

RESEARCH PAPER

 OPEN ACCESS



# M6A-BiNP: predicting N<sup>6</sup>-methyladenosine sites based on bidirectional position-specific propensities of polynucleotides and pointwise joint mutual information

Mingzhao Wang<sup>a,b</sup>, Juanying Xie<sup>b</sup>, and Shengquan Xu<sup>a</sup>

<sup>a</sup>College of Life Sciences, Shaanxi Normal University, Xi'an, China; <sup>b</sup>School of Computer Science, Shaanxi Normal University, Xi'an, China

## ABSTRACT

N<sup>6</sup>-methyladenosine (m<sup>6</sup>A) plays an important role in various biological processes. Identifying m<sup>6</sup>A site is a key step in exploring its biological functions. One of the biggest challenges in identifying m<sup>6</sup>A sites is how to extract features comprising rich categorical information to distinguish m<sup>6</sup>A and non-m<sup>6</sup>A sites. To address this challenge, we propose bidirectional dinucleotide and trinucleotide position-specific propensities, respectively, in this paper. Based on this, we propose two feature-encoding algorithms: Position-Specific Propensities and Pointwise Mutual Information (PSP-PMI) and Position-Specific Propensities and Pointwise Joint Mutual Information (PSP-PJMI). PSP-PMI is based on the bidirectional dinucleotide propensity and the pointwise mutual information, while PSP-PJMI is based on the bidirectional trinucleotide position-specific propensity and the proposed pointwise joint mutual information in this paper. We introduce parameters  $\alpha$  and  $\beta$  in PSP-PMI and PSP-PJMI, respectively, to represent the distance from the nucleotide to its forward or backward adjacent nucleotide or dinucleotide, so as to extract features containing local and global classification information. Finally, we propose the M6A-BiNP predictor based on PSP-PMI or PSP-PJMI and SVM classifier. The 10-fold cross-validation experimental results on the benchmark datasets of non-single-base resolution and single-base resolution demonstrate that PSP-PMI and PSP-PJMI can extract features with strong capabilities to identify m<sup>6</sup>A and non-m<sup>6</sup>A sites. The M6A-BiNP predictor based on our proposed feature encoding algorithm PSP-PJMI is better than the state-of-the-art predictors, and it is so far the best model to identify m<sup>6</sup>A and non-m<sup>6</sup>A sites.

## ARTICLE HISTORY

Received 24 February 2021

Revised 1 May 2021

Accepted 10 May 2021

## KEYWORDS




N<sup>6</sup>-methyladenosine (m<sup>6</sup>A); predictive model; feature representation; nucleotide position-specific propensities; pointwise joint mutual information


## Introduction


Epigenetics refers to the study of genetic variations in gene expression under the condition that the nucleotide sequence composition of genes remains unchanged [1]. RNA methylation is the most important epigenetic modification of ~150 chemical modifications. It is the process to transfer methyl catalytic from an active methyl compound, such as S-adenosine methionine, to different positions of an RNA molecule and make the chemical modification to form methylated products [2,3]. The common RNA methylation patterns include N<sup>6</sup>-methyladenosine (m<sup>6</sup>A), N<sup>1</sup>-methyladenosine (m<sup>1</sup>A) and C<sup>5</sup>-methylcytidine (m<sup>5</sup>C) etc., where m<sup>6</sup>A modification exists in Bacteria [4], Homo sapiens [5], Arabidopsis thaliana [6], etc. The m<sup>6</sup>A is a dynamic reversible modification regulated by a series of methyltransferases, such as MettL3/14, WTAP and YTHDF2 [7–9], and the demethylases, such as FTO and ALKBH5 [10,11]. It plays an important role in many molecular processes, such as protein translation and localization [12], splicing [13], RNA stability [14], mRNA longevity control and degradation [12], and cell differentiation promotion [15]. It is also associated with the occurrence of complex diseases [16], such as Glioblastoma formation [17], breast cancer [18] and obesity [11]. Therefore,

identifying m<sup>6</sup>A will benefit the diagnosis and treatment of complex diseases, even understanding their mechanism. It has valuable scientific and applicable value in personal medicine and drug development.

With the development of second-generation sequencing technology, a number of non-single-base resolution m<sup>6</sup>A site identification protocols, such as m<sup>6</sup>A-seq [19] and MeRIP-Seq [13], and single-base resolution m<sup>6</sup>A site identification protocols, such as miCLIP [20], m<sup>6</sup>A-CLIP [21] and m<sup>6</sup>A-REF-seq [22], were proposed based on high-throughput sequencing technology. At present, the m<sup>6</sup>A sites of *Saccharomyces cerevisiae* [23], *Arabidopsis Thaliana* [6], *Oryza sativa* [24], *Mus Musculus* [5] and *Homo sapiens* [5] have been identified at the full-transcriptome level. The study results show that the distribution of m<sup>6</sup>A sites is highly conservative, and most of them have a common consensus motif DRACH (A = m<sup>6</sup>A; D = A or G or U; R = A or G; H = A or C or U) [5,13,20]. This lays a theoretical base for identifying m<sup>6</sup>A sites using machine learning techniques. However, the high-throughput sequencing technology-based m<sup>6</sup>A site identification methods are time-consuming and inaccuracy, such that they cannot be used on large-scale genomic data. Therefore, many m<sup>6</sup>A site predictive models have been proposed in recent years based on various feature representation methods of sequence and

**CONTACT** Juanying Xie  [xiejuany@snnu.edu.cn](mailto:xiejuany@snnu.edu.cn)  School of Computer Science, Shaanxi Normal University, Xi'an, China; Shengquan Xu  [xushengquan@snnu.edu.cn](mailto:xushengquan@snnu.edu.cn)

 College of Life Sciences, Shaanxi Normal University, Xi'an China

 Supplemental data for this article can be accessed [here](#)

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

traditional machine learning algorithms or deep learning framework [25–48]. The latest several predictors, such as Gene2Vec [38], DeepPromise [39], WHISTLE [40], im6A-TS-CNN [42], iRNA-m6A [43] and HSM6AP [44] etc., were developed to identify and predict the m<sup>6</sup>A sites with the golden standard datasets at the single-base resolution level.

Although there are so many computable models to identify m<sup>6</sup>A sites, it is still a challenging task to distinguish m<sup>6</sup>A from non-m<sup>6</sup>A sites accurately. The most key issue is how to extract features containing more categorical information from RNA sequences. Therefore, this paper proposes two new feature encoding algorithms named Position-Specific Propensities and Pointwise Mutual Information (PSP-PMI) and Position-Specific Propensities and Pointwise Joint Mutual Information (PSP-PJMI), respectively. The bidirectional dinucleotide and trinucleotide position-specific propensities are, respectively, proposed in PSP-PMI and PSP-PJMI based on Pointwise Mutual Information (PMI) and our Pointwise Joint Mutual Information (PJMI) theories, respectively. The parameters  $\alpha$  and  $\beta$  are introduced to represent the distance between nucleotides in a pair of nucleotides in PSP-PMI, and the distance from the nucleotide to its forward or backward consecutive dinucleotide in PSP-PJMI, respectively, so as to extract more discriminative features from RNA sequences. The features corresponding to different  $\alpha$  and  $\beta$  are, respectively, concatenated to comprise a high-dimensional feature vector embodying both local and global position-specific information of nucleotides between m<sup>6</sup>A and non-m<sup>6</sup>A sites. Finally, the novel m<sup>6</sup>A site predictor named as M6A-BiNP is proposed based on aforementioned contributions and Support Vector Machine (SVM) classifier. We test our M6A-BiNP models on a number of non-single-base resolution and single-base resolution m<sup>6</sup>A benchmark datasets of different species. The 10-fold cross-validation experimental results demonstrate that our PSP-PMI and PSP-PJMI algorithms can extract features with much more discriminative capability for identifying m<sup>6</sup>A sites from RNA sequences. The M6A-BiNP predictor based on our feature encoding algorithm PSP-PJMI is superior to the state-of-the-art predictive models, and it is so far the best model for identifying m<sup>6</sup>A site.

**Table 1.** The detailed information of the non-single-base resolution benchmark datasets.

Species	# positive samples	# negative samples	#Total samples	Sequence length (nt)
<i>Arabidopsis thaliana</i>	394	394	788	25
<i>Musculus</i>	725	725	1450	41
<i>Homo sapiens</i>	1130	1130	2260	41
<i>Saccharomyces cerevisiae</i>	1307	1307	2614	51

## Materials and methods

### Datasets

There are two types of benchmark datasets used to test our feature encoding algorithms PSP-PMI and PSP-PJMI, and our M6A-BiNP predictors. The first type is non-single-base resolution data that across four species of *Arabidopsis thaliana* [27,49], *Musculus* [5,34], *Homo Sapiens* [50] and *Saccharomyces cerevisiae* [25] were generated from the low-resolution level technique MeRIP-Seq. The detailed information of the non-single-base resolution datasets is shown in Table 1. The second type is the single-base resolution data including three species of human, mouse and rat, which were generated from two single-base resolution m<sup>6</sup>A sequencing techniques miCLIP or m<sup>6</sup>A-REF-seq. The three species datasets with different tissues based on m<sup>6</sup>A-REF-seq technique are downloaded from Dao's study in [42], and the dataset of human species based on miCLIP technique is obtained from Xing's study in [31]. The dataset of human species from Xing's study is denoted as Human51. The detailed information of the single-base resolution datasets is shown in Table 2. These m<sup>6</sup>A benchmark datasets have been used to test the m<sup>6</sup>A site predictive models [30–32,35,37,41–43,50–52].

### PMI and PJMI theory

Mutual information  $I(\mathbf{X}; \mathbf{Y})$  is to measure the correlation between two random variables  $\mathbf{X}(x_1, x_2, \dots, x_n)$  and  $\mathbf{Y}(y_1, y_2, \dots, y_n)$  [53–55]. It is calculated in (1) when  $\mathbf{X}$ ,  $\mathbf{Y}$  are discrete random variables.

**Table 2.** The detailed information of the single-base resolution benchmark datasets.

Species	Tissues	Name	Training dataset		Independent dataset		Identification method	Sequence length (nt)
			# positive	# negative	# positive	# negative		
Rat	Brain	RB	2352	2352	2351	2351	m <sup>6</sup> A-REF-seq	41
	Kidney	RK	3433	3433	3432	3432		
	Liver	RL	1762	1762	1762	1762		
Mouse	Brain	MB	8025	8025	8025	8025		
	Heart	MH	2201	2201	2200	2200		
	Kidney	MK	3953	3953	3952	3952		
	Liver	ML	4133	4133	4133	4133		
	Testis	MT	4707	4707	4706	4706		
Human	Brain	HB	4605	4605	4604	4604		
	Kidney	HK	4574	4574	4573	4573		
	Liver	HL	2634	2634	2634	2634		
	–	Human51	8366	8366	–	–		

$$I(\mathbf{X}; \mathbf{Y}) = \sum_{x_i \in \mathbf{X}} \sum_{y_j \in \mathbf{Y}} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad (1)$$

where  $p(x_i)$  and  $p(y_j)$  are the marginal probability distribution functions of  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively, and  $p(x_i, y_j)$  is the joint probability distribution function of  $\mathbf{X}$  and  $\mathbf{Y}$ .

The mutual information between random variables  $\mathbf{X}(x_1, x_2, \dots, x_n)$ ,  $\mathbf{Y}(y_1, y_2, \dots, y_n)$  and  $\mathbf{Z}(z_1, z_2, \dots, z_n)$  is calculated in (2).

$$I(\mathbf{X}; \mathbf{Y}, \mathbf{Z}) = \sum_{x_i \in \mathbf{X}} \sum_{y_j \in \mathbf{Y}} \sum_{z_k \in \mathbf{Z}} p(x_i, y_j, z_k) \log \frac{p(x_i, y_j, z_k)}{p(x_i)p(y_j)p(z_k)} \quad (2)$$

where  $p(x_i, y_j, z_k)$  is the joint probability distribution function of  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{Z}$ .

Pointwise mutual information  $PMI(x_i; y_j)$  is a special case of  $I(\mathbf{X}; \mathbf{Y})$ . It is to record the amount of uncertainty reduction in  $x_i$  when giving  $y_j$  in information theory. It is also used to measure the correlation between  $x_i$  and  $y_j$ . It is calculated in (3).

$$PMI(x_i; y_j) = \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad (3)$$

The domain of  $PMI(x_i; y_j)$  is  $(-\infty, +\infty)$ .  $PMI(x_i; y_j) = 0$  iff  $x_i$  and  $y_j$  are independent to each other. In addition,  $PMI(x_i; y_j)$  is symmetric, that is,  $PMI(x_i; y_j) = PMI(y_j; x_i)$ . Proof S1 proves this symmetry in supplementary material.

Inspired by pointwise mutual information, we propose and define the pointwise joint mutual information  $PJMI(x_i; y_j, z_k)$  in (4) to measure the amount of uncertainty reduction of  $x_i$  when giving  $y_j$  and  $z_k$ . It can also measure the correlation between  $x_i$ ,  $y_j$  and  $z_k$ . The  $x_i$ ,  $y_j$  and  $z_k$  are the specific events of random variables  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{Z}$ .

$$PJMI(x_i; y_j, z_k) = \log \frac{p(x_i, y_j, z_k)}{p(x_i)p(y_j)p(z_k)} \quad (4)$$

The domain of  $PJMI(x_i; y_j, z_k)$  is also  $(-\infty, +\infty)$ .  $PJMI(x_i; y_j, z_k) = 0$  iff  $x_i$ ,  $y_j$  and  $z_k$  are independent to each other.  $PJMI(x_i; y_j, z_k)$  is symmetric and is independent to the order of  $y_j$  and  $z_k$ , that is,  $PJMI(x_i; y_j, z_k) = PJMI(y_j, z_k; x_i)$  and  $PJMI(x_i; y_j, z_k) = PJMI(x_i; z_k, y_j)$  both hold. The proofs of these two properties of  $PJMI$  are Proof S2 and Proof S3 in supplementary material.

### Sequence encoding algorithms

Position-specific propensity has been applied to mine and identify the functional sites of biological sequences [29,31,36,56–58]. The basic principle is to calculate the frequencies of each nucleotide or amino acid of all sequences, and convert the input sequences into feature vectors using the difference between frequencies of positive and negative datasets. To extract features containing rich category information from RNA sequences using position-specific propensity, we propose bidirectional dinucleotide and trinucleotide position-specific propensities, and

feature-encoding algorithms PSP-PMI and PSP-PJMI by combing PMI and PJMI, respectively. To extract both local and global position-specific information of nucleotides from RNA sequences, we introduce parameters  $\alpha$  and  $\beta$  in PSP-PMI and PSP-PJMI, respectively, to represent the spacing between nucleotides.

We formalize the m<sup>6</sup>A benchmark datasets in Table 1 and 2 as following mathematics. Let  $\mathbf{D}$  represent the m<sup>6</sup>A dataset,  $\mathbf{D}^+$  the positive dataset, that is, the true m<sup>6</sup>A dataset, and  $\mathbf{D}^-$  the negative dataset, that is the non-m<sup>6</sup>A dataset. The relationship between  $\mathbf{D}$ ,  $\mathbf{D}^+$  and  $\mathbf{D}^-$  is  $\mathbf{D} = \mathbf{D}^+ \cup \mathbf{D}^-$ .

For RNA sequence  $\mathbf{R} = N_1 N_2 \dots N_i \dots N_l$  in  $\mathbf{D}$ , where  $l$  is its length, and  $N_i \in \{A, C, G, U\}$  is the nucleotide at position  $i$  ( $i = 1, \dots, l$ ). The position-specific occurrence frequency of four nucleotides at position  $i$  ( $1 \leq i \leq l$ ) in  $\mathbf{D}^+$  is denoted as vector  $\mathbf{f}_i^+ = (f_{A,i}^+, f_{C,i}^+, f_{G,i}^+, f_{U,i}^+)^T$ , where the elements of  $\mathbf{f}_i^+$  are the occurrence frequencies of nucleotides A, C, G and U at position  $i$  in  $\mathbf{D}^+$ , respectively. We define the nucleotide position-specific propensity matrix  $\mathbf{M}_S^+$  in (5) to represent the statistic information of four nucleotides in  $\mathbf{D}^+$ .

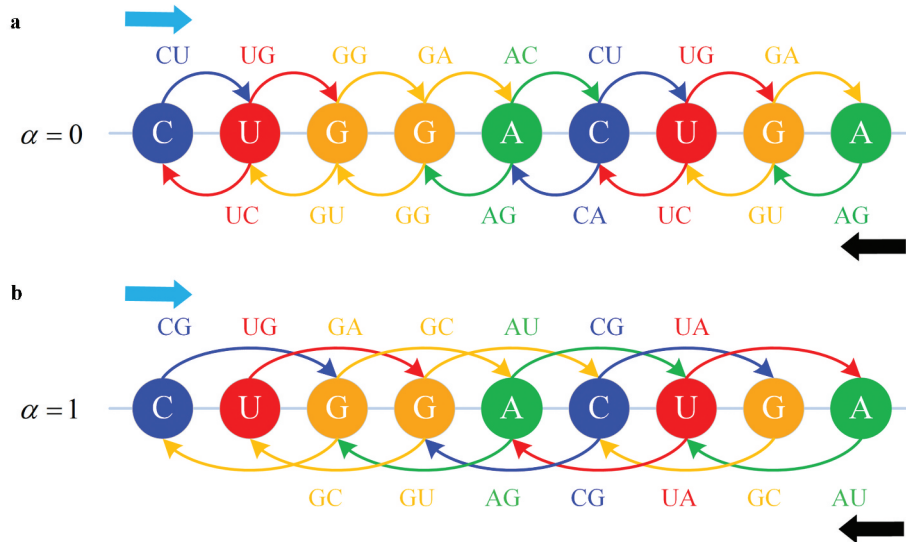
$$\mathbf{M}_S^+ = \begin{bmatrix} f_{A,1}^+ & f_{A,2}^+ & \dots & f_{A,i}^+ & \dots & f_{A,l}^+ \\ f_{C,1}^+ & f_{C,2}^+ & \dots & f_{C,i}^+ & \dots & f_{C,l}^+ \\ f_{G,1}^+ & f_{G,2}^+ & \dots & f_{G,i}^+ & \dots & f_{G,l}^+ \\ f_{U,1}^+ & f_{U,2}^+ & \dots & f_{U,i}^+ & \dots & f_{U,l}^+ \end{bmatrix} \quad (5)$$

### PSP-PMI algorithm

The bidirectional dinucleotide position-specific propensity is proposed in PSP-PMI algorithm in this paper, so as to extract more position-specific information of nucleotide from forward and backward directions. Furthermore, to extract both local and global category information from RNA sequences, we introduce parameter  $\alpha$  to represent the distance between two nucleotides in a pair of nucleotides. The  $\alpha = 0$  means that the two nucleotides are adjacent.

We take the RNA sequence with length  $l = 9$  to describe the idea of our bidirectional dinucleotide position-specific propensity in Figure 1, where Figure 1(a,b) correspond to  $\alpha = 0$  and  $\alpha = 1$ , respectively.

We first take  $\mathbf{D}^+$  into consideration. The frequency of the positional-specific propensity of forward dinucleotide at position  $i$  ( $1 \leq i \leq l - \alpha - 1$ ,  $0 \leq \alpha \leq (l - 3)/2$ ) is the vector  $\vec{\mathbf{f}}_i^+ = (\vec{f}_{AA,i}^+, \vec{f}_{AC,i}^+, \dots, \vec{f}_{UU,i}^+)^T$  of 16 elements. Its elements are frequencies of the dinucleotides of  $\{A, C, G, U\} \times \{A, C, G, U\}$ , respectively. Such as  $\vec{f}_{AA,i}^+$  in  $\vec{\mathbf{f}}_i^+$  represents the frequency of the dinucleotide pair AA in  $\mathbf{D}^+$ , where the nucleotides A and A appear at positions  $i$  and  $i + \alpha + 1$ , respectively. Then, we define the positional-specific propensity matrix  $\vec{\mathbf{M}}_d^+$  in (6) for forward dinucleotides to represent the statistic information of 16 types of dinucleotides in  $\mathbf{D}^+$ .



**Figure 1.** The bidirectional dinucleotide position-specific propensity. (a) for  $\alpha = 0$ , (b) for  $\alpha = 1$ .

$$\vec{\mathbf{M}}_d^+ = \begin{bmatrix} \vec{f}_{AA,1}^+ & \vec{f}_{AA,2}^+ & \cdots & \vec{f}_{AA,i}^+ & \cdots & \vec{f}_{AA,l-\alpha-1}^+ \\ \vec{f}_{AC,1}^+ & \vec{f}_{AC,2}^+ & \cdots & \vec{f}_{AC,i}^+ & \cdots & \vec{f}_{AC,l-\alpha-1}^+ \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \vec{f}_{UU,1}^+ & \vec{f}_{UU,2}^+ & \cdots & \vec{f}_{UU,i}^+ & \cdots & \vec{f}_{UU,l-\alpha-1}^+ \end{bmatrix} \quad (6)$$

Similarly, the frequency of the position-specific propensity of the backward dinucleotide at position  $i$  ( $\alpha + 2 \leq i \leq l$ ,  $0 \leq \alpha \leq (l-3)/2$ ) of  $\mathbf{D}^+$  can be represented as the vector  $\overleftarrow{\mathbf{f}}_i^+ = (\overleftarrow{f}_{AA,i}^+, \overleftarrow{f}_{AC,i}^+, \dots, \overleftarrow{f}_{UU,i}^+)^T$  comprising

16 elements. The element  $\overleftarrow{f}_{AA,i}^+$  of  $\overleftarrow{\mathbf{f}}_i^+$  denotes the frequency of the dinucleotide pair AA, where these nucleotides A and A appear at positions  $i$  and  $i - \alpha - 1$  of  $\mathbf{D}^+$ , respectively. We define  $\overleftarrow{\mathbf{M}}_d^+$  in (7) as the backward dinucleotide position-specific propensity matrix for  $\mathbf{D}^+$ .

$$\overleftarrow{\mathbf{M}}_d^+ = \begin{bmatrix} \overleftarrow{f}_{AA,\alpha+2}^+ & \overleftarrow{f}_{AA,\alpha+3}^+ & \cdots & \overleftarrow{f}_{AA,i}^+ & \cdots & \overleftarrow{f}_{AA,l}^+ \\ \overleftarrow{f}_{AC,\alpha+2}^+ & \overleftarrow{f}_{AC,\alpha+3}^+ & \cdots & \overleftarrow{f}_{AC,i}^+ & \cdots & \overleftarrow{f}_{AC,l}^+ \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \overleftarrow{f}_{UU,\alpha+2}^+ & \overleftarrow{f}_{UU,\alpha+3}^+ & \cdots & \overleftarrow{f}_{UU,i}^+ & \cdots & \overleftarrow{f}_{UU,l}^+ \end{bmatrix} \quad (7)$$

Assume that the nucleotides at positions  $i$  ( $\alpha + 2 \leq i \leq l - \alpha - 1$ ,  $0 \leq \alpha \leq (l-3)/2$ ),  $i - \alpha - 1$  and  $i + \alpha + 1$  are A, G and C, respectively, then the forward PMI value  $\overrightarrow{v}_i^+$  for the nucleotide at position  $i$  can be calculated in (8), and its backward PMI value  $\overleftarrow{v}_i^+$  is calculated in (9).

$$\overrightarrow{v}_i^+ = PMI(A; C) = \log \frac{p(A, C)}{p(A)p(C)} = \log \frac{\overrightarrow{f}_{AC,i}^+}{f_{A,i}^+ f_{C,i+\alpha+1}^+} \quad (8)$$

$$\overleftarrow{v}_i^+ = PMI(A; G) = \log \frac{p(A, G)}{p(A)p(G)} = \log \frac{\overleftarrow{f}_{AG,i}^+}{f_{A,i}^+ f_{G,i-\alpha-1}^+} \quad (9)$$

The  $\overrightarrow{f}_{AC,i}^+$  in (8) and  $\overleftarrow{f}_{AG,i}^+$  in (9) can be obtained from matrixes  $\vec{\mathbf{M}}_d^+$  and  $\overleftarrow{\mathbf{M}}_d^+$ , respectively. The  $f_{A,i}^+$ ,  $f_{G,i-\alpha-1}^+$  and  $f_{C,i+\alpha+1}^+$  come from matrix  $\mathbf{M}_S^+$ . They are, respectively, the occurrence probabilities of nucleotides A, G and C at positions  $i$ ,  $i - \alpha - 1$  and  $i + \alpha + 1$  of  $\mathbf{D}^+$ .

The PMI encoding value  $v_i^+$  for nucleotide at position  $i$  of RNA sequences is defined as the average of its forward PMI value  $\overrightarrow{v}_i^+$  and its backward PMI value  $\overleftarrow{v}_i^+$ , that is,  $v_i^+ = (\overrightarrow{v}_i^+ + \overleftarrow{v}_i^+)/2$ . Therefore, the PMI feature encoding vector  $\mathbf{V}^+$  of the RNA sequence with length  $l$  in  $\mathbf{D}^+$  is the feature vector  $\mathbf{V}^+ = (v_{\alpha+2}^+, v_{\alpha+3}^+, \dots, v_i^+, \dots, v_{l-\alpha-1}^+)$  containing  $l - 2\alpha - 2$  elements.

Similarly, we can obtain the nucleotide position-specific propensity matrix  $\mathbf{M}_S^-$ , position-specific propensity matrix  $\overleftarrow{\mathbf{M}}_d^-$  of forward dinucleotide and  $\overrightarrow{\mathbf{M}}_d^-$  of backward dinucleotide for  $\mathbf{D}^-$ . Then, we calculate its forward PMI value  $\overrightarrow{v}_i^-$ , backward PMI value  $\overleftarrow{v}_i^-$  and PMI encoding value  $v_i^-$  of the nucleotide at position  $i$  ( $\alpha + 2 \leq i \leq l - \alpha - 1$ ,  $0 \leq \alpha \leq (l-3)/2$ ). The PMI feature encoding vector  $\mathbf{V}^-$  for the RNA sequence with length  $l$  in  $\mathbf{D}^-$  is the feature vector  $\mathbf{V}^- = (v_{\alpha+2}^-, v_{\alpha+3}^-, \dots, v_i^-, \dots, v_{l-\alpha-1}^-)$  comprising  $l - 2\alpha - 2$  elements.

Finally, we encode the RNA sequence with length  $l$  into a feature vector  $\mathbf{V}$  containing  $l - 2\alpha - 2$  elements by feature vector  $\mathbf{V}^+$  minus  $\mathbf{V}^-$  as follows in (10).

$$\mathbf{V} = \mathbf{V}^+ - \mathbf{V}^- = (V_{\alpha+2}, V_{\alpha+3}, \dots, V_i, \dots, V_{l-\alpha-1}) \quad (10)$$

where  $V_i = v_i^+ - v_i^-$  and  $i \in [\alpha + 2, l - \alpha - 1]$ ,  $\alpha \in [0, (l-3)/2]$ .

It should be noted that the PMI in (8) and (9) is not the strict PMI in theory due to not satisfying the symmetry property of theoretic PMI. As we know that the nucleotides in RNA sequences have their own orders such that  $PMI(A;C) \neq PMI(C;A)$  when encoding RNA sequences. The Proof S4 in supplementary material addresses this fact.

We summarize our PSP-PMI in Figure S1 in supplemental material. We first partition dataset  $\mathbf{D}$  into positive dataset  $\mathbf{D}^+$  and negative dataset  $\mathbf{D}^-$ . Then, the mononucleotide position-specific propensity matrix and the bidirectional dinucleotide position-specific propensity matrices for  $\mathbf{D}^+$  and  $\mathbf{D}^-$  are calculated, respectively. The PMI values of nucleotides are calculated based on above six matrices. Finally, the RNA sequence with length  $l$  is encoded into a feature vector  $\mathbf{V}$  with  $l - 2\alpha - 2$  variables. We introduce parameter  $\alpha (0 \leq \alpha \leq (l-3)/2)$  to represent the distance between two nucleotides in a pair of nucleotides, so as to get both local and global categorical information from RNA sequences. The encoded feature vectors corresponding to different  $\alpha$  are concatenated to comprise one final feature vector containing  $(l-1)^2/4$  elements.

#### PSP-PJMI algorithm

To extract much more meaningful information from RNA sequences, we further propose PSP-PJMI feature encoding algorithm. PSP-PJMI proposes bidirectional trinucleotide position-specific propensity. It calculates the trinucleotide position-specific propensity matrices of forward and backward for  $\mathbf{D}^+$  and  $\mathbf{D}^-$ , respectively, and utilizes our proposed PJMI in (4) to encode RNA sequences. We introduce parameter  $\beta (0 \leq \beta \leq (l-5)/2)$  into bidirectional trinucleotide position-specific propensity to represent the distance from the nucleotide to its forward or backward successive dinucleotide. It is worth noting that  $\beta = 0$  means that the three nucleotides are successive. Here, we adopt the RNA

sequence of length  $l = 9$  to show our bidirectional trinucleotide position-specific propensity for  $\beta = 0$  and  $\beta = 1$  in Figure 2(a,b), respectively.

The forward trinucleotide position-specific propensity frequency for RNA sequences in  $\mathbf{D}^+$  at position  $i (1 \leq i \leq l - \beta - 2, 0 \leq \beta \leq (l-5)/2)$  can be expressed as vector  $\vec{\mathbf{f}}_i^+ = (\vec{f}_{AAA,i}^+, \vec{f}_{AAC,i}^+, \dots, \vec{f}_{UUU,i}^+)^T$  with 64 elements. It represents the frequencies of the trinucleotides of  $\{A, C, G, U\} \times \{A, C, G, U\} \times \{A, C, G, U\}$ . The element  $\vec{f}_{AAA,i}^+$  in  $\vec{\mathbf{f}}_i^+$  represents the frequency of the trinucleotide AAA. These nucleotides A, A and A are at positions  $i, i + \beta + 1$  and  $i + \beta + 2$  of  $\mathbf{D}^+$ , respectively. The trinucleotide position-specific propensity matrix  $\vec{\mathbf{M}}_i^+$  of forward direction of  $\mathbf{D}^+$  is shown in (11).

$$\vec{\mathbf{M}}_i^+ = \begin{bmatrix} \vec{f}_{AAA,1}^+ & \vec{f}_{AAA,2}^+ & \cdots & \vec{f}_{AAA,i}^+ & \cdots & \vec{f}_{AAA,l-\beta-2}^+ \\ \vec{f}_{AAC,1}^+ & \vec{f}_{AAC,2}^+ & \cdots & \vec{f}_{AAC,i}^+ & \cdots & \vec{f}_{AAC,l-\beta-2}^+ \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \vec{f}_{UUU,1}^+ & \vec{f}_{UUU,2}^+ & \cdots & \vec{f}_{UUU,i}^+ & \cdots & \vec{f}_{UUU,l-\beta-2}^+ \end{bmatrix} \quad (11)$$

The frequency of trinucleotide position-specific propensity of backward direction at position  $i (\beta + 3 \leq i \leq l, 0 \leq \beta \leq (l-5)/2)$  of  $\mathbf{D}^+$  is  $\overleftarrow{\mathbf{f}}_i^+ = (\overleftarrow{f}_{AAA,i}^+, \overleftarrow{f}_{AAC,i}^+, \dots, \overleftarrow{f}_{UUU,i}^+)^T$ . The first element  $\overleftarrow{f}_{AAA,i}^+$  of  $\overleftarrow{\mathbf{f}}_i^+$  represents the frequency of trinucleotide AAA, where the nucleotides A, A and A appear at positions  $i, i - \beta - 1$  and  $i - \beta - 2$  of  $\mathbf{D}^+$ , respectively. The backward trinucleotide position-specific propensity matrix  $\overleftarrow{\mathbf{M}}_i^+$  of  $\mathbf{D}^+$  is defined in (12).

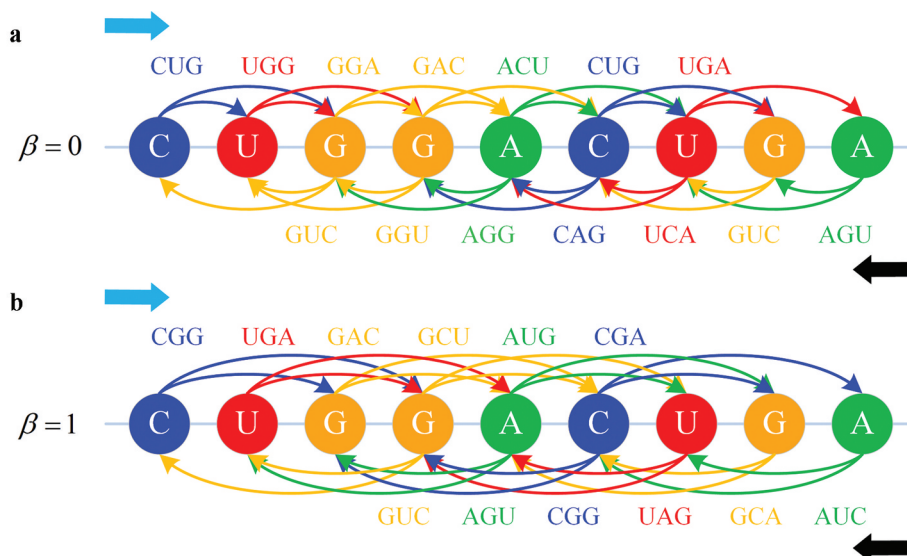


Figure 2. The bidirectional trinucleotide position-specific propensity. (a) for  $\beta = 0$ , (b) for  $\beta = 1$ .

$$\overleftarrow{\mathbf{M}}_t^+ = \begin{bmatrix} \overleftarrow{f}_{AAA,\beta+3}^+ & \overleftarrow{f}_{AAA,\beta+4}^+ & \cdots & \overleftarrow{f}_{AAA,i}^+ & \cdots & \overleftarrow{f}_{AAA,l}^+ \\ \overleftarrow{f}_{AAC,\beta+3}^+ & \overleftarrow{f}_{AAC,\beta+4}^+ & \cdots & \overleftarrow{f}_{AAC,i}^+ & \cdots & \overleftarrow{f}_{AAC,l}^+ \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \overleftarrow{f}_{UUU,\beta+3}^+ & \overleftarrow{f}_{UUU,\beta+4}^+ & \cdots & \overleftarrow{f}_{UUU,i}^+ & \cdots & \overleftarrow{f}_{UUU,l}^+ \end{bmatrix} \quad (12)$$

Assume that the nucleotide at position  $i(\beta + 3 \leq i \leq l - \beta - 2, 0 \leq \beta \leq (l - 5)/2)$  is A, the nucleotides at positions  $i - \beta - 1$  and  $i - \beta - 2$  are both G, and the nucleotides at positions  $i + \beta + 1$  and  $i + \beta + 2$  are C and U, respectively. The forward PJMI value  $\overrightarrow{v}_i^+$  and the backward PJMI value  $\overleftarrow{v}_i^+$  for the nucleotide at position  $i$  are calculated in (13) and (14), respectively.

$$\begin{aligned} \overrightarrow{v}_i^+ &= PJMI(A; C, U) = \log \frac{p(A, C, U)}{p(A)p(C, U)} \\ &= \log \frac{\overrightarrow{f}_{ACU,i}^+}{f_{A,i}^+ \overrightarrow{f}_{CU,i+\beta+1}^+} \end{aligned} \quad (13)$$

$$\begin{aligned} \overleftarrow{v}_i^+ &= PJMI(A; G, G) = \log \frac{p(A, G, G)}{p(A)p(G, G)} \\ &= \log \frac{\overleftarrow{f}_{AGG,i}^+}{f_{A,i}^+ \overleftarrow{f}_{GG,i-\beta-1}^+} \end{aligned} \quad (14)$$

The  $\overrightarrow{f}_{ACU,i}^+$  and  $\overrightarrow{f}_{CU,i+\beta+1}^+$  in (13) are obtained from  $\overrightarrow{\mathbf{M}}_t^+$  and  $\overrightarrow{\mathbf{M}}_d^+$ . The  $\overleftarrow{f}_{AGG,i}^+$  and  $\overleftarrow{f}_{GG,i-\beta-1}^+$  in (14) come from  $\overleftarrow{\mathbf{M}}_t^+$  and  $\overleftarrow{\mathbf{M}}_d^+$ . The  $f_{A,i}^+$  in (13) and (14) is from  $\mathbf{M}_S^+$ .

The PJMI value  $v_i^+$  of the nucleotide at position  $i$  of an RNA sequence is defined as  $v_i^+ = (\overrightarrow{v}_i^+ + \overleftarrow{v}_i^+)/2$ , that is, the average of the forward and backward PJMI values  $\overrightarrow{v}_i^+$  and  $\overleftarrow{v}_i^+$ . The PJMI feature vector  $\mathbf{V}^+$  of the RNA sequence with length  $l$  in  $\mathbf{D}^+$  is as  $\mathbf{V}^+ = (v_{\beta+3}^+, v_{\beta+4}^+, \dots, v_i^+, \dots, v_{l-\beta-2}^+)$ .

Similarly, we can calculate the forward trinucleotide position-specific propensity matrix  $\overrightarrow{\mathbf{M}}_t$  and the backward trinucleotide position-specific propensity matrix  $\overleftarrow{\mathbf{M}}_t^-$  of  $\mathbf{D}^-$ . Then the forward and backward PJMI values  $\overrightarrow{v}_i^-$  and  $\overleftarrow{v}_i^-$  are calculated using our PJMI in (4). The average of  $\overrightarrow{v}_i^-$  and  $\overleftarrow{v}_i^-$  is the PJMI encoding value  $v_i^-$  of the nucleotide at position  $i(\beta + 3 \leq i \leq l - \beta - 2, 0 \leq \beta \leq (l - 5)/2)$ . The PJMI feature encoding vector  $\mathbf{V}^-$  in  $\mathbf{D}^-$  is as  $\mathbf{V}^- = (v_{\beta+3}^-, v_{\beta+4}^-, \dots, v_i^-, \dots, v_{l-\beta-2}^-)$ .

Finally, we encode the given RNA sequence with length  $l$  of  $\mathbf{D}^-$  into a feature vector  $\mathbf{V}$  containing  $l - 2\beta - 4$  elements in (15) by  $\mathbf{V}^+$  minus  $\mathbf{V}^-$ .

$$\mathbf{V} = \mathbf{V}^+ - \mathbf{V}^- = (V_{\beta+3}, V_{\beta+4}, \dots, V_i, \dots, V_{l-\beta-2}) \quad (15)$$

where  $V_i = v_i^+ - v_i^-$ ,  $i \in [\beta + 3, l - \beta - 2]$ ,  $\beta \in [0, (l - 5)/2]$ .

It should be noted that the PJMI in (13) and (14) is not a strict theoretic PJMI because the nucleotides in an RNA sequence have their own orders. The PJMI in (13) and (14) does not satisfy the symmetry and the order of  $y_j$  and  $z_k$  being not irrelevant. The detail proof of this is shown in Proof S5 in supplementary material.

Figure S2 in supplemental material shows the schematic of our PSP-PJMI. It introduces bidirectional trinucleotide position-specific propensity matrixes for  $\mathbf{D}^+$  and  $\mathbf{D}^-$ . For the given RNA sequence with length  $l$ , the PJMI values  $v_i^+$  and  $v_i^-$  of nucleotide at position  $i(\beta + 3 \leq i \leq l - \beta - 2, 0 \leq \beta \leq (l - 5)/2)$  are calculated using our PJMI theory. The RNA sequence is converted into a feature vector  $\mathbf{V}$  comprising  $l - 2\beta - 4$  elements using PJMI feature vectors  $\mathbf{V}^+$  minus  $\mathbf{V}^-$ . We introduce parameter  $\beta$  for extracting both local and global categorical information from RNA sequences. The feature vectors of different  $\beta$  are concatenated into a  $(l - 3)^2/4$ -dimensional vector.

### Support vector machine

SVM is proposed by Cortes and Vapnik [59]. It maps non-linear separation samples in low-dimensional input space to high-dimensional feature space using kernel functions such that samples become linearly separable in it. SVM has got excellent learning and generalization capability, and has been widely used in complex disease diagnoses, biological function site predictions and other bioinformatics fields [60–64].

We adopt the LibSVM toolbox (<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>) developed by Chang et al [65] to train our m<sup>6</sup>A predictive model. The radial basis function (RBF) is as the kernel function. The grid search is adopted to find the optimal parameter pair  $(C, \gamma)$ , so as to get the best predictive model. The penalty factor  $C$  and the parameter  $\gamma$  of RBF are, respectively, as  $\log_2 C \in [-5, 15]$  and  $\log_2 \gamma \in [-15, 5]$  with both steps of 1.

### Metrics to evaluate m<sup>6</sup>A predictors

To test the power of our PSP-PMI and PSP-PJMI in extracting features with rich categorical information, we evaluate the performance of our M6A-BiNP predictor built on features extracted by PSP-PMI or PSP-PJMI in terms of very popular metrics, such as Accuracy (Acc), Sensitivity (Sn), Specificity (Sp) and Mathew's correlation coefficient (MCC) [31,38–44,52,62,66,67] and other two comprehensive indexes Area under the receiver operating characteristic curve (AUROC) and Area under the precision recall curve (AUPRC). AUROC and AUPRC are to value the performance of a binary classifier [28,41–43,52,68,69]. AUROC is the area under ROC (Receiver operating characteristic). ROC curve [70] is plotted by multiple pairs of false-positive rate (FPR) and true-positive rate (TPR) corresponding to different thresholds in a two-dimensional space with FPR as  $x$ -axis and TPR as  $y$ -axis. AUPRC is the area under the Precision-Recall (P-R) curve [71]. The P-R curve is better than ROC curve when dealing with imbalanced binary classification problems [72].

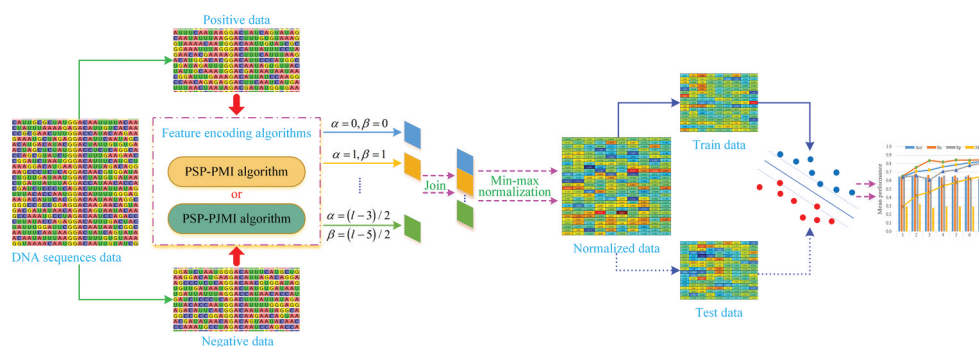


Figure 3. Framework of our M6A-BiNP predictor.

### Framework of M6A-BiNP

Figure 3 shows the framework of our M6A-BiNP predictor. We first encode RNA sequence into feature vectors using our PSP-PMI and PSP-PJMI algorithms, respectively, and concatenate feature vectors of different  $\alpha$  and  $\beta$ , respectively, to form the final feature vector and normalize it in min-max normalization. The 10-fold cross-validation experiments are done to train SVM classifiers to build our M6A-BiNP predictor. The results of 10-fold cross-validation experiments are used to evaluate our M6A-BiNP predictor.

### Results

Since the m<sup>6</sup>A sequencing data of non-single-base resolution were popular in available studies, we first test our PSP-PMI and PSP-PJMI on four species non-single-base resolution datasets in Table 1. Then, we carry out experiments to test them on single-base resolution datasets in Table 2. We compare the performance of our M6A-BiNP predictors with state-of-the-art predictive models on non-single-base resolution and single-base resolution datasets, respectively.

#### Performance evaluation on non-single-base resolution m<sup>6</sup>A datasets

##### Analysis to position-specific propensities of nucleotide

To reveal the nucleotide position-specific propensities of four species m<sup>6</sup>A datasets, we adopt Two Sample Logo [73] tool

(t-test,  $p = 0.05$ ) to calculate whether there is a significant difference in the distribution of nucleotides at each site in m<sup>6</sup>A dataset between its positive and negative samples and to visualize the significant distribution difference of the nucleotide at each site using a nucleotide symbol proportional to the significant difference. The result is shown in Figure 4.

The results in Figure 4 show that there is the consensus motif AC at positions 0 and 1 in *Arabidopsis thaliana* and *Musculus* sequences, and the consensus motif GAC at positions -1 to 1 in *Saccharomyces cerevisiae* sequences, and the consensus motif A at position 0 in *Homo sapiens* sequences. The nucleotide position-specific propensity exists in both upstream and downstream of the m<sup>6</sup>A site of the four species of m<sup>6</sup>A datasets, such as the nucleotide A is enriched while nucleotide U is depleted at both upstream and downstream of m<sup>6</sup>A site in *Saccharomyces cerevisiae*. The results in Figure 4 also show that the closer to the m<sup>6</sup>A site, the more significant difference exists in the nucleotide distribution, such as nucleotides A, G and U at positions -4, -2 and 4 are significantly enriched while nucleotide A at position -2 is significantly depleted in *Saccharomyces cerevisiae*; nucleotides A and C are significantly enriched while nucleotide U is significantly depleted at position 2 in *Arabidopsis thaliana*, *Musculus* and *Homo sapiens*, and nucleotides U at position -2 is significantly enriched in *Musculus* and *Homo sapiens*.

The above analyses discover that the nucleotide distributions are various in sequence positions in four specific species. This means the nucleotide position-specific propensity is the

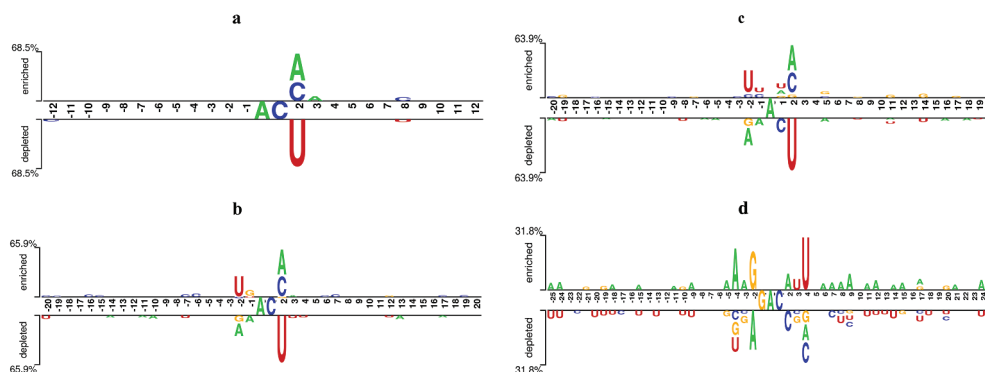
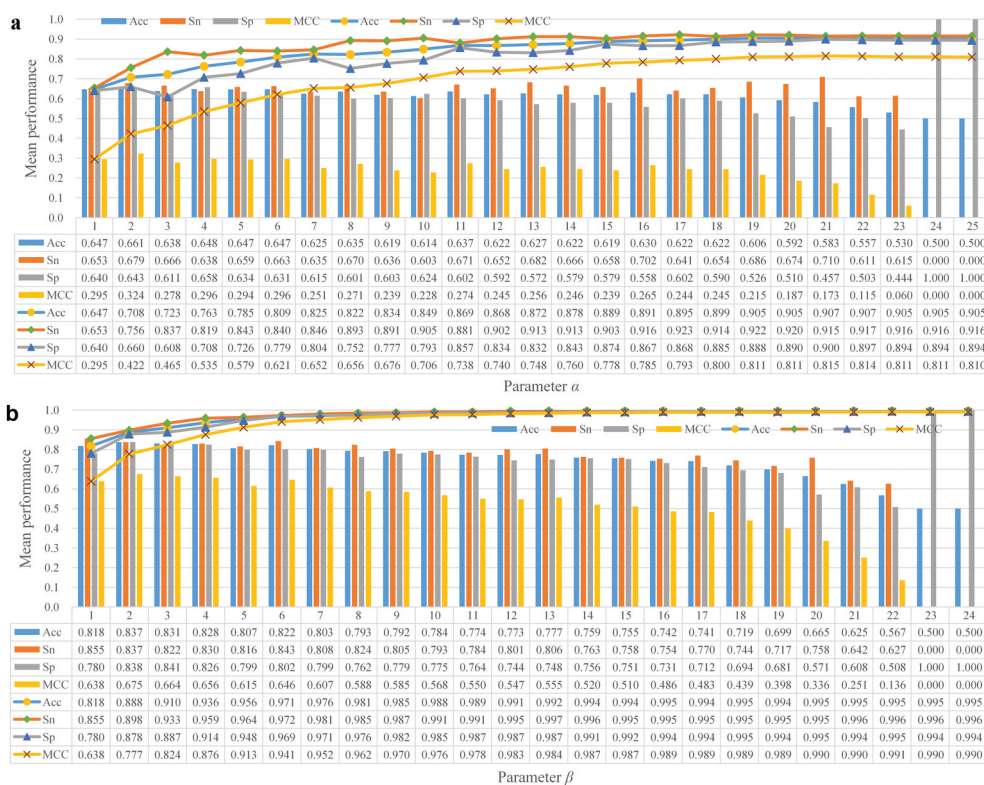


Figure 4. The nucleotide position-specific propensity. (a) *Arabidopsis thaliana*, (b) *Musculus*, (c) *Homo sapiens* and (d) *Saccharomyces cerevisiae*. The nucleotide A at position 0 is m<sup>6</sup>A site in positive sequence and non-m<sup>6</sup>A site in negative sequence. The nucleotide symbols in the upper of a picture indicate that the corresponding nucleotide is enriched in positive dataset. The nucleotide symbols in the lower indicate that the corresponding nucleotide is depleted in positive dataset. The nucleotide symbols in the middle indicate that the corresponding nucleotide is the consensus motif in both datasets.



**Figure 5.** The performance of SVM built on features encoded by (a) PSP-PMI and (b) PSP-PJMI via varying parameters  $\alpha$  and  $\beta$  on *Saccharomyces cerevisiae*. The bar chart represents the performance of the SVM built on features corresponding to different  $\alpha$  and  $\beta$ . The line chart represents the performance of the SVM built on concatenating features.

key predictive information to distinguish m<sup>6</sup>A from non-m<sup>6</sup>A samples. This guarantees the correctness of our PSP-PMI and PSP-PJMI, and the capability of features encoded by them in recognizing m<sup>6</sup>A sites.

### Effects of parameters $\alpha$ and $\beta$

To extract the features with local and global categorical information from RNA sequences, the parameters  $\alpha$  and  $\beta$  are introduced into PSP-PMI and PSP-PJMI, respectively, to represent the distance from the nucleotide to its forward or backward nucleotides in PSP-PMI, or to its forward or backward dinucleotides in PSP-PJMI. The features encoded by PSP-PMI or PSP-PJMI are various when parameters  $\alpha$  and  $\beta$  are different.

We take *Saccharomyces cerevisiae* in Table 1 as an example to investigate the impact of parameters  $\alpha$  and  $\beta$  in PSP-PMI and PSP-PJMI, respectively. The features corresponding to different  $\alpha$  and  $\beta$  are concatenated to train SVM models. Figure 5 displays the 10-fold cross-validation experimental results. The experimental results on *Arabidopsis thaliana*, *Musculus* and *Homo sapiens* are shown in Figure S3 in supplemental materials.

The results in Figure 5 show that the SVM model built on features encoded by our PSP-PMI or PSP-PJMI performs worse and worse when parameters  $\alpha$  and  $\beta$  going up, till the worst when  $\alpha$  is up to 24 in PSP-PMI and  $\beta$  up to 23 in PSP-PJMI. The Acc is 0.5 under the worst condition, which means it randomly classifies sequences of m<sup>6</sup>A and non-m<sup>6</sup>A in *Saccharomyces cerevisiae*. This is due to the distance between

two nucleotides in dinucleotides and trinucleotides becoming larger and larger as parameters  $\alpha$  and  $\beta$  going up, such that the number of nucleotides encoded by PSP-PMI and PSP-PJMI becomes less and the dimensionality of encoded features decrease, even the useful categorical information cannot be extracted from RNA sequences.

The results in Figure 5 also show that the performance of the SVM model goes up built on the concatenating features encoded by PSP-PMI or PSP-PJMI at various  $\alpha$  or  $\beta$ , respectively. This fact discloses that the features with local classification information are extracted when giving different values to parameters  $\alpha$  and  $\beta$ , and these features can be concatenated to comprise the features with global categorical information to maximize the performance of the SVM classifier. This further validates the correctness of our introducing parameters  $\alpha$  and  $\beta$  into PSP-PMI and PSP-PJMI, respectively.

Moreover, it can be seen from the results in Figure 5 that the SVM predictor built on the features encoded by PSP-PJMI performs better than that built on the features encoded by PSP-PMI on *Saccharomyces cerevisiae* for nearly 11%. This fact demonstrates that PSP-PJMI can extract features containing far more categorical information than PSP-PMI.

### Comparison with other feature encoding algorithms

To test the performance of our PSP-PMI and PSP-PJMI, we compare them with other seven feature encoding algorithms on four species non-single-base resolution datasets from Table 1, including position-specific nucleotide propensities (PSNP) [29], position-specific dinucleotide propensities (PSDP) [29],

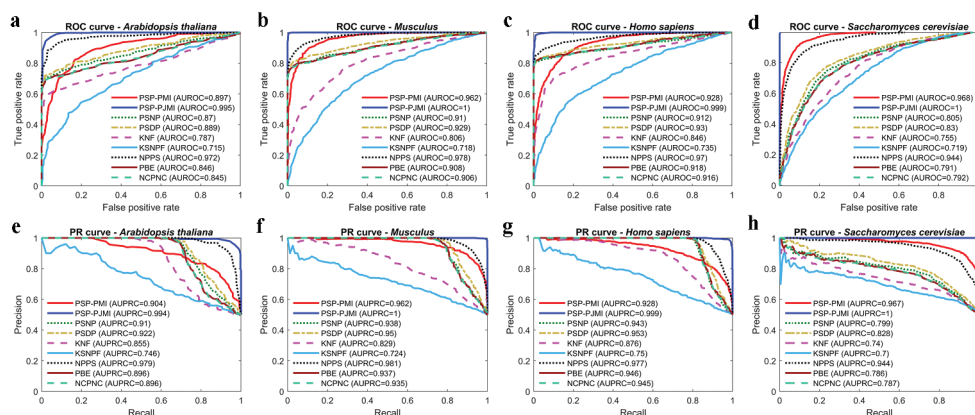


**Table 3.** Performance comparison between our PSP-PMI, PSP-PJMI and other seven feature representation algorithms on four non-single-base resolution datasets.

Algorithms	<i>Arabidopsis thaliana</i>				<i>Musculus</i>			
	Acc	Sn	Sp	MCC	Acc	Sn	Sp	MCC
PSP-PMI	0.830	0.825	0.835	0.661	0.901	0.909	0.894	0.804
PSP-PJMI	0.961 <sup>-</sup>	0.960 <sup>-</sup>	0.962 <sup>-</sup>	0.924 <sup>-</sup>	0.994 <sup>-</sup>	0.996 <sup>-</sup>	0.992 <sup>-</sup>	0.988 <sup>-</sup>
PSNP	0.840 <sup>=</sup>	0.683 <sup>+</sup>	0.998 <sup>-</sup>	0.717 <sup>-</sup>	0.856 <sup>+</sup>	0.712 <sup>+</sup>	1.000 <sup>-</sup>	0.745 <sup>+</sup>
PSDP	0.843 <sup>=</sup>	0.685 <sup>+</sup>	1.000 <sup>-</sup>	0.723 <sup>=</sup>	0.885 <sup>=</sup>	0.771 <sup>+</sup>	1.000 <sup>-</sup>	0.793 <sup>=</sup>
KNF	0.787 <sup>+</sup>	0.622 <sup>+</sup>	0.952 <sup>-</sup>	0.609 <sup>=</sup>	0.732 <sup>+</sup>	0.657 <sup>+</sup>	0.808 <sup>+</sup>	0.472 <sup>+</sup>
KSNPF	0.666 <sup>+</sup>	0.617 <sup>+</sup>	0.715 <sup>+</sup>	0.334 <sup>+</sup>	0.663 <sup>+</sup>	0.648 <sup>+</sup>	0.678 <sup>+</sup>	0.329 <sup>+</sup>
NPPS	0.925 <sup>-</sup>	0.891 <sup>-</sup>	0.959 <sup>-</sup>	0.854 <sup>-</sup>	0.918 <sup>-</sup>	0.855 <sup>=</sup>	0.981 <sup>-</sup>	0.844 <sup>=</sup>
PBE	0.840 <sup>=</sup>	0.683 <sup>+</sup>	0.998 <sup>-</sup>	0.717 <sup>-</sup>	0.885 <sup>=</sup>	0.771 <sup>+</sup>	1.000 <sup>-</sup>	0.793 <sup>=</sup>
NCPNC	0.843 <sup>=</sup>	0.685 <sup>+</sup>	1.000 <sup>-</sup>	0.723 <sup>=</sup>	0.881 <sup>=</sup>	0.763 <sup>+</sup>	1.000 <sup>-</sup>	0.786 <sup>=</sup>
+ / - / =	2/4/2	6/0/2	1/0/7	1/5/2	3/4/1	6/1/1	2/0/6	3/4/1

Algorithms	<i>Homo sapiens</i>				<i>Saccharomyces cerevisiae</i>			
	Acc	Sn	Sp	MCC	Acc	Sn	Sp	MCC
PSP-PMI	0.849	0.858	0.841	0.700	0.905	0.916	0.894	0.810
PSP-PJMI	0.986 <sup>-</sup>	0.982 <sup>-</sup>	0.989 <sup>-</sup>	0.972 <sup>-</sup>	0.995 <sup>-</sup>	0.996 <sup>-</sup>	0.994 <sup>-</sup>	0.990 <sup>-</sup>
PSNP	0.902 <sup>-</sup>	0.804 <sup>+</sup>	1.000 <sup>-</sup>	0.821 <sup>-</sup>	0.747 <sup>+</sup>	0.751 <sup>+</sup>	0.743 <sup>+</sup>	0.495 <sup>+</sup>
PSDP	0.903 <sup>-</sup>	0.806 <sup>+</sup>	1.000 <sup>-</sup>	0.822 <sup>-</sup>	0.766 <sup>+</sup>	0.764 <sup>+</sup>	0.769 <sup>+</sup>	0.534 <sup>+</sup>
KNF	0.797 <sup>+</sup>	0.695 <sup>+</sup>	0.899 <sup>-</sup>	0.607 <sup>+</sup>	0.692 <sup>+</sup>	0.741 <sup>+</sup>	0.643 <sup>+</sup>	0.387 <sup>+</sup>
KSNPF	0.680 <sup>+</sup>	0.612 <sup>+</sup>	0.749 <sup>+</sup>	0.365 <sup>+</sup>	0.651 <sup>+</sup>	0.712 <sup>+</sup>	0.591 <sup>+</sup>	0.307 <sup>+</sup>
NPPS	0.908 <sup>-</sup>	0.817 <sup>+</sup>	0.999 <sup>-</sup>	0.830 <sup>-</sup>	0.874 <sup>+</sup>	0.884 <sup>+</sup>	0.864 <sup>=</sup>	0.749 <sup>+</sup>
PBE	0.908 <sup>-</sup>	0.817 <sup>+</sup>	1.000 <sup>-</sup>	0.831 <sup>-</sup>	0.727 <sup>+</sup>	0.727 <sup>+</sup>	0.728 <sup>+</sup>	0.456 <sup>+</sup>
NCPNC	0.909 <sup>-</sup>	0.818 <sup>+</sup>	1.000 <sup>-</sup>	0.832 <sup>-</sup>	0.731 <sup>+</sup>	0.735 <sup>+</sup>	0.726 <sup>+</sup>	0.463 <sup>+</sup>
+ / - / =	2/0/6	7/0/1	1/0/7	2/0/6	7/0/1	7/0/1	6/1/1	7/0/1

**Figure 6.** ROC and P-R curves of nine feature representation algorithms on four datasets. (a) – (d) AUROC, (e) – (h) AUPRC.

K-nucleotide frequencies (KNF) [32], K-spaced nucleotide pair frequencies (KSNPF) [32], nucleotide pair position specificity (NPPS) [31], positional binary encoding (PBE) [74] and nucleotide chemical property and nucleotide composition (NCPNC) [27]. The parameter  $\xi$  of NPPS belongs to  $[0, l - 1]$ . The performance of each feature encoding algorithm is shown in Table 3 in terms of Acc, Sn, Sp and MCC of SVM classifier. Figure 6 displays the AUROC and AUPRC of each algorithm. The mean value of the optimal parameters of SVM classifier obtained by grid search for each feature encoding algorithm in 10-fold cross-validation experiments is shown in Table S1 in supplemental material.

Furthermore, we adopt paired two-tailed t-test method ( $p = 0.05$ ) to carry out the statistical test of PSP-PMI, PSP-PJMI and other seven feature encoding algorithms via their 10-fold cross-validation experimental results on four species of m<sup>6</sup>A datasets, so as to verify whether or not there is significant difference between these feature encoding algorithms. We adopt the symbols ‘+’, ‘=’ and ‘-’ to denote that PSP-PMI has a significant difference and is better than compared algorithm, has no significant difference, and has

significant difference and is worse than compared algorithm, respectively, at the 5% significant level. We count the number of symbols ‘+’, ‘=’ and ‘-’ of each algorithm in terms of Acc, Sn, Sp and MCC, so as to compare the performance of PSP-PMI with that of PSP-PJMI and other seven algorithms on four species m<sup>6</sup>A datasets. The statistic test results are also shown in Table 3.

The results in Table 3 show that PSP-PJMI defeats seven compared algorithms and PSP-PMI in terms of Acc, Sn and MCC, especially on *Saccharomyces cerevisiae*, it is superior to any compared feature encoding algorithms including PSP-PMI in terms of Acc, Sn, Sp and MCC. Our PSP-PMI outperforms seven compared feature encoding algorithms on *Saccharomyces cerevisiae*. It is statistically better than or equal to other seven feature encoding algorithms on *Musculus*, and other six feature encoding algorithms except for NPPS on *Arabidopsis thaliana* in terms of Acc, Sn and MCC. It performs poor on *Homo sapiens* with only superior to KSNPF in terms of four metrics and to KNF in terms of Acc, Sn and MCC, but it can defeat seven compared feature encoding algorithms in terms of Sn.

The paired two-tailed t-test results in Table 3 show us that PSP-PMI performs best on *Saccharomyces cerevisiae* when comparing to seven compared feature encoding algorithms, and worst on *Homo sapiens*. Its capability to extract features to identify non-m<sup>6</sup>A sites on *Arabidopsis Thaliana* and *Homo Sapiens* is inferior to compared feature encoding algorithms except for KSNPF, so is on *Musculus* except for KSNPF and KNF.

We are more interested in the capability to identify the true m<sup>6</sup>A sites, that is, the bigger the Sn, the better is the algorithm. Therefore, although our PSP-PMI is inferior to our PSP-PJMI algorithm, it can extract more useful features from RNA sequences compared to other seven algorithms.

The results in Figure 6 show that PSP-PJMI is far better than PSP-PMI and other seven compared algorithms. Its AUROC and AUPRC obtain the maximal value 1 on *Musculus* and *Saccharomyces cerevisiae*. Our PSP-PMI defeats other seven feature encoding algorithms on *Saccharomyces cerevisiae* in terms of AUROC and AUPRC. However, it is inferior to NPPS in terms of AUROC on *Arabidopsis thaliana*,

*Musculus* and *Homo sapiens*, and also inferior to PSDP on *Homo sapiens*. The results in Figure 6 about each algorithm's AUPRC show that our PSP-PMI is inferior to NPPS, PSNP and PSDP on *Arabidopsis thaliana*, and inferior to NPPS on *Musculus*. Its performance is poor on *Homo sapiens* in terms of AUPRC only superior to KNF and KSNPF.

From all above analyses, we can say that PSP-PJMI is definitely superior to PSP-PMI and other seven compared feature encoding algorithms. It can extract features with strong categorical discernibility from RNA sequences. Although PSP-PMI is inferior to PSP-PJMI, it is superior to other seven compared algorithms in encoding features to identify m<sup>6</sup>A sites.

### Comparison with the state-of-the-art predictors

To comprehensively compare the performance of our M6A-BiNP predictors based on the features encoded by proposed PSP-PMI or PSP-PJMI, we try our best to collect the state-of-the-art predictive models based on traditional machine learning algorithms and deep learning framework on four species

**Table 4.** Performance comparison between our M6A-BiNP and the state-of-the-art predictors on four species m<sup>6</sup>A benchmark datasets.

Datasets	Predictors	Classifiers	Experiment methods	Evaluation criteria					
				Acc	Sn	Sp	MCC	AUROC	AUPRC
<i>Arabidopsis thaliana</i>	M6ATH [27]	SVM	jackknife	0.844	0.688	<b>1.000</b>	0.720	0.846	0.870
	RAM-NPPS [31]	SVM	jackknife	0.895	0.873	0.916	0.790	–	–
	m6A-word2vec [75]	CNN	10-fold cross-validation	0.905	0.950	0.859	0.810	0.928	–
	M6A-BiNP	SVM (PSP-PMI)	10-fold cross-validation	0.830	0.825	0.835	0.661	0.897	0.904
<i>Musculus</i>		SVM (PSP-PJMI)	10-fold cross-validation	<b>0.961</b>	<b>0.960</b>	0.962	<b>0.924</b>	<b>0.995</b>	<b>0.994</b>
	iN6-Methyl [37]	CNN	10-fold cross-validation	0.895	0.789	<b>1.000</b>	0.808	0.913	–
	M6AMRFS [34]	XGBoost	10-fold cross-validation	0.793	0.828	0.758	0.588	–	–
	MethylRNA [50]	SVM	jackknife	0.884	0.778	<b>1.000</b>	–	–	–
	iMRM [41]	XGboost	jackknife	0.890	0.783	0.996	0.779	0.820	–
	m6A-NeuralTool [76]	CNN	10-fold cross-validation	0.958	0.915	<b>1.000</b>	0.912	0.960	–
	pm6A-CNN [77]	CNN	10-fold cross-validation	0.938	0.904	0.972	0.880	0.970	–
	Second order-MM [78]	Markov model	10-fold cross-validation	0.883	0.875	0.889	0.775	–	–
	SRAMP [28]	RF	10-fold cross-validation	0.889	0.778	<b>1.000</b>	0.798	–	–
	M6A-BiNP	SVM (PSP-PMI)	10-fold cross-validation	0.901	0.909	0.894	0.804	0.962	0.962
		SVM (PSP-PJMI)	10-fold cross-validation	<b>0.994</b>	<b>0.996</b>	0.992	<b>0.988</b>	<b>1.000</b>	<b>1.000</b>
	<i>Homo sapiens</i>	M6AMRFS [34]	XGBoost	10-fold cross-validation	0.910	0.820	<b>1.000</b>	0.834	–
MethylRNA [50]		SVM	jackknife	0.904	0.817	0.991	–	–	–
iRNA-Methyl [25]		SVM	jackknife	0.672	0.575	0.769	–	–	–
iN6-Methyl [37]		CNN	10-fold cross-validation	0.911	0.821	<b>1.000</b>	0.835	0.903	–
iMRM [41]		XGboost	jackknife	0.910	0.825	0.996	0.820	0.940	–
m6A-NeuralTool [76]		CNN	10-fold cross-validation	0.960	0.920	<b>1.000</b>	0.882	0.950	–
pm6A-CNN [77]		CNN	10-fold cross-validation	0.936	0.886	0.986	0.878	0.960	–
m6A-word2vec [75]		CNN	10-fold cross-validation	0.927	0.981	0.882	0.850	0.951	–
Second order-MM [78]		Markov model	10-fold cross-validation	0.906	0.865	0.947	0.814	–	–
SRAMP [28]		RF	10-fold cross-validation	0.898	0.797	<b>1.000</b>	0.814	–	–
M6A-BiNP		SVM (PSP-PMI)	10-fold cross-validation	0.849	0.858	0.841	0.700	0.928	0.928
		SVM (PSP-PJMI)	10-fold cross-validation	<b>0.986</b>	<b>0.982</b>	0.989	<b>0.972</b>	<b>0.999</b>	<b>0.999</b>
<i>Saccharomyces cerevisiae</i>	M6APredict-EL [35]	EL	10-fold cross-validation	0.808	0.807	0.810	0.620	0.902	0.901
	RAM-NPPS [31]	SVM	10-fold cross-validation	0.799	0.790	0.808	0.598	–	–
	M6AMRFS [34]	XGBoost	10-fold cross-validation	0.743	0.752	0.733	0.485	–	–
	M6A-HPCS [79]	SVM	jackknife	0.724	0.774	0.674	0.450	0.782	–
	iRNA-Methyl [25]	SVM	jackknife	0.656	0.706	0.606	0.290	0.705	–
	pRNA-PC [30]	SVM	jackknife	0.697	0.697	0.698	0.400	0.763	–
	RAM-ESVM [80]	SVM	jackknife	0.748	0.789	0.778	0.570	–	–
	BERMP [81]	DL and RF	independent	0.713	0.730	0.696	0.430	0.800	–
	iMethyl-STTNC [52]	SVM	10-fold cross-validation	0.698	0.703	0.682	0.380	–	–
	iN6-Methyl [37]	CNN	10-fold cross-validation	0.754	0.762	0.746	0.508	0.803	–
	M6A-PXGB [51]	XGBoost	10-fold cross-validation	0.771	0.764	0.760	0.535	0.839	–
	DeepM6APred [36]	SVM	10-fold cross-validation	0.805	0.795	0.815	0.610	–	–
	iMRM [41]	XGboost	jackknife	0.778	0.770	0.785	0.555	0.85	–
	m6A-NeuralTool [76]	CNN	10-fold cross-validation	0.790	0.783	0.796	0.614	–	–
	pm6A-CNN [77]	CNN	10-fold cross-validation	0.850	0.846	0.855	0.703	0.920	–
	m6A-word2vec [75]	CNN	10-fold cross-validation	0.832	0.865	0.799	0.660	0.901	–
	iMethyl-deep [82]	CNN	10-fold cross-validation	0.892	0.885	0.899	0.780	0.931	–
	DNN-m6A [83]	DNN	10-fold cross-validation	0.785	0.787	0.783	0.571	–	–
	M6A-BiNP	SVM (PSP-PMI)	10-fold cross-validation	0.905	0.916	0.894	0.810	0.968	0.967
		SVM (PSP-PJMI)	10-fold cross-validation	<b>0.995</b>	<b>0.996</b>	<b>0.994</b>	<b>0.990</b>	<b>1.000</b>	<b>1.000</b>

benchmark m<sup>6</sup>A datasets, and compare them to our M6A-BiNP predictors. The performances of the state-of-the-art predictive models and M6A-BiNP predictors on four datasets are shown in Table 4. The bold fonts mean the best results. The CNN, EL, RF and DNN represent convolutional neural networks, ensemble learning, random forest and deep neural network, respectively.

The results in Table 4 show that our M6A-BiNP predictor built on features encoded by our PSP-PJMI is far superior to the state-of-the-art predictors on four species of m<sup>6</sup>A benchmark datasets in terms of Acc, Sn, MCC, AUROC and AUPRC, especially on *Saccharomyces Cerevisiae*, it is superior to all available predictors in terms of all metrics including Sp. Although it is not the best one on *Arabidopsis Thaliana*, *Musculus* and *Homo Sapiens* in terms of Sp, it is the best in terms of Acc, Sn, Mcc, AUROC and AUPRC. The accuracy of this M6A-BiNP is higher 6.19%, 3.76%, 2.71% and 11.55% than that of the best m6A-word2vec, m6A-NeuralTool, m6A-NeuralTool and iMethyl-deep predictors on *Arabidopsis Thaliana*, *Musculus*, *Homo Sapiens* and *Saccharomyces Cerevisiae*, respectively.

The results in Table 4 also show that M6A-BiNP predictor built on features encoded by PSP-PMI defeats all the state-of-the-art models on *Saccharomyces cerevisiae* in terms of Acc, Sn, MCC, AUROC and AUPRC, except for a little inferior to iMethyl-deep in terms of Sp. However, this M6A-BiNP predictor does not perform well on *Arabidopsis Thaliana*, *Musculus* and *Homo Sapiens*. It is just superior to iRNA-Methyl and inferior to other predictive models on *Homo Sapiens* in terms of Acc. It is the worst one on *Arabidopsis thaliana* in terms of Acc, Sp and MCC. It can defeat compared models except for m6A-NeuralTool and pm6A-CNN on *Musculus* in terms of Acc. Although PSP-PMI-based M6A-BiNP predictor is not as good as the one based on PSP-PJMI, it is still a comparatively good predictive model in identifying m<sup>6</sup>A sites.

### Performance evaluation on the single-base resolution m<sup>6</sup>A datasets

This section will test the performance of our PSP-PMI and PSP-PJMI feature encoding algorithms and M6A-BiNP models based on PSP-PMI and PSP-PJMI respectively, on the single-base resolution datasets in Table 2. We first carry out experiments on the Human51 data which is based on miCLIP technique, then on the other single-base resolution data based on m<sup>6</sup>A-REF-seq technique in Table 2.

### Performance comparison with RAM-NPPS model on Human51 dataset

The reference [31] only provided the experimental results in terms of AUROC and AUPRC of RAM-NPPS model on Human51. To obtain the results of this model in terms of other evaluation metrics, we re-implement the RAM-NPPS prediction model. The experimental results of M6A-BiNP and RAM-NPPS models on Human51 are shown in Table 5. The best value of each criterion is displayed in bold fonts.

As can be seen from the experimental results in Table 5 that our M6A-BiNP model based on PSP-PJMI algorithm

**Table 5.** Comparison of M6A-BiNP models and RAM-NPPS on Human51 dataset.

Model	Acc	Sn	Sp	MCC	AUROC	AUPRC
M6A-BiNP (PSP-PMI)	0.711	0.733	0.689	0.423	0.782	0.772
M6A-BiNP (PSP-PJMI)	<b>0.851</b>	<b>0.856</b>	<b>0.845</b>	<b>0.702</b>	<b>0.927</b>	<b>0.927</b>
RAM-NPPS	0.722	0.733	0.710	0.443	0.794	0.785

obtained the best performance on the single-base resolution Human51 dataset no matter using any evaluation criterion. Although the performance of the M6A-BiNP model based on our PSP-PMI algorithm is the worst among three compared models in most cases on Human51 dataset, it obtains very similar performance as RAM-NPPS model.

### Performance comparison with existing models on datasets of Human, Mouse and Rat

The base-resolution data based on m<sup>6</sup>A-REF-seq technique in Table 2 contain training and independent data. We trained the M6A-BiNP models on the 11 training datasets using 10-fold cross-validation experiments, and compared the performance of our M6A-BiNP models to that of predictors iRNA-m6A [43], im6A-TS-CNN [42] and DNN-m6A [83]. The results are shown in Table 6. After that the M6A-BiNP models are tested on independent datasets and compared with the models of iRNA-m6A, im6A-TS-CNN and DNN-m6A. The results are shown in Table 7. The best values of each criterion in Tables 6 and 7 are shown in bold fonts.

The results in Table 6 show that our M6A-BiNP model based on PSP-PJMI algorithm performs best in most cases except for on MB dataset. Its performance is better than that of SVM-based predictor iRNA-m6A and deep learning framework-based predictors im6A-TS-CNN and DNN-m6A in most cases. However, our M6A-BiNP model based on PSP-PMI algorithm is not as good as the one based on PSP-PJMI algorithm on these 11 training datasets. It only outperforms the iRNA-m6A, im6A-TS-CNN and DNN-m6A predictors in MH, RB and RL datasets.

The experimental results in Table 7 show that our M6A-BiNP model based on PSP-PJMI algorithm is superior to the models of iRNA-m6A, im6A-TS-CNN, DNN-m6A and our M6A-BiNP model based on PSP-PMI algorithm on 8 among 11 datasets in terms of most criteria. Its performance on MB dataset is not good, nor on HK and RK datasets.

The above analyses shown that our proposed M6A-BiNP model based on the PSP-PJMI algorithm has obtained the best performance on the datasets based on two different base-resolution techniques. Its performance is better than the m<sup>6</sup>A sites prediction models based on deep learning framework. This not only demonstrates the effectiveness of the PSP-PJMI feature encoding algorithm proposed in this paper but also proves the correctness of our using SVM classifier to build the prediction model.

Although our M6A-BiNP model based on PSP-PMI algorithm does not perform as good as our M6A-BiNP model based on our proposed PSP-PJMI algorithm, it is still good enough and better than other compared prediction models in most cases, which shows that our proposed PSP-PMI is a useful feature encoding algorithm. Furthermore, PSP-PMI

**Table 6.** Performance comparison of our M6A-BiNP with iRNA-m6A, im6A-TS-CNN and DNN-m6A models on the training datasets.

Species	Tissues	Name	Methods	Acc	Sn	Sp	MCC	AUROC
Human	Brain	HB	M6A-BiNP (PSP-PMI)	0.720	0.711	0.729	0.440	0.793
			M6A-BiNP (PSP-PJMI)	<b>0.820</b>	<b>0.810</b>	<b>0.831</b>	<b>0.641</b>	<b>0.900</b>
			iRNA-m6A	0.713	0.748	0.662	0.410	0.776
			im6A-TS-CNN	0.725	0.754	0.697	0.452	0.803
			DNN-m6A	0.738	0.785	0.691	0.480	0.817
	Kidney	HK	M6A-BiNP (PSP-PMI)	0.746	0.755	0.738	0.493	0.832
			M6A-BiNP (PSP-PJMI)	<b>0.816</b>	0.809	<b>0.823</b>	<b>0.633</b>	<b>0.896</b>
			iRNA-m6A	0.790	0.809	0.763	0.570	0.863
			im6A-TS-CNN	0.800	0.817	0.783	0.601	0.878
			DNN-m6A	0.805	<b>0.836</b>	0.774	0.610	0.884
	Liver	HL	M6A-BiNP (PSP-PMI)	0.775	0.769	0.781	0.550	0.856
			M6A-BiNP (PSP-PJMI)	<b>0.874</b>	<b>0.874</b>	<b>0.874</b>	<b>0.748</b>	<b>0.951</b>
			iRNA-m6A	0.801	0.813	0.781	0.590	0.874
			im6A-TS-CNN	0.802	0.797	0.799	0.599	0.881
			DNN-m6A	0.813	0.822	0.804	0.630	0.891
Mouse	Brain	MB	M6A-BiNP (PSP-PMI)	0.732	0.744	0.720	0.464	0.818
			M6A-BiNP (PSP-PJMI)	0.772	0.768	<b>0.775</b>	0.544	0.858
			iRNA-m6A	0.788	0.793	0.769	0.580	0.870
			im6A-TS-CNN	0.787	0.815	0.759	0.575	0.871
			DNN-m6A	<b>0.794</b>	<b>0.818</b>	0.770	<b>0.590</b>	<b>0.878</b>
	Heart	MH	M6A-BiNP (PSP-PMI)	0.794	0.807	0.780	0.588	0.880
			M6A-BiNP (PSP-PJMI)	<b>0.937</b>	<b>0.937</b>	<b>0.936</b>	<b>0.873</b>	<b>0.984</b>
			iRNA-m6A	0.728	0.752	0.690	0.440	0.795
			im6A-TS-CNN	0.730	0.784	0.676	0.463	0.812
			DNN-m6A	0.762	0.775	0.748	0.520	0.844
	Kidney	MK	M6A-BiNP (PSP-PMI)	0.775	0.795	0.754	0.550	0.859
			M6A-BiNP (PSP-PJMI)	<b>0.848</b>	<b>0.842</b>	<b>0.853</b>	<b>0.696</b>	<b>0.929</b>
			iRNA-m6A	0.800	0.826	0.773	0.600	0.873
			im6A-TS-CNN	0.805	0.799	0.810	0.609	0.884
			DNN-m6A	0.820	0.832	0.807	0.640	0.895
Liver	ML	M6A-BiNP (PSP-PMI)	0.728	0.755	0.701	0.456	0.813	
		M6A-BiNP (PSP-PJMI)	<b>0.851</b>	<b>0.856</b>	<b>0.845</b>	<b>0.702</b>	<b>0.927</b>	
		iRNA-m6A	0.706	0.749	0.656	0.410	0.774	
		im6A-TS-CNN	0.713	0.724	0.702	0.429	0.795	
		DNN-m6A	0.736	0.776	0.696	0.470	0.814	
Testis	MT	M6A-BiNP (PSP-PMI)	0.743	0.777	0.709	0.487	0.824	
		M6A-BiNP (PSP-PJMI)	<b>0.850</b>	<b>0.848</b>	<b>0.852</b>	<b>0.701</b>	<b>0.930</b>	
		iRNA-m6A	0.744	0.781	0.700	0.480	0.816	
		im6A-TS-CNN	0.754	0.752	0.756	0.509	0.838	
		DNN-m6A	0.766	0.810	0.723	0.530	0.849	
Rat	Brain	RB	M6A-BiNP (PSP-PMI)	0.785	0.784	0.786	0.570	0.869
			M6A-BiNP (PSP-PJMI)	<b>0.926</b>	<b>0.918</b>	<b>0.935</b>	<b>0.853</b>	<b>0.980</b>
			iRNA-m6A	0.760	0.770	0.735	0.500	0.828
			im6A-TS-CNN	0.766	0.790	0.742	0.538	0.847
			DNN-m6A	0.783	0.791	0.775	0.570	0.868
	Kidney	RK	M6A-BiNP (PSP-PMI)	0.781	0.792	0.771	0.563	0.868
			M6A-BiNP (PSP-PJMI)	<b>0.875</b>	<b>0.867</b>	<b>0.883</b>	<b>0.750</b>	<b>0.946</b>
			iRNA-m6A	0.818	0.825	0.801	0.630	0.888
			im6A-TS-CNN	0.825	0.842	0.808	0.650	0.902
			DNN-m6A	0.834	0.843	0.825	0.670	0.910
	Liver	RL	M6A-BiNP (PSP-PMI)	0.826	0.826	0.826	0.653	0.912
			M6A-BiNP (PSP-PJMI)	<b>0.950</b>	<b>0.946</b>	<b>0.954</b>	<b>0.900</b>	<b>0.989</b>
			iRNA-m6A	0.809	0.831	0.763	0.600	0.877
			im6A-TS-CNN	0.806	0.816	0.796	0.613	0.883
			DNN-m6A	0.826	0.842	0.811	0.650	0.899

algorithm can be combined with other encoding algorithms to enhance its capability to extract informative features.

## Conclusions

Two feature encoding algorithms named PSP-PMI and PSP-PJMI are proposed in this paper to extract features with more nucleotide position information and strong categorical information from RNA sequences. The bidirectional dinucleotide and trinucleotide position-specific propensities are proposed in PSP-PMI and PSP-PJMI based on PMI and our PJMI theories, respectively. The parameters  $\alpha$  and  $\beta$  are introduced to represent the distance between nucleotides in a pair of nucleotides in PSP-PMI, and the distance from the nucleotide to its forward or backward consecutive

dinucleotide in PSP-PJMI, respectively. The features corresponding to different  $\alpha$  and  $\beta$  are, respectively, concatenated to comprise the high-dimensional features containing both local and global categorical information in PSP-PMI and PSP-PJMI.

The SVM-based M6A-BiNP predictors are built on features encoded by PSP-PMI or PSP-PJMI. The 10-fold cross-validation experimental results on the m<sup>6</sup>A benchmark datasets including four species non-single-base resolution datasets and three species single-base resolution datasets using two different m<sup>6</sup>A sites detection techniques demonstrate that parameters  $\alpha$  and  $\beta$  in PSP-PMI and PSP-PJMI are helpful to extract features with much more categorical information from RNA sequences. There is few redundant feature existing in

**Table 7.** Performance comparison of our M6A-BiNP with iRNA-m6A, im6A-TS-CNN and DNN-m6A models on the independent datasets.

Species	Tissues	Name	Methods	Acc	Sn	Sp	MCC	AUROC
Human	Brain	HB	M6A-BiNP (PSP-PMI)	0.708	0.746	0.670	0.417	0.779
			M6A-BiNP (PSP-PJMI)	<b>0.767</b>	0.580	<b>0.954</b>	<b>0.576</b>	<b>0.894</b>
			iRNA-m6A	0.711	0.695	0.730	0.420	0.785
			im6A-TS-CNN	0.727	<b>0.752</b>	0.702	0.454	0.806
			DNN-m6A	0.733	0.750	0.715	0.470	0.815
	Kidney	HK	M6A-BiNP (PSP-PMI)	0.694	0.883	0.506	0.419	0.807
			M6A-BiNP (PSP-PJMI)	0.682	<b>0.964</b>	0.400	0.441	<b>0.879</b>
			iRNA-m6A	0.778	0.771	0.784	0.560	0.857
			im6A-TS-CNN	0.792	0.800	<b>0.785</b>	0.585	0.873
			DNN-m6A	<b>0.799</b>	0.832	0.766	<b>0.600</b>	0.878
	Liver	HL	M6A-BiNP (PSP-PMI)	0.739	0.650	<b>0.829</b>	0.487	0.824
			M6A-BiNP (PSP-PJMI)	<b>0.862</b>	<b>0.920</b>	0.805	<b>0.730</b>	<b>0.948</b>
			iRNA-m6A	0.790	0.782	0.799	0.580	0.868
			im6A-TS-CNN	0.799	0.848	0.750	0.601	0.881
			DNN-m6A	0.810	0.818	0.801	0.620	0.885
Mouse	Brain	MB	M6A-BiNP (PSP-PMI)	0.719	0.596	<b>0.842</b>	0.451	0.815
			M6A-BiNP (PSP-PJMI)	0.756	0.838	0.674	0.518	0.849
			iRNA-m6A	0.783	0.772	0.794	0.570	0.861
			im6A-TS-CNN	0.785	<b>0.862</b>	0.707	<b>0.577</b>	0.872
			DNN-m6A	<b>0.786</b>	0.751	0.821	0.570	<b>0.876</b>
	Heart	MH	M6A-BiNP (PSP-PMI)	0.774	0.651	0.898	0.566	0.881
			M6A-BiNP (PSP-PJMI)	<b>0.838</b>	0.681	<b>0.996</b>	<b>0.712</b>	<b>0.983</b>
			iRNA-m6A	0.713	0.705	0.721	0.430	0.788
			im6A-TS-CNN	0.736	0.758	0.714	0.472	0.816
			DNN-m6A	0.751	<b>0.773</b>	0.730	0.500	0.834
	Kidney	MK	M6A-BiNP (PSP-PMI)	0.765	0.707	<b>0.822</b>	0.533	0.854
			M6A-BiNP (PSP-PJMI)	<b>0.832</b>	<b>0.906</b>	0.758	<b>0.672</b>	<b>0.925</b>
			iRNA-m6A	0.793	0.784	0.803	0.590	0.870
			im6A-TS-CNN	0.808	0.805	0.810	0.615	0.886
			DNN-m6A	0.809	0.812	0.806	0.620	0.889
	Liver	ML	M6A-BiNP (PSP-PMI)	0.735	0.676	0.795	0.474	0.817
			M6A-BiNP (PSP-PJMI)	<b>0.828</b>	0.699	<b>0.957</b>	<b>0.680</b>	<b>0.937</b>
			iRNA-m6A	0.688	0.678	0.699	0.380	0.762
			im6A-TS-CNN	0.716	0.756	0.676	0.433	0.793
			DNN-m6A	0.730	<b>0.764</b>	0.695	0.460	0.808
	Testis	MT	M6A-BiNP (PSP-PMI)	0.746	0.836	0.657	0.501	0.832
			M6A-BiNP (PSP-PJMI)	<b>0.851</b>	<b>0.857</b>	<b>0.845</b>	<b>0.702</b>	<b>0.928</b>
			iRNA-m6A	0.735	0.722	0.751	0.470	0.818
			im6A-TS-CNN	0.762	0.835	0.689	0.529	0.847
			DNN-m6A	0.771	0.801	0.742	0.540	0.854
			M6A-BiNP (PSP-PMI)	0.766	0.612	<b>0.920</b>	0.559	0.883
			M6A-BiNP (PSP-PJMI)	<b>0.866</b>	<b>0.988</b>	0.744	<b>0.755</b>	<b>0.982</b>
Rat	Brain	RB	iRNA-m6A	0.751	0.739	0.765	0.500	0.827
			im6A-TS-CNN	0.770	0.781	0.758	0.539	0.852
			DNN-m6A	0.780	0.777	0.783	0.560	0.862
			M6A-BiNP (PSP-PMI)	0.777	0.786	0.767	0.553	0.861
			M6A-BiNP (PSP-PJMI)	0.771	<b>0.966</b>	0.575	0.588	<b>0.936</b>
	Kidney	RK	iRNA-m6A	0.814	0.802	<b>0.828</b>	0.630	0.897
			im6A-TS-CNN	0.827	0.849	0.806	0.655	0.908
			DNN-m6A	<b>0.830</b>	0.853	0.807	<b>0.660</b>	0.911
			M6A-BiNP (PSP-PMI)	0.829	0.857	0.801	0.659	0.912
			M6A-BiNP (PSP-PJMI)	<b>0.887</b>	<b>0.989</b>	0.786	<b>0.791</b>	<b>0.986</b>
	Liver	RL	iRNA-m6A	0.799	0.777	<b>0.823</b>	0.600	0.876
			im6A-TS-CNN	0.802	0.845	0.759	0.607	0.885
			DNN-m6A	0.816	0.828	0.805	0.630	0.896

final features by concatenating features corresponding to different  $\alpha$  in PSP-PMI and various  $\beta$  in PSP-PJMI, respectively. Our PSP-PMI and PSP-PJMI are superior to the state-of-the-art feature encoding algorithms in extracting features with much better capability to identify m<sup>6</sup>A sites from RNA sequences. The PSP-PJMI is better than PSP-PMI. The M6A-BiNP predictor based on PSP-PJMI feature encoding algorithm outperforms the existing predictors for identifying m<sup>6</sup>A sites.

## Acknowledgments

We acknowledge those researchers who published the datasets for us to use in this research.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the [National Natural Science Foundation of China] under Grant [number 61673251, 62076159, 12031010]; [National Key Research and Development Program of China] under Grant [number 2016YFC0901900]; [Fundamental Research Funds for the Central Universities] under Grant [number GK202105003, GK201701006, 2018TS078]; [Scientific and Technological Achievements Transformation and Cultivation Funds] under Grant [number GK201806013]; and [Innovation Funds of Graduate Programs at Shaanxi Normal University] under Grant [number 2015CXS028, 2016CSY009].

## Data availability statement

The source code and datasets in this study are freely available in the GitHub repository <https://github.com/Mingzhao2017/M6A-BiNP>.

## ORCID

Juanying Xie  <http://orcid.org/0000-0001-6540-4397>

## Reference

- [1] Wolffe AP, Matzke MA. Epigenetics: regulation through repression. *Science*. 1999;286(5439):481–486.
- [2] Motorin Y, Helm M. RNA nucleotide methylation. Wiley Interdiscip Rev-RNA. 2011;2(5):611–631.
- [3] Cantara WA, Crain PF, Rozenski J, et al. The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Res*. 2011;39(suppl\_1):D195–D201.
- [4] Deng X, Chen K, Luo G, et al. Widespread occurrence of N6-methyladenosine in bacterial mRNA. *Nucleic Acids Res*. 2015;43(13):6557–6567.
- [5] Dominissini D, Moshitch-moshkovitz S, Schwartz S, et al. Topology of the human and mouse m<sup>6</sup>A RNA methylomes revealed by m<sup>6</sup>A-seq. *Nature*. 2012;485(7397):201–206.
- [6] Luo G, Macqueen A, Zheng G, et al. Unique features of the m<sup>6</sup>A methylome in *Arabidopsis thaliana*. *Nat Commun*. 2014;5(1):5630.
- [7] Wu R, Jiang D, Wang Y, et al. N<sup>6</sup>-Methyladenosine (m<sup>6</sup>A) Methylation in mRNA with a dynamic and reversible epigenetic modification. *Mol Biotechnol*. 2016;58(7):450–459.
- [8] Liu J, Yue Y, Han D, et al. A METTL3-METTL14 complex mediates mammalian nuclear RNA N<sup>6</sup>-adenosine methylation. *Nat Chem Biol*. 2014;10(2):93–95.
- [9] Ping X, Sun B, Wang L, et al. Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Res*. 2014;24(2):177–189.
- [10] Zheng G, Dahl JA, Niu Y, et al. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol Cell*. 2013;49(1):18–29.
- [11] Jia G, Fu Y, Zhao X, et al. N<sup>6</sup>-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat Chem Biol*. 2011;7(12):885–887.
- [12] Meyer KD, Jaffrey SR. The dynamic epitranscriptome: N<sup>6</sup>-methyladenosine and gene expression control. *Nat Rev Mol Cell Biol*. 2014;15(5):313–326.
- [13] Meyer KD, Saletore Y, Zumbo P, et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3'UTRs and near stop codons. *Cell*. 2012;149(7):1635–1646.
- [14] Nilsen TW. Internal mRNA methylation finally finds functions. *Science*. 2014;343(6176):1207–1208.
- [15] Geula S, Moshitchmoshkovitz S, Dominissini D, et al. m<sup>6</sup>A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science*. 2015;347(6225):1002–1006.
- [16] Lan Q, Liu P, Haase J, et al. The critical role of RNA m<sup>6</sup>A methylation in cancer. *Cancer Res*. 2019;79(7):1285–1292.
- [17] Cui Q, Shi H, Ye P, et al. m<sup>6</sup>A RNA methylation regulates the self-renewal and tumorigenesis of glioblastoma stem cells. *Cell Rep*. 2017;18(11):2622–2634.
- [18] Zhang C, Samanta D, Lu H, et al. Hypoxia induces the breast cancer stem cell phenotype by HIF-dependent and ALKBH5-mediated m<sup>6</sup>A-demethylation of NANOG mRNA. *Proc Natl Acad Sci U S A*. 2016;113(14):E2047–E2056.
- [19] Dominissini D, Moshitchmoshkovitz S, Salmonddivon M, et al. Transcriptome-wide mapping of N<sup>6</sup>-methyladenosine by m<sup>6</sup>A-seq based on immunocapturing and massively parallel sequencing. *Nat Protoc*. 2013;8(1):176–189.
- [20] Linder B, Grozhik AV, Olarerin-George AO, et al. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods*. 2015;12(8):767–772.
- [21] Ke S, Alemu EA, Mertens C, et al. A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev*. 2015;29(19):2037–2053.
- [22] Zhang Z, Chen L, Zhao Y, et al. Single-base mapping of m6A by an antibody-independent method. *Sci Adv*. 2019;5(7):eaax0250.
- [23] Schwartz S, Agarwala SD, Mumbach MR, et al. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*. 2013;155(6):1409–1421.
- [24] Li Y, Wang X, Li C, et al. Transcriptome-wide N<sup>6</sup>-methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification. *RNA Biol*. 2014;11(9):1180–1188.
- [25] Chen W, Feng P, Ding H, et al. iRNA-Methyl: identifying N<sup>6</sup>-methyladenosine sites using pseudo nucleotide composition. *Anal Biochem*. 2015;490:26–33.
- [26] Chen W, Tran H, Liang Z, et al. Identification and analysis of the N<sup>6</sup>-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci Rep*. 2015;5:13859.
- [27] Chen W, Feng P, Ding H, et al. Identifying N<sup>6</sup>-methyladenosine sites in the *Arabidopsis thaliana* transcriptome. *Mol Genet Genomics*. 2016;291(6):2225–2229.
- [28] Zhou Y, Zeng P, Li Y-H, et al. SRAMP: prediction of mammalian N<sup>6</sup>-methyladenosine (m<sup>6</sup>A) sites based on sequence-derived features. *Nucleic Acids Res*. 2016;44(10):e91–e91.
- [29] Li G, Liu Z, Shen H, et al. TargetM6A: identifying N<sup>6</sup>-methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. *IEEE Trans Nanobiosci*. 2016;15(7):674–682.
- [30] Liu Z, Xiao X, Yu D, et al. pRNAm-PC: predicting N<sup>6</sup>-methyladenosine sites in RNA sequences via physical-chemical properties. *Anal Biochem*. 2016;497:60–67.
- [31] Xing P, Su R, Guo F, et al. Identifying N<sup>6</sup>-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Sci Rep*. 2017;7:46757.
- [32] Wang X, Yan R. RFATHM6A: a new tool for predicting m<sup>6</sup>A sites in *Arabidopsis thaliana*. *Plant Mol Biol*. 2018;96(3):327–337.
- [33] Zhao Z, Peng H, Lan C, et al. Imbalance learning for the prediction of N<sup>6</sup>-methylation sites in mRNAs. *BMC Genomics*. 2018;19(1):574.
- [34] 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;9:495.
- [35] Wei L, Chen H, Su R. M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol Ther-Nucl Acids*. 2018;12:635–644.
- [36] Wei L, Su R, Wang B, et al. Integration of deep feature representations and handcrafted features to improve the prediction of N<sup>6</sup>-methyladenosine sites. *Neurocomputing*. 2019;324:3–9.
- [37] Nazari I, Tahir M, Tayara H, et al. iN6-Methyl (5-step): identifying RNA N6-methyladenosine sites using deep learning mode via Chou's 5-step rules and Chou's general PseKNC. *Chemometrics Intell Lab Syst*. 2019;193:103811.
- [38] Zou Q, Xing P, Wei L, et al. Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. *RNA*. 2019;25(2):205–218.
- [39] Chen Z, Zhao P, Li F, et al. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. *Brief Bioinformatics*. 2020;21(5):1676–1696.
- [40] Chen K, Wei Z, Zhang Q, et al. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res*. 2019;47(7):e41.
- [41] Liu K, Chen W. iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. *Bioinformatics*. 2020;36(11):3336–3342.
- [42] Liu K, Cao L, Du P, et al. im6A-TS-CNN: identifying the N (6)-Methyladenine site in multiple tissues by using the convolutional neural network. *Mol Ther Nucleic Acids*. 2020;21:1044–1049.

- [43] Dao F, Lv H, Yang Y, et al. Computational identification of N6-methyladenosine sites in multiple tissues of mammals. *Comput Struct Biotechnol J*. 2020;18:1084–1091.
- [44] Li J, He S, Guo F, et al. HSM6AP: a high-precision predictor for the Homo sapiens N6-methyladenosine (m<sup>6</sup>A) based on multiple weights and feature stitching. *RNA Biol*. 2021. DOI:10.1080/15476286.2021.1875180
- [45] Zhai J, Song J, Zhang T, et al. deepEA: a containerized web server for interactive analysis of epitranscriptome sequencing data. *Plant Physiol*. 2021;185(1):29–33.
- [46] Luo X, Li H, Liang J, et al. RMVar: an updated database of functional variants involved in RNA modifications. *Nucleic Acids Res*. 2021;49(D1):D1405–D1412.
- [47] Chen K, Song B, Tang Y, et al. RMDisease: a database of genetic variants that affect RNA modifications, with implications for epitranscriptome pathogenesis. *Nucleic Acids Res*. 2021;49(D1):D1396–D1404.
- [48] Tang Y, Chen K, Song B, et al. m6A-Atlas: a comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome. *Nucleic Acids Res*. 2021;49(D1):D134–D143.
- [49] Chen W, Feng P, Ding H, et al. Benchmark data for identifying N6-methyladenosine sites in the Saccharomyces cerevisiae genome. *Data Brief*. 2015;5:376–378.
- [50] Chen W, Tang H, Lin H. MethyRNA: a web server for identification of N<sup>6</sup>-methyladenosine sites. *J Biomol Struct Dyn*. 2017;35(3):683–687.
- [51] Zhao X, Zhang Y, Ning Q, et al. Identifying N<sup>6</sup>-methyladenosine sites using extreme gradient boosting system optimized by particle swarm optimizer. *J Theor Biol*. 2019;467:39–47.
- [52] Akbar S, Hayat M. iMethyl-STNC: identification of N<sup>6</sup>-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. *J Theor Biol*. 2018;455:205–211.
- [53] Hu L, Gao W, Zhao K, et al. Feature selection considering two types of feature relevancy and feature interdependency. *Expert Syst Appl*. 2018;93:423–434.
- [54] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*. 2005;27(8):1226–1238.
- [55] Xie J, Wang M, Zhou Y, et al. Differential expression gene selection algorithms for unbalanced gene datasets. *Chin J Comput*. 2019;42(6):1232–1251.
- [56] Xu Y, Ding Y, Ding J, et al. Phogly-PseAAC: prediction of lysine phosphoglycylation in proteins incorporating with position-specific propensity. *J Theor Biol*. 2015;379:10–15.
- [57] He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N<sup>4</sup>-methylcytosine sites prediction. *Bioinformatics*. 2019;35(4):593–601.
- [58] Liu B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief Bioinform*. 2019;20(4):1280–1294.
- [59] Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–297.
- [60] Chen W, Feng P, Song X, et al. iRNA-m7G: identifying N<sup>7</sup>-methylguanosine sites by fusing multiple features. *Mol Ther-Nucl Acids*. 2019;18:269–274.
- [61] Wang M, Ding L, Xu M, et al. A novel method detecting the key clinic factors of portal vein system thrombosis of splenectomy & cardia devascularization patients for cirrhosis & portal hypertension. *BMC Bioinformatics*. 2019;20(Suppl 22):720.
- [62] Wei L, Luan S, Nagai LAE, et al. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics*. 2019;35(8):1326–1333.
- [63] Xie J, Wang C. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythematous-squamous diseases. *Expert Syst Appl*. 2011;38(5):5809–5815.
- [64] Lin H, Liang Z, Tang H, et al. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE-ACM Trans Comput Biol Bioinform*. 2019;16(4):1316–1321.
- [65] Chang C, Lin C. LIBSVM: a library for support vector machines. *ACM T Intel Syst Tec*. 2011;2(3):27.
- [66] Sabooh MF, Iqbal N, Khan M, et al. Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. *J Theor Biol*. 2018;452:1–9.
- [67] Chen W, Song X, Lv H, et al. iRNA-m2G: identifying N2-methylguanosine sites based on sequence derived information. *Mol Ther-Nucl Acids*. 2019;18:253–258.
- [68] Yang H, Qiu W, Liu G, et al. iRSpot-Pse6NC: identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. *Int J Biol Sci*. 2018;14(8):883–891.
- [69] Zhang J, Feng P, Lin H, et al. Identifying RNA N6-methyladenosine sites in Escherichia coli genome. *Front Microbiol*. 2018;9:955.
- [70] Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27(8):861–874.
- [71] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006. Association for Computing Machinery, pp 233–240.
- [72] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
- [73] Vacic V, Iakoucheva LM, Radivojac P. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*. 2006;22(12):1536–1537.
- [74] Xiang S, Yan Z, Liu K, et al. AthMethPre: a web server for the prediction and query of mRNA m<sup>6</sup>A sites in Arabidopsis thaliana. *Mol Biosyst*. 2016;12(11):3333–3337.
- [75] Tahir M, Hayat M, Chong KT. Prediction of n6-methyladenosine sites using convolution neural network model based on distributed feature representations. *Neural Netw*. 2020;129:385–391.
- [76] Rehman MU, Hong KJ, Tayara H, et al. m6A-NeuralTool: convolution neural tool for RNA N6-Methyladenosine site identification in different species. *IEEE Access*. 2021;9:17779–17786.
- [77] Alam W, Ali SD, Tayara H, et al. A CNN-based RNA n6-methyladenosine site predictor for multiple species using heterogeneous features representation. *IEEE Access*. 2020;8:138203–138209.
- [78] Pian C, Yang Z, Yang Y, et al. Identifying RNA N6-Methyladenine sites in three species based on a markov model. *Front Genet*. 2021;12:428.
- [79] Zhang M, Sun J, Liu Z, et al. Improving N6-methyladenosine site prediction with heuristic selection of nucleotide physical-chemical properties. *Anal Biochem*. 2016;508:104–113.
- [80] Chen W, Xing P, Zou Q. Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. *Sci Rep*. 2017;7(1):1–8.
- [81] Huang Y, He N, Chen Y, et al. BERM: a cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach. *Int J Biol Sci*. 2018;14(12):1669.
- [82] Mahmoudi O, Wahab A, Chong KT. iMethyl-Deep: N6 methyladenosine identification of yeast genome with automatic feature extraction technique by using deep learning algorithm. *Genes (Basel)*. 2020;11(5):529.
- [83] Zhang L, Qin X, Liu M, et al. DNN-m6A: a cross-species method for identifying RNA N6-methyladenosine sites based on deep neural network with multi-information fusion. *Genes (Basel)*. 2021;12(3):354.