6mA-Pred: identifying DNA N6-methyladenine sites based on deep learning

Qianfei Huang[1,*], Wenyang Zhou[2,*], Fei Guo[1], Lei Xu[3] and Lichao Zhang[4]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] School of Life Science and Technology, Harbin Institute of Technology, Harbin, China
[3] School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China
[4] School of Intelligent Manufacturing and Equipment, Shenzhen Institute of Information Technology, Shenzhen, China
* These authors contributed equally to this work.

ABSTRACT

With the accumulation of data on 6mA modification sites, an increasing number of scholars have begun to focus on the identification of 6mA sites. Despite the recognized importance of 6mA sites, methods for their identification remain lacking, with most existing methods being aimed at their identification in individual species. In the present study, we aimed to develop an identification method suitable for multiple species. Based on previous research, we propose a method for 6mA site recognition. Our experiments prove that the proposed 6mA-Pred method is effective for identifying 6mA sites in genes from taxa such as rice, *Mus musculus*, and human. A series of experimental results show that 6mA-Pred is an excellent method. We provide the source code used in the study, which can be obtained from http://39.100.246.211:5004/6mA_Pred/.

Subjects Bioinformatics, Computational Biology, Data Mining and Machine Learning
Keywords 6mA, LSTM, Attention

Submitted 28 October 2020
Accepted 30 December 2020
Published 3 February 2021

Corresponding authors
Lei Xu, csleixu@szpt.edu.cn
Lichao Zhang, lczhang5354@szu.edu.cn

Academic editor
Joseph Gillespie

Additional Information and Declarations can be found on page 10

DOI **10.7717/peerj.10813**

INTRODUCTION

DNA modification sites play vital roles in multiple biological processes and are attracting increasing research attention. Methylation continues to be a hot topic in epigenetics, and 5mC methylation has been extensively studied (*Liu, Li & Zuo, 2019*). With the advancement of sequencing technology, 6mA methylation has slowly attracted increasing attention. 6mA methylation not only affects gene expression but also regulates development in plants and animals (*Xu et al., 2020a*). Many diseases, including cancer, are related to 6mA methylation (*Chen et al., 2019a*, *2019b*; *Xu et al., 2019a*). With the progress of 6mA methylation-related research, large amounts of data have been collected. However, effective methods for 6mA site identification are lacking.

Methods for identifying modification sites have consistently been a hot spot in bioinformatics. Many methods have been studied and have achieved good results. Although research on 4mC (*He, Jia & Zou, 2019*) and 5mC is mature, research on the identification of 6mA modification sites has just begun. The computational method

i6mA-Pred was used to identify 6mA modification sites in the rice genome with high accuracy. Several methods for identifying 6mA loci in the rice genome have been proposed, such as MM-6mAPred, iDNA-6mA-rice (*Hao et al., 2019*), SDM6A (*Basith et al., 2019*), i6mA-DNCP (*Kong & Zhang, 2019*) and SNNRice6mA (*Yu & Dai, 2019*). In addition, methods for the identification of 6mA sites in *Mus musculus* and humans have gradually emerged, such as iDNA6mA-PseKNC (*Feng et al., 2019*), csDMA (*Liu et al., 2019c*), SICD6mA, and 6mA-Finder (*Xu et al., 2020b*). Several datasets are publicly available, and many desirable features and models have been proposed. Application of the feature algorithms NCP and one-hot, feature fusion and deep learning methods has greatly accelerated the identification of 6mA-modified sites. Among the employed algorithms, SVM and RF exhibit stable performance and perform well on some datasets (*Liu, Gao & Zhang, 2019*; *Shen, Tang & Guo, 2019*; *Sun et al., 2020*; *Wang et al., 2020a*, *2020b*; *Yan et al., 2020*; *Zhou et al., 2018*, *2017*). In addition, the Markov model has achieved excellent results in predicting 6mA sites in the rice genome. In the application of feature methods, most researchers use multiple feature fusion methods and analyze various features. In general, the different methods have achieved good results and provided direction for subsequent research.

In the research mentioned above, most methods have employed machine learning (*Patil & Chouhan, 2019*; *Zou, 2019*; *Zou & Ma, 2019*) and detailed analysis of different feature methods. There are some good models that use deep learning methods, such as SNNRice6mA and SICD6mA. SNNRice6mA employs CNN (*Ren et al., 2019*) to build a network that works well. SICD6mA uses GRU to achieve a good network structure and has been applied extensively to datasets of two species. In this article, through a summary of the previous research work, we found that LSTM+Attention can identify the modification sites very well, and a large number of experimental results suggest that this is a very good method.

MATERIALS AND METHODS

Datasets

Much research has aimed to identify 6mA sites in rice. In reviewing research from the past 2 years, we found that the amount of data on 6mA sites is increasing. We obtained datasets for three species. The first dataset is a rice dataset obtained from 6mA-RicePred (*Huang et al., 2020b*). This dataset was first used in i6mA-Pred (*Chen et al., 2019c*) and was provided by the author (*Hu et al., 2019*). The second dataset is a *Mus musculus* dataset obtained from iDNA-PseKNC, and it has achieved good results with this dataset. The third dataset is a human dataset obtained from SICD6mA and is the largest of the three datasets. Table 1 provides a summary of each dataset. The lengths of their sequences are all the same: 41 bp. Details of these datasets are provided in their source papers. We have organized the datasets, which can be obtained from https://github.com/huangqianfei0916/6ma-rice.

All three data sets use CD-HIT to remove redundancy. Sequences with the similarity above 80% were excluded by using the CD-HIT program. all negative samples were 41 bp

Table 1 All datasets.

Dataset	Positive	Negative	Total
Rice	154,000	154,000	308,000
Mus musculus	1,934	1,934	3,868
Human_Train	491,885	491,885	983,770
Human_Test	122,971	122,971	245,942

in length and the center was A, but not being detected by the SMRT sequencing technology as of 6mA. Moreover the rice dataset collected negative samples based on the ratio of GAGG, AGG and AG motifs in the positive samples. the mouse dataset removed positive samples with modQV greater than 30.

Feature encoding and classification algorithms

One-hot encoding has been used by many researchers for sequence processing with good results (*Cheng, 2019*; *Cheng et al., 2018a*; *Li et al., 2020*; *Liu & Li, 2019*; *Zou et al., 2019*). One-hot encoding encodes each nucleotide separately. A disadvantage of one-hot is the lack of timing. Therefore, we used Kmer word segmentation instead of one-hot to capture the relationship between bases (*Zuo et al., 2017*). The role of Kmer was to help Embedding generate better word vectors. We investigated both normal word segmentation and Kmer word segmentation, and the experimental results showed that Kmer word segmentation achieved superior performance. Figure 1 shows the process of Kmer word segmentation. Our test for the selection of the k value revealed three to be the most suitable value. the experimental results are shown in Fig. 2. When k is 3, the dictionary size is 64; this is not a large parameter. In the feature extraction stage, the embedding layer is used to extract features. we chose the init method for our experiment. The effect of using init or fine-tune is almost the same, and in some cases, the init method is superior. If there is an excellent pretrained model, it is also a good choice. The quality of the features largely determines the effect of the model. Embedding is a very important module in deep learning, and word2vec is one of the best embedding methods. The encoding of features can be learned dynamically, and a method of secondary learning called finetune can be achieved in deep learning. In this paper, we use simple Init embedding and Kmer word segmentation.

Most methods currently employed for 6mA site recognition are machine learning methods, and most of them are only effective for a single species (*Cheng, 2019*; *Cheng et al., 2019*). In reviewing the latest research, we found that there are many similarities between the attention mechanism and the recognition of 6mA sites. Furthermore, LSTM has achieved excellent performance in dealing with sequence problems (*Huang et al., 2020a*). In constructing the model, we did not adopt a particularly complex structure, and the complexity and effect of the model are not directly related. After feature extraction with the embedding layer, bidirectional LSTM is used to process the sequence features (*Xia et al., 2019*). The sequence information obtained after LSTM processing can be used to obtain a good feature vector, and this feature is a representation of the overall sequence

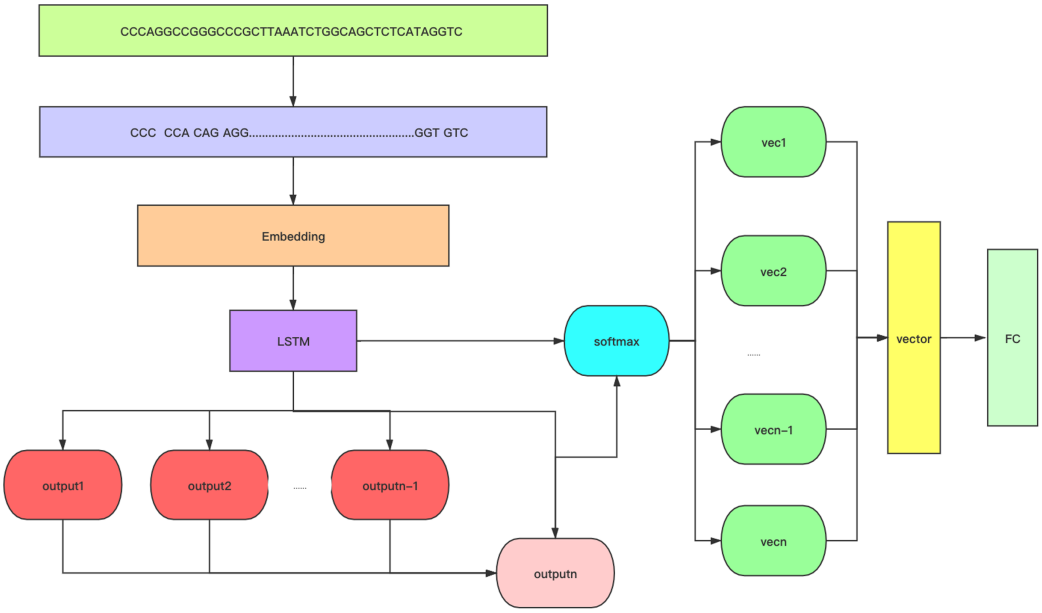

Figure 1 A flow chart of the structure of 6mA-Pred. 6mA-Pred includes kmer word segmentation and attention mechanism. Among them, the attention score uses the dot product method. Optimize features through the attention mechanism. Full-size DOI: 10.7717/peerj.10813/fig-1

information. Each time step of LSTM has an output that represents the sequence information up to the current time. The LSTM algorithm can be formulated as follows:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{1}$$

$$f_t = \sigma\left(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}\right) \tag{2}$$

$$g_t = \tanh\left(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}\right) \tag{3}$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{4}$$

$$c_t = f_t * c_{t-1} + i_t * g_t \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

In general, LSTM can be used to obtain an output at each time step and obtain a feature containing the sequence information (*Liu, Li & Yan, 2020*). We can analyze these features to obtain our expected results. The typical approach is to average this information or take the last one and then apply the fully connected layer to obtain the result. Many scholars have added other layers after LSTM to obtain good features. However, the design of these levels of network structure varies according to the specific application scenarios and problems. 6mA-Pred applies the attention mechanism to the output of LSTM and connects the fully connected layer after the attention layer.

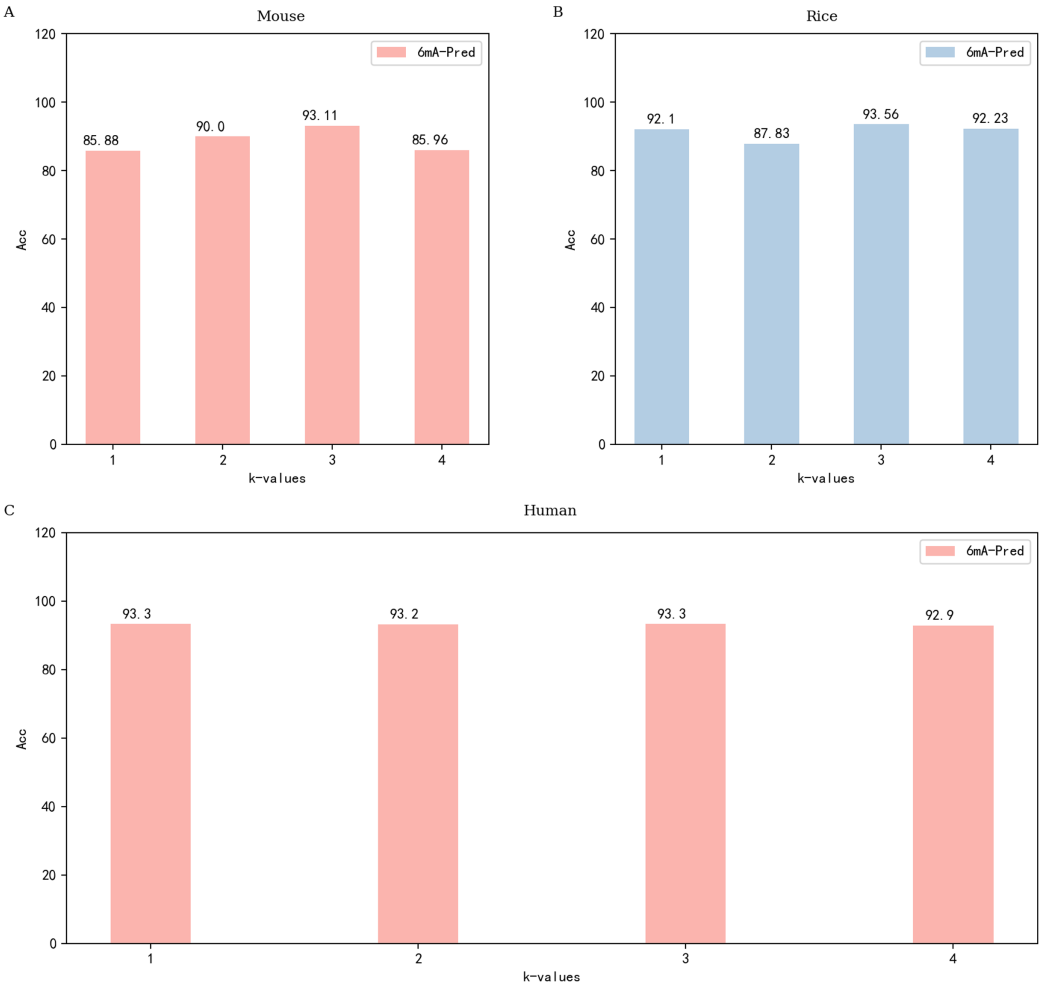

Figure 2 Performance of 6mA-Pred evaluated via independent testing based on different *k*-values. (A) The performance of different *k* values based on the mouse data set; (B) performance of different *k* values based on the rice data set; (C) performance of different *k* values based on the human data set.
Full-size DOI: 10.7717/peerj.10813/fig-2

The attention layer is added after the LSTM, and the output of the LSTM is analyzed with attention. The inner output of the final output of LSTM and the results of the previous time step can be used to generate the corresponding attention score. then, the Softmax layer is added to the attention layer to obtain the weight. The output of LSTM and this weight are weighted to obtain the final context vector. The last layer of the network is the fully connected layer, and this layer can obtain the probability of each category. Figure 1 shows the structure of the entire network and describes the Kmer word segmentation and attention mechanism. The attention mechanism adopted by 6mA-Pred is not complicated and acts directly on the output of LSTM. The purpose of 6mA-Pred is to obtain the final feature through the difference between global information and local information. We know that the feature corresponding to the sequence containing the modification site is very different from the feature corresponding to the sequence not

Table 2 The parameters of each experiment.

Experiment	lr	hidden_dim	dropout	Bach_size
Fig. 2	0.001	100	0.3	64
Table 3	0.001	100	0.3	64
Table 4-cv	0.005	100	0.3	64
Table 4-ind	0.005	100	0.3	64
Table 5	0.001	100	0.3	64

containing the modification site. Because of the differences, their final context vectors differ. We used the inner product method to obtain the attention score to reflect the intersection of global information and local information. The inner product is not the only option; other operations are possible. Self-attention in Transformer is also a good choice, but the network structure of the model is more complicated. Dot product can get the intersection between different sequences. 6mA-Pred uses this structure to increase the amount of local information in the final feature.

PERFORMANCE EVALUATION

A good model evaluation standard is crucial for assessing the utility of a model. Different indicators can be used to reveal the advantages and disadvantages of a model from different perspectives. Sensitivity (Sn), specificity (Sp), accuracy (Acc), and Mathew's correlation coefficient (MCC) are used to evaluate models in machine learning (*Chu et al., 2019*; *Deng et al., 2020*; *Gong et al., 2019*; *Jin et al., 2019*; *Shan et al., 2019*; *Su et al., 2019a*, *2019b*; *Wei et al., 2018a*, *2018b*; *Xu et al., 2018a*, *2018b*, *2018c*; *Zhang et al., 2019a*, *2019b*). These metrics are formulated as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{8}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{9}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}*\mathrm{TN} - \mathrm{FP}*\mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})*(\mathrm{TP} + \mathrm{FN})*(\mathrm{TN} + \mathrm{FP})*(\mathrm{TN} + \mathrm{FN})}} \tag{10}$$

TP, TN, FP and FN represent true positive, true negative, false positive, and false negative, respectively. Sn, Sp, Acc, and MCC can be calculated from these indicators. In addition, AUC (area under the ROC curve) was used to evaluate our model (*Cheng & Hu, 2018*; *Cheng et al., 2018b*; *Ding, Tang & Guo, 2019a*, *2019b*; *Shen et al., 2019*). For further experiments, Table 2 records the hyperparameters of the model.

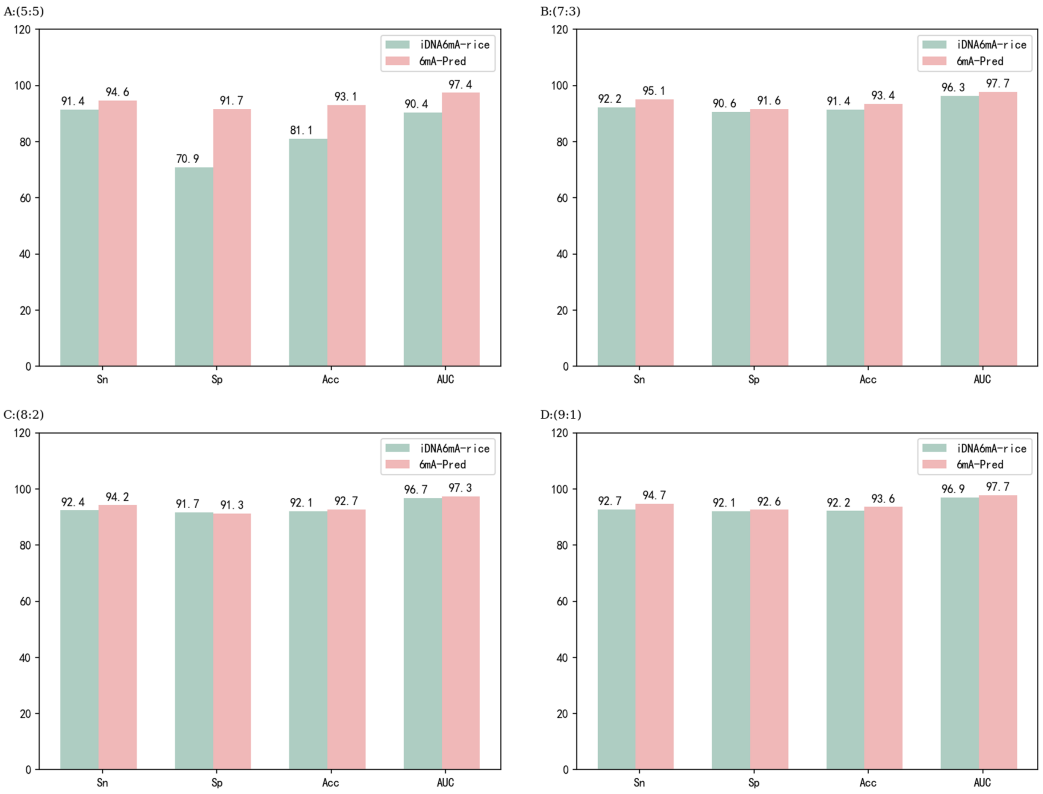

Figure 3 Predictive performance at different ratios for the rice dataset. (A–D) correspond to the performance of the model on different proportions of the rice dataset, respectively.
Full-size DOI: 10.7717/peerj.10813/fig-3

PERFORMANCE COMPARISON WITH DIFFERENT DATASETS

Methods for identifying sites in the rice genome include iDNA6mA-Rice and SNNRice6mA, which are excellent models. After comparing different features in feature extraction, the developers of iDNA6mA-Rice chose binary encoding, and they chose RF (random forest) for the classifier. Both the choice of feature method and the performance of the classifier are excellent. iDNA6mA-Rice was applied to various scale segmentation experiments on a rice dataset and achieved very good results. 6mA-Pred was applied in a similar experiment with the rice dataset. the results are shown in Fig. 3. The performance of 6mA-Pred was better than iDNA6mA-Rice at all ratios. However, iDNA6mA-Rice is also a very good model, and the performance difference between the two models was very small. SNNRice6mA also performs very well for rice genes. Unlike iDNA6mA-Rice, SNNRice6mA uses a deep learning model. SNNRice6mA uses one-hot in the feature encoding stage and has achieved good results. Regarding the overall network structure, SNNRice6mA uses a stack structure of CNN (convolutional neural networks). The network structure of SNNRice6mA was adjusted to derive SNNRice6mA-large, which also achieved good results. SNNRice6mA and SNNRice6mA-large were employed for five-fold cross-validation on the rice dataset. Table 3 shows the results of comparisons

Table 3 Performance comparison between 6mA-Pred and other methods via 5-fold cross validation based on the rice dataset.

Method	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
SNNRice6mA	93.67	86.74	90.20	0.81	0.96
SNNRice6mA-large	94.33	89.75	92.04	0.84	0.97
iDNA6mA-rice	93.00	90.50	91.70	0.84	0.96
6mA-Pred	95.66	92.38	94.02	0.88	0.981

Table 4 Performance of 6mA-Pred evaluated via 5-fold cross validation and independent testing based on the *Mus musculus* dataset.

Method	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
6mA-Pred-cv	93.8	98.5	96.1	0.92	0.981
6mA-Pred-ind	87.8	98.4	93.8	0.861	0.949
IDNA6mA-PseKNC	93.28	100	96.73	0.93	–

among the different models. The performance of 6mA-Pred was excellent compared to that of the other models.

The model also performed well on the *Mus musculus* dataset. iDNA6mA-PseKNC has achieved good results in predicting 6mA loci in the *Mus musculus* genome and uses machine learning methods for analysis. iDNA6mA-PseKNC uses NCP as the feature algorithm, and many experiments have been conducted for this feature. In addition, iDNA6mA-PseKNC employs the SVM classifier and achieved very good results. 6mA-Pred is also effective in identifying 6mA sites in the *Mus musculus* genome. In this study, two experiments were conducted with 6mA-Pred, one involving five-fold cross-validation on the dataset, and one involving independent testing by splitting the dataset. Table 4 shows the results of these two experiments and the results for iDNA6mA-PseKNC. iDNA6mA-PseKNC was evaluated via the jackknife test; for deep learning methods, leave-one-out cross-validation is time consuming and not representative. For evaluation of 6mA-Pred, five-fold cross-validation (*Fang et al., 2019*; *He et al., 2018a*; *Liu, 2019*; *Xiong et al., 2018*; *Xu et al., 2019b*; *Zhu et al., 2019*) and segmentation of the dataset were employed. As shown in Table 4, the performance of 6mA-Pred remained good.

Among the models used for identifying the 6mA sites of human genes, SICD6mA is currently the best model. SICD6mA is a deep learning model and uses GRU as the basic unit. SICD6mA performs well not only for human genes but also for rice genes. The developers of SICD6mA contributed data and performed extensive data processing. We used the training set and test set provided by SICD6mA's developers for our experiments. SICD6mA does not use one-hot for encoding; rather, it uses 3-mer. Two basic units, BGRU and UGRU, are used in the network model structure, and a two-layer fully connected layer and a Softmax layer are used to improve the network. The experimental results revealed that the performance of SICD6mA was very good. Table 5 shows the experimental results for 6mA-Pred, which were very similar to the

Table 5 Performance of 6mA-Pred evaluated via independent testing based on the human dataset.

Method	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
6mA-Pred-ind	93.28	94.2	93.34	0.87	0.98
SICD6mA	93.33	95.00	93.66	0.874	–

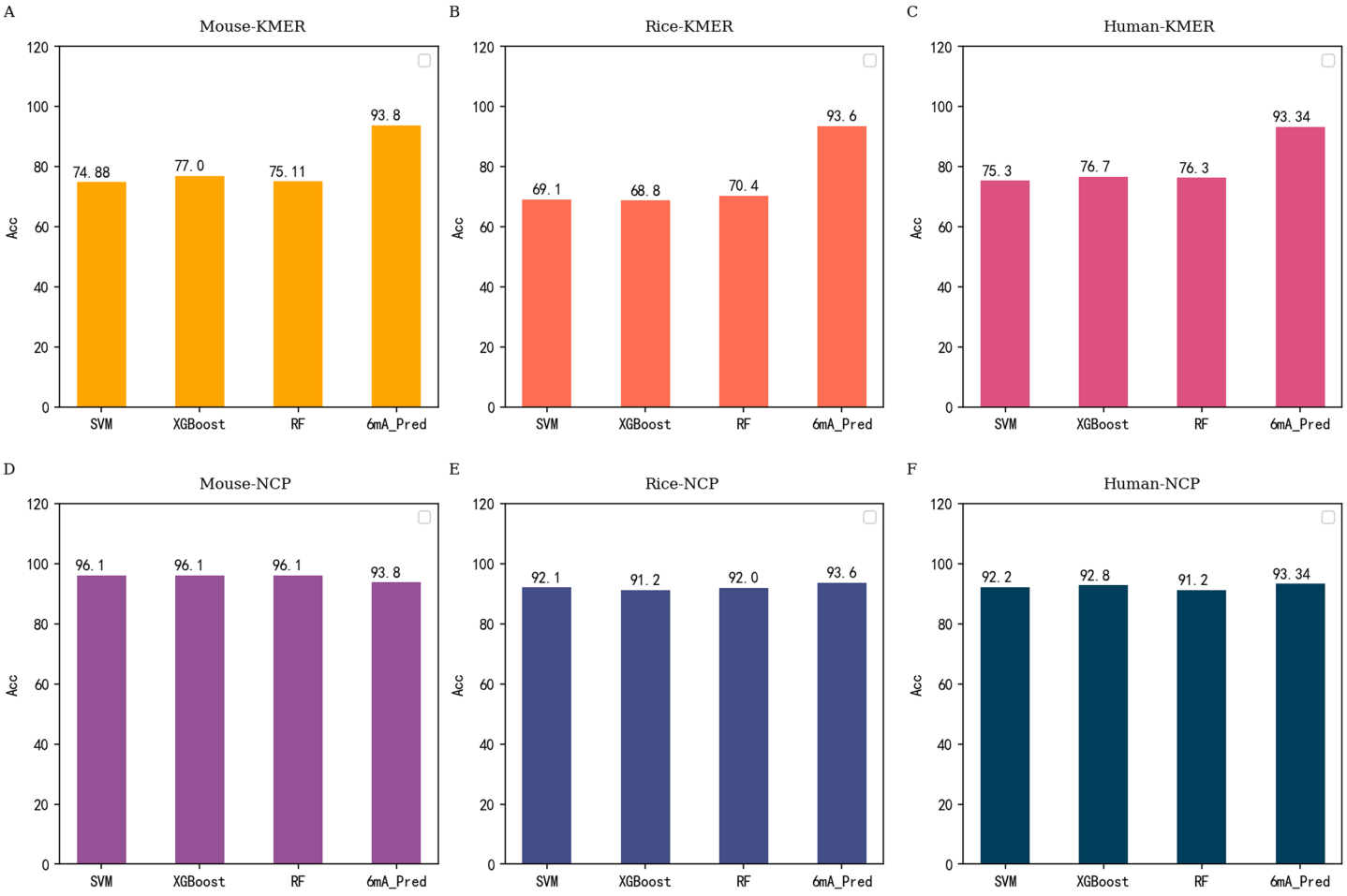

Figure 4 Performance comparison between 6mA-Pred and other machine learning methods independent testing based on all datasets. (A–C) correspond to the performance of commonly used machine learning models under KMER features of different species, respectively. (D–F) are the resulting contrasts under NCP features. Full-size DOI: 10.7717/peerj.10813/fig-4

SICD6mA results. These findings proved that 6mA-Pred is very effective in identifying 6mA sites in human genes.

According to the previous conclusions, we conducted related experiments on traditional machine learning methods. NCP and KMER were used in experiments as excellent feature extraction methods. SVM, RF and XGB were excellent algorithms and performed well in previous studies. Therefore, we use them to carry out further experiments. the experimental results are shown in Fig. 4.

CONCLUSION

Through the analysis of current studies and the performance of a large number of experimental comparisons, we found that 6mA-Pred is an effective method for identifying 6mA sites. LSTM performs well in processing sequence features and can obtain good features. In addition, the attention mechanism we used is effective for identifying 6mA sites. The combination of LSTM and Attention mechanism can produce a theoretically excellent model, and the experiment proves that this conclusion is correct. Related methods will be considered for RNA and protein modification prediction (*Dou et al., 2020*; *He, Wei & Zou, 2018*; *Huang & Li, 2018*) in the future.

The previous studies on this topic are excellent and provide theoretical and experimental support for our research. The attention mechanism in 6mA-Pred can be improved; for example, self-attention or a combination of two attention mechanisms could be used to obtain a better context vector. It is also possible to use a combination of CNN and attention mechanism to obtain an excellent method (*Su et al., 2014*). These possibilities warrant investigation.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Natural Science Foundation of China (No. 61902259) and the Natural Science Foundation of Guangdong province (grant no. 2018A0303130084). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Natural Science Foundation of China: 61902259.
Natural Science Foundation of Guangdong province: 2018A0303130084.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Qianfei Huang conceived and designed the experiments, performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Wenyang Zhou conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Fei Guo analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Lei Xu analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

- Lichao Zhang analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Raw data is available at GitHub: https://github.com/huangqianfei0916/6ma-rice.

Code is also available at GitHub: https://github.com/huangqianfei0916/Attention_Classification/tree/master/lstm_attention.

A Web Server For Predicting 6mA Sites is available at: http://39.100.246.211:5004/6mA_Pred/.

REFERENCES

Basith S, Manavalan B, Shin TH, Lee G. 2019. SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Molecular Therapy-Nucleic Acids* **18**:131–141 DOI 10.1016/j.omtn.2019.08.011.

Chen J, Han G, Xu A, Cai H. 2019a. Identification of multidimensional regulatory modules through multi-graph matching with network constraints. *IEEE Transactions on Biomedical Engineering* **67(4)**:987–998 DOI 10.1109/TBME.2019.2927157.

Chen J, Peng H, Han G, Cai H, Cai J. 2019b. HOGMMNC: a higher order graph matching with multiple network constraints model for gene-drug regulatory modules identification. *Bioinformatics* **35(4)**:602–610 DOI 10.1093/bioinformatics/bty662.

Chen W, Lv H, Nie F, Lin H. 2019c. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* **35(16)**:2796–2800 DOI 10.1093/bioinformatics/btz015.

Cheng L. 2019. Computational and biological methods for gene therapy. *Current Gene Therapy* **19(4)**:210 DOI 10.2174/1566523219041910221133O7.

Cheng L, Hu Y. 2018. Human disease system biology. *Current Gene Therapy* **18(5)**:255–256 DOI 10.2174/1566523218666181010101114.

Cheng L, Hu Y, Sun J, Zhou M, Jiang Q. 2018a. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* **34(11)**:1953–1956 DOI 10.1093/bioinformatics/bty002.

Cheng L, Zhao H, Wang P, Zhou W, Luo M, Li T, Han J, Liu S, Jiang Q. 2019. Computational methods for identifying similar diseases. *Molecular Therapy Nucleic Acids* **18**:590–604 DOI 10.1016/j.omtn.2019.09.019.

Cheng L, Zhuang H, Yang S, Jiang H, Wang S, Zhang J. 2018b. Exposing the causal effect of c-reactive protein on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Frontiers in Genetics* **9**:657 DOI 10.3389/fgene.2018.00657.

Chu Y, Kaushik AC, Wang X, Wang W, Zhang Y, Shan X, Salahub DR, Xiong Y, Wei DQ. 2019. DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Briefings in Bioinformatics* **16(Suppl. 3)**:19 DOI 10.1093/bib/bbz152.

Deng Y, Xu X, Qiu Y, Xia J, Zhang W, Liu S. 2020. A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics* **36(15)**:4316–4322 DOI 10.1093/bioinformatics/btaa501.

Ding Y, Tang J, Guo F. 2019a. Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* **325**:211–224 DOI 10.1016/j.neucom.2018.10.028.

Ding Y, Tang J, Guo F. 2019b. Identification of drug-side effect association via semisupervised model and multiple kernel learning. *IEEE Journal of Biomedical and Health Informatics* **23(6)**:2619–2632 DOI 10.1109/JBHI.2018.2883834.

Dou LJ, Li XL, Ding H, Xu L, Xiang HK. 2020. Is there any sequence feature in the rna pseudouridine modification prediction problem? *Molecular Therapy-Nucleic Acids* **19**:293–303 DOI 10.1016/j.omtn.2019.11.014.

Fang T, Zhang Z, Sun R, Zhu L, He J, Huang B, Xiong Y, Zhu X. 2019. RNAm5CPred: prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition. *Molecular Therapy—Nucleic Acids* **18**:739–747 DOI 10.1016/j.omtn.2019.10.008.

Feng P, Yang H, Ding H, Lin H, Chen W, Chou K-C. 2019. iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* **111(1)**:96–102 DOI 10.1016/j.ygeno.2018.01.005.

Gong Y, Niu Y, Zhang W, Li X. 2019. A network embedding-based multiple information integration method for the MiRNA-disease association prediction. *BMC Bioinformatics* **20(1)**:468 DOI 10.1186/s12859-019-3063-3.

Hao L, Dao F-Y, Guan Z-X, Zhang D, Tan J-X, Zhang Y, Chen W, Lin H. 2019. iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Frontiers in Genetics* **10**:793 DOI 10.3389/fgene.2019.00793.

He J, Fang T, Zhang Z, Huang B, Zhu X, Xiong Y. 2018a. PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics* **19(1)**:306 DOI 10.1186/s12859-018-2321-0.

He W, Jia C, Zou Q. 2019. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* **35(4)**:593–601 DOI 10.1093/bioinformatics/bty668.

He W, Wei L, Zou Q. 2018. Research progress in protein post-translational modification site prediction. *Briefings in Functional Genomics* **18(4)**:220–229 DOI 10.1093/bfgp/ely039.

Hu B, Zheng L, Long C, Song M, Li T, Yang L, Zuo Y. 2019. EmExplorer: a database for exploring time activation of gene expression in mammalian embryos. *Open Biology* **9(6)**:190054 DOI 10.1098/rsob.190054.

Huang GH, Li JC. 2018. Feature extractions for computationally predicting protein post-translational modifications. *Current Bioinformatics* **13(4)**:387–395 DOI 10.2174/1574893612666170707094916.

Huang Q, Zhang Y, Peng H, Dan T, Weng W, Cai H. 2020a. Deep subspace clustering to achieve jointly latent feature extraction and discriminative learning. *Neurocomputing* **404**:340–350 DOI 10.1016/j.neucom.2020.04.120.

Huang QF, Zhang J, Guo F, Zou Q. 2020b. 6mA-RicePred: a method for identifying DNA N6-methyladenine sites in the rice genome based on feature fusion. *Frontiers in Plant Science* **11**:4 DOI 10.3389/fpls.2020.00004.

Jin Q, Meng Z, Pham TD, Chen Q, Wei L, Su R. 2019. DUNet: a deformable network for retinal vessel segmentation. *Knowledge-Based Systems* **178**:149–162 DOI 10.1016/j.knosys.2019.04.025.

Kong L, Zhang L. 2019. i6mA-DNCP: computational identification of DNA N6-methyladenine sites in the rice genome using optimized dinucleotide-based features. *Genes* **10(10)**:828 DOI 10.3390/genes10100828.

Li J, Pu Y, Tang J, Zou Q, Guo F. 2020. DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides. *IEEE Journal of Biomedical and Health Informatics* **1(10)**:1 DOI 10.1109/JBHI.2020.2977091.

Liu B. 2019. BioSeq-analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Briefings in Bioinformatics* **20(4)**:1280–1294 DOI 10.1093/bib/bbx165.

Liu B, Gao X, Zhang H. 2019. BioSeq-analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Research* **47(20)**:e127 DOI 10.1093/nar/gkz740.

Liu B, Li C, Yan K. 2020. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Briefings in Bioinformatics* **21(5)**:1733–1741 DOI 10.1093/bib/bbz098.

Liu B, Li K. 2019. iPromoter-2L2.0: identifying promoters and their types by combining Smoothing Cutting Window algorithm and sequence-based features. *Molecular Therapy-Nucleic Acids* **18**:80–87 DOI 10.1016/j.omtn.2019.08.008.

Liu D, Li G, Zuo Y. 2019. Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Briefings in Bioinformatics* **20(5)**:1826–1835 DOI 10.1093/bib/bby053.

Liu Z, Dong W, Jiang W, He Z. 2019c. csDMA: an improved bioinformatics tool for identifying DNA 6 mA modifications via Chou's 5-step rule. *Scientific Reports* **9(1)**:1–9 DOI 10.1038/s41598-018-37186-2.

Patil K, Chouhan U. 2019. Relevance of machine learning techniques and various protein features in protein fold classification: a review. *Current Bioinformatics* **14(8)**:688–697 DOI 10.2174/1574893614666190204154038.

Ren F, Yang C, Qiu Q, Zeng N, Cai C, Hou C, Zou Q. 2019. Exploiting discriminative regions of brain slices based on 2D CNNs for alzheimer's disease classification. *IEEE Access* **7**:181423–181433 DOI 10.1109/ACCESS.2019.2920241.

Shan X, Wang X, Li CD, Chu Y, Zhang Y, Xiong Y, Wei DQ. 2019. Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method. *Journal of Chemical Information and Modeling* **59(11)**:4577–4586 DOI 10.1021/acs.jcim.9b00749.

Shen Y, Ding Y, Tang J, Zou Q, Guo F. 2019. Critical evaluation of web-based prediction tools for human protein subcellular localization. *Briefings in Bioinformatics* **21(5)**:1628–1640 DOI 10.1093/bib/bbz106.

Shen Y, Tang J, Guo F. 2019. Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. *Journal of Theoretical Biology* **462**:230–239 DOI 10.1016/j.jtbi.2018.11.012.

Su J, Wang Y, Xing X, Liu J, Zhang Y. 2014. Genome-wide analysis of DNA methylation in bovine placentas. *BMC Genomics* **15(1)**:12 DOI 10.1186/1471-2164-15-12.

Su R, Liu X, Wei L, Zou Q. 2019a. Deep-resp-forest: a deep forest model to predict anti-cancer drug response. *Methods* **166**:91–102 DOI 10.1016/j.ymeth.2019.02.009.

Su R, Wu H, Xu B, Liu X, Wei L. 2019b. Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **16(4)**:1231–1239 DOI 10.1109/TCBB.2018.2858756.

Sun J, Zhang Z, Bao S, Yan C, Hou P, Wu N, Su J, Xu L, Zhou M. 2020. Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. *Journal for ImmunoTherapy of Cancer* **8(1)**:e000110 DOI 10.1136/jitc-2019-000110.

Wang H, Ding Y, Tang J, Guo F. 2020a. Identification of membrane protein types via multivariate information fusion with Hilbert—Schmidt Independence Criterion. *Neurocomputing* **383**:257–269 DOI 10.1016/j.neucom.2019.11.103.

Wang Z, He W, Tang J, Guo F. 2020b. Identification of highest-affinity binding sites of yeast transcription factor families. *Journal of Chemical Information and Modeling* **60(3)**:1876–1883 DOI 10.1021/acs.jcim.9b01012.

Wei L, Hu J, Li F, Song J, Su R, Zou Q. 2018a. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Briefings in Bioinformatics* **55(8)**:165 DOI 10.1093/bib/bby107.

Wei L, Zhou C, Chen H, Song J, Su R. 2018b. ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **34**:4007–4016.

Xia W, Xu J, Yu G, Yao G, Xu K, Ma X, Zhang N, Liu B, Li T, Lin Z, Chen X, Li L, Wang Q, Shi D, Shi S, Zhang Y, Song W, Jin H, Hu L, Bu Z, Wang Y, Na J, Xie W, Sun YP. 2019. Resetting histone modifications during human parental-to-zygotic transition. *Science* **365(6451)**:353–360 DOI 10.1126/science.aaw5118.

Xiong Y, Wang Q, Yang J, Zhu X, Wei DQ. 2018. PredT4SE-Stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Frontiers in Microbiology* **9**:2571 DOI 10.3389/fmicb.2018.02571.

Xu A, Chen J, Peng H, Han G, Cai H. 2019a. Simultaneous interrogation of cancer omics to identify subtypes with significant clinical differences. *Frontiers in Genetics* **10**:236 DOI 10.3389/fgene.2019.00236.

Xu B, Liu D, Wang Z, Tian R, Zuo Y. 2020a. Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family. *Cellular and Molecular Life Sciences* **13(2)**:136 DOI 10.1007/s00018-020-03594-9.

Xu H, Hu R, Jia P, Zhao Z. 2020b. 6mA-Finder: a novel online tool for predicting DNA N6-methyladenine sites in genomes. *Bioinformatics* **36(10)**:3257–3259 DOI 10.1093/bioinformatics/btaa113.

Xu L, Liang G, Liao C, Chen G-D, Chang C-C. 2018a. An efficient classifier for alzheimer's disease genes identification. *Molecules* **23(12)**:3140 DOI 10.3390/molecules23123140.

Xu L, Liang G, Liao C, Chen G-D, Chang C-C. 2019b. k-Skip-n-Gram-RF: a random forest based method for alzheimer's disease protein identification. *Frontiers in Genetics* **10**:S168 DOI 10.3389/fgene.2019.00033.

Xu L, Liang G, Shi S, Liao C. 2018b. SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *International Journal of Molecular Sciences* **19(6)**:1773 DOI 10.3390/ijms19061773.

Xu L, Liang G, Wang L, Liao C. 2018c. A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* **9(3)**:158 DOI 10.3390/genes9030158.

Yan C, Zhang Z, Bao S, Hou P, Zhou M, Xu C, Sun J. 2020. Computational methods and applications for identifying disease-associated lncrnas as potential biomarkers and therapeutic targets. *Molecular Therapy—Nucleic Acids* **21**:156–171 DOI 10.1016/j.omtn.2020.05.018.

Yu H, Dai Z. 2019. SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Frontiers in Genetics* **10**:1071 DOI 10.3389/fgene.2019.01071.

Zhang W, Jing K, Huang F, Chen Y, Li B, Li J, Gong J. 2019a. SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug-drug interactions. *Information Sciences* **497**:189–201 DOI 10.1016/j.ins.2019.05.017.

Zhang W, Li Z, Guo W, Yang W, Huang F. 2019b. A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Piscataway: IEEE DOI 10.1109/TCBB.2019.2931546.

Zhou M, Hu L, Zhang Z, Wu N, Sun J, Su J. 2018. Recurrence-associated long non-coding rna signature for determining the risk of recurrence in patients with colon cancer. *Molecular Therapy—Nucleic Acids* **12**:518–529 DOI 10.1016/j.omtn.2018.06.007.

Zhou M, Zhao H, Xu W, Bao S, Cheng L, Sun J. 2017. Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Molecular Cancer* **16(1)**:16 DOI 10.1186/s12943-017-0580-4.

Zhu X, He J, Zhao S, Tao W, Xiong Y, Bi S. 2019. A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae. *Briefings in Functional Genomics* **18(10)**:367–376 DOI 10.1093/bfgp/elz018.

Zou Q. 2019. Latest machine learning techniques for biomedicine and bioinformatics. *Current Bioinformatics* **14(3)**:176–177 DOI 10.2174/1574893614031902201128855.

Zou Q, Ma Q. 2019. The application of machine learning to disease diagnosis and treatment. *Mathematical Biosciences* **320**:108305 DOI 10.1016/j.mbs.2019.108305.

Zou Q, Xing P, Wei L, Liu B. 2019. Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *rna* **25(2)**:205–218 DOI 10.1261/rna.069112.118.

Zuo Y, Li Y, Chen Y, Li G, Yan Z, Yang L. 2017. PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* **33(1)**:122–124 DOI 10.1093/bioinformatics/btw564.