNC	10-fold cross validation	*S. cerevisiae*	2614	51 bp	76.15%	74.62%	75.38%	-	-	-
				H. sapiens	2260	41 bp	82.14%	100.0%	91.11%	-	-	
				M. musculus	1450	41 bp	78.87%	100.0%	89.51%	-	-	
Gene2vec [55]	CNN	Gene2vec	Independent test	*H. sapiens* and *M. musculus*	57516	1001 bp	-	-	-	84.1%	98.0%	https://github.com/jingcheng-du/Gene2vec
DeepM6ASeq [57]	CNN	One-hot	Independent test	*H. sapiens* and *M. musculus*	22012	101 bp	-	-	76.4%	84.4%	83.1%	https://github.com/rreybeyb/DeepM6ASeq
Pm6A-CNN [58]	CNN	One-hot, NCP	10-fold cross validation	*H. sapiens*	2260	41 bp	88.6%	98.6%	93.6%	96.0%	-	https://home.jbnu.ac.kr/NSCL/pm6acnn.htm
				M. musculus	1450	41 bp	90.4%	97.2%	93.8%	97.0%	-	
				S. cerevisiae	2614	51 bp	84.6%	85.5%	85.0%	92.0%	-	
				A. thaliana	4200	101 bp	92.3%	92.6%	92.5%	97.0%	-	
M6Apred- EL [53]	Ensemble SVM	PS (k-mer) NP, PCPs, RFHC- GACs	10-fold cross validation	*S. cerevisiae*	2614	51 bp	80.72%	80.95%	80.83%	90.2%	90.1%	http://lin-group.cn/server/m6Apred
iRNA- Methyl [48]	SVM	PseDNC	10-fold cross validation	*S. cerevisiae*	2614	51 bp	70.55%	60.63%	65.59%	-	-	http://lin-group.cn/server/iRNA-Methyl
M6AMRFS [54]	XGBoost	NCP, ANF	Jackknife	*H. sapiens*	2260	41 bp	82.04%	100%	91.02%	-	-	-
			Jackknife	*M. musculus*	1450	41 bp	82.81%	75.84%	79.33%	-	-	
			10-fold cross validation	*S. cerevisiae*	2614	5 1 bp	75.21%	73.30%	74.25%	-	-	
			10-fold cross validation	*A. thaliana*	2000	101 bp	80.67%	81.43%	81.05%	-	-	
MethyRNA [51]	SVM	NCP, ANF	Jackknife	*H. sapiens*	1130	41 bp	81.68%	99.11%	90.38%	-	-	http://lin-group.cn/server/methyrna
				M. musculus	725	41 bp	77.79%	100.0%	88.39%	-	-	
M6A-word2vec [60]	CNN	Word2vec, Distributed features encoding	10-fold cross validation	*H. Sapiens*	2260	41 bp	79.0%	86.48%	83.17%	-	-	-
				S. cerevisiae	2614	51 bp	88.21%	98.05%	92.69%	-	-	
				M. musculus	788	41 bp	85.89%	95%	90.50%	-	-	
WHISTLE [12]	SVM	Sequence and genome derived features	Independent test	*H. sapiens* Mature RNA	-	-	-	-	-	89.5%	-	https://whistle-epitranscriptome.com
				H. sapiens Full transcript	-	-	-	-	-	96.0%	-	
SRAMP [61]	Random forest	Binary, KNN, Spectrum, RP in transcript	Independent test	*H. sapiens*	73940					78.4%	34.2%	http://www.cuilab.cn/sramp/
				M. musculus								
				S. cerevisiae	-	-	-	-	-	63.3%	25.3%	
EDLm6APred [59]	CNN	One-hot, word2vec, word embedding	Independent test	*H. sapiens*	-	1001 bp	67.50%	-	78.43%	86.6%	-	-
				M. musculus	-	1001 bp	66.39%	-	77.54%	85.8%		
				H. sapiens *M. musculus*	-	1001 bp	71.28%	-	78.62%	86.0%		
m6A Reader [52]	SVM	Sequence and genome derived features	Independent test	*H. sapiens* Mature RNA	30358	41 bp	-	-	-	89.3%	-	-
				H. sapiens Full transcript	32188	41 bp	-	-	-	98.1%		
RNAMethPre [31]	SVM	K-mer, RP in mRNA, MFE score	Independent test	*H. sapiens*	9849	101 bp	-	-	-	84.0%	55.6%	-
				M. musculus	7580	101 bp	-	-	-	89.3%	67.0%	
M6ATH [50]	SVM	NCP, DN	Jackknife	*A. thaliana*	788	25 bp	68.78%	100%	84.39%	84.6%	87.0%	http://lin-group.cn/server/M6ATH
RFAthM6A [62]	Random Forest	KSNPF, KNF, PSDSP, PSNSP	5-fold cross validation	*A. thaliana*	4200	101 bp	-	-	-	93.0%	-	https://github.com/nongdaxiaofeng/RFAthM6A

area under the precision-recall (PR) curve, a plot showing the trade-off between precision and recall. In addition to AUROC and AUPRC, we also report the sensitivity (Sn), specificity (Sp), and overall accuracy (Acc) as indicated below:

$$\begin{cases} Sn = \frac{TP}{TP+FN} \\ Sp = \frac{TN}{TN+FP} \\ Acc = \frac{TP+TN}{TP+FN+TN+FP} \end{cases}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

5. Machine learning approaches for RNA modification sites prediction

5.1. N6-methyladenosine

Among all types of RNA modification, the m^6A site has the most significant number of datasets due to its high presence and early discovery in various types of RNAs. Therefore, several computational methods have been used for m^6A site classification. iRNA-Methyl [48] was the first published algorithm with an available server developed using an SVM model and PseDNC as features. An updated version called M6Apred [49] was introduced to predict m^6A sites in the *S. cerevisiae* genome. Similar to iRNA-Methyl, M6Apred uses SVM for classification. However, features are extracted from RNA fragments based on the physicochemical properties of nucleotides. M6Apred showed intermediate results with an accuracy of 79.21%. A year later, RNA secondary structure and compositional based features were combined using SVM to produce RNAMethPre [31], and resulted in ROC score of 84% for *H. sapiens* and 89.3% for *M. musculus*. Simultaneously, Chen et al. created M6ATH [50] to predict m^6A in *A. thaliana* using SVM with sequence and genome derived features reaching an accuracy of 84.39%. Later, Chen et al. developed another SVM based tool called MethyRNA [51], which predicts m^6A in *H. sapiens* and *M. musculus*. MethyRNA uses the physicochemical properties of nucleotides (NCP and ANF). Using jackknife cross-validation test, MethyRNA produced an accuracy of 90.38% for *H. sapiens* and 88.39% for *M. musculus*. Few years later, Chen et al. and Zhen et al. added more sequence and genome derived features to build SVM models for their tools WHISTLE [12] and m6A Reader [52], respectively.

In addition to SVM, ensemble classifiers are also used to predict m^6A sites. M6Apred-EL [53] used three SVM classifiers and three features, namely RFHC- GACs, PCPs, and PS(1-mer) NPs, achieving an accuracy of 80.83%. In addition, an eXtreme gradient boosting algorithm called M6AMRFS [54] was used with nucleotide chemical properties for four distinct species, namely *S. cerevisiae*, *A. thaliana*, *M. musculus*, and *H. sapiens*, achieving an accuracy of 74.25%, 81.05%, 79.33, and 91.02%, respectively. In addition, a random forest based model called SRAMP gave an accuracy of 78.4% on an independent dataset of combined *H. sapiens* and *M. musculus* sequences.

Recently, multiple deep learning-based approaches have been introduced to predict m^6A, such as Gene2vec [55], iN6-Methyl (5-steps)[56], DeepM6ASeq [57], Pm6A-CNN [58], EDLm6APred [59], and m6A-word2vec [60]. Pm6A-CNN [58], a CNN-based algorithm, achieved an accuracy of 93.6%, 93.8%, 85.0%, and 92.5% on *H. sapiens*, *M. musculus*, *S. cerevisiae*, and *A. thaliana*, respectively. m6A-word2vec [60] is another deep learning based approach that transforms raw RNA sequences into a binary vector before applying a custom CNN architecture, achieving an accuracy of 83.17%. In addition, Gene2Vec [55] used word integration with CNN, achieving a low false positive rate and an ROC score of 84.1%. Table 2 gives an overall summary of available m^6A site prediction tools.

5.2. 2'-O-Methylation

In 2015, Chen et al. [10] were the first to use nucleotide composition and chemical properties to detect 2'-O-methylation sites, with an accuracy of 95.58%. Two years later, iRNA-2OM [63] was introduced for *H. sapiens*. The Features used in iRNA-2OM were filtered using mRMR, and the best 32 features were selected. SVM was used to predict methylation sites, which improved the results to 97.95% accuracy. iRNA-2methyl is another method that integrated PseKNC composition with a set of sequence coupled variables. The model was generated by combining 12 simple random forest classifiers into four ensemble predictors, achieving 93% accuracy [45]. Another tool called NmSEER V2.0 [28] uses a new random forest approach with an introduced combination of one-hot encoding, PSDSP, and KNF; they achieved an ROC of 86.2% [28]. Deep learning approaches include Deep-2-O-Me [64] and iRNA- PsKNC(2methyl) [46]. The latter gave better results with an accuracy of 98.27%. It used the 5-step rules of Chou and CNN to detect 2'-O-Methylation sites considering the raw genome sequence data. Deep-2-O-Me employed word2vec for encoding and a larger dataset but its accuracy was 85.36%. Table 3 gives an overall summary of some popular 2OMe site prediction tools.

5.3. N7-methylguanosine

Using the immuno–histochemical m7G-methylated RNA sequencing method (MeRIP-seq), Zhang et al. [65] created a benchmark dataset of m^7G sites in *H. sapiens* and *M. Musculus*, in addition, produced the location of 801 m^7G in Human Hela and HepG2 [65]. In late 2019, Wei Chen et al. built an m^7G predictor called iRNA-M7G [12]. It focused on combining sequence and structure-based features to extract the best features, followed by an SVM model,

Table 3
Performance of $2\prime - O - Me$ site prediction tools in *H. sapiens* dataset.NS refers to the number of samples and WS represents the word size.

Predictor	ML-Algorithm	Features	Testing	NS	WS	Sn	Sp	ACC	ROC	PRC	Link
Chen et al. Approach [10]	SVM	NCP, ANF	Jackknife	294	41 bp	92.52%	98.64%	95.58%	-	-	-
iRNA-2OM [63]	SVM	NCP, ANF Type 2 PseKNC	5-fold cross validation	294	41 bp	97.27%	98.63%	97.95%	99.5%	-	http://lin-group.cn/server/iRNA-2OM/
NmSEER V2.0 [28]	Random forest	one-hot PSDSP, KNF	Independent test	33020	50 bp	-	-	-	86.2%	25.4%	www.rnanut.net/nmseer-v2/
iRNA-2methyl [45]	Random forest	PseKNC	Jackknife	147	-	-	-	93.00%	-	-	-
iRNA-PsKNC (2methyl) [46]	CNN by 5-step rules of Chouis	One-hot, MMI	5-fold cross validation	294	41 bp	96.29%	100.0%	98.27%	-	-	-
Deep-2-O-Me [64]	CNN	word2vec	Independent test	20160	25 bp	-	-	85.36%	92.0%	93.0%	-

Table 4
Performance of *m^7G* site prediction tools in *H. sapiens* dataset. NS refers to the number of samples and WS represents the word size.

Predictor	ML-Algorithm	Features	Testing	NS	WS	Sn	Sp	ACC	ROC	Link
iRNA-m7G [12]	SVM	NPF,SSC, PseDNC	10-fold cross validation	1482	41 bp	88.66%	90.96%	89.81%	94.6%	http://lin-group.cn/server/iRNA-m7G/
XG-m7G [29]	XGBoost	CKS-P, E-C, NCP,ND	10-fold cross validation	1482	41 bp	91.48%	90.96%	91.22%	97.2%	-
m7g model [66]	SVM	One hot, NCP, NC, k-mer, PseKNC	10-fold cross validation	1482	41 bp	95.11%	93.74%	94.67%	98.2%	https://github.com/MapFM/m7g_model
m7GHub [27]	SVM	Sequence and genome derived features.	Independent test	-	30 bp	84.20%	71.00%	76.00%	85.5%	-
m7GPredictor [68]	SVM	NP, K-mer, PseDNC, Ksnpf, PseKNC	Independent test	300	50 bp	84.00%	88.00%	86.00%	93.3%	https://github.com/NWAFU-LiuLab/m7Gpredictor

Table 5
Performance of *m^5C* site prediction tools. NS refers to the number of samples and WS represents the word size.

Predictor	ML-Algorithm	Features	Testing	Species	NS	WS	Sn	Sp	ACC	ROC	PRC	Link
iRNAm5C-PseDNC [69]	Random Forest	PseDNC	Jackknife	*H. sapiens, M. musculus*	1900	41 bp	69.86%	99.86%	92.37%	-	-	http://www.jci-bioinfo.cn/iRNAm5C-PseDNC
PEA-m5C [70]	Random Forest	Binary encoding, k-mer, PseDNC	10-fold cross validation	*A. thaliana*	158	43 bp	86.0%	90.0%	88.0%	93.9%	94.5%	https://github.com/cma2015/PEA-m5C
RNAm5CPred [72]	SVM	KNF, KSNPFs, and pseDNC	10-fold cross validation	*H. sapiens*	240	41 bp	90.83%	94.17%	92.5%	95.7%	-	-
iRNA-PseColl [71]	SVM	PseKNC	Jackknife	*H. sapiens*	240	41 bp	75.83%	79.17%	77.50%	-	-	http://lin-group.cn/server/iRNA-PseColl/
M5C–HPCR [57]	Ensemble of SVM	PseDNC, HPCR	Jackknife	*H. sapiens*	240	41 bp	90.83%	95%	92.92%	96.2%	-	-
IRNAm5C_NB [79]	NB,RF, SVM, and AdaBoost	BPB, k-mer, ENAC, EIIP, PseEIIP	Jackknife	*H. sapiens*	240	41 bp	82.81%	81.11%	82.20%	91%.0	-	-
m5C-PseDNC [74]	SVM	PSNP, KSPSDP, CPD, PseDNC	10-fold cross validation	*A. thaliana* *H. sapiens* *M. musculus*	12578 538 11126	41 bp 41 bp 41 bp	68.1 85.5 75.7	75.5 80 72.8	71.8 82.8 74.3	- - -	- - -	-
iRNA-m5C_SVM [73]	SVM	KNFC, MNBE, NV	10-fold cross validation	*A. thaliana*	10578	41 bp	79.40%	80.90%	80.15%	-	88%	-
m5CPred-SVM [75]	SVM	PSNP,4NF, 5SNPF, PseDNC, 5SPSDP	Independent test	*H. sapiens* *M. musculus* *A. thaliana*	2000 2000 138	41 bp 41 bp 41 bp	75.4% 79.9% 75.5%	79.9% 74.9% 76.1%	77.5% 71.4% 75.8%	- - -	85.8% 77.5% 83.6%	-
iRNA5hmC [76]	SVM	K-mer	5-fold cross validation	*D. melanogaster*	1324	41 bp	67.67%	63.29%	65.48%	-	-	http://server.malab.cn/iRNA5hmC/
iRNA5hmC-PS [77]	LR	Ps-Mono (G-gap) DiMer	5-fold cross validation	*D. melanogaster*	1192	41 bp	80%	79.5%	78.3%	-	-	https://github.com/zahid6454/iRNA5hmC-PS
iRhm5CNN [78]	CNN	One hot, NCP	5-fold cross validation	*D. melanogaster*	1324	41 bp	82%	80%	81%	-	-	http://nsclbio.jbnu.ac.kr/rightarrow ols/iRhm5CNN/

reaching an accuracy of 89.81%. M7g_model [66] is another algorithm that used feature selection methods such as mRMR and Relief [66] and produces an accuracy of 94.67%. In addition, m7GHub [27] is an online platform for decoding location and regulating the internal N7-methylguanosine in mRNA. It used SVM and m7GFinder [66], a feature extraction approach based on likelihood ratio (LR), reaching an accuracy of 76% on an independent dataset obtained using three approaches (m7G-Seq, m7G-MeRIP-Seq, and m7G-miCLIP-Seq) [67]. m7GPredictor [68] is another SVM based approach that combined NP, K-mer, PseDNC, Ksnpf,

Table 6
Performance of ψ site prediction tools. NS represents the number of samples and WS is to the window size.

Predictor	ML-Algorithm	Features	Testing	Species	NS	WS	Sn	Sp	ACC	ROC	PRC	Link
iPseU-NCP [82]	Random Forest	NCP	5-fold cross validation	*H. sapiens*	990	31 bp	58.79%	65.05%	62.92%	-	-	https://github.com/ngphubinh/iPseU-NCP
			5-fold cross validation	*S. cerevisiae*	628	21 bp	66.36%	70.45%	69.59%	-	-	
			5-fold cross validation	*M. musculus*	944	21 bp	67.37%	76.27%	71.82%	-	-	
iRNA-PseU [43]	SVM	NCP ANF PseKNC	Jackknife	*H. sapiens*	990	31 bp	61.01%	59.80%	60.40%	-	-	-
				S. cerevisiae	628	21 bp	64.65%	64.33%	64.49%	-	-	
				M. musculus	944	21 bp	73.31%	64.83%	69.07%	-	-	
PseUI [88]	SVM	PSNP DC	Jackknife	*H. sapiens*	990	31 bp	64.85 %	63.64 %	64.24 %	-	-	-
		DC PSNP pseDNC	Jackknife	*S. cerevisiae*	628	21 bp	64.97%	66.88%	65.92%	-	-	
		PSNP + DC	Jackknife	*M. musculus*	944	21 bp	74.58%	66.31%	70.44%	-	-	
XG-PseU [83]	XGBoost	NC, DNC TNC, NCP One-hot	10-fold cross validation	*H. sapiens*	990	31 bp	67.24%	63.64%	65.44%	70.0%	-	http://www.bioml.cn/
				S. cerevisiae	628	21 bp	66.84%	69.45%	68.15%	74.0%		
				M. musculus	944	21 bp	76.48%	76.48%	72.03%	77.0%		
iPseU-CNN [86]	CNN	n-gram MMI	5-fold cross validation	*H. sapiens*	990	31 bp	61.01%	59.80%	60.40%	-	-	-
				S. cerevisiae	628	21 bp	64.65%	64.33%	64.49%	-	-	
				M. musculus	944	21 bp	73.31%	64.83%	69.07%	-	-	
MU-PseUDeep [87]	CNN	NCP ND	10-fold cross validation	*H. sapiens*	-	51pb	70.9%	81%	72.6%	-	-	https://github.com/smk5g5/MU-PseUDeep
				M. musculus	-	51pb	80%	73%	76%	-	-	
				S. cerevisiae	-	51pb	74.2%	79.8%	76.8%	-	-	
EnsemPseU [84]	SVM, RF XGBoost NB, KNN	K-mer ENAC NCP, ND	10-fold cross validation	*H. sapiens*	990	31 bp	63.46%	60.09%	66.28%	70.0%	-	https://github.com/biyue1026/EnsemPseU
				S. cerevisiae	628	21 bp	73.88%	74.45%	74.16%	78.6%	-	
				M. musculus	944	21 bp	75.43%	72.25%	73.84%	77.5%	-	
PSI-MOUSE [85]	SVM	Genome and sequence derived features	5-fold cross validation	*S. cerevisiae*	628	21 bp	74.12%	77.64%	71.94%	-	-	-
				M. musculus	944	21 bp	86.62%	97.31%	91.97%			

Table 7
Performance of *ac4C* site prediction tools using 5-fold cross-validation on a *H. sapiens* dataset of 12015 samples and a window size of 415 bp.

Predictor	ML-Algorithm	Features	ROC	PRC
XG-ac4C [34]	XGBoost	NCP, DN, Kmer, one-hot, EIIP, PseEIIP	91%	65.3%
PACES [89]	Random Forest	one-hot, PSNSP, PSDSP, KNF,KSNPF, PseKNC	88.5%	55.96%

Table 8
Performance of m^1A site prediction tools. NS represents the number of samples and the window size is equal to 41 bp.

Predictor	ML-Algorithm	Features	Testing	Species	NS	Sn	Sp	ACC	RPC	PRC	Link
RAMPred [90]	SVM	NCP ANF	Jackknife	*H. sapiens*	12732	98.38%	99.89%	99.13%	98%	98%	http://lin-group.cn/server/RAMPred
				M. musculus	2128	97.46%	100%	98.73%	99%	99%	
				S. cerevisiae	966	95.65%	100%	97.83%	98%	98%	
ISGm1A [91]	Random Forest	NCP ANF	5-fold cross validation	*H. sapiens*	31283	83.5%	83.8%	83.2%	90.9%	90.3%	https://github.com/lianliu09/m1a_predictionm

and PseKNC features to reach an accuracy of 86%. XG-m7G [29] is a technique that uses the XGboost algorithm along with multiple features (ND, NCP, ENAC, CKSNAP), followed by SHAP (Shapley additive interpretations) to predict the modification sites. XG-m7G showed promising results in terms of true positive rate with an accuracy of 91.22%. Table 4 gives an overall summary of popular m^7G site prediction tools.

5.4. 5-Methylcytosine and 5-Hydroxymethylcytosine

Computational techniques advantageously complement traditional sequencing methods and are a suitable alternative for further research on m^5C RNA modification. In mammalian genomes, iRNAm5C-PseDNC [69] is one of the few available algorithms that predict m^5C sites. iRNAm5C-PseDNC uses PseDNC features and Random Forest, reaching an accuracy of 92.37%. PEA-m5C [70] is another Random Forest based algorithm trained on a dataset with a highly unbalanced positive/negative ratio, making it robust while excluding false positives but indifferent to true positives, resulting in an accuracy of 88%.

Several algorithms produce performances that sacrifice sensitivity for specificity and vice versa. RNAi-PseColl [71] showed a specificity of 79.17%, meaning that the algorithm could predict most m^5C sites but simultaneously resulted in a large number of false positives. M5C–HPCR [57] balanced the true and false positives rates, producing a sensitivity of 90.83% and a specificity of 95.00%. RNAm5CPred [72] produced a an accuracy of 92.5% due to its different feature extraction approaches. RNAm5CPred proposed three different types of features: KNFs (K-nucleotide frequencies), pseDNC (pseudo-dinucleotide composition), KSNPFs (K-spaced nucleotide pair frequencies) where K random nucleotides separate a nucleotide pair.

The majority of m^5C predictors use SVM algorithm. For example, pM5CS-Comp-mRMR [73], m5C-PseDNC [74], iRNA-m5C_SVM [73] and m5CPred-SVM [75], used different datasets and feature extraction approaches.

Liu et al. developed a prediction method called iRNA5hmC [76]. In this method, an SVM classifier used k-mers as features, producing an accuracy of 65.48%. Similarly, another SVM-based predictor called iRNA5hmC-PS [77] was established. It proposed a new feature extraction method called Position-Specific Gapped k-mer and used Position-Specific k-mer to retain most of the characteristic information of RNA sequences. This approach achieved an accuracy of 78.3%. iRhm5CNN is another tool based on CNN for 5hmC site prediction that uses primary RNA sequences [78]. The CNN architecture derives the most relevant information from the primary RNA-seq representations, resulting in an accuracy of 81%. Table 5 gives an overall summary of some popular m^5C site prediction tools.

5.5. Pseudouridine (ψ)

In 2015, Li et al. were the first to develop a ψ modification sites algorithm based on the SVM algorithm called PPUS [80]. Later in the year, PseUI [81] was introduced and used SVM and multiple features, including NC, DC, pseDNC, position-specific nucleotide propensity (PSNP), and position-specific dinucleotide propensity (PSDP). One year later, Chen et al. established iRNA-PseU [43] using SVM and PseKNC features in *S. cerevisiae*, *M. musculus*, and *H. sapiens*. IPseU-NCP [82] was constructed in 2019 using Random Forest as a classifier and NCP as features, resulting in an accuracy of 62.92%, 69.59%, and 71.82% for *H. sapiens*, *S. cerevisiae*, and *M. musculus*, respectively. Similarly, an XGboost based predictor called XG-PseU [83] was published in 2020. Lui

Table 9
Performance of D site prediction tools using Jackknife cross-validation. NS represents the number of samples and WS is to the window size.

Predictor	ML-Algorithm	Features	Species	NS	WS	Sn	Sp	ACC	ROC	Link
D-pred [92]	Ensemble SVM	NPCP, PseKNC	*S. cerevisiae*	136	41 bp	76.47%	89.71%	83.09%	-	-
iRNAD [93]	SVM	NCP, ANF, SSC	*H. sapiens*, M. musculus, *D. melanogaster*, *S. cerevisiae*, and *E. coli*	551	23 bp	92.05	98.13	96.18	98.39	http://lin-group.cn/server/iRNAD/

Table 10
Performance of A-to-I site prediction tools.NS represents the number of samples and the window size is equal to 51 bp.

Predictor	ML-Algorithm	Features	Testing	Species	NS	Sn	Sp	ACC	Link
iRNA-AI [95]	SVM	NCP, ANF, PseKNC, DN	Independent set	*H. sapiens*	6486	84.19%	84.19%	93.81%	http://lin-group.cn/server/iRNA-AI/
PAI [94]	SVM	PseDNC	Jackknife	*D. melanogaster*	244	85.60%	73.11%	79.51%	http://lin-group.cn/server/PAI
EPAI-NC [96]	SVM	PseKNC, l-mers, n-gapped-lmers	Jackknife	*D. melanogaste*	244	85.60%	73.11%	79.51%	-
PAI-SAE [30]	SVM	DACC, PseDNC	Jackknife	*D. melanogaste*	224	87.20%	76.47%	81.97%	-
PRESa2i [36]	Hoeffding tree	K-mer, Gapped k-mers, and other sequence based features	Independent set	*D. melanogaste*	300	95.20%	77.31%	86.48%	http://brl.uiu.ac.bd/presa2i/

et al. improved the accuracy on *H. sapiens*, *S. cerevisiae*, and *M. musculus* datasets to 65.44%, 68.15%, and 72.03%, respectively. Recently, EnsemPseU [84] combined known classification algorithms such as SVM, NB, KNN, XGBoost, and RF into an ensemble predictor. This approach improved the state-of-the-art accuracy to 66.28% for *H. sapiens*, 74.16% for *S. cerevisiae*, and 73.84% for *M. musculus*. PSI-MOUSE [85] is an SVM based tool that combined genome and sequence derived features such as clustering information, secondary structure and density nucleotide. PSI-MOUSE performed the best on *M. musculus*, reaching an accuracy of 91.97% using 5-fold cross validation. iPseU-CNN [86] is a predictor that was introduced by Tahir et al. in 2019 based on a CNN architecture, allowing them to automatically extract features from ψ sites. MU-PseUDeep [87] is another CNN-based predictor that used a novel approach based on the secondary structure context of an mRNA sample as an input feature, resulting in an accuracy of 72.6% for *H. sapiens*, 76.8% for *S. cerevisiae*, and 76% for *M. musculus*. Table 6 gives an overall summary of popular ψ site prediction tools.

5.6. N4-acetylcytidine (ac4C)

PACES [89] was the first algorithm to predict *ac*4C modification sites in human mRNA. The authors created a reference dataset using data generated by Arango et al. [8]. Site-specific dinucleotide sequence profiles, K-nucleotide frequencies, and two random forest classifiers are included in PACES, which resulted in an ROC score of 88.5%. PACES results were further improved by Tahir et al., who used a model called XG-ac4C [34] based on the eXtreme Gradient Boosting (XGboost) as classifier and nucleotide chemical features (NCP), hot coding, nucleotide density (DN), Kmer, pseudo-electron–ionic interaction (PseEI), and pseudo-nucleotide triple interactions (PseEI) as features. This predictor improved

Table 11
Summary of multi- modification type prediction tools. NS represents the number of samples and WS is to the window size.

Predictor	ML-Algorithm	Features	Species	RNA modification	NS	WS	Link
iMRM [98]	XGboost	k-tuple, one-hot, DBE, ND, NCP, DPCP	*H. sapiens*	m^1A	12732	41	http://www.bioml.cn/XG_iRNA/home
				m5C	240	41	
				m^6A	2260	41	
				ψ	990	21	
				A-to-I	6000	51	
			S. cerevisiae	m^1A	966	41	
				m5C	422	41	
				m^6A	2614	51	
				ψ	627	31	
			M. musculus	m^1A	2128	41	
				m5C	194	41	
				m^6A	1450	41	
				ψ	944	21	
iRNA-3typeA [99]	SVM	PseKNC, PseAAC, NCP, one hot	*H. sapiens*	m1A	12732	41	http://lin-group.cn/server/iRNA-3typeA
				m^6A	2260	41	
				A-to-I	6000	41	
			M. musculus	m1A	2128	41	
				m^6A	1450	41	
				A-to-I	1662	41	
iRNA-PseColl [71]	SVM	PseKNC, PseAAC, DN	*H. sapiens*	m1A	12732	41	http://lin-group.cn/server/iRNA-PseColl/
				m^6A	2260	41	
				m5C	240	41	
DeepMRMP [47]	BGRU and transfer learning	One-hot	*H. sapiens*	m1A	2574	41	https://github.com/Chenyb939/DeepMRMP
				ψ	4128	41	
				m5C	680	41	
			S. cerevisiae	m1A	1220	41	
				ψ	2122	41	
				m5C	211	41	
			M. musculus	m1A	1052	41	
				ψ	3320	41	
				m5C	97	41	

Table 12
Performance of multi-modification type prediction tools.

Predictor	Testing	Species	RNA modification	Sn	Sp	ACC	ROC	PRC
iMRM [98]	Jackknife	*H. sapiens*	m1A	99.04%	99.78%	99.41%	100%	-
			m5C	90.83%	93.33%	92.08%	96%	-
			m^6A	82.48%	99.56%	90.38%	94%	-
			ψ	62.00%	67.11%	64.24%	71%	-
			A-to-I	87.33%	95.80%	90.71%	98%	-
		S. cerevisiae	m^1A	97.72%	100%	98.86%	99%	-
			m5C	99.05%	100%	99.52%	100%	-
			m^6A	77.04%	78.50%	77.77%	85%	-
			ψ	68.69%	73.48%	71.08%	76%	-
		M. musculus	m^1A	98.49%	99.90%	99.20%	100%	-
			m5C	97.94%	98.97%	98.45%	100%	-
			m^6A	76.90%	69.28%	73.09%	82%	-
			ψ	78.34%	99.57%	88.97%	79%	-
iRNA-3typeA [99]	Jackknife	*H. sapiens*	m1A	98.38%	99.89%	99.13%	-	-
			m^6A	81.68%	99.11%	90.38%	-	-
			A-to-I	86.18%	95.23%	90.71%	-	-
		M. musculus	m1A	97.46%	100.0%	98.73%	-	-
			m^6A	77.79%	100.0%	88.39%	-	-
			A-to-I	96.75%	100.0%	98.38%	-	-
iRNA-PseColl [71]	Jackknife	*H. sapiens*	m1A	98.38%	99.89%	99.13%	99.8%	-
			m6A	81.86%	99.11%	90.38%	84.9%	-
			m5c	75.83%	79.17%	77.50%	91.1%	-
DeepMRMP [47]	10-fold cross-validation	*H. sapiens,*	m1A	98.87%	98.86%	99.27%	100%	99%
		S. cerevisiae,	m5C	75.80%	84.69%	66.32%	89.0%	90%
		M. musculus	ψ	66.75%	74.92%	62.64%	70.0%	69%

the previous algorithm's ROC score to 91%. Table 7 gives an overall summary of some popular *ac4C* site prediction tools.

5.7. N1-Methyladenosine

To our knowledge, only two predictors have been introduced to identify N1-Methyladenosine sites, namely RAMPred and ISGm1A. RAMPred [90] was developed to predict m^1A modification sites in *M. musculus*, *H. sapiens*, and *S. cerevisiae*. It is based on PC, CP, and CFB features and uses the SVM classifier. In contrast, ISGm1A used a random forest algorithm and typical sequence properties, namely physical and chemical properties of nucleotides and cumulative frequencies, and 75 additional properties derived from genome annotations. Table 8 gives an overall summary of some popular m^1A site prediction tools.

5.8. Dihydrouridine (D)

An SVM-based classifier, namely D-pred [92], was implemented to detect D-modification sites in *S. cerevisiae*. The algorithm incorporates sequence-based heterogeneous features, resulting in an accuracy of 83.09%. Likewise, iRNAD [93] used the chemical properties of nucleotides and nucleotide density to encode RNA samples of five different species, namely *H. sapiens*, *M. musculus*, *S. cerevisiae*, *E. coli*, and *D. melanogaster*. The classification was done using SVM and reached an accuracy of 96.18%. Table 9 gives an overall summary of some popular D site prediction tools.

5.9. A-to-I modification

In 2016, Chen et al. introduced the first algorithm used to detect A-to-I modification sites in *D. melanogaster* called PAI [94]. The algorithm is based on SVM and PseDNC features and produced an accuracy of 79.51%. Another SVM-based algorithm called iRNA-AI [95] was proposed in the following year using a different dataset. iRNA-AI achieved promising results with an accuracy of 93.81%. Xia et al. improved the accuracy of PAI using an approach called PAI-SAE [30]. The algorithm used new hybrid features by incorporating DACC and PseDNC density, nucleotide density, and a backup autoencoder with the SVM algorithm. EPAI-NC [96] also used SVM with K-mer, reaching an accuracy of 93.90% on an independent test set. The Hoeffding tree, an incremental decision tree for sample identification, along with BFS + LF for feature selection, was used by PRESa2i [36]. This method achieved an accuracy of 86.48% in the independent dataset. Table 10 gives an overall summary of popular m^1A site prediction tools.

5.10. N2-methylguanosine

iRNA-m2G [97] is the only available predictor proposed for m^2G sites for eukaryotes. The models were trained using SVM, and RNA sequences were encoded using the chemical property of nucleotides and the accumulated replication of nucleotides. The results are promising, with an accuracy of 94.56% for *H. sapiens*, 100.00% for *M. musculus*, and 96.27% for *S. cerevisiae*. iRNA-m2G reached a sensitivity of 89.13% for *H. sapiens*, 100.00% for *M. musculus*, 92.53% for *S. cerevisiae*, and specificity of 100.00% for *H. sapiens*, *M. musculus*, and *S. cerevisiae*.

5.11. Simultaneous identification of more than one RNA modification

Numerous RNA modification data have been collected as a result of the advances in high-throughput sequencing techniques. However, most of these approaches cannot distinguish between different RNA modifications in the same RNA sequence. For example, Adenosine typically undergoes m^1A, m^6A, and A-to-I modifications. Unfortunately, it is not easy to determine whether Adenosine modifications have occurred simultaneously using the methods reviewed so far. Consequently, the development of computational methods for solving this problem is critical. iRNA-PseColl [71] is the first platform developed to simultaneously identify three different RNA modification sites, namely m^1A, m^6A and m^5C, based solely on the sequence information reaching an accuracy of 99.13% 90.38%,77.50%, respectively. iMRM [98] is an XGboost-based prediction tool that can concurrently classify m^6A, m^5C, m^1A, ψ, and A-to-I modifications in *H. sapiens*, *M. muscu-*

lus, and *S. cerevisiae* using sequence information and nucleotide physicochemical properties. Similarly, Chen et al. developed an SVM-based predictor named iRNA-3typeA [99] that considered nucleotide chemical properties and lingering density. It can detect m^1A, m^6A, and A-to-I RNA modifications in both *H. sapiens* and *M. musculus* transcriptomes.

Another multi-type predictor is DeepMRMP [47] for m^1A, ψ, m^5C. It is the first deep learning-based tool for multiple types of RNA modification site prediction. The model used large-scale m^6A data to pre-train a deep learning model and then employed a transfer-learning strategy to fine-tune its network parameters for the targeted types of RNA modifications, reaching an accuracy of 99.27% for m^1A, 62.64% for ψ, and 66.32% for m^5C. Table 11 summarizes the available multi-type predictors and Table 12 provides the performance measures for each tool and dataset.

6. Summary and outlook

The study of RNA modification has gained high interest as it reveals the importance of RNA modifications in regulating gene expression and disease pathogenesis. As epitranscriptome sequencing data increases, more RNA modification benchmark datasets become available. The recent availability of large datasets and the advances in computational biology through machine learning have transformed research in the area. As a result, these technologies have ultimately improved our understanding of the biological significance of RNA modifications. This paper reviewed and updated recent advances and emerging machine learning-based approaches to RNA modification prediction. Despite the rapid progress in this field, several limitations and problems exist and should be addressed.

Perhaps the most important observation is that most approaches across all modification types share common techniques, features, and classification algorithms. However, their reported accuracy varies within the same or between different modification types. Our review shows that the performance correlates with the benchmark dataset's quality and size. For example, m7g-model [66] and m7GPredictor [24] tested several classifiers and features. When using SVM with the PseKNC feature, m7g-model reported specificity of 81.49% on an independent dataset of 300 samples, while m7GPredictor reported a specificity of 88% on the same independent dataset. m7GPredictor extracted 801 m7G sites from high-confident internal sites as a positive dataset and randomly selected 801 negative examples. Conversely, m7g-model created a positive dataset from 741 experimentally validated m^7G sites and 741 random negative samples.

We also observe a trade-off between sensitivity and specificity among different approaches, and in several cases, the false positives and false negatives rates have an inverse causal relationship. For example, some algorithms have a higher false positive rate, such as DeepM6ASeq [57], while others have a higher false negative rate, such as Gene2vec [55]. Some tools attempt to circumvent this problem using imbalanced dataset. For example, DeepM6ASeq has a 16:10 positive-to-negative ratio, while for Gene2vec it is 1:1 for training and 1:10 for testing.

Therefore, the most critical step is to create a single comprehensive Benchmark dataset for each modification type. Current databases include small epitranscriptomics data with matching RNA modification profiles. In addition, the lack of a standard benchmark directly impacts the performance comparison between different approaches. In most cases, new approaches use custom-built datasets that make the comparison biased. Therefore, more enriched experimentally verified data is required to train better machine learning models and compare emerging tools fairly. These benchmark datasets should consider the confidence of the modification sites, the correct size of the sequences, the best positive-to-negative ratio, and the best way to select negative samples.

Most reviewed approaches use similar features and explore multiple features and classification models before concluding with the best combination. However, future algorithms should explore other biological-related features, such as structure-based features. In addition, features specific to a particular species or tissue could be introduced to facilitate the functional investigation of modifications in other species such as plants [70].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank Dr. Morad M. Mokhtar and Dr. Alsamman M. Alsamman for reviewing the manuscript.

References

[1] Xuan JJ, Sun WJ, Lin PH, Zhou KR, Liu S, Zheng LL, Qu LH, Yang JH. Rmbase v2.0: Deciphering the map of rna modifications from epitranscriptome sequencing data. Nucleic Acids Res 2018;46:D327–34. https://doi.org/10.1093/nar/gkx934.

[2] W.A. Cantara, P.F. Crain, J. Rozenski, J.A. Mccloskey, K.A. Harris, X. Zhang, F.A.P. Vendeix, D. Fabris, P.F. Agris, The rna modification database, rnamdb: 2011 update, Nucleic Acids Research doi:10.1093/nar/gkq1028..

[3] Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. Single-nucleotide-resolution mapping of m6a and m6am throughout the transcriptome. Nat Methods 2015;12:767–72. https://doi.org/10.1038/nmeth.3453.

[4] Hauenschild R, Tserovski L, Schmid K, Thüring K, Winz ML, Sharma S, Entian KD, Wacheul L, Lafontaine DL, Anderson J, Alfonzo J, Hildebrandt A, Jäschke A, Motorin Y, Helm M. The reverse transcription signature of n-1-methyladenosine in rna-seq is sequence dependent. Nucleic Acids Res 2015;43:9950–64. https://doi.org/10.1093/nar/gkv895.

[5] Tserovski L, Marchand V, Hauenschild R, Blanloeil-Oillo F, Helm M, Motorin Y. High-throughput sequencing for 1-methyladenosine (m1a) mapping in rna. Methods 2016;107:110–21. https://doi.org/10.1016/j.ymeth.2016.02.012.

[6] K.E. Bohnsack, C. Höbartner, M.T. Bohnsack, Eukaryotic 5-methylcytosine (m 5 c) rna methyltransferases: Mechanisms, cellular functions, and links to disease, Genes 10. doi:10.3390/genes10020102..

[7] Lin I-H, Chen Y-F, Hsu M-T. Correlated 5-hydroxymethylcytosine (5hmc) and gene expression profiles underpin gene and organ-specific epigenetic regulation in adult mouse brain and liver. PLOS ONE 2017;12:. https://doi.org/10.1371/journal.pone.0170779e0170779.

[8] Arango D, Sturgill D, Alhusaini N, Dillman AA, Sweet TJ, Hanson G, Hosogane M, Sinclair WR, Nanan KK, Mandler MD, Fox SD, Zengeya TT, Andresson T, Meier JL, Coller J, Oberdoerffer S. Acetylation of cytidine in mrna promotes translation efficiency. Cell 2018;175:1872–1886.e24. https://doi.org/10.1016/j.cell.2018.10.030.

[9] Ayadi L, Galvanin A, Pichot F, Marchand V, Motorin Y. Rna ribose methylation (2-o-methylation): Occurrence, biosynthesis and biological functions. Biochimica et Biophysica Acta - Gene Regulatory Mechanisms 1862;2019:253–69. https://doi.org/10.1016/j.bbagrm.2018.11.009.

[10] Chen W, Feng P, Tang H, Ding H, Lin H. Identifying 2-o-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions. Genomics 2016;107:255–8. https://doi.org/10.1016/j.ygeno.2016.05.003.

[11] Y. Motorin, V. Marchand, Detection and analysis of rna ribose 2-o-methylations: Challenges and solutions, Genes 9. doi:10.3390/genes9120642..

[12] K. Chen, Z. Wei, Q. Zhang, X. Wu, R. Rong, Z. Lu, J. Su, J.P. de Magalhães, D.J. Rigden, J. Meng, WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach, Nucleic Acids Research 47 (7) (2019) e41–e41. arXiv: https://academic.oup.com/nar/article-pdf/47/7/e41/28467841/gkz074.pdf, doi:10.1093/nar/gkz074. URL:https://doi.org/10.1093/nar/gkz074..

[13] M. Sprinzl, K.S. Vassilenko, Compilation of trna sequences and sequences of trna genes, Nucleic Acids Research 33. doi:10.1093/nar/gki012..

[14] Krogh N, Nielsen H. Sequencing-based methods for detection and quantitation of ribose methylations in rna. Methods 2019;156:5–15. https://doi.org/10.1016/j.ymeth.2018.11.017.

[15] Z.W. Dong, P. Shao, L.T. Diao, H. Zhou, C.H. Yu, L.H. Qu, Rtl-p: A sensitive approach for detecting sites of 2-o-methylation in rna molecules, Nucleic Acids Research 40. doi:10.1093/nar/gks698..

[16] Li X, Xiong X, Yi C. Epitranscriptome sequencing technologies: Decoding rna modifications. Nat. Methods 2016;14:23–31. https://doi.org/10.1038/nmeth.4110.

[17] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis, A. Soboleva, Ncbi geo: Archive for functional genomics data sets - update, Nucleic Acids Research 41. doi:10.1093/nar/gks1193..

[18] P. Boccaletto, M.A. Machnicka, E. Purta, P.L.P. Atkowski, B.L.Z.B. Nski, T.K. Wirecki, V.D. Crécy, C. Crécy-Lagard, R. Ross, P.A. Limbach, A. Kotter, M. Helm, J.M. Bujnicki, Modomics: a database of rna modification pathways. 2017 update, Nucleic Acids Research 46 (2017) 303–307. doi:10.1093/nar/gkx1030..

[19] Kiran A, Baranov PV. Darned: a database of rna editing in humans. Bioinformatics 2010;26:1772–6. https://doi.org/10.1093/bioinformatics/btq285.

[20] Y. Tang, K. Chen, B. Song, J. Ma, X. Wu, Q. Xu, Z. Wei, J. Su, G. Liu, R. Rong, Z. Lu, J. de Magalhães, D.J. Rigden, J. Meng, m6A-Atlas: a comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome, Nucleic Acids Research 49 (D1) (2020) D134–D143. arXiv: https://academic.oup.com/nar/article-pdf/49/D1/D134/35364836/gkaa692.pdf, doi:10.1093/nar/gkaa692. URL:https://doi.org/10.1093/nar/gkaa692..

[21] Picardi E, D'Erchia AM, Lo Giudice C, Pesole G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res 2017;45 (D1):D750–7.

[22] G. Ramaswami, J.B. Li, RADAR: a rigorously annotated database of A-to-I RNA editing, Nucleic Acids Res 42 (Database issue) (2014) D109–113..

[23] Liu H, Wang H, Wei Z, Zhang S, Hua G, Zhang SW, Zhang L, Gao SJ, Meng J, Chen X, Huang Y. Met-db v2.0: Elucidating context-specific functions of n 6 -methyl-adenosine methyltranscriptome. Nucleic Acids Res 2018;46:D281–7. https://doi.org/10.1093/nar/gkx1080.

[24] Liu S, Zhu A, He C, Chen M. Repic: A database for exploring the n 6-methyladenosine methylome. Genome Biol. 2020;21:100. https://doi.org/10.1186/s13059-020-02012-4.

[25] Y. Han, J. Feng, L. Xia, X. Dong, X. Zhang, S. Zhang, Y. Miao, Q. Xu, S. Xiao, Z. Zuo, L. Xia, C. He, CVm6A: A Visualization and Exploration Database for m6As in Cell Lines, Cells 8 (2)..

[26] B. Song, K. Chen, Y. Tang, Z. Wei, J. Su, J.P. de Magalhães, D.J. Rigden, J. Meng, ConsRM: collection and large-scale prediction of the evolutionarily conserved RNA methylation sites, with implications for the functional epitranscriptome, Brief Bioinform..

[27] Song B, Tang Y, Chen K, Wei Z, Rong R, Lu Z, Su J, de Magalhães JP, Rigden DJ, Meng J. m7ghub: deciphering the location, regulation and pathogenesis of internal mrna n7-methylguanosine (m7g) sites in human. Bioinformatics 2020;36:3528–36. https://doi.org/10.1093/bioinformatics/btaa178.

[28] Y. Zhou, Q. Cui, Y. Zhou, Nmseer v2.0: A prediction tool for 2-o-methylation sites based on random forest and multi-encoding combination, BMC Bioinformatics 20 690. doi:10.1186/s12859-019-3265-8..

[29] Bi Y, Xiang D, Ge Z, Li F, Jia C, Song J. An interpretable prediction model for identifying n7-methylguanosine sites based on xgboost and shap. Mol Therapy - Nucleic Acids 2020;22:362–72. https://doi.org/10.1016/j.omtn.2020.08.022.

[30] X. Xiao, P. Wang, Z. Xu, W. Qiu, X. Fang, Pai-sae: Predicting adenosine to inosine editing sites based on hybrid features by using spare auto-encoder, in: test, Vol. 170, Institute of Physics Publishing, 2018, p. 52018. doi:10.1088/1755-1315/170/5/052018..

[31] Xiang S, Liu K, Yan Z, Zhang Y, Sun Z. RNAMethPre: A Web Server for the Prediction and Query of mRNA m6A Sites. PLoS One 2016;11(10):e0162707.

[32] A.T.G. Bari, M.R. Reaz, H.J. Choi, B.S. Jeong, Dna encoding for splice site prediction in large dna sequence, in: test, Vol. 7827 LNCS, Springer, Berlin, Heidelberg, 2013, pp. 46–58. doi:10.1007/978-3-642-40270-8_4..

[33] Al-Ajlan A, Allali AE. Cnn-mgp: Convolutional neural networks for metagenomics gene prediction, Interdisciplinary Sciences: Computational. Life Sci 2019;11:628–35. https://doi.org/10.1007/s12539-018-0313-4.

[34] Alam W, Tayara H, Chong KT. Xg-ac4c: identification of n4-acetylcytidine (ac4c) in mrna using extreme gradient boosting with electron-ion interaction pseudopotentials. Sci Rep 2020;10:1–10. https://doi.org/10.1038/s41598-020-77824-2.

[35] Han S, Liang Y, Ma Q, Xu Y, Zhang Y, Du W, Wang C, Li Y. Lncfinder: an integrated platform for long non-coding rna identification utilizing sequence intrinsic composition, structural information and physicochemical property. Briefings Bioinformatics 2019;20:2009–27. https://doi.org/10.1093/bib/bby065.

[36] A. Choyon, A. Rahman, M. Hasanuzzaman, D.M. Farid, S. Shatabda, Presa2i: incremental decision trees for prediction of adenosine to inosine rna editing sites, F1000Research 9 (2020) 262. doi:10.12688/f1000research.22823.1..

[37] I.T. Jollife, J. Cadima, Principal component analysis: A review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374. doi:10.1098/rsta.2015.0202..

[38] H. Peng, F. Long, 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..

[39] Al-Ajlan A, Allali AE. Feature selection for gene prediction in metagenomic fragments. BioData Mining 2018;11:9. https://doi.org/10.1186/s13040-018-0170-z.

[40] Allali AA-AAE. Cnn-mgp: Convolutional neural networks for metagenomics gene prediction, Interdisciplinary Sciences: Computational. Life Sci. 2019;11:628–35. https://doi.org/10.1007/s12539-018-0313-4.

[41] S.K. Singhi, H. Liu, Feature subset selection bias for classification learning, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 849–856. doi:10.1145/1143844.1143951. URL:https://doi.org/10.1145/1143844.1143951..

[42] Li L, Neal RM, Zhang J. A method for avoiding bias from feature selection with application to naive Bayes classification models. Bayesian Analysis 2008;3 (1):171–96. https://doi.org/10.1214/08-BA307. URL:https://doi.org/10.1214/08-BA307.

[43] Chen W, Tang H, Ye J, Lin H, Chou KC. irna-pseu: Identifying rna pseudouridine sites. Mol Therapy - Nucleic Acids 2016;5:. https://doi.org/10.1038/mtna.2016.37e332.

[44] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13–17-August-2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

[45] W.-R. Qiu, S.-Y. Jiang, B.-Q. Sun, X. Xiao, X. Cheng, K.-C. Chou, irna-2methyl: Identify rna 2-o-methylation sites by incorporating sequence-coupled effects into general pseknc and ensemble classifier, Medicinal Chemistry 13. doi:10.2174/1573406413666170623082245..

[46] Tahir M, Tayara H, Chong KT. irna-pseknc(2methyl): Identify rna 2'-o-methylation sites by convolution neural network and chou's pseudo components. J. Theor. Biol. 2019;465:1–6. https://doi.org/10.1016/j.jtbi.2018.12.034.

[47] Sun P, Chen Y, Liu B, Gao Y, Han Y, He F, Ji J. Deepmrmp: A new predictor for multiple types of rna modification sites using deep learning. Math Biosciences Eng 2019;16:6231–41. https://doi.org/10.3934/mbe.2019310.

[48] Chen W, Feng P, Ding H, Lin H, Chou KC. Irna-methyl: Identifying n6-methyladenosine sites using pseudo nucleotide composition. Anal Biochem 2015;490:26–33. https://doi.org/10.1016/j.ab.2015.08.021.

[49] Chen W, Tran H, Liang Z, Lin H, Zhang L. Identification and analysis of the n6-methyladenosine in the saccharomyces cerevisiae transcriptome. Sci Rep 2015;5:1–8. https://doi.org/10.1038/srep13859.

[50] Chen W, Feng P, Ding H, Lin H. methyladenosine sites in the Arabidopsis thaliana transcriptome. Mol Genet Genomics 2016;291(6):2225–9.

[51] Chen W, Tang H, Lin H. Methyrna: a web server for identification of n6-methyladenosine sites. J Biomol Struct Dyn 2017;35:683–7. https://doi.org/10.1080/07391102.2016.1157761.

[52] Zhen D, Wu Y, Zhang Y, Chen K, Song B, Xu H, Tang Y, Wei Z, Meng J. m6a reader: Epitranscriptome target prediction and functional characterization of n6-methyladenosine (m6a) readers. Front Cell Dev Biol 2020;8:741. https://doi.org/10.3389/fcell.2020.00741. URL:https://www.frontiersin.org/article/10.3389/fcell.2020.00741.

[53] Wei L, Chen H, Su R. M6apred-el: A sequence-based predictor for identifying n6-methyladenosine sites using ensemble learning. Mol Therapy - Nucleic Acids 2018;12:635–44. https://doi.org/10.1016/j.omtn.2018.07.004.

[54] Qiang X, Chen H, Ye X, Su R, Wei L. M6amrfs: Robust prediction of n6-methyladenosine sites with sequence-based features in multiple species. Front Genetics 2018;9:495. https://doi.org/10.3389/fgene.2018.00495.

[55] Zou Q, Xing P, Wei L, Liu B. Gene2vec: gene subsequence embedding for prediction of mammalian n 6 -methyladenosine sites from mrna. RNA 2019;25:205–18. https://doi.org/10.1261/rna.069112.118.

[56] Nazari I, Tahir M, Tayara H, Chong KT. in6-methyl (5-step): Identifying rna n6-methyladenosine sites using deep learning mode via chou's 5-step rules and chou's general pseknc. Chemometrics and Intelligent Laboratory Systems 2019;193:. https://doi.org/10.1016/j.chemolab.2019.103811103811.

[57] Zhang Y, Hamada M. Deepm6aseq: Prediction and characterization of m6a-containing sequences using deep learning. BMC Bioinformatics 2018;19:524. https://doi.org/10.1186/s12859-018-2516-4.

[58] Alam W, Ali SD, Tayara H, Chong KT. A cnn-based rna n6-methyladenosine site predictor for multiple species using heterogeneous features representation. IEEE Access 2020;8:138203–9. https://doi.org/10.1109/ACCESS.2020.3002995.

[59] L. Zhang, G. Li, X. Li, H. Wang, S. Chen, H. Liu, Edlm6apred: ensemble deep learning approach for mrna m6a site prediction, BMC Bioinformatics 22..

[60] Tahir M, Hayat M, Chong KT. Prediction of n6-methyladenosine sites using convolution neural network model based on distributed feature representations. Neural Networks 2020;129:385–91. https://doi.org/10.1016/j.neunet.2020.05.027.

[61] Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res 2016;44(10):e91.

[62] Wang X, Yan R. A sites in Arabidopsis thaliana. Plant Mol Biol 2018;96 (3):327–37.

[63] Yang H, Lv H, Ding H, Chen W, Lin H. irna-2om: A sequence-based predictor for identifying 2-o-methylation sites inhomo sapiens. J Comput Biol 2018;25:1266–77. https://doi.org/10.1089/cmb.2018.0004.

[64] Mostavi M, Salekin S, Huang Y. Deep-2-o-me: Predicting 2-o-methylation sites by convolutional neural networks. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS. p. 2394–7. https://doi.org/10.1109/EMBC.2018.8512780.

[65] Cui X, Zhang L, Meng J, Rao MK, Chen Y, Huang Y. Metdiff: A novel differential rna methylation analysis for merip-seq data. IEEE/ACM Trans Comput Biol Bioinf 2018;15:526–34. https://doi.org/10.1109/TCBB.2015.2403355.

[66] Yang YH, Ma C, Wang JS, Yang H, Ding H, Han SG, Li YW. Prediction of n7-methylguanosine sites in human rna based on optimal sequence features. Genomics 2020;112:4342–7. https://doi.org/10.1016/j.ygeno.2020.07.035.

[67] Zhang LS, Liu C, Ma H, Dai Q, Sun HL, Luo G, Zhang Z, Zhang L, Hu L, Dong X, He C. Transcriptome-wide mapping of internal n7-methylguanosine methylome in mammalian mrna. Mol Cell 2019;74:1304–1316.e8. https://doi.org/10.1016/j.molcel.2019.03.036.

[68] Liu X, Liu Z, Mao X, Li Q. m7gpredictor: An improved machine learning-based model for predicting internal m7g modifications using sequence properties. Anal Biochem 2020;609:. https://doi.org/10.1016/j.ab.2020.113905113905.

[69] Qiu WR, Jiang SY, Xu ZC, Xiao X, Chou KC. iRNAm 5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget 2017;8(25):41178–88.

[70] Song J, Zhai J, Bian E, Song Y, Yu J, Ma C. Transcriptome-wide annotation of m5c rna modifications using machine learning. Front Plant Sci 2018;9:519. https://doi.org/10.3389/fpls.2018.00519.

[71] Feng P, Ding H, Yang H, Chen W, Lin H, Chou KC. irna-psecoll: Identifying the occurrence sites of different rna modifications by incorporating collective effects of nucleotides into pseknc. Mol Therapy - Nucleic Acids 2017;7:155–63. https://doi.org/10.1016/j.omtn.2017.03.006.

[72] Fang T, Zhang Z, Sun R, Zhu L, He J, Huang B, Xiong Y, Zhu X. Rnam5cpred: Prediction of rna 5-methylcytosine sites based on three different kinds of nucleotide composition. Mol Therapy - Nucleic Acids 2019;18:739–47. https://doi.org/10.1016/j.omtn.2019.10.008.

[73] Dou L, Li X, Ding H, Xu L, Xiang H. Prediction of m5c modifications in rna sequences by combining multiple sequence features. Mol Therapy - Nucleic Acids 2020;21:332–42. https://doi.org/10.1016/j.omtn.2020.06.004.

[74] Feng P, Ding H, Chen W, Lin H. Identifying rna 5-methylcytosine sites: Via pseudo nucleotide compositions. Mol BioSyst 2016;12:3307–11. https://doi.org/10.1039/c6mb00471g.

[75] Chen X, Xiong Y, Liu Y, Chen Y, Bi S, Zhu X. m5cpred-svm: a novel method for predicting m5c sites of rna. BMC Bioinformatics 2020;21:489. https://doi.org/10.1186/s12859-020-03828-4.

[76] Liu Y, Chen D, Su R, Chen W, Wei L. irna5hmc: The first predictor to identify rna 5-hydroxymethylcytosine modifications using machine learning. Front Bioeng Biotechnol 2020;8:227. https://doi.org/10.3389/fbioe.2020.00227.

[77] Ahmed S, Hossain Z, Uddin M, Taherzadeh G, Sharma A, Shatabda S, Dehzangi A. Accurate prediction of rna 5-hydroxymethylcytosine modification by utilizing novel position-specific gapped k-mer descriptors, Computational and Structural. Biotechnol J 2020;18:3528–38. https://doi.org/10.1016/j.csbj.2020.10.032.

[78] Ali SD, Kim JH, Tayara H, Chong KT. Prediction of rna 5-hydroxymethylcytosine modifications using deep learning. IEEE Access 2021;9:8491–6. https://doi.org/10.1109/ACCESS.2021.3049146.

[79] Dou L, Li X, Ding H, Xu L, Xiang H. Irna-m5c_nb: A novel predictor to identify rna 5-methylcytosine sites based on the naive bayes classifier. IEEE Access 2020;8:84906–17. https://doi.org/10.1109/ACCESS.2020.2991477.

[80] Li Y-H, Zhang G, Cui Q. Ppus: a web server to predict pus-specific pseudouridine sites: Table 1. Bioinformatics 2015;31:3362–4. https://doi.org/10.1093/bioinformatics/btv366.

[81] Y. Furuichi, Discovery of m7g-cap in eukaryotic mrnas, Proceedings of the Japan Academy Series B: Physical and Biological Sciences 91 (2015) 394–409. doi:10.2183/pjab.91.394..

[82] Nguyen-Vo TH, Nguyen QH, Do TT, Nguyen TN, Rahardja S, Nguyen BP. Ipseu-ncp: Identifying rna pseudouridine sites using random forest and ncp-encoded features. BMC Genomics 2019;20:971. https://doi.org/10.1186/s12864-019-6357-y.

[83] Liu K, Chen W, Lin H. Xg-pseu: an extreme gradient boosting based method for identifying pseudouridine sites. Mol Genet Genomics 2020;295:13–21. https://doi.org/10.1007/s00438-019-01600-9.

[84] Bi Y, Jin D, Jia C. Ensempseu: Identifying pseudouridine sites with an ensemble approach. IEEE Access 2020;8:79376–82. https://doi.org/10.1109/ACCESS.2020.2989469.

[85] Song B, Chen K, Tang Y, Ma J, Meng J, Wei Z. PSI-MOUSE: Predicting Mouse Pseudouridine Sites From Sequence and Genome-Derived Features. Evol Bioinform Online 2020;16. 1176934320925752.

[86] Tahir M, Tayara H, Chong KT. ipseu-cnn: Identifying rna pseudouridine sites using convolutional neural networks. Mol Therapy - Nucleic Acids 2019;16:463–70. https://doi.org/10.1016/j.omtn.2019.03.010.

[87] Khan SM, He F, Wang D, Chen Y, Xu D. Mu-pseudeep: A deep learning method for prediction of pseudouridine sites, Computational and Structural. Biotechnol J 2020;18:1877–83. https://doi.org/10.1016/j.csbj.2020.07.010.

[88] He J, Fang T, Zhang Z, Huang B, Zhu X, Xiong Y. Pseui: Pseudouridine sites identification based on rna sequence information. BMC Bioinformatics 2018;19:1–11. https://doi.org/10.1186/s12859-018-2321-0.

[89] Zhao W, Zhou Y, Cui Q, Zhou Y. Paces: prediction of n4-acetylcytidine (ac4c) modification sites in mrna. Sci Rep 2019;9:11112. https://doi.org/10.1038/s41598-019-47594-7.

[90] W. Chen, P. Feng, H. Tang, H. Ding, H. Lin, Rampred: Identifying the n1-methyladenosine sites in eukaryotic transcriptomes, Scientific Reports 6. doi:10.1038/srep31080..

[91] Liu L, Lei X, Meng J, Wei Z. Isgm1a: Integration of sequence features and genomic features to improve the prediction of human m1a rna methylation sites. IEEE Access 2020;8:81971–7. https://doi.org/10.1109/ACCESS.2020.2991070.

[92] P. Feng, Z. Xu, H. Yang, H. Lv, H. Ding, L. Liu, Identification of d modification sites by integrating heterogeneous features in saccharomyces cerevisiae, Molecules 24. doi:10.3390/molecules24030380..

[93] Xu Z-C, Feng P-M, Yang H, Qiu W-R, Chen W, Lin H. irnad: a computational tool for identifying d modification sites in rna sequence. Bioinformatics 2019;35:4922–9. https://doi.org/10.1093/bioinformatics/btz358.

[94] W. Chen, P. Feng, H. Ding, H. Lin, Pai: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions open, IOP Conference Series: Earth and Environmental Science doi:10.1038/srep35123..

[95] Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. Irna-ai: Identifying the adenosine to inosine editing sites in rna sequences. Oncotarget 2017;8:4208–17. https://doi.org/10.18632/oncotarget.13758.

[96] Ahmad A, Shatabda S. Epai-nc: Enhanced prediction of adenosine to inosine rna editing sites using nucleotide compositions. Anal Biochem 2019;569:16–21. https://doi.org/10.1016/j.ab.2019.01.002.

[97] Chen W, Song X, Lv H, Lin H. irna-m2g: Identifying n2-methylguanosine sites based on sequence-derived information. Mol Therapy - Nucleic Acids 2019;18:253–8. https://doi.org/10.1016/j.omtn.2019.08.023.

[98] Liu K, Chen W. imrm: a platform for simultaneously identifying multiple kinds of rna modifications. Bioinformatics 2020;36:3336–42. https://doi.org/10.1093/bioinformatics/btaa155.

[99] Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. irna-3typea: Identifying three types of modification at rna's adenosine sites. Mol Therapy - Nucleic Acids 2018;11:468–74. https://doi.org/10.1016/j.omtn.2018.03.012.