PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2023, 50(12): 3032~3044
www.pibb.ac.cn

Prediction of m^{6}A Methylation Sites in Mammalian Tissues Based on a Double-layer BiGRU Network*

LI Hui-Min**, CHEN Peng-Hui, TANG Yi**, XU Quan-Feng, HU Meng, WANG Yu
(*School of Mathematics and Computer Science, Yunnan Minzu University, Kunming* 650504, China)

Abstract **Objective** N^6-methyladenosine (m^{6}A) is the most common and abundant chemical modification in RNA and plays an important role in many biological processes. Several computational methods have been developed to predict m^{6}A methylation sites. However, these methods lack robustness when targeting different species or different tissues. To improve the robustness of the prediction performance of m^{6}A methylation sites in different tissues, this paper proposed a double-layer bidirectional gated recurrent unit (BiGRU) network model that combines reverse sequence information to extract higher-level features of the data. **Methods** Some representative mammalian tissue m^{6}A methylation site datasets were selected as the training datasets. Based on a BiGRU, a double-layer BiGRU network was constructed by collocation of the model network, the model structure, the number of layers and the optimizer. **Results** The model was applied to predict m^{6}A methylation sites in 11 human, mouse and rat tissues, and the prediction performance was compared with that of other methods using the same tissues. The results demonstrated that the average area under the receiver operating characteristic curve (*AUC*) predicted by the proposed model reached 93.72%, equaling that of the best prediction method at present. The values of accuracy (*ACC*), sensitivity (*SN*), specificity (*SP*) and Matthews correlation coefficient (*MCC*) were 90.07%, 90.30%, 89.84% and 80.17%, respectively, which were higher than those of the current methods for predicting m^{6}A methylation sites. **Conclusion** Compared with that of existing research methods, the prediction accuracy of the double-layer BiGRU network was the highest for identifying m^{6}A methylation sites in the 11 tissues, indicating that the method proposed in this study has an excellent generalizability.

Key words N^6-methylated adenosine site, bidirectional gated recurrent unit, base sequence, deep learning
DOI: 10.16476/j.pibb.2023.0011

RNA methylation is a new field of epigenetic regulation[1-2]. m^{6}A methylation is the most common and abundant chemical modification in RNA, accounting for approximately 80% of RNA methylation modifications[3-4]. It plays an important role in regulating RNA maturation, cleavage, transport, degradation and translation[5-7]. Many enzymes involved in m^{6}A methylation can be modified at the m^{6}A methylation sites[8]. Therefore, the accurate identification of m^{6}A methylation sites from RNA sequences is crucial for understanding the biological function of RNA methylation modifications.

Early detection methods of m^{6}A methylation sites were mainly based on biological experiments, such as two-dimensional cellulose thin chromatography, high-performance liquid chromatography and mass spectrometry[9]. However, due to the limitations of experimental conditions, these methods generally have many problems, such as being time-consuming, having a high cost and having a small detection scale. The emergence of high-throughput sequencing technology has provided strong technical support for methylation research[10-12] and generated a large amount of m^{6}A methylation site data, which has led to the identification of m^{6}A

* This work was supported by a grant from The National Natural Science Foundation of China (61866040).
** Corresponding author.
LI Hui-Min. Tel: 86-13888183685, E-mail: lihuimin_1980@126.com
TANG Yi. Tel: 86-18314589836, E-mail: yitang.math@gmail.com
Received: January 9, 2023 Accepted: May 15, 2023

methylation sites from biological experiments and computational research. Using high-throughput experimental data and traditional machine learning methods, some models for predicting m^{6}A methylation sites have been developed. Examples include iRNA-Methyl[13] and pRNAm-PC predictors[14] based on base resolution technology, SRAMP[15] based on random forest (RF), and models based on support vector machine (SVM), such as RAM-NPPS[16], M6APred-EL[17], iMethyl-STTNC[18] and iRNA(m6A) -PseDNC[19]. Traditional machine learning algorithms require more professional knowledge to manually extract features from datasets, reduce the features' dimensions and transfer the best features to the model. The process of feature extraction is very complicated. In recent years, many researchers have proposed m^{6}A methylation site prediction algorithms based on deep learning algorithms[20], which can automatically obtain high-level features based on sample datasets, and developed methods for cross-species prediction of m^{6}A methylation sites.

Researchers have mainly targeted m^{6}A methylation sites in different species, such as *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Mus musculus* (mouse), *Rattus norvegicus* (rat) and *Homo sapiens* (human), to make macroscopic predictions. However, less attention has been given to m^{6}A methylation sites in more microscopic biological tissues. As an example, the expression levels of m^{6}A methylation were found to be different between diseased and unaffected tissues[21-23], while few methods have predicted m^{6}A methylation sites in different tissues. In recent years, some researchers have refined m^{6}A methylation site prediction to tissue sites[24-29]. For example, Dao *et al*. [26] and Wang *et al.* [27] proposed iRNA-m6A[26] and M6A-BiNP[27], respectively, which mainly rely on SVMs, to predict m^{6}A methylation sites in 11 tissues of 3 species (human, mouse and rat). Liu *et al.*[28] developed im6A-TS-CNN based on a single-layer convolutional neural network to further improve the values of the area under the receiver operating characteristic (ROC) curve (*AUC*). Zhang *et al.*[29] developed a tool named DNN-m6A using deep neural networks to identify m^{6}A methylation sites in multiple human, mouse and rat tissues and showed an excellent generalizability. Although there have been an increasing number of computational methods for m^{6}A methylation site prediction and some progress has been made in the prediction of tissue m^{6}A methylation sites, the following problems remain. (1) The predicted regions are generally not sufficiently refined. Only a few algorithms, such as iRNA-m6A, im6A-TS-CNN, DNN-m6A and M6A-BiNP, subdivide the predicted regions into various tissues. (2) Most algorithms have low prediction accuracy in some tissues, and the prediction accuracy is generally below 80%.

m^{6}A methylation site prediction is based on nucleotide sequences, in which nucleotides are associated with each other. As one of the classical deep learning algorithms, a recurrent neural network (RNN) has excellent performance in processing sequence data. In particular, a bidirectional RNN can combine the reverse characteristics of sequences. Therefore, based on a bidirectional gating recurrent unit (BiGRU), which is a variant of the bidirectional RNN, and selected representative mammalian tissue m^{6}A methylation site datasets as training data, we constructed a double-layer BiGRU network. m^{6}A methylation sites in 11 mammalian tissues were predicted using our method, and the predicted results show that the proposed method is superior to existing methods.

1 Materials and methods

1.1 Materials

The datasets used in this research were from the m^{6}A methylation site benchmark datasets constructed by Dao *et al.*[26] and downloaded from their paper. The datasets contain m^{6}A methylation sites in 11 mammalian tissues from 3 species: human (brain, liver and kidney), mouse (brain, liver, heart, testis and kidney) and rat (brain, liver and kidney). Each dataset of the above 11 tissues contained two parts: a training dataset used to train the model and an independent test dataset used to test the performance of the model. In each training dataset and independent test dataset, the same sequence numbers of positive samples (m^{6}A sites) and negative samples (non-m^{6}A sites) were included. The length of each sequence in the positive and negative samples was 41 nt, with adenine (A) in the center of a sequence. The detailed sample sizes in the datasets are shown in Table 1[26]. It was observed that the sample size of human brain tissue was at a medium level in all datasets, so we used it to debug the model parameters.

Table 1 Benchmark datasets of m^{6}A methylation sites

Species	Tissues	Abbreviations	Training datasets		Independent test datasets	
			Positive	Negative	Positive	Negative
Human	Brain	H_B	4 605	4 605	4 604	4 604
	Kidney	H_K	4 574	4 574	4 573	4 573
	Liver	H_L	2 634	2 634	2 634	2 634
Mouse	Brain	M_B	8 025	8 025	8 025	8 025
	Heart	M_H	2 201	2 201	2 200	2 200
	Kidney	M_K	3 953	3 953	3 952	3 952
	Liver	M_L	4 133	4 133	4 133	4 133
	Testis	M_T	4 707	4 707	4 706	4 706
Rat	Brain	R_B	2 352	2 352	2 351	2 351
	Kidney	R_K	3 433	3 433	3 432	3 432
	Liver	R_L	1 762	1 762	1 762	1 762

To make the original data acceptable to the model, the sample RNA sequences were processed by one-hot encoding. Let A=(1, 0, 0, 0)^T, U=(0, 1, 0, 0)^T, C=(0, 0, 1, 0)^T and G=(0, 0, 0, 1)^T; then, each RNA sequence can be represented as a numerical matrix that contains only 1s and 0s with 4 rows and 41 columns.

1.2 Methods

1.2.1 Construction of the double-layer BiGRU prediction model

The core model of our method is a gated recurrent unit (GRU). The GRU model can better and more automatically capture the dependence relationship in a sequence[30], and it is suitable for predicting m^{6}A methylation sites in a sequence. The GRU controls the flow of information by resetting and updating the gate, which can effectively solve the gradient disappearance problem in RNNs, and the model has fewer parameters and is more concise. The network structure is shown in Figure 1. The model mainly includes two bidirectional GRU (BiGRU) layers. The first BiGRU layer (BiGRU_layer1) processes the data transformed by the input layer to obtain the initially extracted feature vector, and the second BiGRU layer (BiGRU_layer2) further extracts the features obtained from the previous layer. Hence, the function of BiGRU_layer2 is to capture more advanced information and make the model obtain more useful data characteristics.

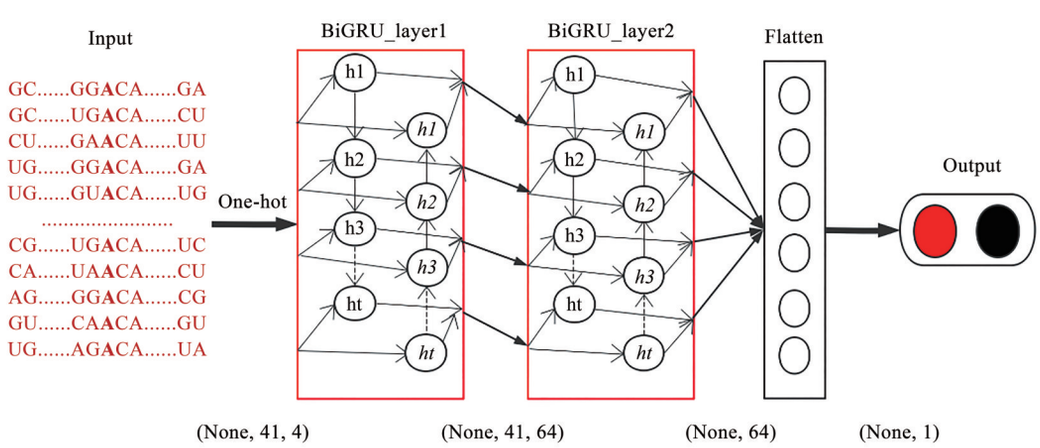

Fig. 1 Model structure diagram of bidirectional gated recurrent unit (BiGRU) network

(None, 41, 4) represents the data dimension from the input layer to BiGRU_layer1, (None, 41, 64) represents the data dimension entered into BiGRU_layer2 after the BiGRU_layer1 operation, (None, 64) represents the data dimension entered into the flatten layer after the BiGRU_layer2 operation, and (None, 1) is the data dimension calculated by flattening to the output layer. The regular fonts h1, h2, h3, ..., ht indicate the hidden status of the forward GRU network, and the italics fonts *h1, h2, h3, ..., ht* indicate the hidden status of the backward GRU network. The length of each sequence is 41 nt, with adenine (A) in the center of a sequence.

1.2.2 Detailed algorithm procedure

(1) The nucleotide sequence data were converted into the form of one-hot encoding, and each sample RNA sequence with dimension (4, 41) was fed into the model.

(2) Two BiGRU layers were added using the Python library Keras. Since the previous data input dimension is (4, 41), we set ‘input_shape’ to (4, 41) in BiGRU_layer1 and the number of neurons in both BiGRU layers to 32.

(3) The results of BiGRU_layer1 and BiGRU_layer2 were passed to the ‘Flatten layer’, and a high-dimensional data input vector was converted into a one-dimensional output vector.

(4) In the output layer, ‘sigmoid’ was selected as the activation function, and its formula is given in Equation (1):

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

In Equation (1), x is the output value of the previous flattened layer processing, and the range of $f(x)$ is [0, 1], which is similar to the probability value. The prediction was positive samples (m^{6}A sites) when $f(x) > 0.5$, and the prediction was negative samples (non-m^{6}A sites) when $f(x) \leqslant 0.5$.

1.2.3 Design of model and parameters

The prediction of m^{6}A methylation sites in this work was treated as a classification problem; therefore, the loss function of the model was the binary cross-entropy function, as shown in Equation (2):

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i - \left[y_i \cdot \ln p_i + \left(1 - y_i\right) \cdot \ln\left(1 - p_i\right) \right] \tag{2}$$

where y_i represents the label of the sample i, the positive class is 1, and the negative class is 0; p_i represents the probability of the sample i being predicted to be a positive class.

In the model, the epoch number was set to 150, the batch size was set to 32, and the ‘Adam’ optimizer was used. When the initial learning rate was not applicable, the accuracy of the model did not improve after a certain number of epoch iterations. Therefore, the callback function ‘ReduceLROnPlateau’ was added to optimize the learning rate. The monitoring variable in the callback function was ‘Val_loss’, and ‘patience’ was set to 20. That is, when the model loss value did not decrease after 20 epochs, the mechanism of learning rate reduction in the callback function was triggered. A ‘factor’ value of 0.1 was used to reduce the learning rate in the training process, thus improving the accuracy of the model. Because the callback function ‘ReduceLROnPlateau’ needs several iterations to optimize the learning rate to make the model reach the best state, to achieve higher accuracy and accelerate the model training, the ‘EarlyStopping’ strategy was added to stop the model training in advance. The monitoring variable in ‘EarlyStopping’ was ‘val_binary_accuracy’, and ‘patience’ was set to 30. Training was stopped when the accuracy of the model after 30 epochs had not changed. In this situation, ‘EarlyStopping’ will not be triggered early, so the model can be fully trained while avoiding overfitting.

A 10-fold cross-validation test was used in the experiments. That is, the datasets were randomly divided into 10 subsets. In turn, 8 of them were used as a training set, 1 of them was used as a validation set, and the remaining one was used as a test set. In each experiment, a correct rate was obtained, and finally, the average correct rate of the 10 results was used as the estimation of the accuracy of the model or algorithm.

1.2.4 Evaluation metrics

Four classical evaluation metrics, including sensitivity (*SN*), specificity (*SP*), accuracy (*ACC*), and Matthews correlation coefficient (*MCC*) [31], were implemented to assess the performance of the model. The corresponding metrics can be expressed as formulas (3)−(6):

$$SN = \frac{TP}{TP + FN} \tag{3}$$

$$SP = \frac{TN}{TN + FP} \tag{4}$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}} \tag{6}$$

where *TP*, *FP*, *FN* and *TN* represent the number of correctly predicted positive samples, incorrectly predicted negative samples, incorrectly predicted positive samples and correctly predicted negative

samples[32], respectively. The *AUC*[33] was also introduced to evaluate the overall performance of the model[34]. The value range of *AUC* is [0, 1], and the *AUC* is positively correlated with the prediction performance. The larger the *AUC* value is, the better the overall performance of the predictor. In the aspect of code implementation, the encapsulation function for the prediction model of Chen *et al*.[35-36] was used.

2 Results and discussion

Our method was compared with several existing methods[26-29]. These methods include iRNA-m6A and M6A-BiNP based on SVM, im6A-TS-CNN based on a single-layer convolutional neural network, and DNN-m6A based on a deep neural network. At present, these 4 methods have achieved good performance in m^{6}A methylation site prediction in mammalian tissues. Since M6A-BiNP had the better comprehensive performance among the methods, we only reported comparison results with M6A-BiNP in each tissue.

2.1 Prediction results on human tissues

For the human independent test datasets (Table 2), our method showed the best *ACC* values (H_B: 87.48%, H_K: 90.87% and H_L: 90.72%) and *MCC* values (H_B: 74.96%, H_K: 81.76% and H_L: 81.47%). The *AUC* value of our method for H_L showed the best performance simultaneously with that of M6A-BiNP (our method: 94.04% and M6A-BiNP: 94.80%); for H_B and H_K, our method had the best performance (H_B: 91.97% and H_K: 94.51%). Although the model *SN* values for H_K and H_L were lower than those of M6A-BiNP (H_K: 96.4% and H_L: 92%), both were higher than 90% (H_K: 90.94% and H_L: 90.77%). The *SP* value of our method for H_B was lower than that of M6A-BiNP (95.40%), but it reached 87.46%. Compared with the test results of the other methods, the performance of our method was stable for different tissues. For example, although M6A-BiNP achieved a better *SP* value (95.4%) for H_B, its *SP* value for H_K was only 40%. One of the greatest advantages of our method is its universality. It had high *ACC* and *AUC* values for all 3 human tissues, but the other methods did not have such high performance. For example, although the *ACC* value of M6A-BiNP for H_L was 86.20%, *ACC* values of only 76.70% and 68.20% were obtained for H_B and H_K, respectively.

Table 2 Evaluation metrics on human independent test datasets

Tissue	Model	*ACC*/%	*SN*/%	*SP*/%	*MCC*/%	*AUC*/%
H_B	**Our method**	**87.48**	**87.49**	87.46	**74.96**	**91.97**
	M6A-BiNP [27]	76.70	58.00	**95.40**	57.60	89.40
	iRNA-m6A [26]	71.10	69.50	73.00	42.00	78.50
	im6A-TS-CNN [28]	72.70	75.20	70.20	45.40	80.60
	DNN-m6A [29]	73.30	75.00	71.50	47.00	81.50
H_K	**Our method**	**90.87**	90.94	**90.79**	**81.76**	**94.51**
	M6A-BiNP [27]	68.20	**96.40**	40.00	44.10	87.90
	iRNA-m6A [26]	77.80	77.10	78.40	56.00	85.70
	im6A-TS-CNN [28]	79.20	80.00	78.50	58.50	87.30
	DNN-m6A [29]	79.90	83.20	76.60	60.00	87.80
H_L	**Our method**	**90.72**	90.77	**90.66**	**81.47**	94.04
	M6A-BiNP [27]	86.20	**92.00**	80.50	73.00	**94.80**
	iRNA-m6A [26]	79.00	78.20	79.90	58.00	86.80
	im6A-TS-CNN [28]	79.90	84.80	75.00	60.10	88.10
	DNN-m6A [29]	81.00	81.80	80.10	62.00	88.50

2.2 Prediction results on mouse tissues

For the mouse independent test datasets (Table 3), our method showed a better prediction effect. Except for the lower *SP* and *AUC* values for M_H and M_L than those of M6A-BiNP, our method achieved the best *ACC*, *SN*, *SP*, *MCC* and *AUC* values for the other tissues. At the same time, compared with those of M6A-BiNP, the evaluation metrics of our method for different tissues had smaller fluctuation ranges. For example, for 5 mouse tissues, the *SP* values of

M6A-BiNP ranged from 67.40% to 99.60%, and those of our method ranged from 88.09% to 90.96%. The *ACC* values of M6A-BiNP and our method ranged from 75.60% to 85.10% and 87.81% to 91.18%, respectively. On the one hand, these results demonstrate that our method has higher accuracy; on the other hand, it has smaller fluctuation in terms of each evaluation criterion. Therefore, it was concluded that the prediction performance of our method was more stable and more universal for mouse tissue m^6A methylation site prediction.

Table 3　Evaluation metrics on mouse independent test datasets

Tissue	Model	*ACC*/%	*SN*/%	*SP*/%	*MCC*/%	*AUC*/%
M_B	**Our method**	**91.02**	**91.06**	**90.96**	**82.04**	**94.52**
	M6A-BiNP [27]	75.60	83.80	67.40	51.80	84.90
	iRNA-m6A [26]	78.30	77.20	79.40	57.00	86.10
	im6A-TS-CNN [28]	78.50	86.20	70.70	57.70	87.20
	DNN-m6A [29]	78.60	75.10	82.10	57.00	87.60
M_H	**Our method**	**88.05**	**87.99**	88.09	**76.11**	92.05
	M6A-BiNP [27]	83.80	68.10	**99.60**	71.20	**98.30**
	iRNA-m6A [26]	71.30	70.50	72.10	43.00	78.80
	im6A-TS-CNN [28]	73.60	75.80	71.40	47.20	81.60
	DNN-m6A [29]	75.10	77.30	73.00	50.00	83.40
M_K	**Our method**	**91.18**	**91.85**	**90.50**	**82.37**	**94.67**
	M6A-BiNP [27]	83.20	90.60	75.80	67.20	92.50
	iRNA-m6A [26]	79.30	78.40	80.30	59.00	87.00
	im6A-TS-CNN [28]	80.80	80.50	81.00	61.50	88.60
	DNN-m6A [29]	80.90	81.20	80.60	62.00	88.90
M_L	**Our method**	**87.81**	**87.71**	87.90	**75.66**	91.98
	M6A-BiNP [27]	82.80	69.90	**95.70**	68.00	**93.70**
	iRNA-m6A [26]	68.80	67.80	69.90	38.00	76.20
	im6A-TS-CNN [28]	71.60	75.60	67.60	43.30	79.30
	DNN-m6A [29]	73.00	76.40	69.50	46.00	80.80
M_T	**Our method**	**88.83**	**89.05**	**88.61**	**77.69**	**92.94**
	M6A-BiNP [27]	85.10	85.70	84.50	70.20	92.80
	iRNA-m6A [26]	73.50	72.20	75.10	47.00	81.80
	im6A-TS-CNN [28]	76.20	83.50	68.90	52.90	84.70
	DNN-m6A [29]	77.10	80.10	74.20	54.00	85.40

2.3　Prediction results on rat tissues

The model prediction results of m^6A methylation sites on three independent test datasets of rat tissues were compared with those of other methods (Table 4). Our method achieved the best prediction *AUC* value for R_K tissue, and although the *AUC* values were lower than those of M6A-BiNP for R_B and R_L tissues, they exceeded 92%. However, the *ACC* values of our method were highest for 3 tissues. This indicated that our method could improve the prediction accuracy of m^6A methylation sites on rat datasets. Moreover, similar to those for the mouse tissues, the 9 prediction results of our method for 3 different rat tissues also had smaller fluctuation. For example, in the mentioned 3 tissues, the *SP* values of M6A-BiNP ranged from 57.50% to 90.78%, and those of our method ranged from 89.22% to 0.78%. The *ACC* values of M6A-BiNP and our method ranged 77.10%−88.70% and 89.51%−91.47%, respectively. This further demonstrates the universality of the proposed method.

2.4　Model summary

To illustrate the overall comparison between our method and other state-of-the-art methods, the prediction results of 11 tissues were averaged. It can be seen from the results of the training datasets (Figure 2a) that the *AUC* value of our method was almost equal to that of M6A-BiNP and higher than

Table 4　Evaluation metrics on rat independent datasets

Tissue	Model	*ACC*/%	*SN*/%	*SP*/%	*MCC*/%	*AUC*/%
R_B	**Our method**	**89.51**	89.06	**89.95**	**79.08**	92.73
	M6A-BiNP [27]	86.60	**98.80**	74.40	75.50	**98.20**
	iRNA-m6A [26]	75.10	73.90	76.50	50.00	82.70
	im6A-TS-CNN [28]	77.00	78.10	75.80	53.90	85.20
	DNN-m6A [29]	78.00	77.70	78.30	56.00	86.20
R_K	**Our method**	**91.47**	92.15	**90.78**	**82.99**	**94.88**
	M6A-BiNP [27]	77.10	**96.60**	57.50	58.80	93.60
	iRNA-m6A [26]	81.40	80.20	82.80	63.00	89.70
	im6A-TS-CNN [28]	82.70	84.90	80.60	65.50	90.80
	DNN-m6A [29]	83.00	85.30	80.70	66.00	91.10
R_L	**Our method**	**90.04**	90.86	**89.22**	**80.13**	93.91
	M6A-BiNP [27]	88.70	**98.90**	78.60	79.10	**98.60**
	iRNA-m6A [26]	79.90	77.70	82.30	60.00	87.60
	im6A-TS-CNN [28]	80.20	84.50	75.90	60.70	88.50
	DNN-m6A [29]	81.60	82.80	80.50	63.00	89.60

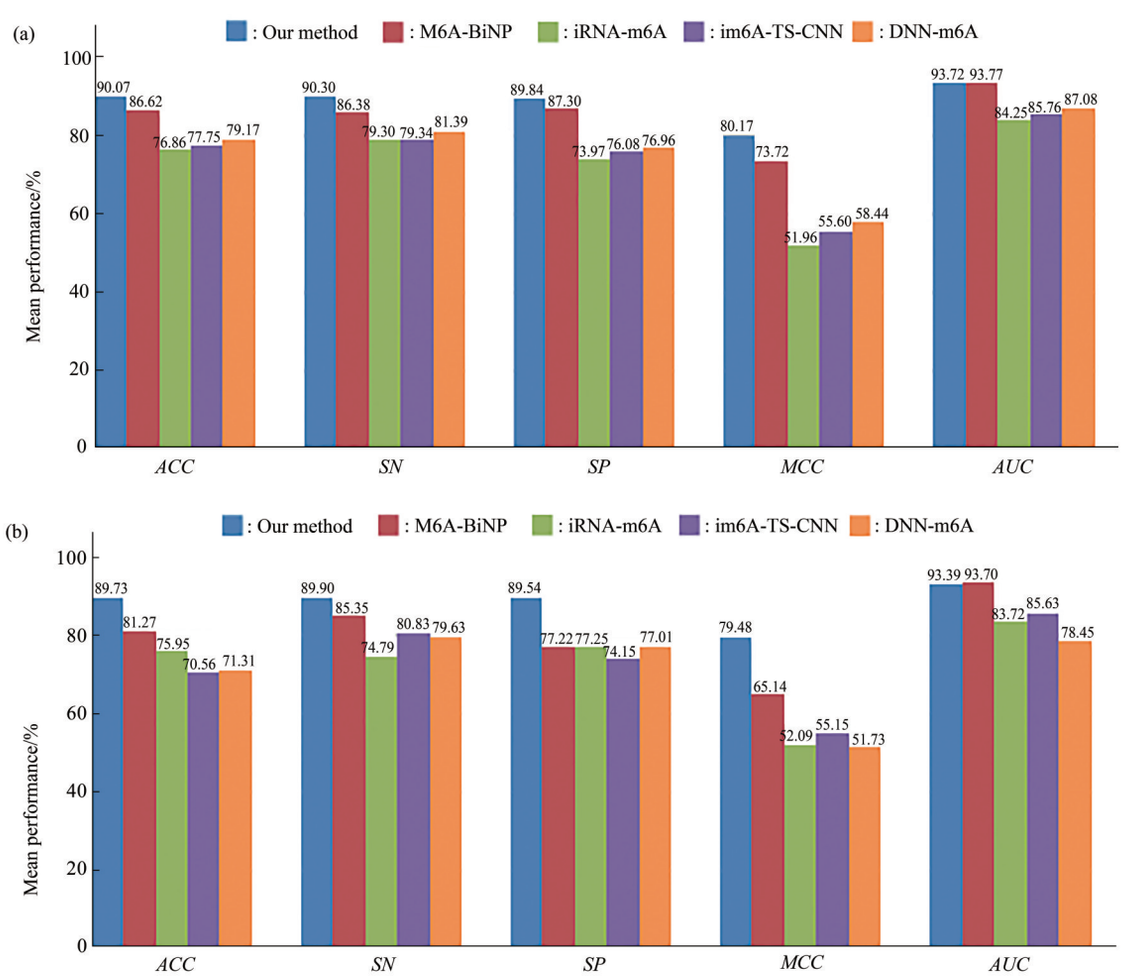

Fig. 2　The overall performance of different methods on 11 tissues

(a) The performance of different methods on the training datasets. (b) The performance of different methods on the independent test datasets. The bars represent the evaluation metrics under different methods, which are averaged by the same evaluation metrics in the 11 tissues.

that of the other methods. The values of *ACC*, *SN*, *SP*, *MCC* of our method were higher than those of M6A-BiNP, iRNA-m6A, im6A-TS-CNN and DNN-m6A. The results for 11 tissues in the independent test datasets were also averaged (Figure 2b). The *AUC* value of our method was also equal to that of M6A-BiNP, but the other prediction results of our method were significantly higher than those of the other 4 methods. This demonstrates that our method can more effectively predict m^{6}A methylation sites than other state-of-the-art methods.

2.5　Ten−fold cross validation ROC curves

To visually show the prediction effect of each cross-validation, the 10-fold cross-validation results of the independent test datasets were plotted as ROC curves (Figure 3−5). As shown in Figure 3−5, for human tissues, the average *AUC* values of our method exceeded 92%. For example, for the H_B, H_K and H_L tissues, the model *AUC* values on the independent test datasets were (92±4)%, (94±3)% and (94±3)% , respectively. For the mouse tissues, the average *AUC* values of our method also exceeded 92%. The *AUC* values of our method with the independent test datasets for the M_B, M_H, M_K, M_L, and M_T tissues were (95±2)% , (92±4)% , (95±2)% , (92±4)% and (93±3)% , respectively. The average *AUC* values of our method were greater than 93% for rat tissues. They were (93±3)%, (95±2)% and (94±2)% for the R_B, R_K and R_L tissues, respectively.

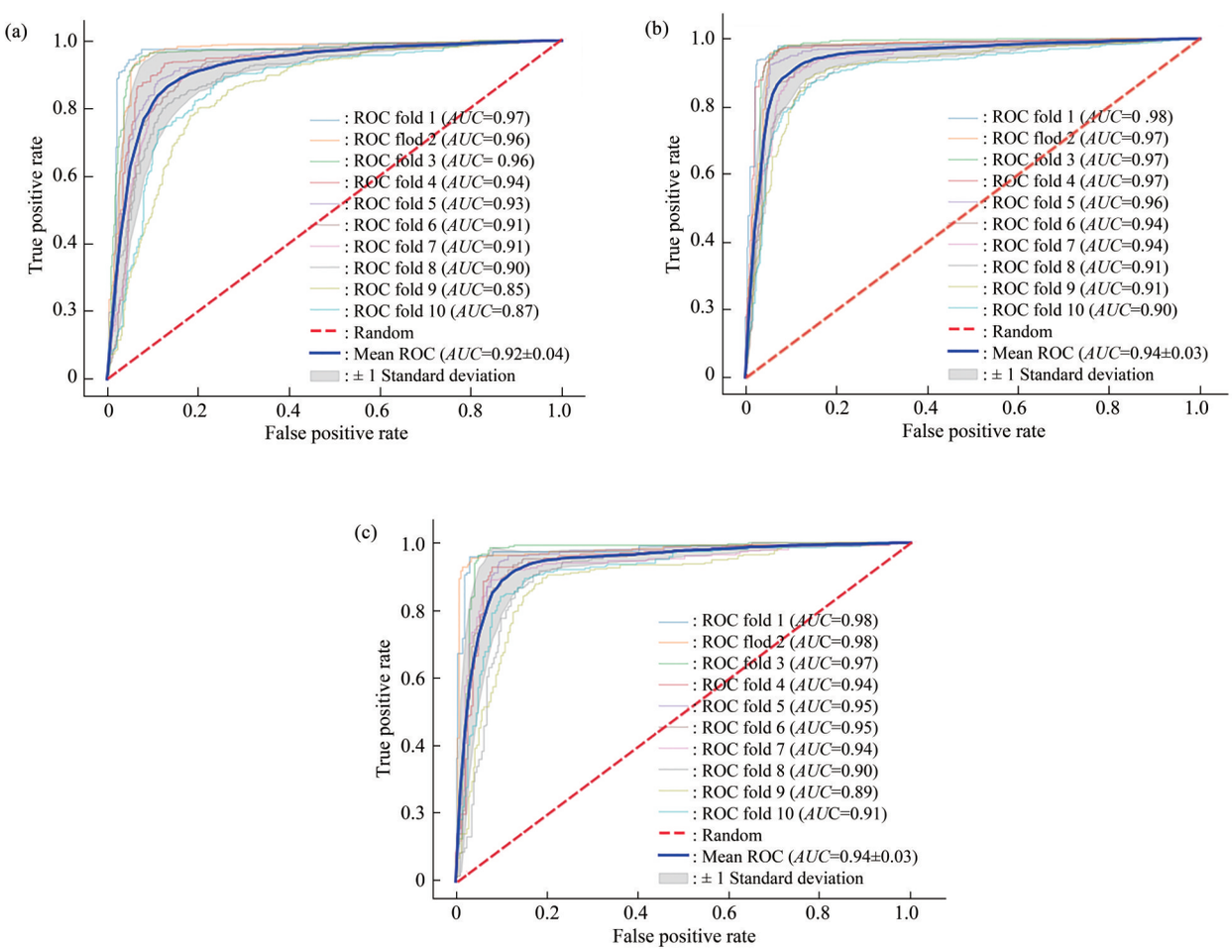

Fig. 3　The 10−fold cross−validation receiver operating characteristic（ ROC ）curves on the independent test datasets of human tissues

(a−c) represent 10-fold cross-validation ROC curves on the human brain (H_B), human kidney (H_K) and human liver (H_L) independent test datasets with our method, respectively. The horizontal axis represents false positive rate and the vertical axis represents true positive rate.

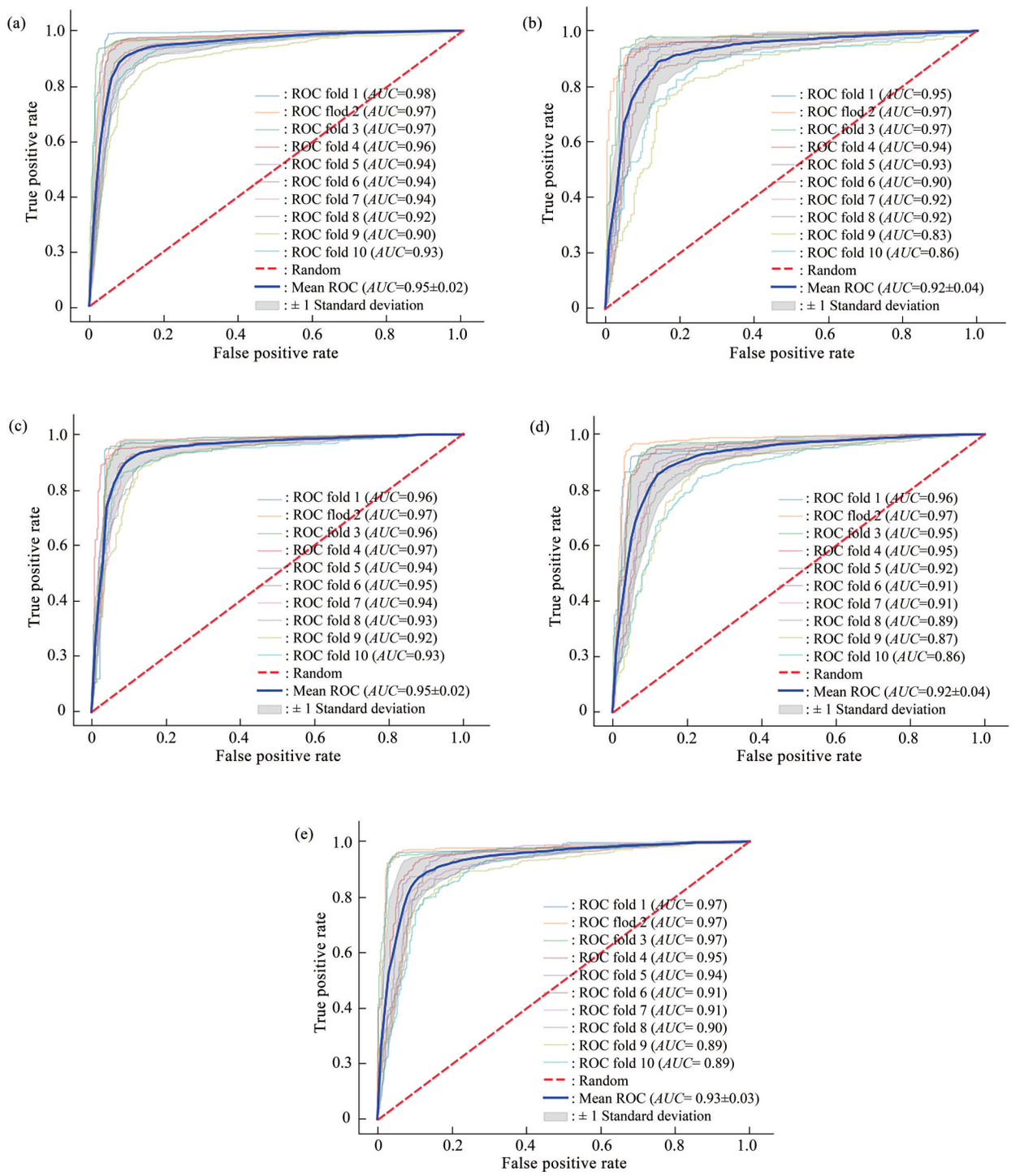

Fig. 4　The 10−fold cross−validation receiver operating characteristic (ROC) curves on the independent test datasets of mouse tissues

(a−e) represent 10-fold cross-validation ROC curves on the mouse brain (M_B), mouse heart (M_H), mouse kidney (M_K), mouse liver (M_L) and mouse testis (M_T) independent test datasets with our method, respectively. The horizontal axis represents false positive rate and the vertical axis represents true positive rate.

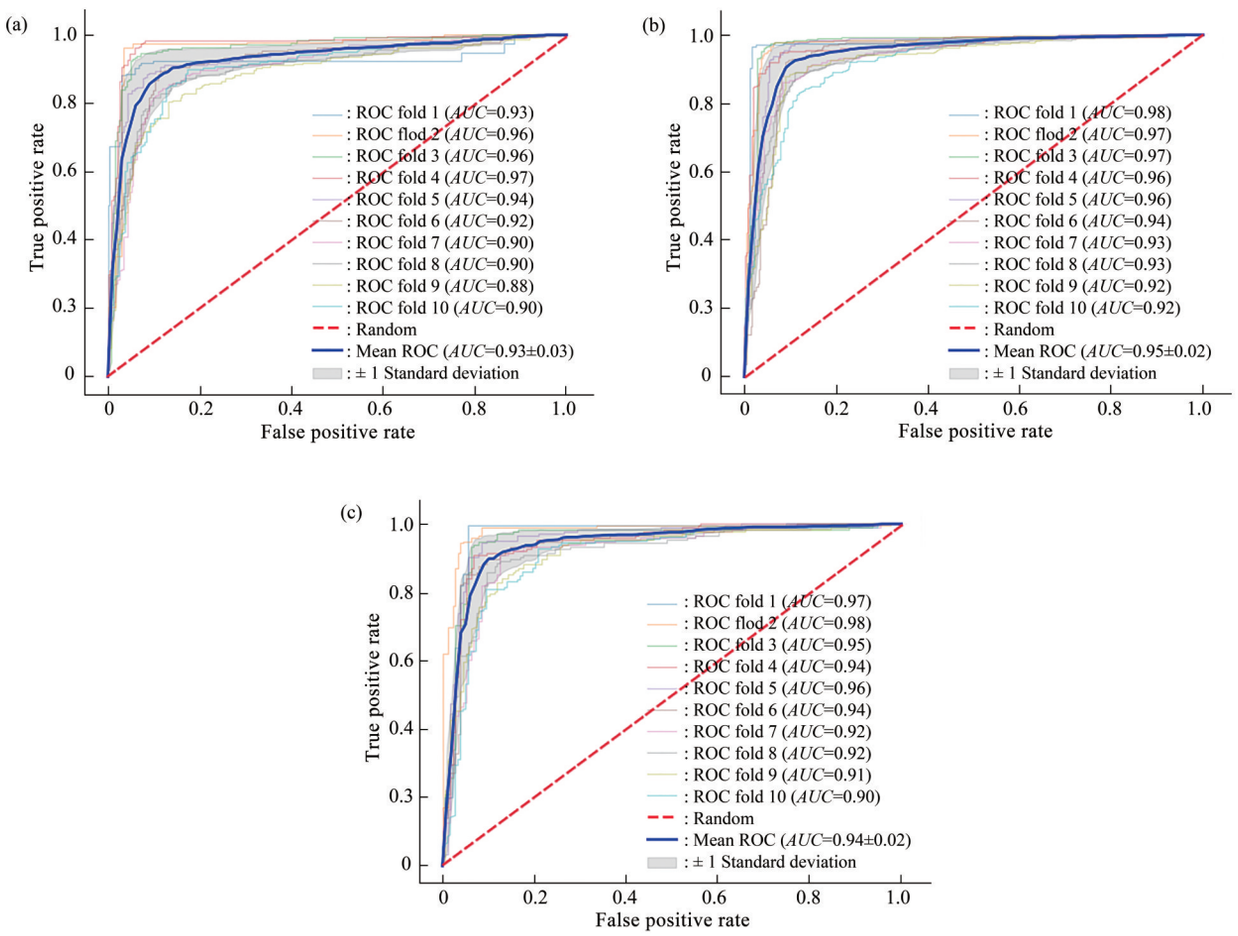

Fig. 5　The 10−fold cross−validation receiver operating characteristic（ROC）curves on the independent test datasets of rat tissues

(a−c) represent 10-fold cross-validation ROC curves on the rat brain (R_B), rat kidney (R_K) and rat liver (R_L) independent test datasets with our method, respectively. The horizontal axis represents false positive rate and the vertical axis represents true positive rate.

Based on the above analysis, it can be seen that the predicted *AUC* values ranged from (92±3)% to (95±2)% . That is, under 10-fold cross-validation, our method can stably predict m^{6}A methylation sites among different tissues.

3　Conclusion

Since m^{6}A plays an important role in many biological processes, the accurate prediction of m^{6}A methylation sites is an essential task in research on RNA methylation modification. Although a large number of state-of-the-art prediction methods for m^{6}A methylation sites have been developed in previous studies, most of them have widely varying predictive performance across different tissues.

In this work, based on a double layer bidirectional gate recurrent network, we developed a model that can simultaneously and effectively predict m^{6}A methylation sites in 11 mammalian tissues. The overall prediction performance of the proposed method was superior to that of the other state-of-the-art methods. For example, the proposed model achieved relatively excellent *ACC* or *AUC* values for each tissue, and the average *ACC* and *AUC* values on the independent test sets were 89.73% and 93.39%, respectively. Compared with the best model, M6A-BiNP, on the training datasets and independent test datasets, although the average *AUC* values of the proposed method were almost equal to those of M6A-BiNP, the average *ACC* values were increased by 3.45% and 8.46%, respectively. Compared with those of the remaining methods (iRNA-m6A, im6A-TS-CNN and DNN-m6A), the average *ACC* values on the

training datasets or independent test datasets were improved by 10.36%−19.13%, and the prediction *ACC* values were 87.27%−92.08%. Our method not only has excellent prediction performance but also has good generalizability. The source code and datasets in this study are freely available in the GitHub repositoryhttps://github.com/cph222/Predict-m6A-methylation-sites-a-double-layer-BiGRU.git.

Although the proposed method is capable of predicting m^6A methylation sites in 11 mammalian tissues, it is currently restricted to humans, mice and rats. It would be intriguing to test the performance of the proposed method on other species, such as *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. Even with the increase in biological data and the development of intelligent computing, it is necessary to establish a model that is applicable to more species, more tissues and even more RNA modification sites. In future studies, we will attempt to make efforts in this direction and establish a more generalized RNA modification site identification method.

References

[1] Hong K. Emerging function of N6-methyladenosine in cancer. Oncol Lett, 2018, **16**(5): 5519-5524

[2] Deng X, Su R, Feng X, *et al*. Role of N^6-methyladenosine modification in cancer. Curr Opin Genet Dev, 2018, **48**: 1-7

[3] Luo G Z, MacQueen A, Zheng G, *et al*. Unique features of the m6A methylome in *Arabidopsis thaliana*. Nat Commun, 2014, **5**: 5630

[4] Wei C M, Gershowitz A, Moss B. Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. Cell, 1975, **4**(4): 379-386

[5] Roundtree I A, Luo G Z, Zhang Z, *et al*. YTHDC1 mediates nuclear export of N^6-methyladenosine methylated mRNAs. Elife, 2017, **6**: e31311

[6] Xiao W, Adhikari S, Dahal U, *et al*. Nuclear m(6)A reader YTHDC1 regulates mRNA splicing. Mol Cell, 2016, **61**(4): 507-519

[7] Hsu P J, Zhu Y, Ma H, *et al*. Ythdc2 is an N^6-methyladenosine binding protein that regulates mammalian spermatogenesis. Cell Res, 2017, **27**(9): 1115-1127

[8] Zhang X, Li M J, Xia L, *et al*. The biological function of m6A methyltransferase KIAA1429 and its role in human disease. PeerJ, 2022, **10**: e14334

[9] Kellner S, Burhenne J, Helm M. Detection of RNA modifications. RNA Biol, 2010, **7**(2): 237-247

[10] Ma S Q, Peng J Y, Yi C Q. RNA modification detection technology. Chinese Science Bulletin, 2018, **30**(4): 440-446
马士清, 彭金英, 伊成器. 生命科学, 2018, **30**(4): 440-446

[11] Meyer K D, 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

[12] Schwartz S, Agarwala S D, Mumbach M R, *et al*. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. Cell, 2013, **155**(6): 1409-1421

[13] Chen W, Feng P, Ding H, *et al*. iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal Biochem, 2015, **490**: 26-33

[14] Liu Z, Xiao X, Yu D J, *et al*. pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences *via* physical-chemical properties. Anal Biochem, 2016, **497**: 60-67

[15] 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

[16] Xing P, Su R, Guo F, *et al*. Identifying N^6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. Sci Rep, 2017, **7**: 46757

[17] Wei L, Chen H, Ran S. M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. Mol Ther Nucleic Acids, 2018, **12**: 635-644

[18] Akbar S, Hayat M. iMethyl-STTNC: identification of N^6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. J Theor Biol, 2018, **455**: 205-211

[19] Chen W, Ding H, Zhou X, *et al*. iRNA(m6A)-PseDNC: identifying N^6-methyladenosine sites using pseudo dinucleotide composition. Anal Biochem, 2018, **561-562**: 59-65

[20] 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

[21] Fang F, Wang X, Li Z, *et al*. Epigenetic regulation of mRNA N6-methyladenosine modifications in mammalian gametogenesis. Mol Hum Reprod, 2021, **27**(5): gaab025

[22] Ju W, Liu K, Ouyang S, *et al*. Changes in N6-methyladenosine modification modulate diabetic cardiomyopathy by reducing myocardial fibrosis and myocyte hypertrophy. Front Cell Dev Biol, 2021, **9**: 702579

[23] Han W, Wang S, Qi Y, *et al*. Targeting HOTAIRM1 ameliorates glioblastoma by disrupting mitochondrial oxidative phosphorylation and serine metabolism. iScience, 2022, **25**(8): 104823

[24] Cheng W, Liu F, Ren Z, *et al*. Parallel functional assessment of m^6A sites in human endodermal differentiation with base editor screens. Nat Commun, 2022, **13**(1): 478

[25] Zhang Q, Zhang Y, Chen H, *et al*. METTL3-induced DLGAP1-AS2 promotes non-small cell lung cancer tumorigenesis through m^6A/c-Myc-dependent aerobic glycolysis. Cell Cycle, 2022, **21**(24): 2602-2614

[26] Dao FY, Lv H, Yang YH, *et al*. Computational identification of N6-Methyladenosine sites in multiple tissues of mammals. Comput Struct Biotechnol J, 2020, **18**: 1084-1091

[27] Wang M, Xie J, Xu S. M6A-BiNP: predicting N^6-methyladenosine

sites based on bidirectional position-specific propensities of polynucleotides and pointwise joint mutual information. RNA Biol, 2021, **18**(12): 2498-2512

[28] 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

[29] 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

[30] Cho K, van Merrienboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. EMNLP//Moschitti A, Pang B, Daelemans. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1724-1734

[31] Hanley J A. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging, 1989, **29**(3): 307-335

[32] Basith S, Manavalan B, Hwan Shin T, *et al*. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. Med Res Rev, 2020, **40**(4): 1276-1314

[33] Wu Y, He K. Group normalization. Int J Comput Vis, 2020, **128**(3): 742-755

[34] Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982, **143**(1): 29-36

[35] Chen W, Lv H, Nie F, *et al*. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics, 2019, **35**(16): 2796-2800

[36] Lv H, Dao F Y, Guan Z X, *et al*. iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. Front Genet, 2019, **10**: 793

基于双层BiGRU网络的哺乳动物组织m^{6}A甲基化位点预测*

李慧敏** 陈鹏辉 唐 轶** 徐权峰 胡 梦 王 煜
（云南民族大学数学与计算机科学学院，昆明 650504）

摘要　目的　N^6-甲基化腺苷（N^6-methyladenosine，m^{6}A）是RNA中最常见、最丰富的化学修饰，在很多生物过程中发挥着重要作用。目前已经发展了一些预测m^{6}A甲基化位点的计算方法。然而，这些方法在针对不同物种或不同组织时，缺乏稳健性。为了提升对不同组织中m^{6}A甲基化位点预测的稳健性，本文提出一种能结合序列反向信息来提取数据更高级特征的双层双向门控循环单元（bidirectional gated recurrent unit，BiGRU）网络模型。**方法**　本文选取具有代表性的哺乳动物组织m^{6}A甲基化位点数据集作为训练数据，通过对模型网络、网络结构、层数和优化器等进行搭配，构建双层BiGRU网络。**结果**　将模型应用于人类、小鼠和大鼠共11个组织的m^{6}A甲基化位点预测上，并与其他方法在这11个组织上的预测能力进行了全面的比较。结果表明，本文构建的模型平均预测接受者操作特征曲线下面积（area under the receiver operating characteristic curve，*AUC*）达到93.72%，与目前最好的预测方法持平，而预测准确率（accuracy，*ACC*）、敏感性（sensitivity，*SN*）、特异性（specificity，*SP*）和马修斯相关系数（Matthews correlation coefficient，*MCC*）分别为90.07%、90.30%、89.84%和80.17%，均高于目前的m^{6}A甲基化位点预测方法。**结论**　和已有研究方法相比，本文方法对11个哺乳动物组织的m^{6}A甲基化位点的预测准确性均达到最高，说明本文方法具有较好的泛化能力。

关键词　N^6-甲基化腺苷位点，双向门控循环单元，碱基序列，深度学习
中图分类号　TP391，Q52　　**DOI**：10.16476/j.pibb.2023.0011

* 国家自然科学基金（61866040）资助项目。
** 通讯联系人。
李慧敏 Tel：13888183685，E-mail：lihuimin_1980@126.com
唐轶 Tel：18314589836，E-mail：yitang.math@gmail.com
收稿日期：2023-01-09，接受日期：2023-05-15