RNA Biology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/krnb20

HSM6AP: a high-precision predictor for the Homo *sapiens* N6-methyladenosine (m^6 A) based on multiple weights and feature stitching

Jing Li, Shida He, Fei Guo & Quan Zou

To cite this article: Jing Li, Shida He, Fei Guo & Quan Zou (2021) HSM6AP: a high-precision predictor for the Homo *sapiens* N6-methyladenosine (m^6 A) based on multiple weights and feature stitching, RNA Biology, 18:11, 1882-1892, DOI: 10.1080/15476286.2021.1875180

To link to this article: https://doi.org/10.1080/15476286.2021.1875180

RNA BIOLOGY
2021, VOL. 18, NO. 11, 1882–1892
https://doi.org/10.1080/15476286.2021.1875180

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

HSM6AP: a high-precision predictor for the Homo *sapiens* N6-methyladenosine (m^6 A) based on multiple weights and feature stitching

Jing Li[a], Shida He[a], Fei Guo[a], and Quan Zou[b]

[a]Institute of computational biology, College of Intelligence and Computing, Tianjin University, Tianjin, China; [b]Bioinformatics Laboratory, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China

ABSTRACT

Recent studies have shown that RNA methylation modification can affect RNA transcription, metabolism, splicing and stability. In addition, RNA methylation modification has been associated with cancer, obesity and other diseases. Based on information about human genome and machine learning, this paper discusses the effect of the fusion sequence and gene-level feature extraction on the accuracy of methylation site recognition. The significant limitation of existing computing tools was exposed by discovered of new features. (1) Most prediction models are based solely on sequence features and use SVM or random forest as classification methods. (2) Limited by the number of samples, the model may not achieve good performance. In order to establish a better prediction model for methylation sites, we must set specific weighting strategies for training samples and find more powerful and informative feature matrices to establish a comprehensive model. In this paper, we present HSM6AP, a high-precision predictor for the *Homo sapiens* N6-methyladenosine (m^6A) based on multiple weights and feature stitching. Compared with existing methods, HSM6AP samples were creatively weighted during training, and a wide range of features were explored. Max-Relevance-Max-Distance (MRMD) is employed for feature selection, and the feature matrix is generated by fusing a single feature. The extreme gradient boosting (XGBoost), an integrated machine learning algorithm based on decision tree, is used for model training and improves model performance through parameter adjustment. Two rigorous independent data sets demonstrated the superiority of HSM6AP in identifying methylation sites. HSM6AP is an advanced predictor that can be directly employed by users (especially non-professional users) to predict methylation sites. Users can access our related tools and data sets at the following website: *http://lab.malab.cn/~lijing/HSM6AP.html* The codes of our tool can be publicly accessible at *https://github.com/lijingtju/HSm6AP.git*

ARTICLE HISTORY
Received 24 July 2020
Revised 2 December 2020
Accepted 8 January 2021

KEYWORDS
Methylation site; sequence-derived feature; gene-derived features; feature stitching; XGBoost

Introduction

In recent years, with the continuous development of high-throughput detection technology and improvement in prediction methods, themachine learning method for the recognition of modification sites on protein RNA and DNA sequences has become a hotspot in computational biology [1–5]. Considering its advantages of automatization, rapidity, precision and relatively low cost, the machine learning is a suitable choice for predicting methylation sites. Therefore, the establishment of more accurate, efficient and strong generalization ability of the model becomes critical.

Due to the key role of N6-methyl adenosine (m^6A) in a series of biological processes, a method for constructing data sets based on the Euclidean distance is proposed. IRNA (m^6A)-PseDNC encodes the RNA sequence by using pseudo nucleotide components and identifies the m^6A site in the genome of brewer yeast [6]. High-throughput experiments have shown that only a small fraction of the m^6A common motif is modified in the mammalian transcriptome. The m^6A computational predictor (SRAMP) for mammals, which combines three feature extraction methods, random forest and voting strategies is established [7]. Consideration of the limitation of the precise location of m^6A decorations, an optical cross-lining-assisted m^6A sequencing strategy, which is expected to more accurately localize m^6A modification sites and is primarily used to predict human methylation sites, is proposed [8]. Considering that high-throughput experiments have produced m^6A peaks in the transcriptome of *A.thaliana* without assigning specific methylation sites, the understanding of m 6 A functions in plants is hampered. AthMethPre employed support vector machines (SVM) to construct the classifier and predict the methylation sites of *A. thaliana* [9]. Based on high-resolution experimental data of *Homo sapiens* and *M. musculus*, MethyRNA is proposed to identify the m^6A site by using the chemical properties of nucleotides and frequency-coding RNA sequences and support vector machine [10]. A new feature representation algorithm and two feature descriptors (dual nucleotide binary coding and local position-specific dinucleotide frequency) are applied to encode the RNA sequence in M6AMRFS. The F-score algorithm and

CONTACT Quan Zou zouquan@nclab.net Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China; Fei Guo fguo@tju.edu.cn College of Intelligence and Computing, Tianjin University, Tianjin, China
Supplemental data for this article can be accessed here.

Sequential Forward Search are employed to enhance the capabilities of feature representation. In addition, the XGBoost algorithm is applied to train predictive models [11–13]. To identify whether RNA sequences are methylated, the RNA sequence of the pseudo dinucleotide (PseDNC) characteristics of the samples is extracted by physical chemistry of the covariance and cross-covariance matrix transformation in pRNAm-PC [14]. Due to the limitations of the non-random distribution and biological functions in the rice genome, a two-layer integrated model is developed for the prediction of methylation sites in the rice genome, which is refered to as SDM6A [15]. Based on the data of miCLIP-Seq, a deep learning framework (DeepM6ASeq) that detects m^6A sites using single-base resolution and visualizes the location of m^6A sites, is proposed [16]. M6ASNP is mainly used to identify genetic variations that are specific to m^6A modification sites, and the random forest model is used to predict whether the methylation status of m^6A is altered by the surrounding variants [16,17,18].

DeepPromise, a computational approach based on deep learning, was proposed that uses three feature encoders (including enhanced nucleic acid composition, one-key encoding, and RNA embedding) as the input to seven successive layers of a convolutional neural network (CNNs) [19]. To complete the identification of m^6A sites, a web-based method named WHISTLE is proposed. WHISTLE [20] is designed to explore hypothetical biological processes that affect individual m^6A changes; integrate RNA methylation maps, gene expression maps, and protein–protein interaction data and implement a web-based m^6A site recognition method based on the principle of association penalty. The features based on sequences and genes information, which are combined with the SVM or random forest to identify the final m^6A site, are extracted. A recent review of RNA modification documents the mainstream method of unravelling the epitranscriptome in the field of bioinformatics [21].

WHISTLE is a state-of-the-art model for m^6A recognition that has some limitations in several aspects. (1) The accuracy of WHISTLE is not sufficiently high, which is a fatal flaw for classification problems. (2) The generalization ability of WHISTLE is poor. The performance of WHISTLE for independent data sets is not satisfactory. Therefore, a new m^6A site recognition method that can effectively solve the problem of weak prediction capacity and poor generalization ability is proposed. In addition, the direct use of m^6A site recognition is established, and users can employ the corresponding model according to their needs. The innovations of our model are described as follows.

(1) Creatively add weights to the positive samples to build a more credible prediction model. According to the explanation of different single-base resolutions for the same site, positive samples are given different weights, which greatly improves the predictive ability of the model.

(2) Integration of negative samples of m^6A. On a chromosome, the number of non-m^6A sites are considerably greater than that of m^6A. Therefore, the original ratio of positive samples to negative samples in the training set is 1:10. To build better performance models, negative samples are clustered, and the same number of negative samples were randomly selected from the clustering to establish the model.

(3) Crucial features are identified. In this experiment, a total of 23 features extraction methods are applied to obtain features from sequence and genomic coordinate information.

(4) The model has high prediction accuracy and strong generalization ability. The HSM6AP has better prediction and generalization ability than WHISTLE. HSM6AP achieves excellent performance for both the fivefold cross-validation and the independent data sets.

2. Methods

The construction of the HSM6AP model involves the re-integration of training data sets, the extraction and selection of sequence-derived features and gene-derived features, the fusion of features and the selection of classifiers (Figures 1–3), which are mainly divided into the following four steps. (1) The original training sequence and the original gene data are extracted from the gene model. The original genetic data are assigned different weights, which are 2, 3, 4, 5, and the weights of each negative samples is 1 (Figure 1 and Figure 2). The negative samples are clustered using the Gaussian

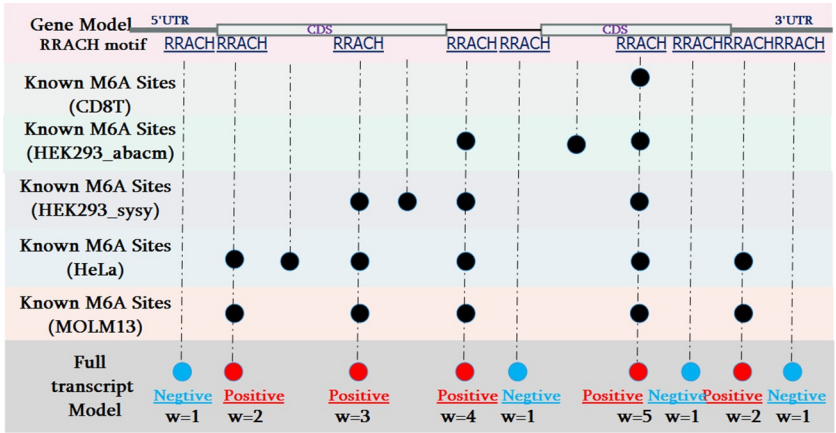

Figure 1. Data generation principle of training dataset and independent dataset 1 and weight setting.

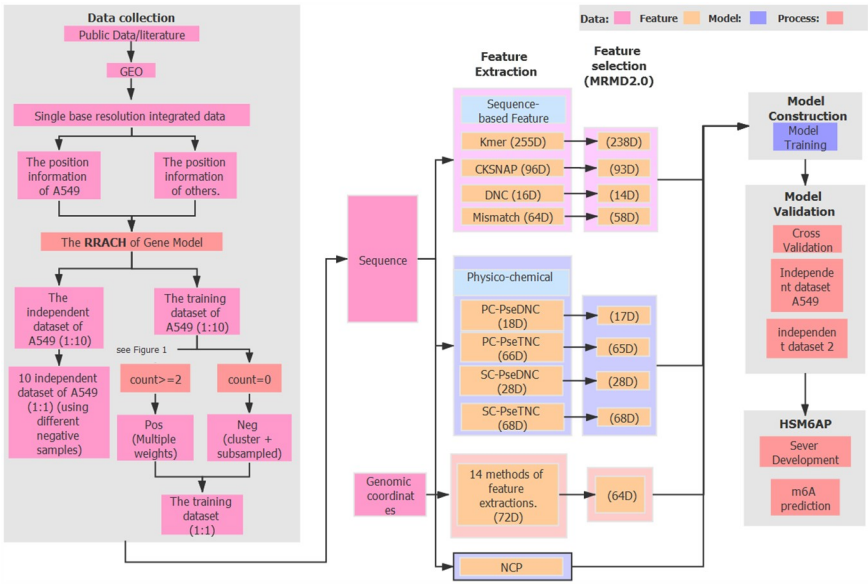

Figure 2. The flowchart of HSM6AP development.

Mixed Model (GMM), and randomly sampled, which constitutes the true training sequence and gene data sets (positive: negative = 1:1). (2) The training sequence is used to extract features with five sequence-based methods and four physico-chemical methods, while the training gene is extracted by using 14 gene-based feature extraction methods. (3) All the features but Nucleotide Chemical Property (NCP) are used for feature selection by using MRMD. The selected features and NCP are spliced to generate the final feature vectors. (4) XGBoost is used to train the model by fivefold cross-validation. The experimental results show that HSM6AP is superior to other models. Accurate identification of HSM6AP can help to understand life development and disease formation. The training data included two modes (full transcript and mature mRNA) from humans, which means that our web has two models. Depending on the type of test data set, users can choose the model they want.

2.1 Data collection

The data used for training and benchmarking in m6A site prediction includes six single-base resolution (derived from the cell type of the species), which are A549, CD8T, HEK293_abacm, HEK293_sysy, HeLa and MOLM13. The data derived from WHISTLE [20], which contains data of two modes (full transcript and mature mRNA). In full transcription, positive and negative samples may be located in the exon and intron regions. In contrast, the exon region was the only source of mature mRNA.

In model training and performance evaluation, A549 of the basic resolution data is used as independent testing data, while the remaining five data sets are used as training data. The positive training data were determined as the m6A sites under RRACH consensus motifs that have been reproduced in 2–5 of the five training data sets, which allows the weights of 2–5 to be obtained. Initially, the number of randomly selected negative samples was ten times the number of positive samples. All the negative samples were integrated and divided into five categories through GMM. Then, random down-sampling was conducted for each category. The final ratio of positive samples and negative samples in the training data is 1:1. The negative data was also generated similarly on testing data (A549). The ratio of positive testing data to negative testing data was also kept as 1:10. The testing performances from the ten independent sessions were averaged. In short, the training data with a positive and negative sample ratio of 1:1 were obtained by taking down-samples of all negative samples, while ten testing data sets (each with 1:1 positive-to-negative ratio), are constructed using different negative samples. To facilitate illustrating, this testing data is named Independent data set 1.

To illustrate the generalization power of our model, the independent 2 and 3 are introduced into the paper, which also derived from human A549 cells and HeLa cells, respectively. In each independent data set, the ratio of positive and negative samples is 1:10. Ten independent data sets, each with 1:1 positive-to-negative ratio, were built utilizing different negative samples. The positive sample size and negative sample size of independent data set 2 are 40,742 and 40,742, respectively (at the beginning, the number of positive samples and the negative samples are 40,742 and 407,420. After using different negative samples to build 1:1 positive-to-negative ratio, 40,742 positive samples and 40,742 negative samples above are obtained in each data set.) [22]. Independent data set 3 (the GEO accession is GSE134103) contains 15,696 positive samples and 15,696 negative samples (initially, the number of positive samples and the negative samples are 15,696 and 156,960. After utilizing different negative samples to establish 1:1 positive-to-negative ratio, 15,696 positive samples and 156,960 negative samples above are obtained in each data set.). There is no crossover between independent data sets 1–3. In addition, each sequence sample consists of 40 nucleotides. In this research, all negative training data (non-

m6A sites) was randomly selected from the non-positive RRACH adenosines.

In view of the single feature extraction method cannot obtain better performance, sequence and genome coordinate data are obtained simultaneously to ensure a more informative and powerful feature matrix (Fig. 2 and Fig. 3). All data set can be downloaded directly from *http://lab.malab.cn/~lijing/HSM6AP.html*

2.2 Feature extraction

The supplementary Tables S1 and S2 show that the performance of single sequence feature and genomic feature is not satisfactory, the combined feature matrix can achieve better performance [23]. Therefore, the features of sequence-based, physico-chemical properties and genome-derived are fused to produce the characteristic matrix.

2.2.1 Sequence-based features

The difference among the nucleotides can be directly reflected by the sequences. Thus, four types of sequence-based features are employed: Kmer [24–27], composition of *k*-spaced nucleic acid pairs (CKSNAP), Di-Nucleotide Composition (DNC), and Mismatch (Fig. 2 and Fig. 3). Kmer generates a 255-dimensional feature vector by characterizing the occurrence frequencies of k neighbouring nucleic acids, CKSNAP generates a 400-dimensional features vector by calculating the frequency of nucleic acid pairs separated by any k nucleic acid [28], and Mismatch generates a 64-dimensional feature vector by counting the occurrences of k-length neighbouring nucleic acids that differ by at most m mismatches [29]. In this process, AAAT, AACC, AACG, GACC, TAGC, the gap of AG, CG, GA, GC, and CG, the proportion of GC, GA, CT, TA, CC have a crucial role in determining the category of a site (Table 1, Table 2, Table S4, Fig. 2 and Fig. 3).

2.2.2 Physico-chemical properties

Physico-chemical properties have been extensively and successfully applied in numerous prediction tasks for DNA and RNA attributes [30–32]. In this experiment, five types of physico-chemical property-based features are included to describe the global compositions of nucleic acids' properties in a gene sequence: Nucleotide Chemical Property (NCP) [28], Parallel Correlation Pseudo Dinucleotide Composition (PCPseDNC) [33], Parallel Correlation Pseudo Trinucleotide Composition (PCPseTNC), Series Correlation Pseudo Dinucleotide Composition (SCPseDNC), and Series Correlation Pseudo Trinucleotide Composition (SCPseTNC)

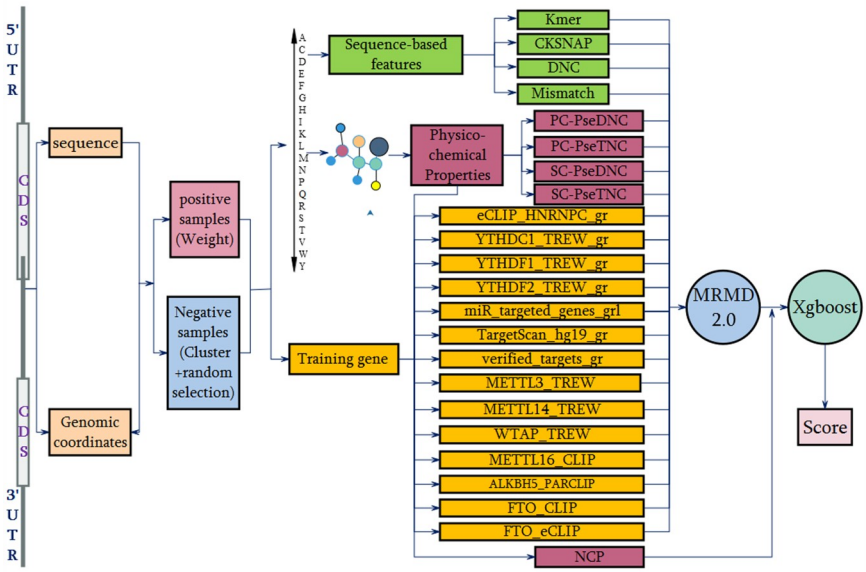

Figure 3. Detailed procedures for constructing the prediction models within HSM6AP architecture.

Table 1. The sizes of different data sets used in our work (independent data set 1).

ID	Cell	Note	Inde Pos	Inde Neg	Train Pos	Train Neg	*w* = 2	*w* = 3	*w* = 4	*w* = 5
1	A549	–	23,478	23,478	14,025	14,025	11,065	2346	511	103
2	CD8T	–	19,677	19,677	17,389	17,389	13,614	2953	686	136
3	HEK293	abacm antibody	9536	9536	20,829	20,829	14,647	4923	1085	174
4	HEK293	sysy antibody	12,051	12,051	20,937	20,937	14,701	4992	1039	205
5	HeLa	–	37,183	37,183	11,241	11,241	8802	1911	438	90
6	MOLM13	–	11,746	11,746	20,047	20,047	14,213	4769	919	146
1	A549	–	22,566	22,566	13,755	13,755	10,817	2327	509	102
2	CD8T	–	17,228	17,228	17,093	17,093	13,341	2932	685	135
3	HEK293	abacm antibody	9013	9013	23,417	23,417	14,288	4874	1082	173
4	HEK293	sysy antibody	7168	7168	20,525	20,525	14,341	4945	1035	204
5	HeLa	–	35,210	35,210	11,059	11,059	8637	1895	438	89
6	MOLM13	–	10,703	10,703	19,669	19,669	13,881	4728	915	145

Table 2. The 64 genome-derived features used for m6A site prediction.

ID	Name	Description	
1	UTR5	5′ UTR	Dummy variables indicating whether the site is overlapped to the topological region on the major RNA transcript
2	UTR3	3′ UTR	
3	cds	CDS	
4	Start_codons	Start codons flanked by 100bp	
5	pos_cds	Relative position on CDS	
6	Stop_codons	Stop codons	
7	constitutive_exon	Constitutive exons	
8	internal_exon	Internal exons	
9	last_exon	Last exon	
10	last_exon_400bp	5′ 400 bp of the last exons	
11	intron	Intros	
12	exon_stop	Exons containing stop codons	
13	alternative_exon	Alternative exon	
14	last_exon_sc400	Sc400 of the last exons	
15	pos_UTR5	Relative position on 5′UTR Relative	Relative position on the region
16	pos_UTR3	Relative position on 3′UTR	
17	pos_exons	Relative position on exon	
18	length_gene_ex	Mature transcript length	The region length in bp
19	length_tx_full	Full transcript length	
20	length_cds	CDS length	
21	length_UTR5	5′UTR length	
22	length_gene_full	Full transcript length	
23	length_UTR3	3′UTR length	
24	length_tx_exon	Mature transcript length	
25	PC_1bp	phastCons scores of the nucleotide	Scores related to evolutionary conservation
26	PC_101bp	Average phastCons scores within the flanking 101bp region	
27	FC_101bp	aAverage fitCons scores within the flanking 101bp region	
28	GC_cont_101bp	Average GC scores within the flanking 101bp	
29	GC_cont_genes	CG proportion of the gene at the site	
30	struct_loop	Predict RNA loop region	RNA secondary structures
31	struct_hybridize	Predict RNA hybridized region	
32	isoform_num	Isoform number	Attribute of the genes or transcripts
33	exon_num	Exon number	
34	HK_genes	Housekeeping gene	
35	miR_targeted_genes	miRNA targeted genes	RNA annotations related to M6A biology
36	Verified_miRtargets	miRNA targeted sites verified by experiment	
37	TargetScan	Predicted miRNA targeted sites by TargetScan	
38	lncRNA	Long non-conding RNA	
39	FTO_CLIP	CLIP data of FTO RNA binding sites	
40	ALKBH5_PARCLIP	PARCLIP data of ALKBH5 RNA binding sites	
41	METTL3_TREW	METTL3-binding region	RNA-binding protein annotation from MeTDB database
42	METTL14_TREW	METTL14-binding region	
43	YTHDC1_TREW	YTHDC1-binding region	
44	YTHDF1_TREW	YTHDF1 -binding region	
45	YTHDF2_TREW	YTHDF2 -binding region	
46	WTAP_TREW	WTAP -binding region	
47	dist_sj_3_p2000	Distance to the 3′ splicing junction	Nucleotide distances towards the splicing junctions or the nearest neighbouring sites.
48	dist_DRACH_p200	The nearest DRACH in 200	
49	dist_sj_5_p2000	Distance to the 5′ splicing junction	
50	dist_DRACH_p2000	The nearest DRACH in 2000	
51	AGACA	Particular motif	Particular motif
52	GAACA	Particular motif	
53	AAACA	Particular motif	
54	GGACA	Particular motif	
55	AAACC	Particular motif	
56	GGACT	Particular motif	
57	AAACT	Particular motif	
58	GAACT	Particular motif	
59	AGACT	Particular motif	
60	GAACC	Particular motif	
61	GGACC	Particular motif	
62	AGACC	Particular motif	
63	clust_DRACH_f100	DRACH motif within 100bp	Contains the DRACH
64	clust_DRACH_f100	DRACH motif within 1000bp	

(Fig. 2 and Fig. 3). All feature-encoding methods categorize the four primary nucleic acids into main classes according to specific types of physico-chemical properties, which produces 123-, 18-, 66-, 28-, and 68-dimensional feature vectors based on different formulas, respectively. NCP generates a 123-dimensional feature vector by establishing different codes for adenine (A), guanine (G), cytosine (C) and uracil (U). Four different kinds of codes (A, G, C, and U) have various chemical structures and chemical bindings. Based on the chemical properties, A, G, C and U can be expressed as (1, 1, 1), (0,

1, 0), (1, 0, 0) and (0, 0, 1), respectively (Table 2, Fig. 2 and Fig. 3). In this stage, the composition of GGA, GCA, TAG, ACG, CGA, ACA, TAA, TGA, AGC, TG, TT, AG, GA, CT, CC, TT, AG, TC, TA are considered as crucial features (Table 2, Table S4, Fig. 2 and Fig. 3).

2.2.3 Genome-derived characteristics

Almost all existing m^6 A prediction algorithms contain only sequence-derived features. To some extent, a single sequence feature is not capable of capturing complete site information. The R package was used to extract genomic features. Seventy-two features are generated, which are feature selected to retain 64 features by MRMD. Genomic features 1–14 represent dummy variables indicating whether the site is overlapped to the topological region on the major RNA transcript. Genomic features 15–16 are m6A relative position on 3′UTR and 5′UTR. Genomic features 25–29 show that the scores related to evolutionary conservation, while the genomic features 18–24 indicate the region length in bGenomic features 32–34 show the attribute of the genes or transcripts. The genomic features 30–32, 35–40 and 41–46 represent RNA secondary structures, RNA annotations related to M6A biology, and RNA-binding protein annotation from MeTDB database, respectively. Genomic features 47–50 reveal the nucleotide distances towards the splicing junctions or the nearest, while 63–64 represent the DRACH within different bGenomic features 51–62 list particular motif [34–39]. Among all properties, AGACH and AAACA are treated as the most important motifs; the internal exon and intro are the vital properties to indicate the degree of overlap between the site and the topological region of the main RNA transcript, and the isoform num is the critical transcription property (Table 2, Fig. 2 and Fig. 3). The original gene-derived feature comes from WHISTLE. MRMD was used for feature selection, and 63 features were eventually left for methylation site identification. By using MRMD2.0 to rank sequencing genomic features, AGACA, internal exon, intron, isoform num and AAACA are taken for crux to identify m6A (Table 2, Table S4, Fig. 2 and Fig. 3).

2.3 Max-Relevance-Max-Distance (MRMD)

The essence of machine learning is the statistical analysis of data and the establishment of corresponding models. Conversely, the nature of feature selection is to measure the superiority of a given subset of features by a specific evaluation criterion. By feature selection, irrelevant features of the original set are removed, while useful features are retained. The arrival of the information age marked that more data is readily available, and feature selection has become a priority in machine learning. Considering the number of training sets and the diversity of feature extraction methods, feature selection is considered necessary. When selecting a feature selection method, the stability of the prediction effect is regarded the most important indicator. MRMD [40] is given priority. MRMD filters non-informative data by sorting high-dimensional features, which can balance the accuracy and stability of feature sorting and prediction. Compared with other feature selection algorithms, a great advantage of MRMD is its stability, which can ensure that the feature after dimension reduction achieves excellent performance. Excellent features represented that the performance of the model can be promoted, and the underlying structure of the data can be understood more easily, and the model and algorithm can be improved more conveniently. In our experiment, MRMD was used for all features except NCP.

In this paper, we use MRMD2.0 (Maximum-Relevance-Maximum-Distance 2.0) to assess critical features. MRMD2.0 is a feature sorting algorithm, which can comprehensively use multiple feature selection algorithms to conduct special 'vote' on the features of the data set. In this algorithm, seven feature selection methods are included, which are used to calculate the importance of each feature. In this way, the results of the seven features are sorted, which can generate a directed graph to form the pointing relationship in the features (this relationship is similar to the link pointing relationship of web pages.). This directed graph is fed into the PageRank algorithm to calculate the score for each feature. Finally, the maximum score strategy is used to rank the features from the largest to the smallest according to the score of each feature. According to the feature selection principle of MRMD2.0, the feature with the highest score is considered as the crucial and important feature. Considering the importance of critical features, we listed the top five features corresponding to each feature selection method in Table S4 for readers' reference.

2.4 Gaussian mixed model (GMM)

In the paper, negative samples are obtained from five different single-base resolutions, which means that clustering function needs to solve the case in which the data in the same set contain multiple different classifications. The GMM [41] can fit any type of distribution and be a linear combination of multiple Gaussian distribution functions, which indicates that GMM has a strong clustering ability and suitable experimental data. The advantage of GMM is that the result of clustering contains not only the label of the data point, but also the probability.

The feature vectors obtained by ANF (Accumulated Nucleotide Frequency) [42] are used for clustering input. The ANF encodes nucleotide frequency information by calculating the nucleotide frequency information, the distribution of each nucleotide in the RNA sequence, and the density of any nucleotide position in the RNA sequence. It not only considers the chemical properties of sequences, but also incorporates the sequence information of long programs.

In this experiment, GMM algorithm designed five Gaussian distributions, which calculated the probability of sample belonging to Gaussian distribution through the analysis of feature vectors. The maximum probability value principle is used to determine the category of samples.

2.5 Extreme gradient boosting (XGBoost)

Considering that the main current classification methods for m^6A site recognition is the random forest or SVM, a more

powerful classification algorithm is expected. As an ascending tree model, XGBoost [43] is a strong classifier from the integration of many tree models (CART regression tree model). XGBoost is designed to grow a tree by continually adding trees and splitting the features. When a tree is added, a new function that is used to fit the residual predicted last time. When the training is complete, K trees are obtained. According to the features of the samples, the corresponding leaf nodes and predicted scores are identified in each tree. The sum of the predicted scores for each tree is the predicted value for the sample.

$$\hat{y} = \Phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \quad (1)$$

$$F = \{f(x) = w_{q(x)}\}(q : R^m \to T, w \in R^T) \quad (2)$$

where $w_{q(x)}$ is the score of the leaf node q, and $f(x)$ is one of the regression trees.

XGBoost is employed for m^6A site recognition for several reasons. (1) Many overfitting prevention strategies are applied in XGBoost. If the model overlearns the characteristics of the training set, the model is likely to consider some features of the training sample as a general property, which causes a decline in the generalization ability. For machine learning algorithms, overfitting cannot be completely avoided, which means that the use of a preventing overfitting strategy has great significance in machine learning. (2) By assigning different weights to the samples, important samples can receive more attention. To obtain more accurate models in the training data set, different samples are assigned different weights, and the effect is further improved.

2.6 Model construction

The construction process of the model is detailed as follows (Figs. 1–3):

1. Data processing

A. Generation of sequence data. In this experiment, the corresponding sequence data are extracted from the original genomic coordinate data, which are used to recognize the methylation sites.

B. Set of sample weights. As the positive sample of each training data set is composed of five single-base resolutions, different single-base resolutions have different labels for the same site. Therefore, different weights (2–5 for positive samples and 1 for negative samples) are assigned to each sample according to the performance of the different single-base resolutions at the same site.

C. Generation of negative samples. Considering that the number of non-methylated sites is much larger than the methylated sites on chromosomes, negative samples are selected in the experiment (positive: negative = 1:10). To ensure a better prediction performance and stronger generalization ability of the model, all negative samples are collected for clustering by using the GMM, and are clustered into five classes. Finally, a training set with a ratio of positive and negative samples to 1:1 is generated.

2. Feature extraction

Sequence-based features, physico-chemical properties and gene-derived features, which are NCP, CKSNAP, DNC, Mismatch, PC-PseDNC, PC-PseTNC, SC-PseDNC, SC-PseTNC and 14 gene feature extraction methods.

3. Feature selection and feature stitching

All features, except NCP, are selected using MRMD, which are spliced together to produce the final feature vector. Compared with training for a single feature, the feature selection and splicing can significantly improve the performance of the model in principle.

4. Training model using XGBoost

Based on the weight information of samples and the advanced classification ability, XGBoost is considered a suitable classification algorithm. In this process, fivefold cross-validation is applied for model training and construction. Independent data sets 2 and 3 are used to further demonstrate the classification capacity and generalization ability of the model. See supplementary table S3 for parameter Settings for XGBoost.

2.7 Performance evaluation

To research the contributions of different feature groups and compare the performance of the model with existing advanced methods, the application based on cross-validation and independent testing is defined as

$$\mathrm{SN} = \frac{TP}{TP + FN} \quad (3)$$

$$\mathrm{SP} = \frac{TN}{TN + FP} \quad (4)$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\mathrm{F_1} = \frac{2 * TP}{2TP + FP + FN} \quad (6)$$

$$\mathrm{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (7)$$

where TP, TN, FP, FN represent the number of true positive, true negative, false positive and false negative, respectively [44–54]. In addition, the performance results between two different methods are visualized using the receiver operating

Table 3. Performance of independent data set of 1 on HSM6AP.

Mode	Model	ACC (fivefold)	SN	SP	F_score	ACC	MCC	AUC	Ave AUC
Full transcript	A549	0.973	0.977	0.94	0.959	0.96	0.918	0.988	**0.976**
	CD8T	0.968	0.919	0.932	0.926	0.925	0.851	0.968	
	HEK293_abacm	0.978	0.907	0.958	0.932	0.931	0.866	0.985	
	HEK293_sysy	0.985	0.671	0.98	0.826	0.794	0.685	0.937	
	HeLa	0.984	0.902	0.974	0.938	0.936	0.879	0.99	
	MOLM13	0.968	0.908	0.939	0.924	0.923	0.848	0.97	
Mature mRNA	A549	0.908	0.863	0.823	0.843	0.846	0.687	0.924	**0.899**
	CD8T	0.89	0.757	0.839	0.798	0.789	0.598	0.892	
	HEK293_abacm	0.97	0.483	0.977	0.73	0.641	0.528	0.89	
	HEK293_sysy	0.899	0.813	0.81	0.812	0.812	0.623	0.895	
	HeLa	0.901	0.795	0.834	0.815	0.811	0.63	0.894	
	MOLM13	0.967	0.415	0.975	0.71	0.576	0.476	0.899	

characteristic curve (ROC), and the area under the Curve of ROC (AUC) is the criterion to judge the performance of dichotomy prediction model.

3. Experimental results

3.1 Performance evaluation based on fivefold cross-validation test

Cross-validation, which is a common method for validating models in machine learning [51,55–64], can accurately adjust the super parameters of models and effectively prevent overfitting due to model overcomplexity. Cross-validation [65] is used to evaluate the predictive performance of models, especially in the new data performance, which can reduce overfitting to some extent. Cross-validation can extract as much valid information as possible from limited data. All model training is fivefold cross-validation in the experiment. In the full transcript, the accuracy of cross-validation exceeds 96%, while the accuracy of cross-validation exceeds 89% in the mature mRNA (Table 3). The performance of full transcript and mature mRNA means that the HSM6AP not only has powerful prediction function, but also strong generalization ability.

3.2 Performance evaluation of independent data sets 1–3

In the section of data collection, we describe the independent data set 1 in detail, which contains both full transcription and mature mRNA mode data. After a series of data processing, we obtained that each single-base resolution is an independent data set, and the corresponding remaining five single-base resolutions are training data sets. Take A549 as an example to elaborate. In full transcription, the A549 model refers to the model when A549 is the independent data set and the integration of CD8T, HEK293_abacm, HEK293_sysy, HeLa, and MOLM13 is the training data set.

In the model evaluation section, the performance of ten independent data set is averaged for each type. On independent data set 1, all indicators of A549, CD8T, HEK293_abacm, HeLa and MOLM13 performed well, and

Table 4. Performance of independent data set 2 on HSM6AP.

Mode	Model	SN	SP	F_score	ACC	MCC	AUC	Ave AUC
Full transcript	A549	0.942	0.939	0.942	0.967	0.716	0.989	**0.981**
	CD8T	0.92	0.936	0.921	0.956	0.654	0.983	
	HEK293_abacm	0.903	0.948	0.907	0.947	0.624	0.98	
	HEK293_sysy	0.928	0.942	0.929	0.96	0.678	0.979	
	HeLa	0.894	0.953	0.899	0.942	0.608	0.977	
	MOLM13	0.906	0.94	0.909	0.948	0.625	0.978	
Mature mRNA	A549	0.78	0.958	0.794	0.875	0.448	0.927	**0.914**
	CD8T	0.768	0.935	0.781	0.866	0.422	0.931	
	HEK293_abacm	0.81	0.717	0.809	0.889	0.343	0.879	
	HEK293_sysy	0.769	0.959	0.784	0.867	0.437	0.92	
	HeLa	0.776	0.925	0.789	0.875	0.419	0.907	
	MOLM13	0.83	0.846	0.831	0.9	0.441	0.918	

Table 5. Performance of independent data set 3 on HSM6AP.

Mode	Model	SN	SP	F_score	ACC	MCC	AUC	Ave AUC
Full transcript	A549	0.934	0.885	0.909	0.911	0.823	0.968	**0.967**
	CD8T	0.932	0.855	0.892	0.896	0.8	0.966	
	HEK293_abacm	0.948	0.758	0.843	0.858	0.732	0.969	
	HEK293_sysy	0.964	0.566	0.713	0.773	0.598	0.968	
	HeLa	0.96	0.617	0.751	0.80	0.633	0.967	
	MOLM13	0.933	0.846	0.887	0.893	0.788	0.966	
Mature mRNA	A549	0.955	0.478	0.637	0.728	0.525	0.905	**0.890**
	CD8T	0.951	0.410	0.573	0.694	0.473	0.916	
	HEK293_abacm	0.979	0.138	0.242	0.567	0.263	0.898	
	HEK293_sysy	0.952	0.458	0.618	0.717	0.509	0.881	
	HeLa	0.945	0.431	0.592	0.703	0.483	0.850	
	MOLM13	0.980	0.151	0.261	0.574	0.277	0.889	

the values of SN, SP, F_score, ACC and AUC exceed 0.9. The effect of HEK293_sysy did not satisfy expectations, and the AUC was 0.937. The average AUC is 0.976 in the full transcript, while the average AUC of mature mRNA is 0.899. The positive and negative sample size of independent data set 2 are 40,742 and 3575, respectively, which means that the AUC is more reasonable as a comprehensive evaluation standard. In independent data set 2, the full transcript has an average AUC 0.981, and almost all values are high. Conversely, the mean AUC of mature mRNA is 0.914. On independent data set 3, average AUC in full transcript and mature mRNA is 0.967 and 0.890, respectively (Table 3, Table 4 and Table 5). The performance of HEK293_sysy is not satisfactory in mode of full transcript and the SP of MOLM13 and HEK293_abacm show poor performance in mature mRNA, which means that HSM6AP has a strong ability to recognize methylation sites. The experimental results analysis shows that the model of the full transcript and mRNA has better performance for independent data sets 1–3, which means that HSM6AP has a stronger predictive and generalization ability for the recognition of a methylation site.

The method of leave-one-out verification at the data set level can reduce the risk of overfitting of the model. The performance of independent data set 2 also directly illustrates that the model has strong prediction and generalization ability.

3.3 Performance comparison with other existing tools

The WHISTLE model is constructed by fusing gene-derived features and sequence-derived features, and combining them with SVM. With the advent of the era of big data, deep learning is extensively applied in the field of data mining [66–69]. DeepM6ASeq can utilize sequence information to predict methylation sites, which is a new addition to the field of methylation site identification. The method of DeepM6ASeq and our training data set were used to train the DeepM6ASeq model.

To further demonstrate the superiority of HSM6AP, WHISTLE and DeepM6ASeq are applied in comparative experiments. The results of these methods are shown in Table 6. For independent data set 1, the AUC of HSM6AP is 0.052 and 0.196 higher than that of WHISTLE and DeepM6ASeq in the full transcript mode, and 0.073 and 0.119 higher than that of WHISTLE and DeepM6ASeq in the mature mRNA. For independent data set 2, the AUC value of HSM6AP for the full transcript is 0.981, which comprehensively outperformed all other toolkits by more than 0.042 in WHISTLE and 0.219 in DeepM6ASeq. HSM6AP also achieved better performance in mature mRNA, whose AUC is 0.084 and 0.153 better than WHISTLE and DeepM6ASeq, respectively. HSM6AP (average AUC of 0.967 and 0.914) is higher accuracy than WHISTLE (average AUC of 0.953 and 0.869) and DeepM6ASeq (average AUC of 0.674 and 0.593) (Table 6). The experimental results show that HSM6AP is superior to the most advanced methods in predicting the methylation sites in *Homo sapiens*.

Table 6. Comparison results of HSM6AP, WHISLE and DeepM6ASeq on independent data sets 1–3.

Model	Full transcript (Ave AUC)	Mature mRNA (Ave AUC)
Independent data set 1		
HSM6AP	**0.976**	**0.899**
WHISLE	0.924	0.826
DeepM6ASeq	0.780	0.780
Independent data set 2		
HSM6AP	**0.981**	**0.914**
WHISLE	0.939	0.830
DeepM6ASeq	0.761	0.761
Independent data set 3		
HSM6AP	**0.967**	**0.890**
WHISLE	0.953	0.869
DeepM6ASeq	0.674	0.593

In this paper, we use the method of WHISTLE and DeepM6ASeq and our training data to build WHISTLE and DeepM6ASeq. It is worth mentioning that, independent data sets 1–3 were tested using actual WHISTLE and DeepM6ASeq. The indicators of independent data sets 1–3 use the actual results that we obtained.

4. Conclusion

In this research, a method for identifying methylation sites of *Homo sapiens* based on feature stitching is proposed. Methylation site prediction is one of the difficulties in bioinformatics, but has great significance to the development of targeted therapies for cancer. Although a variety of models for identifying methylation sites have been published, a large number of problems, such as overfitting and poor generalization ability, have not been properly solved. Conversely, HSM6AP shows a unique advantage, which is the result of a special design for *Homo sapiens* methylation site recognition. In addition, the construction process of HSM6AP is crucial to the identification of methylation site recognition. (1) According to the particularity of positive samples for different single-base resolutions, different weights are set to improve the predictive power of HSM6A. (2) Considering that the number of non-methylated sites on a chromosome is much larger than that of methylated sites on a chromosome, more negative samples are obtained from the chromosome. Negative samples are randomly clustered and downsampled to improve the recognition and generalization performance of the model. (3) Critical features of methylation sites are identified, and the combination of sequence-derived features and gene-derived features further improves the effect of the model. (4) Most of the methylation site recognition models are based on the SVM and random forest. Compared with the former, XGBoost based on the tree algorithm has stronger noise resistance and classification ability. The comprehensive performance using independent data sets of 1–3 indicates that HSM6A has strong site recognition ability and robustness. Therefore, HSM6AP can identify methylation sites more accurately and is expected to become a powerful tool. In addition, the identification of methylation sites is closely related to the generation of various diseases. We aim to carry out the related work of diseases and M6A in the future, and strive to discover the internal relationship between M6A and diseases to contribute to epigenetic association analysis and clinical research.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [61771331 ,61922020, 91935302]; the National Key R&D Program of China [2018YFC0910405].

References

[1] Zhu X, et al. A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae. Brief Funct Genomics. 2019;18(6):367–376.
[2] Li YH, Yu CY, Li XX, et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Res. 2018;46(D1):D1121–D1127.
[3] Yin J, Sun W, Li F, et al. VARIDT 1.0: variability of drug transporter database. Nucleic Acids Res. 2020;48(D1):D1042–D1050.
[4] Lv H, et al. Evaluation of different computational methods on 5-methylcytosine sites identification. Brief Bioinform. 2019;21(3):982–995.
[5] Chen W, Feng P, Song X, et al. iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. Mol Ther Nucleic Acids. 2019;18(p):269–274.
[6] Lai HY, Feng C-Q, Zhang Z-Y, et al. A brief survey of machine learning application in cancerlectin identification. Curr Gene Ther. 2018;18(5):257–267.
[7] Zhou Y, Zeng P, Li Y-H, et al. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res. 2016;44(10):e91.
[8] Chen K, Lu Z, Wang X, et al. High-resolution N6-methyladenosine (m6A) map using photo-crosslinking-assisted m6A sequencing. Angew Chem. 2015;54(5):1587–1590.
[9] Xiang S, Yan Z, Liu K, et al. AthMethPre: A web server for the prediction and query of mRNA m 6 A sites in Arabidopsis thaliana. Mol Biosyst. 2016;12(11):3333–3337.
[10] Chen W, Tang H, Lin H. MethyRNA: a web server for identification of N6-methyladenosine sites. J Biomol Struct Dyn. 2017;35(3):683–687.
[11] Qiang X, Chen H, Ye X, et al. M6AMRFS: robust prediction of N6-methyladenosine sites with sequence-based features in multiple species. Front Genet. 2018;9. DOI:10.3389/fgene.2018.00495.
[12] Fu J, Tang J, Wang Y, et al. Discovery of the consistently well-performed analysis chain for SWATH-ms based pharmacoproteomic quantification. Front Pharmacol. 2018;9(p):681.
[13] Tang J, et al. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. Brief Bioinform. 2019;21(2):621–636.
[14] Liu Z, Xiao X, Yu D-J, et al. pRNAm-PC: predicting N6-methyladenosine sites in RNA sequences via physical–chemical properties. Anal Biochem. 2016;497(p):60–67.
[15] Basith S, Manavalan B, Shin TH, et al. SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. Mol Ther Nucleic Acids. 2019;18(p):131–141.
[16] Zhang Y, Hamada M. DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. BMC Bioinformatics. 2018;19(19):524.
[17] Tang J, Fu J, Wang Y, et al. Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. Mol Cell Proteomics. 2019;18(8):1683–1699.
[18] Xue W, Yang F, Wang P, et al. What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. ACS Chem Neurosci. 2018;9(5):1128–1140.
[19] Chen Z, Liu X, Li F, et al. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief Bioinform. 2019;20(6):2267–2290.
[20] Chen K, Wei Z, Zhang Q, et al. WHISTLE: a high-accuracy map of the human N 6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res. 2019;47(7):e41.
[21] Liu L, Song B, Ma J, et al. Bioinformatics approaches for deciphering the epitranscriptome: recent progress and emerging topics. Comput Struct Biotechnol J. 2020;18(p):1587.
[22] Schwartz S, Mumbach M, Jovanovic M, et al. Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites. Cell Rep. 2014;8(1):1.
[23] Wang J, Yang B, Leier A, et al. Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. Bioinformatics. 2018;34(15):2546–2555.
[24] Fang T, et al. RNAm5CPred: prediction of RNA 5-Methylcytosine sites based on three different kinds of nucleotide composition. Mol Ther Nucleic Acids. 2019;18(p):739–747 .
[25] Su ZD, et al. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. 2018;34(24):4196–4204.
[26] Yang Q, Wang S, Dai E, et al. Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. Brief Bioinform. 2019;20(1):168–177.
[27] Yang H, et al. A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. Brief Bioinform. 2019;21(5):1568–1580.
[28] Chen Z, et al. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. Brief Bioinform. 2019.
[29] Song T, Zou Q, Liu X, et al. Asynchronous spiking neural P systems with rules on synapses. Neurocomputing. 2015;151(p):1439–1445.
[30] Feng CQ, Zhang Z-Y, Zhu X-J, et al. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics. 2019;35(9):1469–1477.
[31] Dao FY, Lv H, Wang F, et al. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics. 2019;35(12):2075–2083.
[32] Chen W, Zhang X, Brooker J, et al. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics. 2014;31(1):119–120.
[33] Liu B, et al. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015;43(W1):W65–W71.
[34] Lawrence M, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.
[35] Hsu F, et al. The UCSC genome browser database: update 2006[J]. Nucleic acids research, 2006, 34(suppl_1): D590–D598.
[36] Gruber AR, Bernhart SH, Lorenz R, *The ViennaRNA Web Services*.
[37] Betel D, et al. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. 2010;11:8:R90.
[38] Vikram A, et al. Predicting effective microRNA target sites in mammalian mRNAs. 2015;4:e05005.
[39] Tang H, Su Z-D, Wei -H-H, et al. Prediction of cell-penetrating peptides with feature selection techniques. Biochem Biophys Res Commun. 2016;477(1):150–154.

[40] Zou Q, et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing. 2015;173:346–354.
[41] Hall P, Ormerod JT, Wand MP. Theory of gaussian variational approximation for a poisson mixed model. Statistica Sinica. 2011;21(1):369–389.
[42] Chen Z, Zhao P, Li F, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform. 2020;21(3):1047–1057.
[43] Chen T, Tong H, Benesty M, *xgboost: extreme gradient boosting*. 2016.
[44] Zeng X, Ding N, Rodríguez-Patón A, et al. Probability-based collaborative filtering model for predicting gene–disease associations. BMC Med Genomics. 2017;10(5):76.
[45] Chen X, Pérez-Jiménez MJ, Valencia-Cabrera L, et al. Computing with viruses. Theor Comput Sci. 2016;623(p):146–159.
[46] Zeng X, et al. A consensus community-based particle swarm optimization for dynamic community detection. IEEE Trans Cybern. 2019. DOI:10.1109/TCYB.2019.2938895.
[47] Xu L, Liang G, Shi S, et al. SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. Int J Mol Sci. 2018;19(6):6.
[48] Chen W, Ding H, Zhou X, et al. iRNA (m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition. Anal Biochem. 2018;561(p):59–65.
[49] Xiong Y, et al. PredT4SE-Stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. Front Microbiol. 2018;9(p):2571.
[50] Xu Q, et al. PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. J Theor Biol. 2017;417(p):1–7.
[51] Shen Y, Tang J, Guo F. Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. J Theor Biol. 2019;462(p):230–239.
[52] Ding Y, Tang J, Guo F. Identification of drug-side effect association via multiple information integration with centered kernel alignment. Neurocomputing. 2019;325(p):211–224.
[53] Zhu F, Shi Z, Qin C, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic Acids Res. 2012;40:D1128–36Database issue
[54] Zhang M, Li F, Marquez-Lago TT, et al. MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. Bioinformatics. 2019;35(17):2957–2965.
[55] Zeng X, et al. Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. Brief Bioinform. 2019. DOI:10.1093/bib/bbz080.
[56] Cabarle FGC, de la Cruz RTA, Cailipan DPP, et al. On solutions and representations of spiking neural P systems with rules on synapses. Inf Sci. 2019;501(p):30–49.
[57] Zeng X, Zhu S, Liu X, et al. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics. 2019;35(24):5191–5198.
[58] Lin X, et al. A novel molecular representation with BiGRU neural networks for learning atom. Brief Bioinform. 2019. DOI:10.1093/bib/bbz125.
[59] Xu L, Liang G, Liao C, et al. An efficient classifier for Alzheimer's disease genes identification. Molecules. 2018;23(12):3140.
[60] Xu L, Liang G, Liao C, et al. k-Skip-n-gram-RF: A random forest based method for Alzheimer's disease protein identification. Front Genet. 2019;10:33.
[61] Dou L, Li X, Ding H, et al. Is there any sequence feature in the RNA pseudouridine modification prediction problem? Mol Ther Nucleic Acids. 2020;19(p):293–303.
[62] Shan X, Wang X, Li C-D, et al. Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method. J Chem Inf Model. 2019;59(11):4577–4586.
[63] Chu Y, Kaushik AC, Wang X, et al. DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. Brief Bioinform. 2019. DOI:10.1093/bib/bbz152.
[64] Jia C, Zuo Y, Zou Q. O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. Bioinformatics. 2018;34(12):2029–2036.
[65] Watanabe S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. Arxiv. 2010;11(18):3571–3594.
[66] Zeng X, Lin Y, He Y, et al. Deep collaborative filtering for prediction of disease genes. IEEE/ACM Trans Computat Biol Bioinf. 2019; 1. DOI:10.1109/TCBB.2019.2907536.
[67] Zeng X, Lin W, Guo M, et al. A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput Biol. 2017;13(6):e1005420.
[68] Song T, Rodriguez-Paton A, Zheng P, et al. Spiking neural p systems with colored spikes. IEEE Trans Cognit Dev Syst. 2018;10(4):1106–1115.
[69] Li B, Tang J, Yang Q, et al. NOREVA: normalization and evaluation of MS-based metabolomics data. Nucleic Acids Res. 2017;45(W1):W162–W170.