Methods 203 (2022) 399–421

Contents lists available at ScienceDirect

Methods

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

A brief review of machine learning methods for RNA methylation sites prediction

Hong Wang [1], Shuyu Wang [1], Yong Zhang, Shoudong Bi [*], Xiaolei Zhu [*]

School of Sciences, Anhui Agricultural University, Hefei, Anhui, China

ARTICLE INFO

Keywords:
RNA methylation
Machine learning
Prediction
Bioinformatics tools

ABSTRACT

Thanks to the tremendous advancement of deep sequencing and large-scale profiling, epitranscriptomics has become a rapidly growing field. As one of the most important parts of epitranscriptomics, ribonucleic acid (RNA) methylation has been focused on for years for its fundamental role in regulating the many aspects of RNA function. Thanks to the big data generated in sequencing, machine learning methods have been developed for efficiently identifying methylation sites. In this review, we comprehensively explore machine learning based approaches for predicting 10 types of methylation of RNA, which include m6A, m5C, m7G, 5hmC, m1A, m5U, m6Am, and so on. Firstly, we reviewed three main aspects of machine learning which are data, features and learning algorithms. Then, we summarized all the methods that have been used to predict the 10 types of methylation. Furthermore, the emergent methods which were designed to predict multiple types of methylation were also reviewed. Finally, we discussed the future perspectives for RNA methylation sites prediction.

1. Introduction

The post-transcriptional modification of RNA is one of the important contents of epitranscriptomics, which has become one of the research hotspots of epigenetics. At present, nearly 200 kinds of RNA chemical modifications have been discovered, which greatly expands the chemical diversity of RNA polymers [1,2]. RNA methylation is the most common type of post-transcriptional modification that occurs in a variety of RNAs, such as tRNA, rRNA, mRNA, snoRNA, snRNA and miRNA, etc. There are many types of RNA methylation, mainly including N6-methyladenosine (m6A), 5-methylcytosine (m5C), 2′-O-methylation, N7-methylguanine (m7G), N1-methyladenosine (m1A), Pseudouridine (Ψ), 5-hydroxymethylcytosine (5hmC), Adenosine to inosine editing (A-to-I editing). The distribution of different types of methylated nucleotides in different species is not uniform. For example, m6A is the most common internal mRNA modification, and it was first discovered in eukaryotic mRNA [3]. m6A, m1A, m5C, m7G, 5-methyluridine (m5U) co-exist in archaea, eukaryotes and prokaryotes [4]. 5hmC exists in almost all mammalian cells, but the abundance in different cell types is very different [5].

RNA methylation is involved in a variety of biological processes, such as transcription and RNA splicing, mRNA translation, extensive translation of circular RNA, cell fate transition, circadian cycle, DNA damage response, heat shock response, neuronal function, sex determination, virus infection [6], which widely affects the function of organisms. In addition, RNA methylation is also involved in gene regulation, DNA repair, stress response, deciphering normal and altered genetic codes, and maintaining RNA biophysics, biochemistry, and metabolic stability [7,8]. Therefore, RNA methylation has an important relationship with some human diseases, including obesity, neurodevelopmental disorders, cancer, congenital hypokeratosis, and X-linked mental retardation [4]. Identifying RNA methylation sites will promote the research of RNA methylation-related diseases and further guide the development of targeted treatment drugs for related diseases. Therefore, the identification of RNA methylation sites has important scientific significance for cell biology and disease mechanism research.

Biological experiments such as high performance liquid chromatography [9] and next-generation sequencing technologies [3,10] have been developed to identify RNA methylation sites. However, these experimental methods are time-consuming, labor-intensive, costly, and slow in detection speed. As a supplement to experimental techniques, computational models, especially machine learning based models, have been developed to identify different types of RNA methylation sites in different species. Different aspects of post-transcriptional modification,

* Corresponding authors.
E-mail addresses: bishoudong@163.com (S. Bi), xlzhu_mdl@hotmail.com (X. Zhu).
[1] These authors equally contributed to the work.

https://doi.org/10.1016/j.ymeth.2022.03.001
Received 31 December 2021; Received in revised form 15 February 2022; Accepted 1 March 2022
Available online 3 March 2022

such as database, functional annotation and computational tools have been reviewed previously [11–13]. In this review, we first summarize the databases, features and learning algorithms that have been used for building machine learning models to recognize RNA methylation sites. Then, from the perspective of different types of RNA methylation, the published predictors used to identify RNA methylation sites were introduced briefly. Through this review of the bioinformatics resources related to RNA methylation prediction, it is hoped that readers could have a deeper understanding of RNA methylation and provide ideas for follow-up research.

2. Databases related to RNA methylation

2.1. RNAMDB

The current version of RNAMDB [14] was proposed in 2011. RNAMDB is an RNA modification database, which contains detailed data of 109 known RNA modifications. The database provides a searchable interface and is easy to use. The record of each entry in the database includes the common name and symbol, chemical structure, element composition and quality, CA registration number and index name, phylogenetic origin, RNA types found, and references for the first reported structural determination and synthesis. The current version of the database is now transferred to The RNA Institute at Albany State University, but the data of RNAMDB are still expanding. In 2010, two new members, agmatine and 8-methyladenosine, were added.

2.2. REACTOME

REACTOME [15] is a human pathway and biological process database, in which RNA modification pathways [16] are also recorded. It is an open source and open access database through manual management and peer review. REACTOME includes a variety of reactions, pathways and biological processes. Among them, biological experts and REACTOME editors cooperate to create pathway annotation, and crossly reference proteins, small molecules, main related research literature and GO controlled vocabulary. REACTOME also includes the transformation of nucleic acids, small molecules, proteins (with or without post-translational modifications) and macromolecular complexes, such as the transport between compartments, the formation of complexes by interactions, and the chemical transformation of classical biochemistry.

2.3. DARNED

DARNED [17] database contains the information of RNA editing sites in human, mouse and *Drosophila*. In addition to the most common A-to-I type RNA editing, part of C-to-U RNA editing sites are also included. Its data sources include differences between cDNA sequences and genome sequences analyzed by bioinformatics, SNP analysis data, miRNA analysis data, high-throughput sequencing results of RNA and DNA samples from the same tissue. The DARNED database supports three query methods for RNA editing sites: query based on region, query based on gene name and query based on sequence. In the retrieval results, information such as the type of each RNA editing site, chromosome position, corresponding gene, SNP and whether it is located in the Alu repeat sequence will be given. DARNED can also link to Wikipedia annotations, which is more convenient.

2.4. RADAR

RADAR [18] is an updating database with strictly annotated A-to-I RNA editing sites for humans, mice and flies. Each editing site in RADAR has detailed comments, which are managed manually. In addition, each editing site contains an organization-specific directory of editing levels from published RNA-seq datasets. The main advantages of RADAR are comprehensive editing of A-to-I editing sites, selection of a large number of comments, and collection of organization-specific editing level measurements for each editing site. Currently, RADAR contains 1.4 million manual editing sites.

2.5. MeT-DB

N6-methyladenosine (m^6A) is a very abundant form of methylation in transcripts. Met-DB [19] is the first complete N6-methyladenosine (m^6A) resource in mammalian transcriptome. The predicted m^6A sites from all published MeRIP-Seq (methylated RNA immunoprecipitation sequencing) data sets were collected in MeT-DB which also includes gene expression data of each MeRIP-Seq data sample. In MeT-DB, the m^6A sites were also integrated with other sites, such as splicing factors, several RBP binding sites and miRNAs. MeT-DB can help to elucidate the biological function of m^6A methyl transcriptome. At present, MeT-DB database only covers the methylation of m6A, but MeT-DB database will continue to be updated and improved.

2.6. RMBase

The establishment of RMBase [20] is based on decoding RNA modification from flux sequencing data. The public RNA modification sites generated by high-throughput sequencing technology were integrated on a large scale, so that they can be easily found, analyzed, visualized and annotated in RMBase. In addition, the relationship between miRNA target and RNA modification can be displayed through the web interface in RMBase, and the RNA modification sites related to clinics can also be shown by integrating GWAS data into the database. At present, more than 100 types of RNA modifications have been integrated into RMBase, which are being maintained and updated continuously.

2.7. MODOMICS

MODOMICS [21] has been updated and maintained since its establishment. The latest update was in 2017, which is an RNA modification pathway database. MODOMICS provides comprehensive information on the chemical structure of modified ribonucleosides, their biosynthetic pathways, the location of modified residues in RNA sequences and RNA modifying enzymes. In the database of the current version, many new data and functions are included, which is the most comprehensive information source in all existing RNA modification databases. MODOMICS contains 163 different RNA modifications. MODOMICS data have been linked to the RNA central database of RNA sequences.

2.8. RCAS

RCAS [22], the RNA Center Annotation System, is a R/Bioconductor package designed as a general reporting tool for functional analysis of regions of interest within the scope of the transcriptome obtained through high-throughput experiments, which makes the process more simplified. These transcriptome regions include RNA modification sites, protein-RNA interaction sites, signal peaks at CAGE-tag positions, and any other collection of query regions at the transcriptome level. RCAS can also be used for performing functional enrichment analysis and identifying motifs. The metagene map, gene center annotation, motif analysis, and gene set analysis provided by RCAS offer contextual knowledge, which can help understand the functions of different biological processes involving RNA.

2.9. REDIdb

REDIdb is a web-based interactive database that contains RNA editing modification data of plant organelles. REDIdb has been updated for three versions [23–25]. The most recent REDIdb 3.0 [25] stores 26618 RNA editing events, which are specifically distributed in 281 organisms and 85 complete organelle genomes. And through the whole

genome map and multiple sequence alignment, the genomics, biology and evolution context of RNA editing are described. These editing modifications include substitution, insertion, and deletion. REDIdb 3.0 also has the function of visualizing amino acid changes caused by RNA editing in protein domains or secondary structures, helping to further understand its potential functional consequences.

2.10. CVm6A

CVm6A [26] is a visual m6A database. To analyze the distribution and patterns of m6A in different cell lines, all available MeRIP-Seq and m6A-CLIP-Seq datasets of human and mouse cell lines from the public database GEO were collected in CVm6A. And then according to 23 human and 8 mouse cell lines, 340,950 and 179,201 m6A peaks were identified respectively. These samples contain most cell lines commonly used in research, such as those associated with stem cell research, immune cell lines, and cancer cell lines. The global distribution of these m6A peaks was studied based on gene type (mRNA/lncRNA), subcellular components, gene regions and other information. The CVm6A database, with its cell-dependent m6A data characteristics, can contribute to the study of the function and regulation of m6A in disease and development.

2.11. RNAmod

RNAmod [27] is a very convenient one-stop online platform that can be used for automatic meta-analysis, functional annotation and visualization of mRNA modifications in up to 21 species, including human, mice, rat, zebrafish, fruit fly, yeast and *Arabidopsis*, etc. RNAmod provides an intuitive interface to display the output, including the distribution of RNA modifications, the range of modifications for different gene features, the functional annotation of the modified mRNA, and the comparison between different groups or specific gene sets. In addition, the known RNA modification sites, as well as the binding site data for hundreds of RNA binding proteins (RBPs) have been integrated into RNAmod to help users compare their modification data with known modifications and explore relationships with known RBP binding sites. In addition, users can download all the output using a one-click link.

2.12. m7GHub

m7GDB, the first database to record internal mRNA m7G sites, is proposed in m7GHub [28], in which, 44,058 internal mRNA m7G sites obtained by 8 experiments, a base resolution technology (m7G-seq) and two non-base resolution methods (m7G-MeRIP-Seq and m7G-miCLIP-Seq) were collected for the first time. In addition to mRNA, the known internal m7G sites on tRNA and rRNA were also recorded in m7GDB. m7GHub also has a m7GDiseaseDB section, in which 1218 genetic mutations related to diseases were gathered. Thus, m7GDiseaseDB provides a research bridge for the association analysis between m7G and diseases. In addition, the m7G sites predictor named m7GFinder and the m7GSNPer web server to evaluate the influence of genetic mutations of the epitranscriptome on the internal m7G methylation status are also proposed in m7GHub.

2.13. REPIC

The REPIC [29] database contains m6A modification and epigenome sequencing data from different species. Specifically, 10 million m6A peaks from 61 different cell lines or tissue types in 11 organisms were collected in the REPIC database. These peaks are derived from publicly available m6A-seq and MeRIP-seq data using a uniform analytical method. In addition, these data can be presented in a built-in genome browser to display a comprehensive map of m6A methylation sites, histone modification sites, and chromatin accessibility regions. Users can search for m6A modification sites according to specific cell line or tissue types and the reports for each peak containing information on genomic location, analysis tools to identify the peak, rich indices, and genome feature annotations are also provided. REPIC also provides a genome browser that can visualize m6A peaks, enrichment levels, and gene expression levels.

2.14. REDIportal

REDIportal [30] is the only database that contains comprehensive human A-to-I editing resources, in which 16 million ab initio A-to-I events detected in 9642 GTEX RNAseq data with single nucleotide resolution were collected. The current version of REDIportal stores single A-to-I location as well as the statistical data and related indicators of each GTEX sample. Users can browse and visually edit sites through embedded and updated JBrowser in REDIportal, or browse them in their gene context through our novel gene view function. In addition, the RNA editings in non-human organisms were also collected in REDIportal which contains annotations for 107,094 A-to-I events from the RNAseq data [31] of newborn mice.

2.15. m^6A-Atlas

m^6A-Atlas [32] is a database dedicated to the record of m6A sites. m^6A-Atlas contains 442,162 m6A sites which were obtained from the data generated in 7 base-resolution techniques. It also contains epitranscriptome profiles which were estimated based on 1363 high-throughput sequencing samples. In addition, some novel features were proposed including the conservation of m6A sites among seven vertebrate species, and m6A epitranscriptome profiles of 10 viruses. Furthermore, the potential pathogenesis of human m6A sites can also be inferred by using m^6A-Atlas, which points out the further research directions for related disease mechanism studies. m^6A-Atlas is an easy-to-use database with search, query, visualization and other functions.

2.16. M6A2Target

The M6A2Target database [33] is the first database depositing the target genes corresponding to the three types of m6A methylases. These three types of enzymes are specifically writers, erasers and readers, collectively referred to as WERs. The m6A WER target gene data in 94 published articles and 78 datasets related to humans and mice were collected and sorted out in M6A2Target, and the potential target genes were analyzed and predicted. The WERs related to m6A are divided into 3 categories through manual sorting. The ones that have been verified by experiments and are more reliable are classified as 'Validated Targets'. The ones with different types of binding relationships (DNA-protein、RNA-protein、protein-protein) obtained from preliminary predictions such as RIP-seq are classified as 'Potential-binding'. And those preliminary screened by gene knockout methods such as m6A-Seq, Ribo-Seq, are classified as 'Potential-perturbation'. Through the classification, users can easily check and learn the records of WERs.

2.17. RMVar

RMVar [34] is specifically designed to deposit RNA modification data involving functional variants, especially the related mutations that affect m6A modification. In addition, eight other RNA modifications including m6Am, m1A, m5C, m7G, m5U, ψ, 2′-O-Me and A-to-I were also collected in RMVar. Totally, RMVar contains 941,955 modification sites and 1,678,126 RNA modification-related variants. These modification-related variants are derived from 3 different data, including single nucleotide discrimination experiments (high confidence), prediction based on CNN from AL-sm6A-SEeq and MeRIP-seq data (medium confidence), prediction using CNN from the full transcriptome (low confidence), and the RBP binding region, miRNA-target and splice site related to the variants are also integrated to help users

study the influence of m6A-related variants on post-transcriptional regulation. In addition, RMVar also includes 31,076 disease-related variants associated with RNA modifications from ClinVar and GWAS, providing support for studying the relationship between m6A-related variants and diseases.

2.18. RMDiease

To fully reveal the association between disease-related variations and their epigenetic transcriptome disturbances, the RMDiease [35] database which contains the genetic variations that affect RNA modification was constructed. After integrating the prediction results of 18 different RNA modification prediction tools and 303,426 experimentally validated RNA modification sites, a total of 202,307 single nucleotide polymorphisms (SNPs) were identified that may affect eight types of RNA modifications (m6A, m5C, m1A, m5U, ψ, m6Am, m7G and Nm), which contain 4289 disease-related variants. These variations suggest that the pathogenesis of the disease may occur mainly in the epitranscriptome layer. In addition, some necessary information including post-transcriptional regulatory mechanisms for these SNPs was annotated in RMDisease.

All the databases were also summarized in Table 1.

3. Features for RNA methylation sites prediction

3.1. Features related to the nucleotide distribution of RNA segments

3.1.1. Binary encoding

3.1.1.1. Nucleotide binary encoding (NBE). The four nucleotides (ACGU) can be encoded as (1,0,0,0), (0,1,0,0), (0,0,1,0), and (0,0,0,1) which is also called one-hot encoding.

3.1.1.2. Dinucleotide binary encoding (DBE). There are 16 types of dinucleotides (AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, and UU) of RNA which can be encoded as (0,0,0,0), (0,0,0,1), (0,0,1,0), (0,0,1,1), …, (1,1,1,1).

3.1.1.3. Physical-chemical based binary encoding (PCBE). The four nucleotides have been classified into six categories according to three types of physical–chemical properties which are ring structure, hydrogen bond and keto or amino functional group (Table 2). According to the three types of physical–chemical properties, the four nucleotides (ACGU) can be encoded as (1,1,1), (0,0,1), (1,0,0) and (0,1,0).

3.1.1.4. Position-specific k-mer. Considering k-mers of RNA as different patterns in the sequences, then, for a specific k-mer pattern, binary indicators are generated by matching the pattern with all possible consecutive-residue-groups along an RNA sequence. If the k-mer pattern was matched, 1 would be assigned, otherwise 0 will be assigned. For example, 'GG' is one of the 2-mers for which an RNA segment 'ACGGCGGUG' could be represented by vector (0,0,1,0,0,1,0,0). 'ACG' is one of the 3-mers, for this specific pattern, the RNA segment 'ACGGCGGUG' could be represented by vector (1,0,0,0,0,0,0). Thus, for position-specific 1-mers, 2-mers and 3-mers, we get L*4, (L-1)*16 and (L-2)*64 features, respectively, for an RNA segment with L nucleotides.

3.1.1.5. Position-specific gapped k-mer. Traditional k-mers are consecutive nucleotide patterns, when we consider k-mers with intervals we obtained gapped k-mers. Similar to the position-specific k-mer features, the position-specific gapped k-mer features are also binary vectors for different gapped k-mer patterns.

Table 1
RNA modification related database.

Name	Year	Objection/Description	Link
RNAMDB	2011	A database for RNA modifications	https://rna-mdb.cas.albany.edu/RNAmods/
REACTOME	2011	A database of reactions, pathways and biological processes	https://www.reactome.org
DARNED	2013	A database for RNA editing	https://darned.ucc.ie
RADAR	2014	An update library of A-to-I editing sites for human, mouse and fly	https://RNAedit.com
MeT-DB	2015	A database of transcriptome methylation in mammalian cells	https://compgenomics.utsa.edu/methylation/
RMBase	2016	A database for RNA modification from high-throughput sequencing data	https://mirlab.sysu.edu.cn/rmbase/
MODOMICS	2017	A database for RNA modifications	https://modomics.genesilico.pl
RCAS	2017	An RNA centric annotation system for transcriptome-wide regions of interest	https://bioconductor.org/packages/release/bioc/html/RCAS.html and https://rcas.mdc-berlin.de
REDIdb	2018	A database of RNA editing events in the genome of plant organelles	https://srv00.recas.ba.infn.it/redidb/index.html
CVm6A	2019	A visual database that explores global m6A patterns across cell lines	https://gb.whu.edu.cn:8080/CVm6A
RNAmod	2019	An integrated system for the annotation of mRNA modifications	https://bioinformatics.sc.cn/RNAmod
m7GHub	2020	The first database of internal mRNA m7G sites	https://www.xjtlu.edu.cn/biologicalsciences/m7ghub
REPIC	2020	A database that can jointly analyze m6A data and epigenome data	https://repicmod.uchicago.edu/repic
REDIportal	2021	A database containing millions of new A-to-I RNA editing events from thousands of RNA-seq experiments	https://srv00.recas.ba.infn.it/atlas/index.html
m6A-Atlas	2021	A database for analyzing the m6A epitranscriptome	https://www.xjtlu.edu.cn/biologicalsciences/atlas
M6A2Target	2021	A database for targets of m6A writers, erasers and readers	https://m6a2target.canceromics.org
RMVar	2021	An updated database of functional variants involved in RNA modifications	https://rmvar.renlab.org
RMDisease	2021	A database of genetic variations affecting RNA modification	https://www.xjtlu.edu.cn/biologicalsciences/rmd

Table 2
The six categories of the four nucleotides (ACGU).

Nucleotides	Physical-chemical properties
C, U	Pyrimidine ring (0)
A, G	Purine ring (1)
A, U	Two hydrogen bonds (1)
C, G	Three hydrogen bonds (0)
G, U	Keto functional group (0)
A, C	Amino functional group (1)

3.1.2. Composition encoding

3.1.2.1. K-mer Composition. The distribution of nucleotides in RNA segments has been thought to relate to the methylation of the target site. K-mer composition was calculated as basic features to describe the distribution, which has been commonly used in previous studies [36–39]. Because RNA was composed of 4 types of nucleotides, an RNA sequence

can be encoded as a 4^k-dimensional vector representing the frequency of each k-mer nucleotide composition. Specifically, a 4-dimensional vector for nucleotide composition, a 16-dimensional vector for dinucleotide composition, and a 64-dimensional vector for trinucleotide composition can be achieved for an RNA segment when k is equal to 1, 2 and 3, respectively.

3.1.2.2. K-spaced Nucleotide Pair Frequencies (KSNPF). K-mer composition describes the distribution of nucleotide with consecutive subsequences, as a kind of discrete subsequences, the composition of k-spaced nucleotide pair (or k-spaced nucleotide pair frequencies) was proposed as features for predicting methylation sites [40,41]. K-spaced nucleotide pair frequencies are used in [42] and it is similar to the 2-mer composition feature. Actually, k-spaced nucleotide pair represents the inconsecutive dinucleotide separated by interval k, for which the dimension is $4^2 = 16$ for any k values.

3.1.2.3. Pseudo K-tuple Nucleotide Composition (PseKNC). Pseudo K-tuple nucleotide composition was proposed by Chen et al., which can be used to characterize an RNA segment with a vector but still keeps the global or long-range sequence order information, via the physicochemical properties of its constituent oligonucleotides [43]. Based on the default 6 physicochemical properties for dinucleotides and 12 physicochemical properties for trinucleotides, various feature vectors to represent an RNA segment as pseudo dinucleotide compositions (PseDNC) or pseudo trinucleotide compositions (PseTNC) can be generated by.

$$D_{PseKNC} = [d_1, \cdots, d_{4^k}, d_{4^k+1}, \cdots, d_{4^k+\alpha}]^T (\alpha < L - k) \tag{1}$$

where, L is the length of an RNA segment and.

$$d_u = \begin{cases} \dfrac{f_u^k}{\sum_{i=1}^{4^k} f_i^k + \omega \sum_{j=1}^{\alpha} \theta_j} (1 < u < 4^k) \\ \dfrac{\omega \theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i^k + \omega \sum_{j=1}^{\alpha} \theta_j} (4^k + 1 < u < 4^k + \alpha) \end{cases} \tag{2}$$

where, f_i^k is the normalized occurrence frequency of the i-th k-mer nucleotide in the RNA sequence, α is the number of the total pseudo components used to show the long-range sequential effect, and ω is the weight factor. θ_j is the j-tier correlation factor that reflects the sequence order correlation between all the most contiguous k-mer nucleotides along the RNA sequence as given by:.

$$\theta_j = \frac{1}{L-j-1} \sum_{i=1}^{L-j-1} C_{i,i+j} (j = 1, 2, \cdots, \alpha; \alpha < L - k) \tag{3}$$

where, the correlation function $C_{i,i+j}$ is defined as.

$$C_{i,i+j} = \frac{1}{\mu} \sum_{g=1}^{\mu} \left[PC_t(D_i) - PC_t(D_{i+j}) \right]^2 \tag{4}$$

where $\mu$is the number of the physical chemical properties considered, and $PC_t(D_i)$ represents the value of t-th physical chemical property for the dinucleotides D_i in the RNA sequence. Hence, the dimensionality of PseKNC feature vector is$4^k + \alpha$.

In addition, the parallel correlation pseudo dinucleotide composition (PCPseDNC) and parallel correlation pseudo trinucleotide composition (PCPseTNC) encoding have been defined similarly to the PseKNC but using more default physiochemical indices.

Furthermore, the series correlation pseudo dinucleotide composition (SCPseDNC) and the series correlation pseudo trinucleotide composition (SCPseTNC) are also encoded in the similar way, however, the correlation function is defined as the product of the physical chemical properties of dinucleotides or trinucleotides. Thus, we would obtain a j-tier correlation factor θ_j for each physical chemical property.

Note the sequence order correlation factor θ_j could also be used as feature.

3.1.2.4. Accumulated Nucleotide/Dinucleotide Frequency (ANF/ADF). Accumulated nucleotide frequency (also called local nucleotide density) is used to describe the occurrence of nucleotide in the subsegment with the former *i* residues, which can be calculated as follows:.

$$d_i = \frac{1}{|N_i|} \sum_{j=1}^{i} f(n_j), f(n_j) = \begin{cases} 1 & if n_j = q \\ 0 & if n_j \neq q \end{cases} \tag{5}$$

where *i* is the current position of RNA sequence, $|N_i|$ is the length of the *ith* prefix string $\{n_1, n_2, \cdots, n_i\}$ in the sequence, and *q* is the symbol of {A, U, C, G}.

Similarly, the accumulated dinucleotide frequency can be calculated.

3.1.2.5. Enhanced Nucleic Acid Composition (ENAC). Enhanced Nucleic Acid Composition (ENAC) encoding is built on the basis of a fixed-length-sequence window, and it can be used to encode an equal-length nucleotide sequence. The dimension of ENAC coding is determined by two parameters, namely, sequence length and sliding window size. ENAC code can be defined as follows:.

$$V = [\frac{N_{A,w_1}}{S}, \frac{N_{C,w_1}}{S}, \frac{N_{G,w_1}}{S}, \frac{N_{U,w_1}}{S}, \frac{N_{A,w_2}}{S}, \cdots, \frac{N_{U,w_{L-S+1}}}{S}] \tag{6}$$

where S denotes the sliding window's size, and N_{t,w_r} denotes the number of certain nucleic acids t, in the sliding window, where$t \in \{A, C, G, U\}$,$r = 1, 2, \cdots, L - S + 1$.

3.1.2.6. Split K-mer Composition. Similar to ENAC, to calculate split k-mer composition, the RNA segment was first split into three subsegments which are 5′ subsegment, 3′ subsegment and the remaining segment. Then, the k-mer composition was calculated as per the three subsegments.

3.1.2.7. xxKGap composition. xxKGap composition feature is a kind of variation of K-mer composition implemented in PyFeat package, where the composition of K-mer with k-gaps is used to describe sequences. For example, 'monoMono1Gap' is the same as 1-spaced nucleotide pair frequencies, 'monoDi1Gap' is obtained by calculating the frequencies of X_XX, and 'diMono1Gap' is obtained by counting the frequencies of XX_X, where X is one of the four kinds of nucleotides.

3.1.3. Physical-chemical-properties-based features

3.1.3.1. Physical-chemical properties (PCP). As mentioned above, the physical–chemical properties (PCP) of the dinucleotides have been incorporated to generate PseKNC features. Actually, the physical–chemical properties themselves can also be used as features. In previous studies [44–46], the following 10 physical–chemical properties were mostly considered: (1) PC^1: rise [47]; (2) PC^2: roll [47]; (3) PC^3: shift [47]; (4) PC^4: slide [47]; (5) PC^5: tilt [47]; (6) PC^6: twist [47]; (7) PC^7: enthalpy [48]; (8) PC^8: entropy [49]; (9) PC^9: stack energy [47]; (10) PC^{10}: free energy [49]. A $10 \times (L-1)$ physical–chemical property matrix PC can be defined as follow:.

$$PC = \begin{bmatrix} PC^1(N_1N_2) & PC^1(N_2N_3) & \cdots & PC^1(N_{L-1}N_L) \\ PC^2(N_1N_2) & PC^2(N_2N_3) & \cdots & PC^2(N_{L-1}N_L) \\ \vdots & \vdots & \vdots & \vdots \\ PC^{10}(N_1N_2) & PC^{10}(N_2N_3) & \cdots & PC^{10}(N_{L-1}N_L) \end{bmatrix} \tag{7}$$

where, $PC^i(N_jN_{j+1})$ is the *ith*$(1 \leq i \leq 10)$ property value for the N_jN_{j+1} $(1 \leq j \leq L-1)$ dinucleotide at *jth* position in the RNA sequence and *L* is the length of the RNA segment. Note that the values of physical–chemical property will need normalization.

3.1.3.2. Auto-covariance of physical–chemical properties (ACPCP). Based on physical–chemical properties of the dinucleotides, a series of auto-covariance (AC) and cross-covariance (CC) [50] have been proposed. The auto-covariance reflects the correlation of the *ith* physical chemical property between two dinucleotides speared by λ nucleotide intervals, which can be calculated as:.

$$\mathrm{AC}(i,\lambda)=\frac{\sum_{j=1}^{\xi-\lambda}\left[PC^{i}\left(N_jN_{j+1}\right)-\overline{PC^{i}}\right]\left[PC^{i}\left(N_{j+\lambda}N_{j+1+\lambda}\right)-\overline{PC^{i}}\right]}{\xi-\lambda} \tag{8}$$

where,$\lambda\in[1,\lambda_m]$, andλ_m, which should be predetermined, is the maximum value that λcan take. $\overline{PC^{i}}$ represents the average of the values along the *lth* segment. Thus, we obtained $(10\times\lambda_m)$-dimensional feature vector from the auto-covariance when 10 physical–chemical properties were used.

3.1.3.3. Cross-covariance of physical–chemical properties (CCPCP). Similarly, the cross-covariance, which measures the correlation between different physical chemical properties of two dinucleotides, can be formulated by.

$$\mathrm{CC}(i_1,i_2,\lambda)=\frac{\sum_{j=1}^{L-\lambda}\left[PC^{i_1}\left(N_jN_{j+1}\right)-\overline{PC^{i_1}}\right]\left[PC^{i_2}\left(N_{j+\lambda}N_{j+1+\lambda}\right)-\overline{PC^{i_2}}\right]}{\xi-\lambda} \tag{9}$$

where, i_1 and i_2 refer to the number of properties which are unequal.

Totally, $(10\times 9\times \lambda_m)$-dimensional feature vector from the cross-covariance is yielded by incorporating 10 physical chemical properties.

3.1.4. Secondary structure related features

3.1.4.1. Secondary structure binary encoding (SSBE). The secondary structure of RNA has been thought to be related to post-transcriptional modification. Predicted secondary structures have been encoded as features for RNA methylation sites detection. By using RNAfold tool, the secondary structures of the input RNA segments can be predicted in dot-bracket notation where unpaired and paired positions were denoted as dots and brackets, respectively. More specifically, the secondary structures can be further classified into hairpin loop (H), multiple loop (M), interior loop (I), paired (P) and bulged loop (B). Then, the binary vectors were used to encode the six types of secondary structures and the terminal gap as (1,0,0,0,0,0), (0,1,0,0,0,0), (0,0,1,0,0), (0,0,0,1,0), (0,0,0,0,1) and (0,0,0,0,0,0), respectively.

According to different classifications of RNA secondary structures, the binary representation can be different.

3.1.4.2. Merged-seq "one-hot" encoding. The RNA sequence consisted of four types of symbols: 'A', 'C','G', and 'U', and the predicted RNA secondary structure of RNAfold [51] consisted of three types of symbols: ".", "(" and ")". Thus, we have 12 types of pairs combining both symbols of RNA sequences and predicted RNA secondary structures. When the string of an RNA sequence was merged with the string of the corresponding predicted RNA secondary structure character-by-character, we obtained a new string. The pairs of the merged string can be encoded using "one-hot" technique, specifically, each of the 12 types of pairs is encoded as a 12-dimensional vector with one 1 and eleven 0.

3.1.4.3. Secondary structure composition (SSC). In addition, in Chen et al.'s work [52], the predicted dot-bracket notation has been encoded into the secondary structure composition. In their method, eight possible structure statuses of tri-nucleotide (i.e., "(((," "((.," "(..," "(.(," ".((," ".(.," "..(," and "…") and the type of the first nucleotide were combined, then the frequencies of these 32(4x8) sequence-structure modes were calculated as features:.

$$\mathrm{R}=\left[f^{A}_{(((},f^{A}_{((.},f^{A}_{(..},\cdots,f^{A}_{\cdots},f^{C}_{(((},\cdots,f^{U}_{\cdots}\right] \tag{10}$$

3.1.5. Stability of the local structure

The minimum free energy (MFE) was used to represent the stability of the local structure of an RNA segment, which can be calculated by RNAFold tool [51]. For each RNA segment, the sequence was shuffled 100 times and the MFE was calculated for each of the shuffled sequences. Then, the Z-score of the MFE value for the original RNA segment was calculated to measure the stability.

3.1.6. Word2vec

The distributed feature representation generated by one of the natural language processing (NLP) techniques, word2vec, has been used to build models for predicting RNA methylation sites. The first step to build a word2vec model is corpus construction. We described an example for building the corpus. Firstly, the sequence of each chromosome is divided into sentences with length of 100 nt. Secondly, the sentences were split into words represented by overlapped k-mers, for instance 3-mers. Thus, the constructed corpus has four different nucleotides (A, C, G, and T) and consequently forms 64 unique words. The second step for the word2vec model is to select an architecture. Generally, there are two architectures which are continuous bag-of-words (CBOW) and skip-gram. Based on the language corpus and the selected architecture, the word2vec model can be built. By using word2vec, each word (3-mer) is represented by a 100-dimensional vector and each sequence with length L is represented by an array of shape$(\mathrm{L}-2)\times 100$. This array can be further used as features in different ways.

3.1.7. Electron-Ion interaction pseudopotentials related features

Electron-Ion Interaction Pseudopotentials (EIIP) has been proposed by Nair and Sreenadhan [53] to encode the distribution of electron–ion energies along a DNA sequence. Considering the similarity between RNA and DNA, it was transferred to encode RNA sequence. The EIIP values of the four nucleotides (i.e.$EIIP_A=0.1260$,$EIIP_c=0.1340$,$EIIP_G=0.0806$, $EIIP_{T/U}=0.1335$) themselves can be used directly to encode the RNA sequence. The average EIIP values of trinucleotides have also been used to encode the RNA sequence.

3.1.8. Entropy-derived features

The nucleotide distributions of an RNA segment have been used as features such as the composition features mentioned above, which can also be evaluated by Shannon information theory. Three different features, Shannon entropy (En), relative entropy (REn) and information gain score (IGS) have been proposed which can be calculated as follows:.

$$En(s)=-\sum_{i\in(A,G,U,C)}p_i^s\log_2(p_i^s) \tag{11}$$

$$REn(s)=-\sum_{i\in(A,G,U,C)}p_i^s\log_2(p_i^s/p_0) \tag{12}$$

$$IGS(s)=En(s)-REn(s) \tag{13}$$

where, p_i^s is the frequency of A, G, U, C in segments, and p_0 is the uniform distribution of each nucleotide occurrence, namely$p_0=1/4$.

3.1.9. SNP Features

Single nucleotide polymorphism is a kind of variant at specific sites in genome. The distribution of SNP sites around the target site may regulate the dynamic methylation. By mapping SNP variants in the transcript obtained from Ensembl database, the RNA segment can be converted into a binary vector (i.e., 0 denotes a non-SNP variant position; 1 denotes an SNP variant position).

3.2. Features related to the distribution of RNA segments in the genome

Genome derived features have been used to predict RNA methylation sites, which can be classified into 10 categories. The features in the first category represent dummy variables indicating whether the site is overlapped to the topological region on the major RNA transcript. All the features in this category can be calculated by the GenomicFeatures R/Bioconductor package using the transcript annotations hg19 TxDb package. The features of the second category are the relative position of the target sites on 3'UTR, 5'UTR and exon. The features of the third category indicate the region length in bp. The features in the fourth category show the scores related to evolutionary conservation. The features in the fifth category show the attribute of the genes or transcripts. The sixth and seventh categories represent RNA annotations related to target sites biology and RNA-binding protein annotation from MeTDB database, respectively. The features in the eighth category reveal the nucleotide distances towards the splicing junctions or the nearest neighboring sites. The features in the ninth category list particular motifs. The features in the tenth category represent the clustering effect of the RNA methylation. Table S1 shows the details of these features.

3.3. Features related to evolutionary conservative

3.3.1. Position-specific K-mer Propensity (PSKP)

Position-specific K-mer nucleotide propensity was proposed by Li et al. [36], which are used to depict the frequency difference of k-mer nucleotide in specific location between positive (the target site can be methylated) and negative (the target site cannot be methylated) sequences.

By using position-specific nucleotide (1-mer) propensity (PSNP) as an example, the *ith* column of the PSNP feature matrix can be calculated as the difference of occurrence frequencies of the 4 types of nucleotide at the *ith* position between the positive and negative samples, which is shown in the following formula.

$$F_{PSNP}(1:4,i) = f^{+}(1:4,i) - f^{-}(1:4,i) = \begin{bmatrix} f^{+}_{A,i} - f^{-}_{A,i} \\ f^{+}_{U,i} - f^{-}_{U,i} \\ f^{+}_{C,i} - f^{-}_{C,i} \\ f^{+}_{G,i} - f^{-}_{G,i} \end{bmatrix} \tag{14}$$

where, i represents the *ith* sequence position of the RNA segments and F_{PSNP} represents the PSNP feature matrix. $f^{+}(1:4,i)$ and $f^{-}(1:4,i)$ denote the occurrence frequency of 4 types of nucleotides (A, U, C, G) for the positive and negative samples, respectively.

Thus, if the length of RNA segments is L, we can obtain a $4 \times L$ global feature matrixF_{PSNP}. Then, an RNA segment can be encoded by assigning the values in the matrix according to sequential positions and nucleotide types.

As for position-specific dinucleotide (2-mer) propensity (PSDP), it has the similar coding rule to PSNP as follow:.

$$F_{PSDP}(1:16,j) = f^{+}(1:16,j) - f^{-}(1:16,j) = \begin{bmatrix} f^{+}_{AA,j} - f^{-}_{AA,j} \\ f^{+}_{AU,j} - f^{-}_{AU,j} \\ f^{+}_{AC,j} - f^{-}_{AC,j} \\ f^{+}_{AG,j} - f^{-}_{AG,j} \\ \vdots \\ f^{+}_{GG,j} - f^{-}_{GG,j} \end{bmatrix} \tag{15}$$

where, j represents the *jth* sequence position of the RNA segments and F_{PSDP} represents the PSDP feature matrix. $f^{+}(1:16,j)$ and $f^{-}(1:16,j)$ denote the occurrence frequency of all types of the 16 dinucleotides (AA, AU, AC, …, GG) at the position of *jth* and (j + Interval + 1)-th nucleotide of the RNA sequence in the positive and negative samples, respectively.

Similarly, we can obtain position specific k-mer propensity matrix which can then be used to encode an RNA segment.

3.3.2. Position-specific Condition Propensity (PSCP)

Position-specific condition propensity (also called nucleotide pair position specificity (NPPS)) was first introduced by Xing et al.[54]. Both global and local information can be captured based on multi-interval nucleotide pair position specificity. The PSCP feature is exactly the combination of PSNP and PSDP features by using the conditional probability formula$p(\mathrm{A}|\mathrm{B}) = \frac{p(AB)}{p(B)}$. An example is shown as follow:.

$$F_{PSCP}(AC,j) = p^{+}(AC,j) - p^{-}(AC,j) = \frac{f^{+}_{AC,j}}{f^{+}_{C,j+Interval+1}} - \frac{f^{-}_{AC,j}}{f^{-}_{C,j+Interval+1}} \tag{16}$$

where, $p^{+}(AC,j)$ and $p^{-}(AC,j)$ denote the conditional probabilities of the nucleotide '*A*' at the *j*-th position and '*C*' at the (*j* + Interval + 1)-th position for the positive and negative samples, respectively. F_{PSCP}represents the PSCP matrix and the meaning of $f^{+}_{C,j+Interval+1}$ and $f^{+}_{AC,j}$ is consistent with $f^{+}_{C,i}$ and $f^{+}_{AC,j}$ in equation (14) and equation (15), respectively.

By analogy, conditional probabilities in other cases can also be calculated using the above formula. Ultimately, the RNA sequence is converted into a (L-Interval-1)-dimensional feature vector.

3.3.3. Bi-profile Bayes (BPB)

Bi-profile Bayes (BPB) is based on all positive and negative samples in the training dataset. The posterior probability was calculated as the occurrence probability of each type of nucleotide at each position for positive and negative samples in training dataset, respectively. Then for a given RNA segment with n residues, a probability vector$V = (p_1, p_2, \cdots, p_n, p_{n+1}, \cdots, p_{2n})$, where the former n probabilities were assigned according to the posterior probability obtained from the positive samples and the later n probabilities were assigned according to the posterior probability obtained from the negative samples.

3.3.4. KNN

The k-nearest neighbor score describes the ratio between the nearest positive samples and all the k nearest samples (including positive and negative samples in the training dataset) to a given RNA segment. The distance between two RNA segments can be represented as the similarity score which is calculated as follows:

$$S(A,B) = \sum_{i=1}^{n} Sim(A[i], B[i]), \tag{17}$$

where, A[i] and B[i] stand for the nucleotides at position i in the RNA segments A and B, respectively. The similarity between two nucleotides can be calculated in different ways, as an example, it may be defined as:.

$$\mathrm{Sim(a,b)} = \begin{cases} 2, if a = b \\ -1, if a \neq b \end{cases} \tag{18}$$

Thus, for different *k*s, we can get different KNN scores.

When features based on multiple RNA segments were used to encode RNA segments, attention should be paid to information leaks. For example, the RNA segments in the validation set cannot be used to generate the PSNP or PSDP matrix. The validation samples cannot be used as background samples in KNN score calculation.

4. Machine learning algorithms for predicting RNA modification sites

As shown in Table 2, most of the computational methods for RNA methylation sites prediction are based on SVM (Support Vector Machine), RF (Random Forest), XGBoost (eXtreme Gradient Boosting) and deep learning framework. In addition, algorithms such as LR (logistic regression) and NB (Naïve Bayes) have also been used. According to our statistics, SVM is the most commonly used machine learning algorithm to develop RNA modification site predictors. Recently, deep learning frameworks such as CNN (convolutional neural network) have also been used for RNA modification sites recognition.

4.1. SVM

Support vector machine (SVM) [55,56] is a type of machine learning algorithm based on statistical learning theory, which was first proposed by Vapnik [56]. In this algorithm, the input vector is mapped to a high-dimensional feature space through a certain non-linear mapping selected in advance. Then, the optimal classification hyperplane is identified, so as to maximize the separation boundary between positive and negative samples. Conceptually, support vectors are the data points closest to the decision plane, and they determine the location of the optimal classification hyperplane. The low-dimensional samples are projected onto a high-dimensional space by using kernel functions, so that the problem can be transformed into a linear problem in the high-dimensional space. The kernel functions for SVM include linear kernel, polynomial kernel, Gaussian kernel, Sigmoid kernel and so on. Among them, the Gaussian kernel, also called radial basis function (RBF), is the most commonly used, which can map data to infinite dimensions. The SVM model with RBF has two essential parameters, namely C and gamma. The former represents the tolerance of the model to errors, and the latter implicitly determines the distribution of the data after mapping to the new feature space. In this review, the predictors based on SVM basically choose to use RBF, and use grid search to optimize the parameters C and gamma.

4.2. Random Forest

Random forest (RF) [57] is essentially a classifier with multiple decision trees, and its output category is determined by the mode of the categories output by multiple trees. A decision tree is a tree structure, in which each internal node represents a test on an attribute, each branch represents a test output, each leaf node represents a category, and several leaf nodes represent the final decision [58]. Random forest also has the important function of calculating the importance of features. For building SRAMP, the feature importance scores were calculated based on random forest. The number of decision trees is an important parameter of the RF classifier, which should be selected according to specific biological problems during the modeling process to obtain better prediction performance [59].

4.3. XGBOOST

XGBoost [60] is a tree-based boosting algorithm. The traditional GBDT model [61] is a weighted regression model that can select features by itself and adapt to multiple loss functions (such as quadratic mean, classification loss). Unlike GBDT which only uses the information of the first derivative, XGBoost does a second-order Taylor expansion of the loss function, and adds a regular term to the objective function to weigh the complexity of the objective function and the model to prevent overfitting. As a new type of machine learning algorithm, the operation process of XGBoost consisted of two parts: learning and reasoning [60]. Among them, the goal of learning is to minimize the loss function, that is, the prediction error is required to be as small as possible when the complexity of the decision tree is as low as possible. The reasoning is the decision tree sequence based on learning process. At present, XGBoost is also widely used in the prediction of RNA modification sites [62–64].

4.4. Logistic Regression

Although Logistic regression (LR) [65] has the word “regression” in its name, it is not a regression algorithm, but a classification algorithm. Logistic function (or Sigmoid function) is mainly used to solve the problem of dichotomy (that is, there are only two outputs, representing two categories respectively). LR makes prediction by associating the characteristics of the sample with the probability of sample occurrence, and the probability is a numerical value. In our review, LR is currently used in the prediction of both m6A and 5hmC sites of RNA [66,67].

4.5. Naïve Bayes

Naïve Bayes (NB) [68] model is a classification algorithm based on Bayes’ theorem and the assumption of independence of features. Although this independence assumption is often overturned in practical applications, NB still often provides competitive classification accuracy. In addition, NB has many characteristics such as computational efficiency and low variance, robustness in the face of noise and robustness in the face of missing values, which makes the NB algorithm widely used in practice. Dou et al. constructed the predictor iRNA-m5C_NB to predict the m5C sites based on the NB algorithm [69].

4.6. Deep learning framework

Deep learning [70] is a subcategory of machine learning. It is inspired by the way the human brain works, and is a learning process that uses deep neural networks to solve feature representation. The architecture of a deep learning model has many layers, which are used to map the connection between input or observation features and output [71]. Different deep learning frameworks have been proposed, of which the convolutional neural network (CNN) and recurrent neural network (RNN) have been used to predict RNA methylation modification sites [72–74]. CNN is a type of feedforward neural network that includes convolution calculations and has a deep structure. It is one of the representative algorithms of deep learning [75]. CNN is mainly composed of input layer, convolution layer, ReLU layer, pooling layer and fully connection layer. RNN is a type of recursive neural network that takes sequence data as input, recurs in the evolution direction of the sequence, and all nodes (cyclic units) are connected in a chained manner. With the development of RNN, Long Short-Term Memory networks (LSTM) [76] and Gated Recurrent Unit networks (GRU) [77] have been proposed successively. BERMP [74] uses the bidirectional gated recurrent unit (BGRU) to construct a predictor that can simultaneously identify m6A sites in multiple species.

5. Predictors for identifying RNA methylation sites

5.1. Predictors for identifying m6A sites

About 30 models for predicting RNA m6A sites have been developed, which has been summarized in Table 3. We will briefly introduce each of them as follows.

5.1.1. iRNA-Methyl

iRNA-Methyl [78] is one of the pioneering predictors specifically developed for the identification of m6A sites in RNA. Based on the training dataset with 1307 positive samples and 1307 negative samples, the model was built by using SVM with the PseDNC features. The performance of the model was evaluated by cross-validation and a user-friendly web server was provided (https://lin.uestc.edu.cn/server/iRNA-Methyl).

Table 3
Predictors for identifying RNA methylation sites.

Predictors	Year	Species	Features[a]	Algorithms	URLs
iRNA-Methyl	2015	*S. cerevisiae*	PseDNC	SVM	https://lin.uestc.edu.cn/server/iRNA-Methyl
m6Apred	2015	*S. cerevisiae*	PCBE; ANF	SVM	https://lin.uestc.edu.cn/server/m6Apred.php
pRNAm-PC	2016	*S. cerevisiae*	K-mer composition; PCP; PseKNC; ACPCP; CCPCP	SVM	https://www.jci-bioinfo.cn/pRNAm-PC
RNA-MethylPred	2016	*S. cerevisiae*	BPB; 2-mer composition; KNN	SVM	/
AthMethPre	2016	*A. thaliana*	NBE; K-mer composition	SVM	https://bioinfo.tsinghua.edu.cn/AthMethPre/index.html
RNAMethPre	2016	*H. sapiens*; *M. musculus*	NBE; K-mer composition; Relative position in mRNA; Stability of the local structure	SVM	https://bioinfo.tsinghua.edu.cn/RNAMethPre/index.html
SRAMP	2016	*H. sapiens*; *M. musculus*	NBE; KNN; KSNPF; SSBE	RF	https://www.cuilab.cn/sramp/
TargetM6A	2016	*S. cerevisiae*	1-mer composition; PSNP; PSDP;	SVM	https://csbio.njust.edu.cn/bioinf/TargetM6A
M6A-HPCS	2016	*S. cerevisiae*	PCP; PseDNC; ACPCP; CCPCP	SVM	https://csbio.njust.edu.cn/bioinf/M6A-HPCS
M6ATH	2016	*A. thaliana*	PCBE; ANF	SVM	https://lin.uestc.edu.cn/server/M6ATH
MethyRNA	2017	*H. sapiens M. musculus*	PCBE; ANF	SVM	https://lin.uestc.edu.cn/server/methyrna
RAM-ESVM	2017	*S. cerevisiae*	PseKNC; Motif features	SVM	https://server.malab.cn/RAM-ESVM/
RAM-NPPS	2017	*S. cerevisiae*; *H. sapiens*; *A. thaliana*	PSCP	SVM	https://server.malab.cn/RAM-NPPS/
iMethyl-STTNC	2018	*S. cerevisiae*; *H. sapiens*	PseDNC; PseTNC; Split K-mer Composition	SVM,	/
M6APred-EL	2018	*S. cerevisiae*	PSKP; PCP; ACPCP; CCPCP; PCBE; ANF	ensemble SVM	https://server.malab.cn/M6APred-EL/
RFAthM6A	2018	*A.thaliana*	PSNP; PSDP; K-mer composition; KSNPF	RF	https://github.com/nongdaxiaofeng/RFAthM6A
BERMP	2018	*H. sapiens*; *M. musculus*; *S. cerevisiae*; *A. thaliana*	ENAC	BGRU and RF	https://www.bioinfogo.org/bermp
HMpre	2018	*H. sapiens*	Site location; Entropy-derived features; SNP features; NBE; PCBE; ANF; K-mer composition	XGBoost	/
Zhang et al.'s method	2018	*E. coli*	PCBE; ANF	SVM	/
M6AMRFS	2018	*S. cerevisiae*; *A. thaliana*; *M. musculus*; *H. sapiens*	DBE; ADF	XGBoost	https://server.malab.cn/M6AMRFS/
DeepM6ASeq	2018	*H. sapiens*; *M. musculus*; Zebrafish	NBE	CNN	https://github.com/rreybeyb/DeepM6ASeq
iRNA(m6A)-PseDNC	2018	*S. cerevisiae*	PseKNC	SVM	https://lin-group.cn/server/iRNA(m6A)-PseDNC.php
Gene2vec	2019	*H. sapiens*; *M. musculus*	NBE; Numbers of DRACH motif neighbors within flanking regions; RNA word embedding	CNN	https://server.malab.cn/Gene2vec/
iN6-Methyl (5-step)	2019	*S. cerevisiae*; *H. sapiens*; *M. musculus*	Word2vec	CNN	https://home.jbnu.ac.kr/NSCL/iN6-Methyl.htm
WHISTLE	2019	*H. sapiens*	PCBE; ANF; Genome-derived features	SVM	https://www.xjtlu.edu.cn/biologicalsciences/whistle; https://whistle-epitranscriptome.com
DeepM6APred	2019	*S. cerevisiae*	NBE; PSCP	SVM	https://server.malab.cn/DeepM6APred
iRNA-Freq	2019	*S. cerevisiae*	xxKGap composition	LR	/
LITHOPHONE	2020	*H. sapiens*	PCBE; ANF; Genomic features	RF	https://180.208.58.19/lith/
WITMSG	2020	*H. sapiens*	PCBE; ANF; Genomic features	RF	https://rnamd.com/intron/
m6A-pred	2020	*S. cerevisiae*	PCBE; ANF; K-mer composition	RF	/
HSM6AP	2021	*H. sapiens*	K-mer composition; KSNPF; mismatch; PCPseDNC; PCPseTNC; SCPseDNC; SCPseTNC; PCBE; Genomic features	XGBoost	https://lab.malab.cn/~lijing/HSM6AP.html

[a] PCBE: physical–chemical based binary encoding; ANF: accumulated nucleotide frequency; PCP: physical–chemical property; ACPCP: auto-covariance of physical–chemical properties; CCPCP: cross-covariance of physical–chemical properties; BPB: Bi-profile Bayes; KNN: K-nearest neighbor score; NBE: nucleotide binary encoding; SSBE: secondary structure binary encoding; KSNPF: K-spaced nucleotide pair frequencies; PSNP: position-specific nucleotide propensity; PSDP: position-specific dinucleotide propensity; ENAC: enhanced nucleic acid composition; PSCP: position-specific condition propensity; PseDNC; pseudo dinucleotide composition; PseTNC; pseudo trinucleotide composition; PseKNC: pseudo K-tuple nucleotide composition; PSKP: Position-Specific k-mer Propensity; DBE: Dinucleotide binary encoding; ADF: Accumulated Dinucleotide Frequency.

5.1.2. m6Apred

m6Apred [79] is a predictor specifically designed to identify the m6A sites in the *Saccharomyces cerevisiae* transcriptome. The features of physical–chemical based binary encoding and accumulated nucleotide frequency were extracted from a training dataset containing 832 positive and 832 negative samples. The model was built based on SVM and was evaluated on an independent test dataset containing 475 positive samples and 4750 negative samples. A web server (https://lin.uestc.edu.cn/server/m6Apred.php) was provided for users.

5.1.3. pRNAm-PC

pRNAm-PC [50] was built based on the same dataset as iRNA-Methyl [78], from which 5 kinds of features were extracted including K-mer composition, physical–chemical property, PseDNC, auto-covariance and cross-covariance of physical–chemical properties. Then, SVM was used to build the final model. The results show that pRNAm-PC performs better and users can easily submit the prediction task on the Web server (https://www.jci-bioinfo.cn/pRNAm-PC).

5.1.4. RNA-MethylPred

RNA-MethylPred [2] was built with the same dataset and learning algorithm as iRNA-Methyl and pRNAm-PC. However, two multiple sequence based features including Bi-Profile Bayes (BPB) and KNN Score was used as features. Their results show that the method performs better than other methods and is a powerful tool for predicting m6A sites.

5.1.5. AthMethPre

AthMethPre [80] is the first online predictor (https://bioinfo.tsinghua.edu.cn/AthMethPre/index.html) for identifying RNA m6A sites of *A. thaliana*. A training dataset containing 5081 positive samples and an equal number of negative samples was collected and two kinds of features including K-mer composition and nucleotide binary encoding were

extracted to build the model based on SVM. In addition, an independent test dataset containing 1500 positive samples and 150,000 negative samples was used to evaluate the generalization of the model.

5.1.6. RNAMethPre

RNAMethPre [81] is a tool proposed in 2016 to predict mammalian mRNA m6A sites. The data of human and mouse were collected and nucleotide binary encoding, K-mer composition, relative position in mRNA and stability of the local structure were extracted to build the model based on SVM. RNAMethPre can be accessed through https://bioinfo.tsinghua.edu.cn/RNAMethPre/index.html. Users can not only submit mRNA sequences on the website for prediction, but also query m6A sites within the transcriptome of previous prediction results and experimental data. In addition, m6A sites in the entire transcriptome can be visualized on the website.

5.1.7. SRAMP

This model is also used to predict RNA m6A sites of mammalian. Based on sequence features including nucleotide binary encoding, KNN score, KSNPF and secondary structure binary encoding, Zhou et al. proposed a computational tool for predicting the m6A sites called SRAMP [82] by using random forest. They optimized the window size of the input data and analyzed the extracted features, resulting in excellent performance in both full transcript and mature mRNA.

5.1.8. TargetM6A

Two kinds of multiple RNA segments based features, position-specific nucleotide propensity (PSNP) and position-specific dinucleotide propensity (PSDP), combined with K-mer composition, were introduced for building the TargetM6A model [83]. To reduce the redundancy of the features, an incremental feature selection (IFS) method was adopted to obtain the optimal feature subset. Then, the final model was built by using SVM. A web server (https://csbio.njust.edu.cn/bioinf/TargetM6A) was provided for users.

5.1.9. M6A-HPCS

The innovation of M6A-HPCS [46] is that a heuristic nucleotide physical–chemical properties selection (HPCS) algorithm was used to select the optimized physicochemical properties which were then incorporated into PseDNC. The features were then fed into SVM for building their model. Their results indicated that the M6A-HPCS model was superior to iRNA-Methyl [78] and pRNAm-PC [50] methods. The M6A-HPCS model is available online at https://csbio.njust.edu.cn/bioinf/M6A-HPCS.

5.1.10. M6ATH

M6ATH [84] is a method to specifically identify the m6A site in the *A. thaliana* transcriptome. Based on physical–chemical based binary encoding and accumulated nucleotide frequency, the model was trained using SVM on a benchmark data set containing 394 positive samples and 394 negative samples. Results of jackknife test confirmed that M6ATH is a reliable method. M6ATH is available at https://lin.uestc.edu.cn/server/M6ATH.

5.1.11. MethyRNA

MethyRNA [85] is another tool for identifying m6A sites of *Homo sapiens* and *Mus musculus*. For Homo sapiens, the dataset contains 1130 positive and 1130 negative samples, and for *Mus musculus*, the dataset contains 725 positive and 725 negative samples. Physical-chemical based binary encoding and accumulated nucleotide frequency were used to encode the segment in the datasets and SVM was used to build the model. The model is available at https://lin.uestc.edu.cn/server/methyrna.

5.1.12. RAM-ESVM

RAM-ESVM [86] is a method for identifying m6A sites of *Saccharomyces cerevisiae* RNA transcriptome using an ensemble SVM. This ensemble SVM integrates three SVM classifiers, SVM-PseKNC, SVM-motif and GkmSVM, which are based on the PseKNC, motif features and optimized gapped Kmers features respectively. Their results indicated that RAM-ESVM was superior to M6A-HPCS [46]. RAM-ESVM can be accessed through https://server.malab.cn/RAM-ESVM/.

5.1.13. RAM-NPPS

A novel feature representation algorithm, multi-interval nucleotide pair position specificity (NPPS) as well as SVM was used to build RAM-NPPS [87] which can identify m6A sites in three species (*S. cerevisiae*, *H. Sapiens*, and *A. taliana*). Datasets for three species were downloaded from the relevant literature [78,82,84]. Results showed that the RAM-NPPS was superior to the existing predictors such as M6A-HPCS [46]. The model is available at https://server.malab.cn/RAM-NPPS/.

5.1.14. iMethyl-STTNC

In this method, two kinds of features, split trinucleotide composition (STNC) and split-tetra-nucleotides-composition (STTNC), were proposed. These features and three learning algorithms were compared and the model based on STTNC and SVM showed the best performance. Compared with existing methods [78,85], iMethyl-STTNC performed better.

5.1.15. M6APred-EL

In this method, three types of features were proposed including position specific K-mer propensity, auto-covariance and cross-covariance of physical–chemical properties and physical–chemical based binary encoding and accumulated nucleotide frequency. Three learners based on each type of feature were built with SVM, which were then integrated as an ensemble classifier, M6APred-EL [45]. The experimental results indicated that the ensemble classifier was superior to single learner. This method is available at https://server.malab.cn/M6APred-EL/.

5.1.16. RFAthM6A

Similar to M6APred-EL, RFAthM6A [42] is an ensemble method for *A. thaliana* by linearly combining four basic learners which were built based on four types of features and RF algorithm. The four types of features used in this method are position specific nucleotide propensity (PSNP), position specific dinucleotide propensity (PSDP), k-spaced nucleotide pair frequency (KSNPF) and K-mer composition. Their results indicated that the ensemble model achieved the best performance. Standalone version of the model is available at https://github.com/nongdaxiaofeng/RFAthM6A.

5.1.17. BERMP

Unlike previous approaches developed using traditional machine learning algorithms, BERMP [74] used a deep learning framework based on a bidirectional GRU (BGRU). BGRU performs well when the training dataset is large, but poorly when the training dataset is small. To this end, a novel encoding method ENAC was proposed and another classifier based on RF was constructed. The results showed that the ensemble classifier, BERMP, based on both BGRU deep learning and RF algorithm achieved the best performance on multi-species datasets with different sizes. The cross-species classifier BERMP is available at https://www.bioinfogo.org/bermp.

5.1.18. DeepM6ASeq

In DeepM6ASeq [72], data were collected from miCLIP-Seq for human, mice, and zebrafish. By using nucleotide binary encoding as input features, CNN models were built to predict RNA m6A sites. Experimental results indicated the model based on CNN was superior to other machine learning algorithms. In addition, the location preference of m6A site was obtained with CNN. The DeepM6ASeq code can be obtained from https://github.com/rreybeyb/DeepM6ASeq.

5.1.19. HMPre

HMpre [62] was constructed for tickling imbalanced datasets. In addition to four commonly used feature representations, nucleotide binary encoding, physical–chemical based binary encoding, accumulated nucleotide frequency and K-mer composition, the authors of HMpre introduced three new features: site location, entropy-derived features and SNP features. The model was built by using XGBoost, and the analysis of the importance of features shows that the use of new features significantly improves the performance of the model. The experimental results showed that the model is competitive with existing predictors.

5.1.20. Zhang et al.'s method [88]

This study is the first to develop a predictor to identify m6A site in the microbial (*E. coli*) genome. In this work, physical–chemical based binary encoding and accumulated nucleotide frequency were used as features and SVM was used as classifier to construct the prediction models. The results of 10-fold cross-validation and independent test showed that the proposed method had satisfactory predictive performance.

5.1.21. M6AMRFS

To build M6AMRFS [63], two types of features, dinucleotide binary encoding and accumulated dinucleotide frequency, were proposed, the importance of which was first sorted by F-score. An SFS method was used to select the optimal feature subset by using XGBoost as classifier. M6AMRFS shows strong generalization ability and can be used to identify m6A sites in multiple species, including *Saccharomyces Cerevisiae*, *Arabidopsis thaliana*, *Musculus*, and *Homo sapiens*, which can be accessed at https://server.malab.cn/M6AMRFS/.

5.1.22. iRNA(m6A)-PseDNC

iRNA(m6A)-PseDNC [89] is a tool for predicting m6A sites in the *Saccharomyces cerevisiae* genome. In this method, feature of PseDNC was extracted and SVM was used as classifier. The model is available at https://lin-group.cn/server/iRNA(m6A)-PseDNC.php.

5.1.23. Gene2vec

In the construction of Gene2vec [90], word embedding and deep neural network were used for the first time to predict m6A site in mRNA sequence. Four prediction models were constructed by using four different RNA sequence representation methods and the optimized CNN networks, respectively. The ensemble model by integrating the four models showed the best performance. Gene2vec is available through the online web server https://server. malab.cn/Gene2vec/.

5.1.24. iN6-Methyl (5-step)

In iN6-Methyl (5-step) [91], the natural language processing technique, word2vec, was used to automatically learn sequence features in the genome, and then the extracted features were fed to a CNN for classification. The continuous bag-of-words (CBOW) was used to train word2vec model. A 10-fold cross validation was used to evaluate the performance of the iN6-Methyl (5-step) method. iN6-Methyl (5-step) was available at: https://home.jbnu.ac.kr/NSCL/iN6-Methyl.htm.

5.1.25. WHISTLE

Based on hundreds of high-throughput sequenced samples, WHISTLE [92] was built by combining genome-derived features, physical–chemical based binary encoding and accumulated nucleotide frequency, which have achieved significant improvements in the accuracy of m6A site prediction. The comparison results show that the performance of WHISTLE is not only better than MethyRNA [85] and SRAMP [82], but also better than existing apparent transcriptome databases predicted by simply searching RRACH motifs, for example, MeT-DB [19] and RMBase [20]. In addition, RNA methylation map, gene expression map and protein–protein interaction data were integrated into WHISTLE to implement a network-based server for facilitating query of high-precision map of human m6A sites. WHISTLE is available at https://www.xjtlu.edu.cn/biologicalsciences/whistle and https://whistle-epitranscriptome.com.

5.1.26. DeepM6APred

In DeepM6APred [93], a deep feature algorithm based on deep belief network (DBN) was proposed, which automatically learned meaningful feature representation from the original input sequence. Then, SVM was used to build the model based on the learned depth features and the traditional manually extracted features. The experimental results showed that the performance of models based on different learning algorithms could be significantly improved by combining traditional hand-extracted features and deep features. The cross-validation results showed that DeepM6APred was superior to other existing predictive tools. DeepM6APred is available at: https://server.malab.cn/DeepM6APred.

5.1.27. iRNA-Freq [67]

In this study, a novel feature extraction algorithm based on Frequent Gapped K-mer Pattern (FGKP) was proposed and linear regression was used to construct the final predictor, iRNA-Freq, which is used to identify the m6A site of *Saccharomyces Cerevisiae*. The 10-fold cross-validation proved the reliability of its prediction performance.

5.1.28. LITHOPHONE

LITHOPHONE [94] was proposed for predicting m6A sites in lncRNA. a new computing framework built using integrated predictors. Due to the limited number of known lncRNA m6A sites, both methylation site data of lncRNA and mRNA were used. However, the patterns of methylation site features of extracted lncRNA and mRNA were different, which required different processing methods. Thus, an ensemble predictor was proposed by combining sequence features and genomic features of lncRNA and mRNA data. LITHOPHONE was tested on an independent test set to prove that it's reliability. LITHOPHONE can be accessed at https://180.208.58.19/lith/.

5.1.29. WITMSG

Unlike other predictors, WITMSG [95] has been designed specifically to be a predictive model for large-scale prediction of human intron m6A RNA methylation sites. Based on the known intron m6A site as the training dataset, the traditional sequence features and various genomic features were calculated, and the RF algorithm was used to build the model. It was worth noting that in order to obtain effective features and further optimize the prediction performance, feature selection was carried out in the study. The model was available at: https://rnamd.com/intron/.

5.1.30. m6A-pred

m6A-pred [96] is unique in that it combines the statistical and chemical characteristics of nucleotides. The fusion of multiple features makes the feature dimension higher, so an evolutionary algorithm is used to optimize the feature vector. RF classifier was used to construct the predictor based on the optimal feature subset. Results on the benchmark dataset showed that m6A-pred was reliable in detecting m6A sites in the transcriptome of *Saccharomyces Cerevisiae*.

5.1.31. HSM6AP

For building HSM6AP [64], the dataset was pre-processed. The positive samples were given different weights according to different single-base resolutions, while the negative samples were randomly clustered and down-sampled, which greatly improved the performance of the model. In the study, a variety of features were extracted including sequence-based features and genome-derived features. In addition, Max-Relevance-Max-Distance (MRMD) was used for feature selection, and the selected features were splicing to generate the final feature vector. XGBoost was then used to train the model. Results indicated that

HSM6AP showed high predictive accuracy and strong generalization ability. HSM6AP is available at https://lab.malab.cn/~lijing/HSM6AP.html.

All in all, we summarized 31 models for predicting m6A sites of RNA, most of which were trained on the data of two species, *Homo sapiens* and *Saccharomyces cerevisiae*. Based on the independent test set results reported in different studies, the highest accuracy (0.953) and AUROC (0.981) were achieved by HSM6AP for mRNA of *Homo sapiens*. WITMSG and LITHOPHONE are the best predictors for identifying m6A sites of introns and lncRNAs of *Homo sapiens,* respectively. In terms of identifying the RNA m6A sites of *Saccharomyces cerevisiae*, m6A-pred, as the latest predictor (subject to the preparation time of this review), is claimed to be superior to other methods.

As summarized in Table 3, different features have been extracted to build these models. The suitable feature combinations and the corresponding effective learning algorithms determine the performances of the models. We counted the occurrence frequency of each kind of feature used in these models (Fig. 1), and physical–chemical based binary encoding (PCBE) and accumulated nucleotide frequency (ANF) are the most commonly used features which were used 11 and 10 times, respectively. Note that genome-derived features have been used for developing the latest models such as HSM6AP, WITMSG and LITHOPHONE.

5.2. Predictors for identifying m5C sites

M5C is another extensively studied RNA modification. The models for predicting RNA m5C sites were summarized in Table 4, which were briefly described as follows.

5.2.1. m5C-PseDNC

m5C-PseDNC [97] is a model built by using SVM with pseudo-dinucleotide compositions which incorporated three physicochemical properties of dinucleotides to extract RNA sequence information.

5.2.2. iRNAm5C-PseDNC

iRNAm5C-PseDNC [98] is a predictive tool for identifying RNA m5C sites of human. The model was built by using RF with pseudo-dinucleotide compositions which incorporated ten types of physicochemical properties of dinucleotides to extract RNA sequence information. The model is available at: https://www.jci-bioinfo.cn/iRNAm5C-PseDNC.

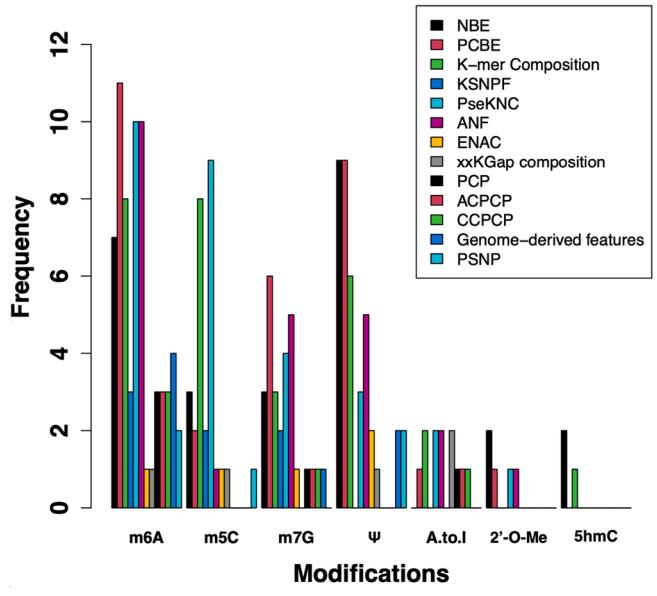

Fig. 1. Occurrence frequencies of different features for the models to predict different modifications of RNA. Only features with total frequency larger than 4 are shown.

5.2.3. pM5CS-Comp-mRMR

In the work of pM5CS-Comp-mRMR [99], different K-mer composition features were extracted and Minimum Redundancy Maximum Relevance algorithm (mRMR) was used to select the optimal feature subset. Based on the optimal feature subset, SVM was finally used as the classifier to construct the predictor. Jackknife cross-validation tests show that pM5CS-Comp-mRMR has good performance.

5.2.4. M5C-HPCR

In the previous study of m6A sites recognition, the authors optimized the subset of nucleotide physicochemical properties by heuristic nucleotide physicochemical property selection (HPCS) algorithm [46] to encode PseDNC features. In this work, the authors put forward the heuristic nucleotide physicochemical properties reduction (HPCR) algorithm which can obtain multiple physical chemical properties of subset to encode multiple PseDNC features, thus, the complementarity between the chemical and physical properties of multiple optimal subsets can be obtained. Multiple models based on SVM with multiple PseDNC features by incorporating multiple physical chemical optimal subsets were built, which were integrated to an ensemble predictor M5C-HPCR [100]. M5C-HPCR can be accessed at: https://cslab.just.edu.cn:8080/M5C-HPCR/.

5.2.5. PEA-m5C

PEA-m5C [101] is a predictor developed to identify m5C sites in *Arabidopsis thaliana*. In this method, three encoding methods, nucleotide binary encoding, K-mer composition and PseDNC, were used to extract the sequence information and then RF algorithm was used to train the model. Verification results on independent test sets show that PEA-m5C has excellent predictive performance. The codes and data for PEA-m5C can be accessed at https://github.com/cma2015/PEA-m5C.

5.2.6. RNAm5Cfinder

RNAm5Cfinder [102] is a machine learning-based web server that predicts m5C sites in RNA. Data were collected from eight tissue/cell types of mouse and human. After extracting nucleotide binary encoding features, RF was used to build predictors for each tissue/cell type, respectively, for accuracy of prediction. The predictive results on the independent test sets showed that RNAm5Cfinder had the ability to identify tissue-specific m5C sites. The server is available at: https://www.rnanut.net/rnam5cfnder.

5.2.7. RNAm5CPred

In order to overcome the shortcomings of small amount and redundancy of existing datasets (i.e. Met240 and Met1900) [100,103,104], the authors first developed a new unbalanced dataset Met935. Then, three feature representation methods, K-mer composition, KSNPF and PseDNC, were used to extract features. Different feature combinations were used as input for SVM algorithm to build models on three datasets, met240, Met1900 and Met935, respectively. The results of jackknife test show that the three prediction models have better or equivalent prediction performance compared with existing methods. In order to determine the final prediction model, an independent test set Test1157 was constructed. The comparison results on Test1157 show that the model trained on Met240 dataset has the best performance and was determined as the final prediction model, RNAm5CPred [105]. RNAm5CPred can be accessed at: https://zhulab.ahu.edu.cn/RNAm5CPred/.

5.2.8. iRNA-m5C

iRNA-m5C [69] was built to predict RNA m5C sites for four species. Four types of features including PseKNC, nucleotide binary encoding, K-mer composition and natural vector, were extracted from RNA

Table 4
Predictors for identifying m5C sites.

Predictors	Year	Species	Features[a]	Algorithms	URLs
m5C-PseDNC	2016	*H. sapiens*	PseDNC	SVM	/
iRNAm5C-PseDNC	2017	*H. sapiens*	K-mer composition; PseDNC	RF	https://www.jci-bioinfo.cn/iRNAm5C-PseDNC
pM5CS-Comp-mRMR	2018	*H. sapiens*	K-mer composition	SVM	/
M5C-HPCR	2018	*H. sapiens*	PseDNC	SVM	https://cslab.just.edu.cn:8080/M5C-HPCR/
PEA-m5C	2018	*A. thaliana*	NBE; K-mer composition; PseDNC	RF	https://github.com/cma2015/PEA-m5C
RNAm5Cfinder	2018	*M. musculus*; *H. sapiens*	NBE; PCBE	RF	https://www.rnanut.net/rnam5cfinder
RNAm5CPred	2019	*H. sapiens*	K-mer composition; KSNPF; PseDNC	SVM	https://zhulab.ahu.edu.cn/RNAm5CPred/
iRNA-m5C	2020	*H. sapiens*; *S. cerevisiae*; *M. musculus*; *A. thaliana*	PseKNC; K-mer composition; NBE; Natural vector	RF	https://lin-group.cn/server/iRNA-m5C/service.html
iRNA-m5C_NB	2020	*H. sapiens*	BPB; K-mer composition; ENAC; xxKGap; EIIP related features; PCPseDNC	NB	/
m5CPred-SVM	2020	*Sapiens*; *Musculus*; *Thaliana*	K-mer composition; KSNPF; PSNP; KSPSDP; PseDNC; PCBE; ANF;	SVM	https://zhulab.ahu.edu.cn/m5CPred-SVM
Staem5	2021	*M. musculus*; *A. Thaliana*	PCPseDNC; PSKP; K-mer composition; PseEIIP	stacking ensemble learning strategy	https://github.com/Cxd-626/Staem5.git

[a] The full name of the abbreviations are the same as Table 3. PCPseDNC: parallel correlation pseudo dinucleotide composition; KSPSDP: K-spaced position-specifc dinucleotide propensity; PseEIIP: electron–ion interaction pseudo potentials of trinucleotide.

sequence. The optimized feature combinations were determined and RF algorithm was used to construct the final models. Users can access iRNA-m5C through the server at: https://lin-group.cn/server/iRNA-m5C/service.html.

5.2.9. iRNA-m5C_NB

iRNA-m5C_NB [69] is a predictor for identifying m5C sites of Homo sapiens based on NB algorithms. BPB was first used for preliminary experiments to determine the processing strategy of the imbalance problem of the dataset and the classification algorithm of the model. Then, a variety of feature extraction methods were used to generate 285 feature combinations from the original data, and the first 57 features were selected as the optimal feature subset by ANOVA F-value. Finally, SMOTEENN was used to deal with the problem of data imbalance, and NB algorithm was used as classifier to train the model. Both the jackknife test and an independent test show that iRNA-m5C_NB is a reliable predictor.

5.2.10. iRNA-m5C_SVM

By using SVM as learning algorithm, iRNA-m5C_SVM [106] was built to recognize the m5C sites in *Arabidopsis thaliana*. Eight widely used features were extracted from the sequences, and then the well-performing feature combination (PSP + Kmer + PseEIIP + SCPseDNC) was selected by RF. Based on the feature combination, the final model was trained by using SVM as the classifier. iRNA-m5C_SVM has achieved good performance on both 10-fold cross-validation and independent test sets.

5.2.11. m5CPred-SVM

m5CPred-SVM [107] was developed to detect m5C sites in *H. sapiens*, *M. Musculus* and *A. Thaliana*. The datasets of the three species were collected, and then six types of sequence features were extracted. The sequential forward feature selection strategy was used to select the optimal feature subset. Based on the optimal feature subset, SVM was used as the classification algorithm to construct the model. Comparison with other methods on independent test sets shows that m5CPred-SVM is a reliable prediction for predicting m5C sites in three species. m5CPred-SVM is available at: https://zhulab.ahu.edu.cn/m5CPred-SVM.

5.2.12. Staem5

Staem5 [108] is a newly proposed tool to identify m5C sites in *M. musculus* and *A. Thaliana*. Four types of encoding methods (PSP, PCPseDNC, K-mer composition, PseEIIP) were used to extract the sequence features. Then F-score was used for feature selection to generate the optimal feature subset. Based on the optimal subset, a stacking algorithm was used to build Staem5, in which the output of the 5 basic classifiers (SVM, GBDT, XGBoost, LightGBM, and ExtraTree) was used as the input of a meta-classifier (LR). The source code is available at: https://github.com/Cxd-626/Staem5.git.

In summary, we reviewed 12 models for predicting m5C sites of RNA, most of which were trained on the data of three species, *Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana*. Based on the independent test set results reported in different studies, Staem5 performed best in predicting RNA m5C sites of *Mus musculus* and *Arabidopsis thaliana*, and iRNA-m5C-NB achieved the best performance for *Homo sapiens*. We counted the occurrence frequency of each kind of feature used in the 12 models (Fig. 1), and PseKNC (including PseDNC) and K-mer Composition were the two most commonly used features which were used in 9 and 8 models, respectively.

5.3. Predictors for identifying m7G sites

Several models have been developed to predict m7G sites as shown in Table 5.

5.3.1. iRNA-m7G

iRNA-m7G [109] is the first predictor to identify RNA m7G sites in the human transcriptome. Four types of features were extracted including physical–chemical based binary encoding, accumulated nucleotide frequency, PseDNC and secondary structure composition, which were fused as input to SVM classifier to build the model. The

Table 5
Predictors for identifying m7G sites.

Predictors	Year	Species	Features [a]	Algorithms	URLs
iRNA-m7G	2019	*H. sapiens*	PCBE; ANF; PseDNC; SSC	SVM	https://lin-group.cn/server/iRNA-m7G/
m7GHub	2020	*H. sapiens*	PCBE; ANF; Genome-derived features	SVM	https://www.xjtlu.edu.cn/biologicalsciences/m7ghub
M7G_model	2020	*H. sapiens*	NBE; PCBE; ANF; K-mer composition; PseKNC	SVM	https://github.com/MapFM/m7g_model.git
XG-m7G	2020	*H. sapiens*	NBE; KSNPF; ENAC; PCBE; ANF; SCPseDNC	XGBoost	https://flagship.erc.mon-ash.edu/XG-m7G/
m7GPredictor	2020	*H. sapiens*	PseKNC; K-mer composition; KSNPF; PCBE	SVM	https://github.com/NWAFU-LiuLab/m7Gpredictor
m7G-IFL	2020	*H. sapiens*	NBE; K-mer composition; PCBE; ANF; PCP; ACPCP; CCPCP	XGBoost	https://server.malab.cn/m7G-IFL/

[a] The full name of the abbreviations are the same as Tables 3 and 4. SSC: secondary structure composition;.

model is available at: https://lin-group.cn/server/iRNA-m7G/.

5.3.2. m7GHub

m7GHub[28] is currently the first and only database of internal mRNA m7G sites which is composed of four parts, m7GDB, m7GFinder, m7GSNPer, and m7GDiseaseDB. m7GDB is a database established by collecting experimentally validated internal mRNA m7G sites and annotated potential post-transcriptional regulation. m7GFinder is a predictor built on data from m7GDB to identify the m7G sites of internal mRNA. For building m7GFinder, both sequence and genome-derived features were extracted and SVM classification algorithm was used. The online platform of m7GHub is at https://www.xjtlu.edu.cn/biologicalsciences/m7ghub.

5.3.3. M7G_model [110]

In the study, five types of features were extracted including nucleotide binary encoding, physical–chemical based binary encoding, accumulated nucleotide frequency, K-mer composition and PseKNC, which were further selected by three feature selection methods, mRMR, F-Score and Relief. The optimal feature subsets were fed to SVM to build the model. The model is available at https://github.com/MapFM/m7g_model.git.

5.3.4. XG-m7G

XG-m7G [111] is a predictor for detecting RNA m7Gssites of human. Six types of features including nucleotide binary encoding, KSNPF, ENAC, physical–chemical based binary encoding, accumulated nucleotide frequency and SCPseDNC were extracted, which were further selected by SHAP (Shapley Additive Explanation). Based on the selected optimal feature subset, XGBoost was used to build the model. Both 10-fold cross validation and jackknife testing demonstrate the superior performance of XG-m7G. XG-m7G is available at: https://flagship.erc.mon-ash.edu/XG-m7G/.

5.3.5. m7GPredictor

The m7GPredictor [112] is another predictor for identifying RNA m7G sites of human. Firstly, PseDNC, PseKNC, K-Mer, KSNPF and physical–chemical based binary encoding were used to encode the sequence information. The RF algorithm was then used to optimize the feature vector and generated the optimal feature subset which contained the 240 features. Finally, SVM was used to train the model. Through 10-fold cross validation, jackknife testing, and independent testing, m7GPredictor is proved to be a competitive predictor. The datasets and source code for building m7GPredictor are available at https://github.com/NWAFU-LiuLab/m7Gpredictor.

5.3.6. m7G-IFL

In m7G-IFL [113], an iterative feature representation algorithm was used to obtain the optimal feature representation from seven types of features including nucleotide binary encoding, K-mer composition, physical–chemical based binary encoding, accumulated nucleotide frequency, physical–chemical properties, auto-covariance and cross-covariance of physical–chemical properties. XGBoost was selected as the classification algorithm to build the model. The results showed that the iterative feature representation algorithm and XGBoost algorithm could improve the prediction performance effectively. The model is available at: https://server.malab.cn/m7G-IFL/.

Overall, we summarized 6 models for predicting m7G sites of RNA for *Homo sapiens*. According to the cross validation results on the same benchmark dataset reported in different studies, M7G_model performed best among all these models. We counted the occurrence frequency of each kind of feature used in the 6 models (Fig. 1), and physical–chemical based binary encoding (PCBE) has been used in all the 6 models.

5.4. Predictors for identifying Ψ sites

Pseudouridine (Ψ) is another modification that has been studied extensively. The models for predicting Ψ sites were summarized in Table 6, which were briefly described as follows.

5.4.1. tRNAmod

TRNAmod [114] is the first predictor for identifying the Ψ sites in tRNA. In this method, the model was trained by SVM based on mixed features with nucleotide binary encoding, K-mer composition, secondary structure binary encoding on the Modomics-2008 dataset. Its effectiveness was proved by both 5-fold cross-validation and independent testing. The model is available at: https://crdd.osdd.net/raghava/trnamod.

5.4.2. PPUS

The data used to build this model was the PUS-specific Ψ sites which were collected from recent works [115,116]. Nucleotide binary encoding was used to encode the RNA segments and then SVM was selected as the classification algorithm. The results showed that PPUS [117] can accurately identify the Ψ sites modified by PUS4 in humans and the Ψ sites modified by PUS1, PUS4 and PUS7 in yeast. PPUS is the first predictor to identify PUS-specific Ψ sites. The model can be accessed at: https://lyh.pkmu.cn/ppus/.

5.4.3. iRNA-PseU

iRNA-PseU [84] is a predictor which can predict Ψ sites for three species, *H. sapiens*, *M. Musculus* and *S. Cerevisiae*. In this work, PseKNC was used to encode RNA sequence and SVM was used to train the model. iRNA-PseU can be freely accessed at:https://lin.uestc.edu.cn/server/iRNA-PseU.

5.4.4. PseUI

Based on the same datasets as iRNA-PseU, four types of features including K-mer composition, PseDNC, PSNP and PSDP were extracted from RNA sequence, from which the optimal feature subsets were

Table 6
Predictors for identifying Ψ sites.

Predictors	Year	Species	Features [a]	Algorithms	URLs
tRNAmod	2014	*B. subtilis*; *E. coli*; *H. volcanii*; *M. capricolum*; *S. cerevisiae*	NBE; K-mer composition; SSBE	SVM	https://crdd.osdd.net/raghava/trnamod
PPUS	2015	*Yeast*; *H. sapiens*	NBE	SVM	https://lyh.pkmu.cn/ppus/
iRNA-PseU	2016	*H. sapiens*; *M. musculus*; *S. cerevisiae*	PCBE; ANF; PseKNC	SVM	https://lin.uestc.edu.cn/server/iRNA-PseU
PseUI	2018	*M. musculus*; *S. cerevisiae*; *H. sapiens*	K-mer composition; PseDNC; PSNP; PSDP	SVM	https://zhulab.ahu.edu.cn/PseUI
iPseU-CNN	2019	*M. musculus*; *S. cerevisiae*; *H. sapiens*	NBE	CNN	/
iPseU-NCP	2019	*M. musculus*; *S. cerevisiae*; *H. sapiens*	PCBE	RF	https://github.com/ngphubinh/iPseU-NCP
RF-PseU	2020	*M. musculus*; *S. cerevisiae*; *H. sapiens*	NBE; DBE; PCBE; ANF; EIIP; ENAC; xxKGap	RF	https://148.70.81.170:10228/rfpseu
iPseU–Layer	2020	*M. musculus*; *S. cerevisiae*; *H. sapiens*	K-mer composition; PSTPss; PCBE	RF	/
XG–PseU	2020	*H. sapiens*; *M. musculus*; *S. cerevisiae*	K-mer composition; PCBE; ANF; NBE	XGBoost	https://www.bioml.cn/
PIANO	2020	*H. sapiens*	PCBE; PSNP; Cluster information; Genome-derived features	SVM	https://piano.rnamd.com
EnsemPseU	2020	*H. sapiens*; *M. musculus*; *S. cerevisiae*	K-mer composition; NBE; ENAC; PCBE; ANF	ENSEMBLE MODEL(SVM,XGBoost, NB, KNN, and RF)	https://github.com/biyue1026/EnsemPseU
PSI-MOUSE	2020	*M. musculus*	PCBE; ANF; Genome-derived features	SVM	https://www.xjtlu.edu.cn/biologicalsciences/psimouse; http://psimouse.rnamd.com
MixedCNN-PseUI	2020	*H. sapiens*; *M. musculus*; *S. cerevisiae*	NBE;	CNN (mixed)	/
PA-PseU	2021	*H. sapiens*; *M. musculus*; *S. cerevisiae*	K-mer composition; SOCF	PA	https://github.com/Jensen-Wang/PA-PseU
Aziz et al.'s model	2021	*H. sapiens*; *M. musculus*; *S. cerevisiae*	NBE; Merged-seq "one-hot" encoding	CNN	https://103.99.176.239/ipseumulticnn/
Porpoise	2021	*H. sapiens*; *M. musculus*; *S. cerevisiae*	NBE; PseKNC; PCBE; PSTNPss	Stacking ensemble learning framework	https://web.unimelb-bioinfortools.cloud.edu.au/Porpoise/

[a] The full name of the abbreviations are the same as Tables 3 and 4. EIIP: electron–ion interaction pseudopotential; SOCF: sequence order correlated factors; PSTNPss: Position-specific trinucleotide propensity based on single strand.

selected by using sequential forward feature selection method. Then, based on the optimal feature subsets, SVM was used to train the model which is named as PseUI [118]. Jackknife test and the independent test proved that PseUI had good predictive power. The model can be accessed at: https://zhulab.ahu.edu.cn/PseUI.

5.4.5. iPseU-CNN

iPseU-CNN [73] is the first model to predict Ψ sites by using CNN. The feature representation by CNN was compared with the model based on hand-crafted features and SVM. The results showed that the deep learning framework has achieved better prediction performance.

5.4.6. iPseU-NCP

For building iPseU-NCP [119], physical–chemical based binary encoding (also known as nucleotide chemistry (NCP)) was used to encode RNA sequence and RF algorithm was used as the classifier to train the model. The codes and data for iPseU-NCP are available at: https://github.com/ngphubinh/iPseU-NCP.

5.4.7. RF-PseU

In order to establish RF-PseU [120], after using different feature extraction methods to obtain high-dimensional feature matrices, the light gradient boosting machine algorithm and incremental feature selection strategy were used to remove redundant features. Then RF was selected as the classification algorithm to train the model. Leave-one-out (LOO) cross-validation and independent testing were used to evaluate the model. The results showed that RF-PseU achieved good performance. RF-PseU was available at: https://148.70.81.170:10228/rfpseu.

5.4.8. iPseU-Layer

The framework of iPseU-Layer [121] contains five machine learning layers including a features selection layer, three feature extraction and fusion layers, and a prediction layer. In the feature selection layer, dimensionality reduction was performed on the extracted features in the data preprocessing stage. In the feature extraction and fusion layer, features were extracted and fused by integrating six machine learning algorithms to obtain intermediate outputs. In the prediction layer, an

ensemble learning model was built based on RF.

5.4.9. XG-PseU

As its name suggests, XG-PseU [122] is an XGBoost-based predictor that identifies Ψ sites. First, datasets for H. sapiens, M. Musculus, and S. Cerevisiae were collected. Then, different features including K-mer composition, physical–chemical based binary encoding, accumulated nucleotide frequency, nucleotide binary encoding were extracted, from which the optimal feature subsets were selected by using forward feature selection strategy. Based on the optimal feature subsets, XGBoost was used to build the models. XG-PseU was available at: https://www.bioml.cn/.

5.4.10. PIANO

PIANO [123] is a network platform that was developed to recognize the Ψ sites and functionally annotate the Ψ sites. Both sequential and genome-derive features were extracted to build the model which achieved high predictive performance for both the whole transcripts and mature mRNA. In addition to predicting Ψ sites, the post-transcriptional regulatory mechanisms were used in PIANO to annotate the potential function of the predicted Ψ sites. Users can use PIANO by logging on: https://piano.rnamd.com.

5.4.11. EnsemPseU

EnsemPseU [124] is an ensemble model for identifying Ψ sites. First, a variety of features including K-mer composition, nucleotide binary encoding, ENAC, physical–chemical based binary encoding, accumulated nucleotide frequency was extracted and then the redundant features were removed by using the chi-square feature selection method. The selected features were fed to SVM, XGBoost, NB, RF and k-nearest neighbor (KNN) to build base learners which were then integrated to establish the ensemble model, EnsemPseU. The source code and datasets used to build EnsemPseU are available at https://github.com/biyue1026/EnsemPseU.

5.4.12. PSI-MOUSE

Similar to PIANO, PSI-MOUSE [125] is a predictor that combines sequence features and genomic features to predict and annotate the Ψ sites of mouse. Firstly, the traditional sequence features and 38 genomic features of mouse genome were obtained. Then the SVM was used to train the model. PSI-MOUSE is available at https://www.xjtlu.edu.cn/biologicalsciences/psimouse and https://psimouse.rnamd.com.

5.4.13. MixedCNN-PseUI [126]

In this study, the deep learning framework of CNN was used to build the model. By using one-hot as input, CNN could automatically extract effective features and predict Ψ sites. Both cross-validation and independent testing proved the reliability of the model.

5.4.14. PA-PseU

In PA-PseU [127], two types of features, K-mer composition and sequence order correlation factor, were extracted. These features were then selected by combining chi-square test and logistic regression. Finally, the Passive-Aggressive algorithm was used to build the model. The predictive performance of PA-PseU was evaluated using 10-fold cross-validation, LOO cross validation, and independent testing. The code and datasets for PA-PseU are available at https://github.com/Jensen-Wang/PA-PseU.

5.4.15. Aziz et al.'s model [128]

In this study, on the basis of the original sequence, the secondary structure predicted by RNAfold was combined to produce the "merged-seq". The merged sequence was then encoded by using merged-seq one-hot encoding technique. Next, the CNN framework was used to construct the model. The experimental results demonstrated the effectiveness of the encoding technique.

5.4.16. Porpoise

Porpoise [129] is a predictor used to identify Ψ sites in H. sapiens, M. musculus and S. cerevisiae. Eighteen feature encoding techniques were used to extract RNA sequence information, and nine machine learning algorithms (AdaBoost, ERT, XGBoost, GBDT, LR, KNN, SVM, RF, GaussianNB) was used to train the models. Four encoding schemes (nucleotide binary encoding, PseKNC, physical–chemical based binary encoding and PSTP) were selected as the input features of the final model according to the prediction performance. Nine base learners were built based on the nine learning algorithms, which were integrated into the stacking model by using LR as the meta-classifier.

Totally, we reviewed 16 models for predicting Ψ sites of RNA mainly for three species, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae*. Based on the results reported in different studies on the same datasets, PA-PseU performed best among all the models for the three species. By counting the occurrence frequency of each kind of feature used in the 16 models (Fig. 1), nucleotide binary encoding (NBE) and physical–chemical based binary encoding (PCBE) were the most commonly used features which were used in 9 models.

5.5. Predictors for identifying A-to-I editing sites

5.5.1. PAI

PAI [130] was built based on a training dataset with 125 positive samples (containing a central A-to-I editing site) and 119 negative samples. PseDNC was used to encode the RNA sequence and SVM was used to build the model. PAI was validated not only by jackknife test, but also on an independent dataset containing 300 positive samples. In addition, a web server has been established for PAI, which can be accessed publicly at: https://lin.uestc.edu.cn/server/PAI.

5.5.2. iRNA-AI

iRNA-AI [131] is the first predictor to identify human A-to-I editing sites based on sequence information. Two types of features, physical–chemical based binary encoding and accumulated nucleotide frequency, were used to encode the RNA sequences, and then SVM was used to train the model. The model is available at https://lin.uestc.edu.cn/server/iRNA-AI/.

5.5.3. PAI-SAE

Based on the same training dataset as PAI, PAI-SAE [132] was trained by using sparse auto-encoder (SAE) to identify the A-to-I editing sites of D. melanogaster. In this method, physical–chemical properties of dinucleotides, auto-covariance of physical–chemical properties, PseDNC and accumulated nucleotide frequency were used to encode RNA sequence.

5.5.4. EPAI-NC

EPAI-NC [133] is another predictor for identifying the A-to-I editing sites of D. melanogaster. Firstly, RNA sequence information was extracted with K-mer compositions and gapped k-mer compositions. Then, based on Pareto principle and Pearson correlation coefficient, the extracted sequence-based features were selected and optimized, and the most effective feature subset was obtained. At last, the model was built by using locally deep SVM (LD-SVM). EPAI-NC is available at: https://epai-nc.info.

5.5.5. PRESa2i

The features used in PRESa2i [134] are similar to EPAI-NC, however, a hybrid feature selection technique was used to calculate the importance of features and obtain the optimal feature subset. And then Hoeffding tree was used to build the model. The related materials of PRESa2i are available on https://github.com/swakkhar/RNA-Editing/, and PRESa2i can be used for free in https://brl.uiu.ac.bd/presa2i/index.php.

All these methods for predicting A-to-I editing sites were summarized

in Table 7. To sum up, five models for predicting A-to-I sites of RNA have been developed. Four of them are for *Drosophila melanogaster*, and the other one is for *Homo sapiens*. Based on the results reported in different studies on the same datasets, EPAI-NC is the best model for the *Drosophila melanogaster*. By counting the occurrence frequency of each kind of feature used in the 5 models (Fig. 1), K-mer Composition, PseDNC, accumulated nucleotide frequency (ANF) and xxKGap composition were the most commonly used features which were used in 2 models.

5.6. Predictors for identifying 2′-O-Me sites

5.6.1. Deep-2′-O-Me

Deep-2′-O-ME [135] is a model for identifying RNA 2′-O-ME sites. For building Deep-2′-O-ME, a sequence embedding method, dna2vec, was used to represent the features of the pre-mRNA sequence, and then CNN framework was used to fine-tune and build the prediction model based on the obtained feature representation. Deep-2′-O-ME has been evaluated on both balanced datasets and imbalanced datasets with different ratios. The results showed that Deep-2′-O-ME could be a practical tool for predicting 2′-O-ME sites.

5.6.2. iRNA-2OM

iRNA-2OM [136] is a SVM-based computational model for predicting 2′-O-Me sites in Homo sapiens. First, physical–chemical based binary encoding, accumulated nucleotide frequency and SCPseDNC were used to extract features from RNA sequence. Then mRMR was used to select the optimal feature subset. Finally, SVM was used to train the model. iRNA-2OM can be freely accessed at: https://lin-group.cn/server/iRNA-2OM/.

5.6.3. iRNA-PseKNC(2methyl)

Based on the same dataset as iRNA-2OM, iRNA-PseKNC(2methyl) [137] was built on the CNN framework for identifying RNA 2′-O-Me sites. In this study, both machine learning and deep learning were used to build prediction models. For the machine learning model, RNA sequence information was extracted using multivariate mutual information (MMI) and K-mer composition, and SVM was used as the classifier. In the deep learning model, CNN was used to automatically extract sequential features and train the model. The cross-validation results showed that the deep learning model, namely iRNA-PseKNC (2methyl), is superior to the machine learning model.

5.6.4. DeepOMe

DeepOMe [138] has adopted CNN and bidirectional long short-term memory (BLSTM) framework to identify 2′-O-Me sites in the human transcriptome. In this method, both RNA sequences and their labels were used to train the model, and this model achieved very high accuracy. DeepOMe was available at: https://deepome.renlab.org.

All these methods for predicting 2′-O-ME sites were summarized in Table 8. Four models for predicting 2′-O-ME sites of RNA for *Homo sapiens* have been developed. Based on the reported results in different studies, DeepOMe achieved the highest AUROC value, however, the comparison is not totally fair because those methods were not tested on the same benchmark datasets. By counting the occurrence frequency of each kind of feature used in the 4 models (Fig. 1), nucleotide binary encoding (NBE) was the most commonly used feature which was used in 2 models.

Table 8
Predictors for identifying 2′-O-Me sites.

Predictors	Year	Species	Features [a]	Algorithms	URLs
Deep-2′-O-Me	2018	*H. sapiens*	rna2vec	CNN	/
iRNA-2OM	2018	*H. sapiens*	PCBE; ANF; Type 2 PseKNC	SVM	https://lin-group.cn/server/iRNA-2OM/
iRNA-PseKNC (2methyl)	2019	*H. sapiens*	NBE	CNN	/
DeepOMe	2021	*H. sapiens*	NBE	CNN; LSTM	https://deepome.renlab.org

[a] The full name of the abbreviations are the same as Tables 3 and 4.

5.7. Predictors for identifying 5hmC sites

5.7.1. iRNA5hmC

iRhm5hmC [139] is the first predictor for identifying 5hmC sites based on machine learning algorithm. In this method, K-mer composition and nucleotide binary encoding was used to extract sequence features which were then selected by the variance analysis (ANOVA) and sequential forward feature selection (SFS) method. Then, SVM was used to train the model. iRNA5hmC is available at: https://server.malab.cn/iRNA5hmC.

5.7.2. iRNA5hmC-PS

In iRNA5hmC-PS [66], three types of features, position-specific gapped k-mer, position-specific k-mer and guanine-cytosine content (GC-content) were extracted, among which position-specific gapped k-mer can retain both short and long-range position-specific information of RNA sequences. In order to reduce the feature dimension and avoid the risk of overfitting, RF was used for feature selection to obtain an optimized feature subset. Finally, LR was used to train the model. The benchmark dataset, source code and other related materials of iRNA5hmC-PS can be obtained at https://github.com/zahid6454/iRNA5hmC-PS. In addition, iRNA5hmC-PS is also available at: https://103.109.52.8:81/iRNA5hmC-PS.

5.7.3. iRhm5CNN

iRhm5CNN [140] is a deep learning model for identifying 5hmC sites. Using nucleotide binary encoding as the input, the CNN framework

Table 7
Predictors for identifying A-to-I editing sites.

Predictors	Year	Species	Features [a]	Algorithms	URLs
PAI	2016	*D. melanogaster*	PseDNC	SVM	https://lin.uestc.edu.cn/server/PAI
iRNA-AI	2017	*H. sapiens*	PCBE; ANF	SVM	https://lin.uestc.edu.cn/server/iRNA-AI/
PAI-SAE	2018	*D. melanogaster*	PCP; ACPCP; CCPCP; PseDNC; ANF	Sparse Auto-Encoder	/
EPAI-NC	2019	*D. melanogaster*	K-mer compositions; xxKGap	LD-SVM	https://epai-nc.info
PRESa2i	2020	*D. melanogaster*	K-mer compositions; xxKGap; Statistical features (ratio of start and end codons and distribution of bases)	Hoeffding tree	https://brl.uiu.ac.bd/presa2i/index.php;https://github.com/swakkhar/RNA-Editing/

[a] The full name of the abbreviations are the same as Tables 3 and 4.

was used to build the model. Experimental results showed that iRhm5CNN was significantly superior to iRNA5hmC. iRhm5CNN is available at https://nsclbio.jbnu.ac.kr/tools/iRhm5CNN/.

All these methods for predicting 5hmC sites were summarized in Table 9. Three models for predicting 5hmC sites of RNA for *D. melanogaster* have been developed. Based on the reported cross-validation results on the same dataset in different studies, iRhm5CNN achieved the highest AUROC. By counting the occurrence frequency of each kind of feature used in the 3 models (Fig. 1) and the results of iRhm5CNN, nucleotide binary encoding (NBE) was considered as the most effective feature.

5.8. Predictors for identifying m1A sites

5.8.1. RAMPred

RAMPred [141] is the first predictor to identify m1A sites of RNA. The data for three species, *H. sapiens*, *M. Musculus* and *S. Cerevisiae* was collected. Physical-chemical based binary encoding and accumulated nucleotide frequency were used to encode RNA sequences, and SVM was used to train the models. RAMPred is available at: https://lin.uestc.edu.cn/server/RAMPred.

5.8.2. ISGm1A

In ISGm1A [142], both traditional sequential features and genome-derived features were extracted. Through comparative analysis, the advantage of combining two categories of features was demonstrated. In addition, RF was selected as the classifier of the final model from five commonly used classifiers (RF, SVM, KNN, LR, and XGBoost). ISGm1A is available at: https://github.com/lianliu09/m1a_prediction.git.

Between the two predictors for identifying m1A sites, ISGm1A shows better performance according to the results reported in the two works.

5.9. Predictors for identifying other RNA methylation sites

5.9.1. m5UPred

m5UPred [143] is the first predictor for identifying RNA m5U sites. The data sets used in this study were collected from data generated by different sequencing technologies and different cell types. Two types of features were extracted from RNA sequence, which are physical–chemical based binary encoding and accumulated nucleotide frequency. SVM was used to train the model. The good generalization was proved based on the predictive results on independent test sets. The model is available at https://www.xjtlu.edu.cn/biologicalsciences/m5u.

5.9.2. m6AmPred

m6AmPred [144] is the first method for predicting m6Am sites of RNA. Three types of features were extracted from RNA which are physical–chemical based binary encoding, accumulated nucleotide frequency, EIIP-derived features. The features were analyzed and EIIP-derived features show better representability. Then, four learning algorithms, SVM, RF, Linear Model (GLM), and eXtreme Gradient Boosting (XgbDart) were adopted to build the models. The model based on XgbDart is superior to others. This is the first time that XgbDart was used to build RNA modification predictors. Both 10-fold cross-validation and independent testing demonstrated the effectiveness of the model. m6AmPred is available at: https://www.xjtlu.edu.cn/biologicalsciences/m6am.

5.10. Predictors for simultaneously identifying multiple RNA methylation sites

5.10.1. iRNA-PseColl

Most of the existing predictors can only recognize a single type of RNA modification, iRNA-PseColl [104] is the first platform developed to identify several different types of RNA modification sites including m1A, m6A and m5C. For building iRNA-PseColl, PseKNC was used to encode RNA sequences and SVM was used to the model. Cross-validation and jackknife test demonstrated the high predictive accuracy of iRNA-PseColl. iRNA-PseColl can be accessed at: https://lin.uestc.edu.cn/server/iRNA-PseColl.

5.10.2. iMRM

iMRM [145] is a predictor that can simultaneously predict m6A, m5C, m1A, ψ, and A-to-I editing sites. iMRM first used six encoding methods to extract RNA sequence features and generate high-dimensional feature vectors. Then feature selection techniques were used to select the optimal feature subset. The specific process was to sort the features based on the F-score obtained by the XGBoost software package, and then use the incremental feature selection (IFS) strategy to determine the top 50 features as the best feature subset. Finally, XGBoost algorithm was determined as the classifier of the model. Both 10-fold cross-validation and jackknife test were used to verify the superior performance of iMRM. The publicly accessible web server of iMRM is at https://www.bioml.cn/XG_iRNA/home.

5.10.3. DeepPromise

DeepPromise [146] is an ensemble model to predict both m1A and m6A sites. In this study, three types of features, ENAC, nucleotide binary encoding and RNA embedding, were extracted from RNA sequences, which were used as inputs of CNN to construct three models, respectively. The three models and the integrated ensemble model were compared and analyzed, which showed that the ensemble model outperformed the other three models. Both cross-validation and independent testing demonstrated the generalization of DeepPromise which can be accessed at https://DeepPromise.erc.monash.edu/.

5.10.4. iRNA-Mod-CNN

iRNA-Mod-CNN [147] is another method that can predict multiple types of methylation sites of RNA. By using nucleotide binary encoding as input and CNN as the deep framework, the deep features were extracted which were concatenated with the K-mer composition to predict the m1A, m6A and m5C sites. The results of 5-fold cross-validation indicated the model was superior to other methods.

5.10.5. MultiRM

MultiRM [148] is a predictor that can identify up to 12 types of RNA modification sites. The benchmarks for building MultiRM contain more than 300,000 sites. Three embedding technologies (Onehot + Conv1D + Pool1D, Word2vec, Hidden Markov Model) were used to extract the sequence information. Experimental results demonstrated the superiority of embeddings obtained by Word2vec. Then multi-label learning

Table 9
Predictors for identifying 5hmC sites.

Predictors	Year	Species	Features [a]	Algorithms	URLs
iRNA5hmC	2020	*D. melanogaster*	K-mer composition; NBE	SVM	https://server.malab.cn/iRNA5hmC
iRNA5hmC-PS	2020	*D. melanogaster*	Position-specific gapped k-mer; Position-specific k-mer; GC-content	LR	https://103.109.52.8:81/iRNA5hmC-PS; https://github.com/zahid6454/iRNA5hmC-PS
iRhm5CNN	2021	*D. melanogaster*	NBE	CNN	https://nsclbio.jbnu.ac.kr/tools/iRhm5CNN/

[a] The full name of the abbreviations are the same as Tables 3 and 4.

method were used to build deep learning model with the attention-based neural network. In addition, the authors have explored to explain the deep model through the integrated gradient (IG) and attention weights. The web server of MultiRM is available at: https://www.xjtlu.edu.cn/biologicalsciences/multirm.

The methods for predicting m1A, m5U, m6Am and multiple methylation sites were summarized in Table 10.

6. Discussion

As shown in Fig. 2, the construction of predictors for identifying RNA methylation sites follow the five-step rules [149]: (1) Obtain an effective benchmark dataset; (2) Extract effective feature representations from RNA sequences; (3) Train models based on effective learning algorithms; (4) Design a reasonable validation or testing strategy to objectively evaluate the performance of the predictor; (5) Establish an open and accessible network platform for the predictor.

Thus, we first reviewed the public databases that could be used to build the benchmark datasets. Thanks to the development of high-throughput sequencing technologies, the "big data" of transcriptome can be obtained so that a bunch of databases related to RNA modification have been built. We reviewed 18 databases in this paper. Different kinds of RNA modification have been collected in these databases for different species. In addition to the information about the modification peaks or locations of RNA, other related information such as functional annotation and modification related diseases has also been recorded in the databases. Moreover, several databases provide tools for visualizing the modification information.

Secondly, designing feature encoding schemes is crucial for the development of predictors for identifying RNA methylation sites based on machine learning. The features used in most of the existing predictors are obtained based on single RNA segment, which mainly describe the distribution of nucleotides in RNA segments, including binary encoding, composition encoding, physical–chemical-properties-based features, secondary structure related features, word2vec, electron–ion interaction pseudopotentials related features, etc. For example, PseKNC was used to extract RNA sequence features in iRNA(m6A)-PseDNC [150]; RNA sequence representation was obtained based on k-mer composition in pM5CS-Comp-mRMR [99]; three coding methods, nucleotide binary encoding, k-mer composition, secondary structure binary encoding, were used in tRNAmod to extract features [114]. The information related to the distribution of a single RNA segment may be not enough, thus, the feature encoding technology based on multiple RNA segments was proposed, which shows the evolutionary relationship between multiple sequences, including PSCP, BPB, and so on. These features have also been used in the recognition of RNA methylation sites. Such as, in RAM-NPPS predictor, PSCP is used to extract features [54]; the BPB feature extraction technology was used both in iRNA-m5C_NB [69] and RNA-MethylPred [2]. Furthermore, according to biological prior, the genomic features which mainly describe the distribution of sequence fragments in the genome, were also used to identify RNA methylation sites. For example, physical–chemical based binary encoding, accumulated nucleotide frequency, genome-derived features were used together for building LITHOPHONE [151], WITMSG [95] and PSI-MOUSE [125]. Thus, in order to improve prediction performance, combining different feature technologies to produce mixed features is a major trend in the development of RNA methylation site identification methods.

The computational models used to identify RNA methylation sites are mainly constructed based on traditional machine learning and deep learning. These traditional machine learning algorithms mainly include support vector machine (SVM), random forest (RF) and extreme gradient enhancement (XGBoost). Machine learning, as a part of the field of artificial intelligence, aims to establish regressors or classifiers through learning training datasets, and then evaluate the performance of regressors or classifiers through test sets [15]. Traditional machine learning (shallow learning) uses manually extracted features as inputs to build computational models for predicting RNA methylation sites. Chen et al. constructed iRNA-Methyl, the first predictive tool for identifying m6A sites in S. cerevisiae, using PseDNC and SVM [78]. Li et al. used one-hot encoding RNA sequence characteristics to construct a prediction tool RNAm5Cfinder for identifying m5C sites in human and mouse based on random forest (RF) [102]. Dai et al. constructed a predictive model m7G-IFL using XGBoost to identify the m7G sites in Homo sapiens based on physical chemical properties, physical–chemical based binary encoding, nucleotide binary encoding and K-mer composition [152]. With the increase of RNA methylation site data, deep learning is considered to be effective algorithms that can improve the ability to identify RNA methylation sites. Deep learning methods have been effectively applied in the field of computational biology, among which CNN is the most widely used network structure for RNA methylation site recognition in deep learning. Ahmed et al. used CNN to develop a method, iRhm5CNN, for predicting modification sites of 5hmC [66]. Zou et al. built Gene2vec, a method for predicting m6A sites in mammalian mRNA based on deep neural networks (DNN) [90]. As reviewed in this

Table 10
Predictors for m1A, m5U, m6Am and multiple methylation sites.

Predictors	Year	Methylation types	Species	Features [a]	Algorithms	URLs
RAMPred	2016	m1A	*H. sapiens*; *M. musculus*; *S. cerevisiae*	PCBE; ANF	SVM	https://lin.uestc.edu.cn/server/RAMPred
ISGm1A	2020	m1A	*H. sapiens*	PCBE; ANF; Genome-derived features	RF	https://github.com/lianliu09/m1a_prediction.git
m5UPred	2020	m5U	*H. sapiens*	PCBE; ANF	SVM	https://www.xjtlu.edu.cn/biologicalsciences/m5u
m6AmPred	2021	m6Am	*H. sapiens*; *M. musculus*	PCBE; ANF; EIIP-derived features	XgbDart	https://www.xjtlu.edu.cn/biologicalsciences/m6am
iRNA-PseColl	2017	m1A; m6A m5C	*H. sapiens*	PseKNC; PCBE; ANF	SVM	https://lin.uestc.edu.cn/server/iRNA-PseColl
iMRM	2020	m6A; m5C; m1A; Ψ; A-to-I editing	*H. sapiens*; *M. musculus*; *S. cerevisiae*	K-mer composition; NBE; DBE; PCBE; ANF; PCP	XGboost	https://www.bioml.cn/XG_iRNA/home
DeepPromise	2020	m1A; m6A	*H. sapiens*; *M. musculus*	ENAC; NBE; RNA embedding	CNN	https://DeepPromise.erc.monash.edu/
iRNA-Mod-CNN	2021	m1A; m6A; m5C	*H. sapiens*	NBE; K-mer composition	CNN	/
MultiRM	2021	m6A; m1A; m5C; m5U; m6Am; m7G; Ψ; I; Am; Cm; Gm; Um	*H. sapiens*	NBE; Hidden Markov Model; Word2vec	CNN, RNN	https://www.xjtlu.edu.cn/biologicalsciences/multirm

[a] The full name of the abbreviations are the same as Tables 3 and 4.

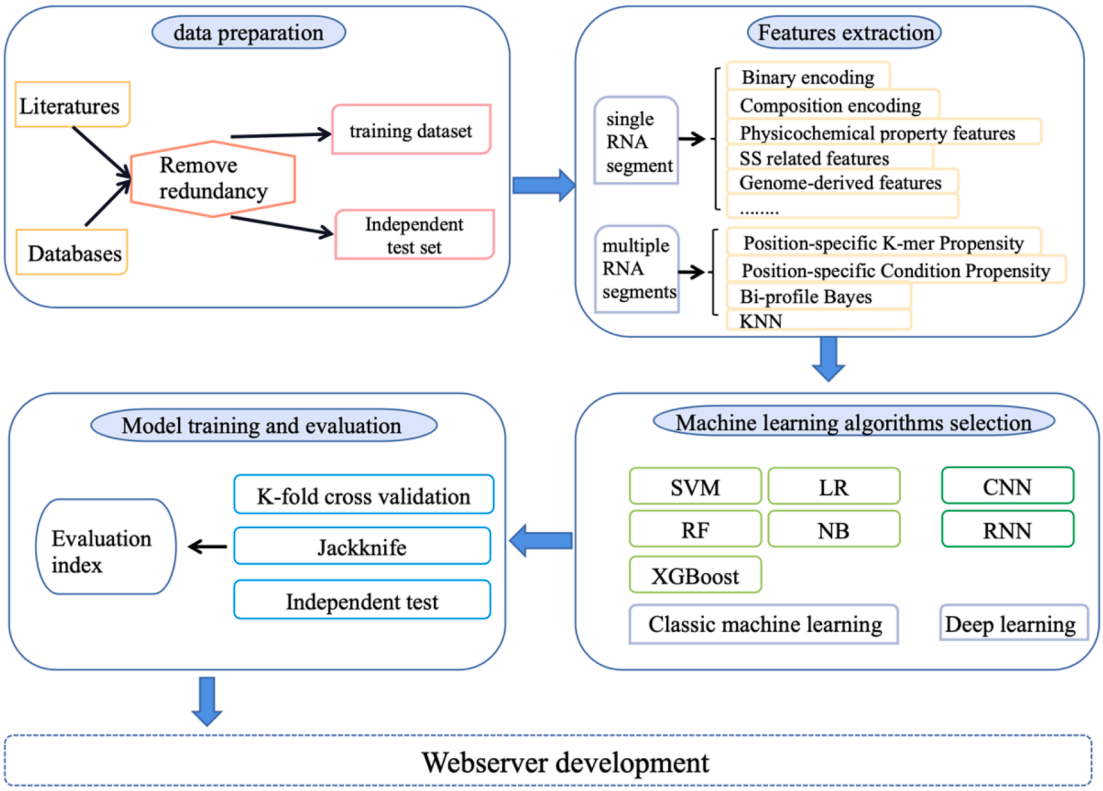

Fig. 2. The pipeline for building a machine learning predictor for identifying RNA methylation sites.

paper, the advanced deep learning algorithms would exert more roles in the development of RNA methylation site identification methods.

In most of the methods, models were built to only target one type of RNA methylation modification of one organism, but the reality is that one organism may contain multiple types of RNA methylation modification, and one type of RNA methylation modification may exist in multiple species. The study on the recognition of m6A sites is the most. In addition, for other common RNA methylation modifications, such as m5C, Ψ, m7G and A-to-I editing, some predictors of site recognition have been developed. However, there are still few studies on sites recognition of some methylation modification types. For example, only one predictor has been developed to identify the m5U and m6Am sites respectively. And for some RNA methylation sites, there even no computational model has been established. Although a few works have attempted to construct predictors that can simultaneously recognize multiple RNA methylation modification sites, for example, iRNA-Psecoll [104] can simultaneously recognize m1A, m6A, m5C sites, and iMRM [145] can simultaneously recognize m6A, m5C, m1A, and A-to-I editing sites. But there are many more types of RNA methylation modification sites. Therefore, it is expected to study different methylation modification sites and establish a technology platform that can simultaneously predict different RNA methylation modification sites in different species. This kind of models can be classified into two categories. For one category, based on the same framework, the models that were built to predict different kinds of methylation sites were integrated into a model, such as iRNA-Psecoll and iMRM. For the other category, based on multi-label learning or multi-task learning, the tasks for predicting different kinds of methylation sites were realized in one model, such as MultiRM. Compared with the models of the former category, the models of the latter category may be more efficient.

7. Conclusion

In this review, we summarized 18 public RNA methylation related databases, three categories of feature encoding methods for extracting information from RNA sequences, and commonly used machine learning algorithms for building models for RNA methylation site recognition. In addition, the state-of-the-art predictors and web servers for identifying different types of RNA methylation sites were also summarized. In the view of feature encoding methods, high dimensional features can now be extracted, but how to select the relevant feature subset is still challenging. In addition, traditional machine learning is still the main force in the construction of computational methods for RNA methylation site recognition. However, with the increase of data, and the application of deep learning framework can be an effective way to improve the performance. Moreover, building the model that can predict multiple methylation sites might be one of the trends in the field. Thus, the main contributions of this review are listed as follows: (i) The databases related to RNA methylation, the features and learning algorithms used for building RNA methylation sites prediction models were comprehensively reviewed in this paper. (ii) About 30 kinds of feature encoding methods were summarized and were classified into three categories, which would be benefit for searching the optimal feature combination to build effective models for predicting RNA methylation sites. (iii) The existing methods for predicting ten types of methylation of RNA were comprehensively reviewed, based on which we prospected the research of the field in the future.

CRediT authorship contribution statement

Hong Wang: Resources, Writing – original draft. **Shuyu Wang:** Resources, Writing – original draft. **Yong Zhang:** Visualization. **Shoudong Bi:** Conceptualization, Supervision, Writing – review & editing. **Xiaolei Zhu:** Conceptualization, Supervision, Writing – review & editing.

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

This work was supported by the National Natural Science Foundation of China (21403002).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymeth.2022.03.001.

References

[1] I. Barbieri, T. Kouzarides, Role of RNA modifications in cancer, Nat Rev Cancer 20 (6) (2020) 303–322.
[2] C.Z. Jia, J.J. Zhang, W.Z. Gu, RNA-MethylPred: a high-accuracy predictor to identify N6-methyladenosine in RNA, Anal Biochem 510 (2016) 72–75.
[3] R. Desrosiers, K. Friderici, F. Rottman, Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells, Proc Natl Acad Sci U S A 71 (10) (1974) 3971–3975.
[4] Y. Motorin, M. Helm, RNA nucleotide methylation, Wiley Interdiscip Rev RNA 2 (5) (2011) 611–631.
[5] D. Globisch, M. Munzel, M. Muller, S. Michalakis, M. Wagner, S. Koch, T. Bruckl, M. Biel, T. Carell, Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates, PLoS ONE 5 (12) (2010), e15367.
[6] X. Chen, Y.Z. Sun, H. Liu, L. Zhang, J.Q. Li, J. Meng, RNA methylation and diseases: experimental results, databases, Web servers and computational models, Brief Bioinform 20 (3) (2019) 896–917.
[7] S. Blanco, M. Frye, Role of RNA methyltransferases in tissue renewal and pathology, Curr Opin Cell Biol 31 (2014) 1–7.
[8] N. Liu, T. Pan, RNA epigenetics, Transl Res 165 (1) (2015) 28–35.
[9] G.Q. Zheng, J.A. Dahl, Y.M. Niu, P. Fedorcsak, C.M. Huang, C.J. Li, C.B. Vagbo, Y. Shi, W.L. Wang, S.H. Song, Z.K. Lu, R.P.G. Bosmans, Q. Dai, Y.J. Hao, X. Yang, W.M. Zhao, W.M. Tong, X.J. Wang, F. Bogdan, K. Furu, Y. Fu, G.F. Jia, X. Zhao, J. Liu, H.E. Krokan, A. Klungland, Y.G. Yang, C. He, ALKBH5 Is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility, Mol Cell 49 (1) (2013) 18–29.
[10] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, R. Sorek, G. Rechavi, Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq, Nature 485 (7397) (2012) 201–206.
[11] The relationship between recall and precision.
[12] S.Y. Zhang, S.W. Zhang, T. Zhang, X.N. Fan, J. Meng, Recent advances in functional annotation and prediction of the epitranscriptome, Comput Struct Biotechnol J 19 (2021) 3015–3026.
[13] J. Ma, L. Zhang, S. Chen, H. Liu, A brief review of RNA modification related database resources, Methods (2021).
[14] W.A. Cantara, P.F. Crain, J. Rozenski, J.A. McCloskey, K.A. Harris, X. Zhang, F.A. Vendeix, D. Fabris, P.F. Agris, The RNA Modification Database, RNAMDB: 2011 update, Nucleic Acids Res 39(Database issue) (2011) D195-201.
[15] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, L. Stein, Reactome: a database of reactions, pathways and biological processes, Nucleic Acids Res 39(Database issue) (2011) D691-7.
[16] M.A. Machnicka, K. Milanowska, O. Osman Oglou, E. Purta, M. Kurkowska, A. Olchowik, W. Januszewski, S. Kalinowski, S. Dunin-Horkawicz, K.M. Rother, M. Helm, J.M. Bujnicki, H. Grosjean, MODOMICS: a database of RNA modification pathways–2013 update, Nucleic Acids Res 41(Database issue) (2013) D262-7.
[17] A.M. Kiran, J.J. O'Mahony, K. Sanjeev, P.V. Baranov, Darned in 2013: inclusion of model organisms and linking with Wikipedia, Nucleic Acids Res 41(Database issue) (2013) D258-61.
[18] G. Ramaswami, J.B. Li, RADAR a rigorously annotated database of A-to-I RNA editing, Nucleic Acids Res 42 (Database issue) (2014) D109–D113.
[19] H. Liu, M.A. Flores, J. Meng, L. Zhang, X. Zhao, M.K. Rao, Y. Chen, Y. Huang, MeT-DB: a database of transcriptome methylation in mammalian cells, Nucleic Acids Res 43(Database issue) (2015) D197-203.
[20] W.J. Sun, J.H. Li, S. Liu, J. Wu, H. Zhou, L.H. Qu, J.H. Yang, RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data, Nucleic Acids Res 44 (D1) (2016) D259–D265.
[21] P. Boccaletto, M.A. Machnicka, E. Purta, P. Piatkowski, B. Baginski, T.K. Wirecki, V. de Crecy-Lagard, R. Ross, P.A. Limbach, A. Kotter, M. Helm, J.M. Bujnicki, MODOMICS: a database of RNA modification pathways. 2017 update, Nucleic Acids Res 46(D1) (2018) D303-D307.
[22] B. Uyar, D. Yusuf, R. Wurmus, N. Rajewsky, U. Ohler, A. Akalin, RCAS: an RNA centric annotation system for transcriptome-wide regions of interest, Nucleic Acids Res 45 (10) (2017), e91.
[23] E. Picardi, T.M. Regina, A. Brennicke, C. Quagliariello, REDIdb: the RNA editing database, Nucleic Acids Res 35(Database issue) (2007) D173-7.
[24] E. Picardi, T.M. Regina, D. Verbitskiy, A. Brennicke, C. Quagliariello, REDIdb: an upgraded bioinformatics resource for organellar RNA editing sites, Mitochondrion 11 (2) (2011) 360–365.
[25] C. Lo Giudice, G. Pesole, E. Picardi, REDIdb 3.0: A comprehensive collection of RNA editing events in plant organellar genomes, Front Plant Sci 9 (2018) 482.
[26] 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 m(6)As in cell lines, Cells 8 (2) (2019).
[27] Q. Liu, R.I. Gregory, RNAmod: an integrated system for the annotation of mRNA modifications, Nucleic Acids Res 47 (W1) (2019) W548–W555.
[28] B. Song, Y. Tang, K. Chen, Z. Wei, R. Rong, Z. Lu, J. Su, J.P. de Magalhaes, D. J. Rigden, J. Meng, m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human, Bioinformatics 36 (11) (2020) 3528–3536.
[29] S. Liu, A. Zhu, C. He, M. Chen, REPIC: a database for exploring the N(6)-methyladenosine methylome, Genome Biol 21 (1) (2020) 100.
[30] L. Mansi, M.A. Tangaro, C. Lo Giudice, T. Flati, E. Kopel, A.A. Schaffer, T. Castrignano, G. Chillemi, G. Pesole, E. Picardi, REDIportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments, Nucleic Acids Res 49 (D1) (2021) D1012–D1019.
[31] K. Licht, U. Kapoor, F. Amman, E. Picardi, D. Martin, P. Bajad, M.F. Jantsch, A high resolution A-to-I editing map in the mouse identifies editing events controlled by pre-mRNA splicing, Genome Res 29 (9) (2019) 1453–1463.
[32] Y. Tang, K. Chen, B. Song, J. Ma, X. Wu, Q. Xu, Z. Wei, J. Su, G. Liu, R. Rong, Z. Lu, J.P. de Magalhaes, D.J. Rigden, J. Meng, m6A-Atlas: a comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome, Nucleic Acids Res 49 (D1) (2021) D134–D143.
[33] S. Deng, H. Zhang, K. Zhu, X. Li, Y. Ye, R. Li, X. Liu, D. Lin, Z. Zuo, J. Zheng, M6A2Target: a comprehensive database for targets of m6A writers, erasers and readers, Brief Bioinform 22 (3) (2021).
[34] X. Luo, H. Li, J. Liang, Q. Zhao, Y. Xie, J. Ren, Z. Zuo, RMVar: an updated database of functional variants involved in RNA modifications, Nucleic Acids Res 49 (D1) (2021) D1405–D1412.
[35] K. Chen, B. Song, Y. Tang, Z. Wei, Q. Xu, J. Su, J.P. de Magalhaes, D.J. Rigden, J. Meng, RMDisease: a database of genetic variants that affect RNA modifications, with implications for epitranscriptome pathogenesis, Nucleic Acids Res 49 (D1) (2021) D1396–D1404.
[36] G.Q. Li, Z. Liu, H.B. Shen, D.J. Yu, Target M6A: identifying N(6)-methyladenosine sites From RNA sequences via position-specific nucleotide propensities and a support vector machine, IEEE Trans Nanobiosci 15 (7) (2016) 674–682.
[37] J. Brayet, F. Zehraoui, L. Jeanson-Leh, D. Israeli, F. Tahi, Towards a piRNA prediction using multiple kernel fusion and support vector machine, Bioinformatics (Oxford, England) 30 (17) (2014) i364–i370.
[38] E.K. Mohamed Hashim, R. Abdullah, Rare k-mer DNA: Identification of sequence motifs and prediction of CpG island and promoter, J Theor Biol 387 (2015) 88–100.
[39] H. Vinje, K.H. Liland, T. Almoy, L. Snipen, Comparing K-mer based methods for improved classification of 16S sequences, BMC Bioinf 16 (2015) 205.
[40] Y.Z. Chen, Y.R. Tang, Z.Y. Sheng, Z. Zhang, Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs, BMC Bioinf 9 (2008) 101.
[41] X. Wang, R. Yan, J. Song, DephosSite: a machine learning approach for discovering phosphotase-specific dephosphorylation sites, Sci Rep 6 (2016) 23510.
[42] X. Wang, R. Yan, RFAthM6A: a new tool for predicting m(6)A sites in Arabidopsis thaliana, Plant Mol Biol 96 (3) (2018) 327–337.
[43] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, K.C. Chou, PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition, Anal Biochem 456 (2014) 53–60.
[44] Z.C. Xu, P. Wang, W.R. Qiu, X. Xiao, iSS-PC: identifying splicing sites via physical-chemical properties using deep sparse auto-encoder, Sci Rep 7 (1) (2017) 8222.
[45] L. Wei, H. Chen, R. Su, M6APred-EL: a sequence-based predictor for identifying n6-methyladenosine sites using ensemble learning, Molecular therapy, Nucleic Acids 12 (2018) 635–644.
[46] M. Zhang, J.W. Sun, Z. Liu, M.W. Ren, H.B. Shen, D.J. Yu, Improving N(6)-methyladenosine site prediction with heuristic selection of nucleotide physical-chemical properties, Anal Biochem 508 (2016) 104–113.
[47] A. Perez, A. Noy, F. Lankas, F.J. Luque, M. Orozco, The relative flexibility of B-DNA and A-RNA duplexes: database analysis, Nucleic Acids Res 32 (20) (2004) 6144–6151.
[48] J.R. Goni, A. Perez, D. Torrents, M. Orozco, Determining promoter location based on DNA structure first-principles calculations, Genome Biol 8 (12) (2007) R263.
[49] S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson, D. H. Turner, Improved free-energy parameters for predictions of RNA duplex stability, PNAS 83 (24) (1986) 9373–9377.
[50] Z. Liu, X. Xiao, D.J. Yu, J. Jia, W.R. Qiu, K.C. Chou, pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties, Anal Biochem 497 (2016) 60–67.
[51] R. Lorenz, S.H. Bernhart, C.H.Z. Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, I. L. Hofacker, ViennaRNA Package 2.0, Algorithms, Mol. Biol. 6 (2011) 14.

[52] W. Chen, P.M. Feng, X.M. Song, H. Lv, H. Lin, iRNA-m7G: identifying N-7-methylguanosine sites by fusing multiple features, Mol Ther-Nucl Acids 18 (2019) 269–274.
[53] A.S. Nair, S.P. Sreenadhan, A coding measure scheme employing electron-ion interaction pseudopotential (EIIP), Bioinformation 1 (6) (2006) 197–202.
[54] P. Xing, R. Su, F. Guo, L. Wei, Identifying N(6)-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine, Sci Rep 7 (2017) 46757.
[55] V. Cherkassky, The nature of statistical learning theory, IEEE Trans Neural Netw 8 (6) (1997) 1564.
[56] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans Neural Netw 10 (5) (1999) 988–999.
[57] L. Breiman, Random forests, Mach Learn 45 (1) (2001) 5–32.
[58] J.H. Frieman, B.E. Popescu, Predictive learning via rule ensembles, Ann Appl Stat 2 (3) (2008) 916–954.
[59] A.L. Boulesteix, S. Janitza, J. Kruppa, I.R. Konig, Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, Wires Data Min Knowl 2 (6) (2012) 493–507.
[60] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
[61] J.H. Friedman, Stochastic gradient boosting, Comput Stat Data An 38 (4) (2002) 367–378.
[62] Z. Zhao, H. Peng, C. Lan, Y. Zheng, L. Fang, J. Li, Imbalance learning for the prediction of N(6)-Methylation sites in mRNAs, BMC Genomics 19 (1) (2018) 574.
[63] X. Qiang, H. Chen, X. Ye, R. Su, L. Wei, M6AMRFS: robust prediction of N6-methyladenosine sites with sequence-based features in multiple species, Front Genet 9 (2018) 495.
[64] J. Li, S. He, F. Guo, Q. Zou, HSM6AP: a high-precision predictor for the Homo sapiens N6-methyladenosine (m^6 A) based on multiple weights and feature stitching, RNA Biol (2021) 1–11.
[65] T.G. Nick, K.M. Campbell, Logistic regression, Methods Mol Biol 404 (2007) 273–301.
[66] S. Ahmed, Z. Hossain, M. Uddin, G. Taherzadeh, A. Sharma, S. Shatabda, A. Dehzangi, Accurate prediction of RNA 5-hydroxymethylcytosine modification by utilizing novel position-specific gapped k-mer descriptors, Comput Struct, Biotechnol J 18 (2020) 3528–3538.
[67] Y.Y. Zhuang, H.J. Liu, X. Song, Y. Ju, H. Peng, A linear regression predictor for identifying N(6)-methyladenosine sites using frequent gapped K-mer Pattern, Mol Ther Nucleic Acids 18 (2019) 673–680.
[68] G.I. Webb, E. Keogh, R.J.E.o.m.l. Miikkulainen, Naïve Bayes, 15 (2010) 713–714.
[69] L.J. Dou, X.L. Li, H. Ding, L. Xu, H.K. Xiang, iRNA-m5C_NB: a novel predictor to identify RNA 5-methylcytosine sites based on the naive bayes classifier, IEEE Access 8 (2020) 84906–84917.
[70] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[71] D. Berrar, W. Dubitzky, Deep learning in bioinformatics and biomedicine, Brief Bioinform 22 (2) (2021) 1513–1514.
[72] Y. Zhang, M. Hamada, DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning, BMC Bioinf 19 (Suppl 19) (2018) 524.
[73] M. Tahir, H. Tayara, K.T. Chong, iPseU-CNN: identifying RNA pseudouridine sites using convolutional neural networks, Mol Ther Nucleic Acids 16 (2019) 463–470.
[74] Y. Huang, N. He, Y. Chen, Z. Chen, L. Li, BERMP: a cross-species classifier for predicting m(6)A sites by integrating a deep learning algorithm and a random forest approach, Int J Biol Sci 14 (12) (2018) 1669–1677.
[75] J.X. Gu, Z.H. Wang, J. Kuen, L.Y. Ma, A. Shahroudy, B. Shuai, T. Liu, X.X. Wang, G. Wang, J.F. Cai, T. Chen, Recent advances in convolutional neural networks, Pattern Recogn 77 (2018) 354–377.
[76] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput 9 (8) (1997) 1735–1780.
[77] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.
[78] W. Chen, P. Feng, H. Ding, H. Lin, K.C. Chou, iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition, Anal Biochem 490 (2015) 26–33.
[79] W. Chen, H. Tran, Z. Liang, H. Lin, L. Zhang, Identification and analysis of the N (6)-methyladenosine in the Saccharomyces cerevisiae transcriptome, Sci Rep 5 (2015) 13859.
[80] S. Xiang, Z. Yan, K. Liu, Y. Zhang, Z. Sun, AthMethPre: a web server for the prediction and query of mRNA m(6)A sites in Arabidopsis thaliana, Mol Biosyst 12 (11) (2016) 3333–3337.
[81] S. Xiang, K. Liu, Z. Yan, Y. Zhang, Z. Sun, RNAMethPre: a web server for the prediction and query of mRNA m6A sites, PLoS ONE 11 (10) (2016), e0162707.
[82] Y. Zhou, P. Zeng, Y.H. Li, Z. Zhang, Q. Cui, SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features, Nucleic Acids Res 44 (10) (2016), e91.
[83] G.Q. Li, Z. Liu, H.B. Shen, D.J. Yu, Target M6A: identifying N-6-methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine, Ieee T Nanobiosci 15 (7) (2016) 674–682.
[84] W. Chen, P. Feng, H. Ding, H. Lin, Identifying N (6)-methyladenosine sites in the Arabidopsis thaliana transcriptome, Mol Genet Genomics 291 (6) (2016) 2225–2229.
[85] W. Chen, H. Tang, H. Lin, MethyRNA: a web server for identification of N(6)-methyladenosine sites, J Biomol Struct Dyn 35 (3) (2017) 683–687.
[86] W. Chen, P.W. Xing, Q. Zou, Detecting N-6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines, Sci Rep-Uk 7 (2017).
[87] P.W. Xing, R. Su, F. Guo, L.Y. Wei, Identifying N-6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine, Sci Rep-Uk 7 (2017).
[88] J. Zhang, P. Feng, H. Lin, W. Chen, Identifying RNA N(6)-methyladenosine sites in escherichia coli genome, Front Microbiol 9 (2018) 955.
[89] W. Chen, H. Ding, X. Zhou, H. Lin, K.C. Chou, iRNA(m6A)-PseDNC: Identifying N (6)-methyladenosine sites using pseudo dinucleotide composition, Anal Biochem 561–562 (2018) 59–65.
[90] Q. Zou, P. Xing, L. Wei, B. Liu, Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA, RNA 25 (2) (2019) 205–218.
[91] I. Nazari, M. Tahir, H. Tayara, K.T. Chong, iN6-Methyl (5-step): Identifying RNA N6-methyladenosine sites using deep learning mode via Chou's 5-step rules and Chou's general PseKNC, Chemometr Intell Lab 193 (2019).
[92] K.Q. Chen, Z. Wei, Q. Zhang, X.Y. Wu, R. Rong, Z.L. Lu, J.L. Su, J.P. de Magalhaes, D.J. Rigden, J. Meng, WHISTLE: a high-accuracy map of the human N-6-methyladenosine (m(6)A) epitranscriptome predicted using a machine learning approach, Nucleic Acids Res 47 (7) (2019).
[93] L.Y. Wei, R. Su, B. Wang, X.T. Li, Q. Zou, X. Gao, Integration of deep feature representations and handcrafted features to improve the prediction of N-6-methyladenosine sites, Neurocomputing 324 (2019) 3–9.
[94] L. Liu, X.J. Lei, Z.Q. Fang, Y.J. Tang, J. Meng, Z. Wei, LITHOPHONE: improving lncRNA methylation site prediction using an ensemble predictor, Front Genet 11 (2020).
[95] L. Liu, X. Lei, J. Meng, Z. Wei, WITMSG: large-scale prediction of human intronic m(6)A RNA methylation sites from sequence and genomic features, Curr Genomics 21 (1) (2020) 67–76.
[96] A. Khan, H.U. Rehman, U. Habib, U. Ijaz, Detecting N6-methyladenosine sites from RNA transcriptomes using random forest, J Comput Sci-Neth 47 (2020).
[97] P. Feng, H. Ding, W. Chen, H. Lin, Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions, Mol Biosyst 12 (11) (2016) 3307–3311.
[98] W.R. Qiu, S.Y. Jiang, Z.C. Xu, X. Xiao, K.C. Chou, iRNAm 5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition, Oncotarget 8 (25) (2017) 41178–41188.
[99] M.F. Sabooh, N. Iqbal, M. Khan, M. Khan, H.F. Maqbool, Identifying 5-methyl-cytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC, J Theor Biol 452 (2018) 1–9.
[100] M. Zhang, Y. Xu, L. Li, Z. Liu, X. Yang, D.J. Yu, Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble, Anal Biochem 550 (2018) 41–48.
[101] J. Song, J.J. Zhai, E.Z. Bian, Y.J. Song, J.T. Yu, C. Ma, Transcriptome-Wide Annotation of m(5)C RNA Modifications Using Machine Learning, Front Plant Sci 9 (2018).
[102] J. Li, Y. Huang, X. Yang, Y. Zhou, Y. Zhou, RNAm 5Cfinder: A Web-server for Predicting RNA 5-methylcytosine (m5C) Sites Based on Random Forest, Sci Rep 8 (1) (2018) 17299.
[103] T. Amort, D. Rieder, A. Wille, D. Khokhlova-Cubberley, C. Riml, L. Trixl, X.Y. Jia, R. Micura, A. Lusser, Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain, Genome Biol 18 (1) (2017) 1.
[104] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, K.C. Chou, iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Mol Ther Nucleic Acids 7 (2017) 155–163.
[105] T. Fang, Z. Zhang, R. Sun, L. Zhu, J. He, B. Huang, Y. Xiong, X. Zhu, RNAm 5CPred: prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition, Mol Ther Nucleic Acids 18 (2019) 739–747.
[106] L. Dou, X. Li, H. Ding, L. Xu, H. Xiang, Prediction of m5C Modifications in RNA sequences by combining multiple sequence features, Mol Ther Nucleic Acids 21 (2020) 332–342.
[107] X. Chen, Y. Xiong, Y. Liu, Y. Chen, S. Bi, X. Zhu, m5CPred-SVM: a novel method for predicting m5C sites of RNA, BMC Bioinf 21 (1) (2020) 489.
[108] D. Chai, C.Z. Jia, J. Zheng, Q. Zou, F.Y. Li, Staem5: A novel computational approach for accurate prediction of m5C site, Mol Ther-Nucl Acids 26 (2021) 1027–1034.
[109] W. Chen, P. Feng, X. Song, H. Lv, H. Lin, iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features, Mol Ther Nucleic Acids 18 (2019) 269–274.
[110] Y.H. Yang, C. Ma, J.S. Wang, H. Yang, H. Ding, S.G. Han, Y.W. Li, Prediction of N7-methylguanosine sites in human RNA based on optimal sequence features, Genomics 112 (6) (2020) 4342–4347.
[111] Y. Bi, D. Xiang, Z. Ge, F. Li, C. Jia, J. Song, An interpretable prediction model for identifying N(7)-methylguanosine sites based on XGBoost and SHAP, Mol Ther Nucleic Acids 22 (2020) 362–372.
[112] X. Liu, Z. Liu, X. Mao, Q. Li, m7GPredictor: An improved machine learning-based model for predicting internal m7G modifications using sequence properties, Anal Biochem 609 (2020), 113905.
[113] C. Dai, P. Feng, L. Cui, R. Su, W. Chen, L. Wei, Iterative feature representation algorithm to improve the predictive performance of N7-methylguanosine sites, Brief Bioinform 22 (4) (2021).
[114] B. Panwar, G.P. Raghava, Prediction of uridine modifications in tRNA sequences, BMC Bioinf 15 (2014) 326.

[115] T.M. Carlile, M.F. Rojas-Duran, B. Zinshteyn, H. Shin, K.M. Bartoli, W.V. Gilbert, Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells, Nature 515 (7525) (2014) 143–146.
[116] S. Schwartz, D.A. Bernstein, M.R. Mumbach, M. Jovanovic, R.H. Herbst, B. X. Leon-Ricardo, J.M. Engreitz, M. Guttman, R. Satija, E.S. Lander, G. Fink, A. Regev, Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA, Cell 159 (1) (2014) 148–162.
[117] W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, K.C. Chou, PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions, Bioinformatics 31 (1) (2015) 119–120.
[118] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, Y. Xiong, PseUI: Pseudouridine sites identification based on RNA sequence information, BMC Bioinf 19 (1) (2018) 306.
[119] T.H. Nguyen-Vo, Q.H. Nguyen, T.T.T. Do, T.N. Nguyen, S. Rahardja, B.P. Nguyen, iPseU-NCP: Identifying RNA pseudouridine sites using random forest and NCP-encoded features, BMC Genomics 20 (Suppl 10) (2019) 971.
[120] Z. Lv, J. Zhang, H. Ding, Q. Zou, RF-PseU: A random forest predictor for RNA pseudouridine sites, Front Bioeng Biotechnol 8 (2020) 134.
[121] Y. Mu, R. Zhang, L. Wang, X. Liu, iPseU-layer: identifying RNA pseudouridine sites using layered ensemble model, Interdiscip Sci 12 (2) (2020) 193–203.
[122] K. Liu, W. Chen, H. Lin, XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites, Mol Genet Genomics 295 (1) (2020) 13–21.
[123] B. Song, Y. Tang, Z. Wei, G. Liu, J. Su, J. Meng, K. Chen, PIANO: a web server for pseudouridine-site (Psi) identification and functional annotation, Front Genet 11 (2020) 88.
[124] Y. Bi, D. Jin, C.Z. Jia, EnsemPseU: identifying pseudouridine sites with an ensemble approach, IEEE Access 8 (2020) 79376–79382.
[125] B.W. Song, K.Q. Chen, Y.J. Tang, J.L. Ma, J. Meng, Z. Wei, PSI-MOUSE: predicting mouse pseudouridine sites from sequence and genome-derived features, Evol Bioinform 16 (2020).
[126] A.Z.B. Aziz, M.A.M. Hasan, A. Mixed Convolution Neural Network for Identifying RNA Pseudouridine sites, in: IEEE Region 10 Symposium (TENSYMP), IEEE, 2020, pp. 799–802.
[127] J.S. Wang, S.L. Zhang, PA-PseU: An incremental passive-aggressive based method for identifying RNA pseudouridine sites via Chou's 5-steps rule, Chemom Intelligent Lab Syst 210 (2021).
[128] A.Z.B. Aziz, M.A.M. Hasan, J. Shin, Identification of RNA pseudouridine sites using deep learning approaches, PLoS ONE 16 (2) (2021), e0247511.
[129] F. Li, X. Guo, P. Jin, J. Chen, D. Xiang, J. Song, L.J.M. Coin, Porpoise: a new approach for accurate prediction of RNA pseudouridine sites, Brief Bioinform 22 (6) (2021).
[130] W. Chen, P. Feng, H. Ding, H. Lin, PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions, Sci Rep 6 (2016) 35123.
[131] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, K.C. Chou, iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, Oncotarget 8 (3) (2017) 4208–4217.
[132] 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. IOP Conference Series: Earth and Environmental Science, IOP Publishing, 2018.
[133] A. Ahmad, S. Shatabda, EPAI-NC: Enhanced prediction of adenosine to inosine RNA editing sites using nucleotide compositions, Anal Biochem 569 (2019) 16–21.
[134] A. Choyon, A. Rahman, M. Hasanuzzaman, D.M. Farid, S. Shatabda, Presa2i: incremental decision trees for prediction of adenosine to inosine RNA editing sites, F1000 Research 9 (262) (2020) 262.
[135] M. Mostavi, S. Salekin, Y. Huang, Deep-2'-O-Me: predicting 2'-O-methylation sites by convolutional neural networks, Annu Int Conf IEEE Eng Med Biol Soc 2018 (2018) 2394–2397.
[136] H. Yang, H. Lv, H. Ding, W. Chen, H. Lin, iRNA-2OM: a sequence-based predictor for identifying 2 '-O-methylation sites in Homo sapiens, J Comput Biol 25 (11) (2018) 1266–1277.
[137] M. Tahir, H. Tayara, K.T. Chong, iRNA-PseKNC(2methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components, J Theor Biol 465 (2019) 1–6.
[138] H. Li, L. Chen, Z. Huang, X. Luo, H. Li, J. Ren, Y. Xie, DeepOMe: A Web Server for the prediction of 2'-O-Me sites based on the hybrid CNN and BLSTM architecture, Front Cell Dev Biol 9 (2021), 686894.
[139] Y. Liu, D. Chen, R. Su, W. Chen, L. Wei, iRNA5hmC: the first predictor to identify RNA 5-hydroxymethylcytosine modifications using machine learning, Front Bioeng Biotechnol 8 (2020) 227.
[140] S.D. Ali, J.H. Kim, H. Tayara, K.T. Chong, Prediction of RNA 5-hydroxymethylcytosine modifications using deep learning, IEEE Access 9 (2021) 8491–8496.
[141] W. Chen, P. Feng, H. Tang, H. Ding, H. Lin, RAMPred: identifying the N(1)-methyladenosine sites in eukaryotic transcriptomes, Sci Rep 6 (2016) 31080.
[142] L. Liu, X.J. Lei, J. Meng, Z. Wei, ISGm1A: integration of sequence features and genomic features to improve the prediction of human m < sub > 1 </sub > A RNA methylation sites, Ieee Access 8 (2020) 81971–81977.
[143] J. Jiang, B. Song, Y. Tang, K. Chen, Z. Wei, J. Meng, m5UPred: a web server for the prediction of RNA 5-methyluridine sites from sequences, Mol Ther Nucleic Acids 22 (2020) 742–747.
[144] J. Jiang, B. Song, K. Chen, Z. Lu, R. Rong, Y. Zhong, J. Meng, m6AmPred: Identifying RNA N6, 2'-O-dimethyladenosine (m(6)Am) sites based on sequence-derived information, Methods (2021).
[145] K. Liu, W. Chen, iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications, Bioinformatics 36 (11) (2020) 3336–3342.
[146] Z. Chen, P. Zhao, F. Li, Y. Wang, A.I. Smith, G.I. Webb, T. Akutsu, A. Baggag, H. Bensmail, J. Song, Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences, Brief Bioinform 21 (5) (2020) 1676–1696.
[147] M. Tahir, M. Hayat, K.T. Chong, A convolution neural network-based computational model to identify the occurrence sites of various RNA modifications by fusing varied features, Chemom Intelligent Lab Syst 211 (2021).
[148] Z. Song, D. Huang, B. Song, K. Chen, Y. Song, G. Liu, J. Su, J.P. Magalhaes, D. J. Rigden, J. Meng, Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications, Nat Commun 12 (1) (2021) 4011.
[149] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J Theor Biol 273 (1) (2011) 236–247.
[150] W. Chen, H. Ding, X. Zhou, H. Lin, K.C. Chou, iRNA(m6A)-PseDNC: Identifying N-6-methyladenosine sites using pseudo dinucleotide composition, Anal Biochem 561–562 (2018) 59–65.
[151] L. Liu, X. Lei, Z. Fang, Y. Tang, J. Meng, Z. Wei, LITHOPHONE: improving lncRNA methylation site prediction using an ensemble predictor, Front Genet 11 (2020) 545.
[152] C. Dai, P. Feng, L. Cui, R. Su, W. Chen, L. Wei, Iterative feature representation algorithm to improve the predictive performance of N7-methylguanosine sites, Brief Bioinform (2020).