SICD6mA: Identifying the 6mA Sites on Human and Rice Genomes using Deep Memory Network

Wenzhong Liu[1,*]

[1] School of Computer Science and Engineering, Sichuan University of Science & Engineering, Zigong, 643000, China;

* To whom correspondence should be addressed. Email: newwenzhong@hotmail.com

ABSTRACT

DNA N6-methyladenine (6mA) is a kind of epigenetic modification in prokaryotes and eukaryotes, which involves multiple biological processes, such as gene regulation and tumorigenesis. Identifying 6mA contributes to understanding its regulatory role. Therefore, to satisfy the needs of large-scale preliminary screening, it is necessary to develop the high-quality computational models for the rapid identification of 6mA sites. However, the existing calculation approaches are mostly specific to rice, and they have not been extensively applied to human genome. This study proposed a classification method of deep learning based on the memory mechanism named SICD6mA. In addition, the large benchmark datasets were constructed for human and rice, respectively, which integrated the recently reported 6mAsites. According to the evaluation results, SICD6mA displayed favorable robustness during cross-validations, which achieved the area under the curve (AUC) values of 0.9824 and 0.9903 for Human and Rice Nip in independent test set evaluation, separately. Besides, the successful prediction rate of 6mAsites on cross-species genomes exhibited higher accuracy than that of the state-of-the-art methods. For the convenience of experimental scientists, the user-friendly tool was developed to predict the cross-species 6mA sites, which allowed for the accelerate computations on the GPU platforms and might be downloaded freely from https://github.com/lwzyb/SICD6mA.

Keywords: Epigenetic Modification, DNA N6-methyladenine, 6mA, 3-mer Nucleic Acid, GRU Network

INTRODUCTION

DNA N6-methyladenine(6mA) refers to methylation at the N6site of adenine[1]. Typically, DNA 6mA modification is the most common DNA modification in prokaryotic genomes[2], which is once thought not to exist in eukaryotes[3] (including humans), because it is not detected in earlier studies. With the development of high-throughput sequencing technology, numerous 6mA modifications have been detected in different multicellular eukaryotes. In recent years, 6mA modification has been found to play an important role in DNA epigenetic modification of eukaryotes. DNA methylation represents a key element in epigenetic regulation, and DNA 6mA modification exerts a vital part in regulating gene expression[4], germ cell differentiation[5] etc. It has been reported in the genomes of ***E. coli***[6], ***Drosophila melanogaster***[7], ***rice***[8], and ***Arabidopsis***[9] etc. Currently, the 6mA sites can be detected in mouse and human cells using the sensitive detection technology. 6mA is extensively distributed throughout the human genome. Of note, [G/C]AGG[C/T] is the most frequent motif at the 6mA modification site[10]. Different types of cancer cells can be distinguished from normal cells based on the heterogeneities in methylation of the CpG islands. It is of great significance to identify the potential function and clinical value of 6mA sites in human tumorigenesis[11]. However, DNA 6mA is lowly expressed, and 6mA sites are

unevenly distributed throughout the genome; particularly, the biological function of 6mA in advanced eukaryotes has not been comprehensively understood yet. To shed more light on its biological function, it is essential to characterize its position on the genome, which is related to understanding the epigenetic modification process through investigating the distribution of DNA 6mA[12].

The recent development of experimental technology is in favor of finding DNA 6mA modification sites. Specifically, the experimental techniques for identifying 6mA sites include mass spectrometry[13], immunoprecipitation[14], and next-generation sequencing[15], among which, single-molecule real-time sequencing (SMRT) [16] is the mainstream experimental technique[17]. Nonetheless, the analytical techniques are costly, complicated, and usually labor-intensive, making it difficult to identify 6mA across the genome. Additionally, the method of antibody detection is susceptible to bacterial contamination, while the SMRT sequencing technique can sometimes hardly distinguish 6mA from N1-adenine (1mA) site[18]. To this end, it is of essential necessity to develop a calculation method for the rapid identification of 6mA site.

Computational methods for predicting 6mA sites are still in their infancy[19], and a number of prediction tools have been reported so far. Notably, iDNA6mA-PseKNC[20] is the first prediction approach to predict 6mA sites in the ***Mus.musculus*** genome, and the prediction accuracy for ***Mus.musculus*** and microbe on the dataset has been verified to be quite high. i6mA-Pred[21] is the first method utilized to identify rice genome through support vector machines (SVM), which employs the chemical characteristics and frequency of nucleotides to encode DNA sequences. In addition, i6mA-Pre provides a benchmark dataset of 6mA-rice-Chen. Various prediction methods are dependent on the dataset. As for MM-6mAPred[22], the Markov model based on the 6mA-rice-Chen dataset, achieves superior prediction capability of 6mA sites to i6mA-Pred. On the other hand, SDM6A[23] is also an approach that integrates coding methods with machine learning on the basis of the 6mA-rice-Chen dataset. After training and evaluation, its performance is likewise better than those of i6mA-Pred and iDNA6mA-PseKNC. csDMA[19] implements distinct algorithms for model optimization using the datasets obtained from iDNA6mA-PseKNC and i6mA-Pred. Meanwhile, i6mA-DNCP[18] employs the dinucleotide composition and dinucleotide-based DNA characteristics to identify 6mA sites in rice genome with high sensitivity, and it represents a novel idea for extracting the sequence features of 6mA sites by *k-mer* encoding. Deep learning is a powerful approach used to predict the 6mA sites. Moreover, Lu et al. proposed another 6mA benchmark dataset for rice genome, namely, 6mA-rice-Lv[24]. iDNA6mA-rice is another calculation tool to predict 6mA sites in rice genomes, which is associated with superior prediction performance on both 6mA-rice-Lv and 6mA-riceChen datasets. At present, SNNRice6mA[25], a convolutional neural network on6mA-rice-Chen and 6mA-rice-Lv datasets, is the optimal approach.

Nonetheless, limited by the computing resources, some web servers of the above-mentioned tools have very limited capabilities to handle the large sequences of online submissions. In the meantime, it is difficult to optimize the existing models and to improve the prediction accuracy in the absence of large datasets for human and rice. Therefore, it is necessary to enhance the performance and accuracy of existing computational methods for the preliminary large-scale screening of eukaryotic genomes. To this end, this study was carried out aiming to predict the 6mA sites in human and rice genomes.

The reported experimental data allow us to build these datasets and to develop the computational models. In this paper, these large datasets of Human and Rice Nip were constructed from the recent reported 6mA sites in human and rice respectively. Moreover, a novel and effective deep learning model SICD6mA (**Figure 1**) was proposed to improve the prediction accuracy of 6mA sites, due to the great insufficiency of calculation methods for humans and plants in this field. SICD6mA directly encoded a sequence based on the serial numbers of three consecutive nucleic acids, which was deemed as the 3-*mer* ID (**Figure 1.A**). In addition, it extracted the information of interest and identified the 6mA sites by the Gated Recurrent Units (GRU) network on the basis of

the memory mechanism (**Figure 1.B**). Also, the SICD6mA was validated repeatedly using datasets. According to the results, the as-proposed model SICD6mA attained very stable performance, with significantly superior prediction accuracy to the current state-of-the-art methods. Furthermore, the user-friendly tool SICD6mA was also developed based on the CPU or GPU platform for biologists, and the codes and tools were completely accessible. Prior to high-throughput screening, SICD6mA was able to quickly identify the potential 6mA sites in human and rice genomes. To sum up, SICD6mA proposed in this study is a primary tool to more accurately predict 6mA sites in human and rice genomes, which accelerates and facilitates future research on human and rice genomes.

RESULT AND DISCUSSION

Overall Framework of SICD6mA

In this study, the SICD6mA model based on GRU networks and memory mechanism was constructed to deal with the *3-Mer* ID feature and to identify the 6mA sites (**Figure 1**). The SICD6mA model contains eight stacked components, among which, the input vector in the first layer of SICD6mA is the *3-Mer* ID feature, an integer vector, the sort numbers of three consecutive nucleic acids on a sequence (**Figure 1. A**). A word-vector embedding layer converts the vector into a two-dimensional float matrix, namely, the embedding features (**Figure 1.B**). The Bi-directional Gated Recurrent Units (BGRU) layer learns the embedding features and extracts the first level feature map. Thereafter, SICD6mA utilizes the transpose operation to rotate the feature map particularly. Meanwhile, the non Bi-directional Gated Recurrent Units (UGRU) layer extracts the instantly second level feature map from the rotated map; subsequently, the Simple Gated Recurrent Units (SGRU) layer extracts the third level abstract features. Finally, the fully connected layers FC1 and FC2 classify the abstract features, while the log_softmax layer generates the score based on the classified features.

Cross-Validation Evaluation on Performance

To evaluate the robustness and sensitivity of SICMD6mA, the 5-fold cross-validation was performed based on the training sets that were comprised of both Human and Rice Nip sets. The redundant sequences of each dataset were removed, which avoided overestimation or underestimation resulted from the homologous sequences. In the 5-fold cross-validation, the training set was randomly divided into 5 groups with the same size. In each step, one group of prediction was invoked as the test set, while the remaining 4 were retained as the training set. Finally, the prediction score of 5 test sets was integrated into one file to draw the receiver operating characteristic (ROC) curve and to calculate the total area under the curve (AUC).

According to **Figure 2.A**, the AUC value of SICD6mA was as high as 0.9813 in human datasets, while that was 0.9897 in the Rice Nip dataset. Besides, the ROC curves of SICD6mA showed no great fluctuations and remained stable across different species, revealing the reliability and high prediction accuracy of the method. Interestingly, the AUC values of SICD6mA in the two datasets were quite high, which indicated that SICD6mA was adequately trained in the large datasets. **Figure 2.B** displays the precision-recall (PR) curves of the training sets. Clearly, the AP values of SICD6mA in human and Rice Nip datasets were 0.9822 and 0.9881, respectively. The two large datasets were balanced without redundancy, accordingly, the two PR curves were also balanced with no sharp fluctuation, suggesting that SICD6mA attained the highest accuracy and recall, with consistent stability between the two species datasets.

Comparison of Performance on Independent Test Sets

Further, the sensitivity and accuracy of the prediction model were also evaluated based on the independent test sets. The deep learning or machine learning approaches, including i6mA-DNCP, MM-6mAPred,

SNNRice6mA and csDMA have been reported recently, which achieved more accurate prediction results. In this study, the performance of SICD6mA was compared with the aforementioned methods thanks to the open source codes. To be specific, each method was trained on the training datasets and evaluated on the independent test sets. Finally, the ROC curves and PR curves of these methods were plotted according to the prediction scores of independent test sets, and then the AUC values, AP values and evaluation metrics were calculated accordingly.

The model prediction accuracy might be reliably improved due to the use of large datasets, and the prediction performance of SICD6mA was also more stable. To prove this, ROC curves were plotted and the AUC values were calculated based on the prediction results of independent test evaluations. **Figure 3.A** and **Figure 3.B** shows the ROC curves of several methods obtained by independent test evaluations. Obviously, the ROC curves of SICD6mA were on the top of other methods in the two figures. As for the Human dataset, the AUC values obtained by SICD6MA, SNNRice6mA, i6mADNCP, csDMA, and MM-6mAPred were 0.9824, 0.9748, 0.9730, 0.9719 and 0.8726, respectively; apparently, SICD6mA attained the highest AUC (**Figure 3.A**). With regard to the Rice Nip dataset, the AUC values acquired by SICD6MA, SNNRice6mA, i6mADNCP, csDMA, and MM-6mAPred were 0.9903, 0.9856, 0.9851, 0.9843, and 0.9456, respectively; clearly, the AUC of SICD6mA was higher than those of other methods (**Figure 3.B**). Overall, SICD6mA achieved quite high prediction performances in two genomes.

To investigate the impacts of several methods according to the distribution of large datasets, PR curves were plotted, as presented in **Figure 3.C** and **Figure 3.D**. Changes in PR curve were insensitive since the large datasets were balanced, and no sharp change was observed. According to these figures, the PR curves of SICD6mA were on the top of other methods. SICD6mA achieved the AP values of 0.9834 and 0.9889 in Human and Rice Nip datasets, respectively, which were the highest compared with those of other methods, indicating the highest prediction accuracy and recall of SICD6mA.Therefore, the above findings revealed that, SICD6mA attained consistent stability on the two species datasets.

To further verify the prediction performance of SICD6mA using the large dataset, these methods were compared through the accuracy and false positive rate. In this study, three thresholds, including high, medium and low, were adopted based on specificity (***Sp***) of 95%, 90%, and 80%, separately. According to **Table 1**, on human data set, SICD6mA obtained accuracy (***Acc***) of 93.66%, sensitivity (***Sn***) of 92.33%, specificity (***Sp***) of 95.00%, precision (***Pre***) of 94.86%, recall (***Rec***) of 92.33%, ***F1-score*** of 93.58% and Matthews Correlation Coefficient (***MCC***) of 87.36% at the high threshold; while ***Acc*** of 92.81%, ***Sn*** of 95.61%, ***Sp*** of 90.00%, ***Pre*** of 90.53%, ***Rec*** of 95.61%, ***F1-score***of 93.00% and ***MCC*** of 85.75% at the medium threshold; ***Acc*** of 88.97%, ***Sn*** of 97.94%, ***Sp*** of 80.00%, ***Pre*** of 83.04%, ***Rec*** of 97.94%, ***F1-score*** of 89.88% and ***MCC*** of 79.23%at the low threshold. Thus, it was obvious that, the large datasets not only intensified the model capabilities to detect the 6mA feature signals, but also increased the ***Sn*** value, thereby reducing the false positive rate. In the meantime, the metrics of SICD6mA were apparently higher than those of other methods, showing that SICD6mA improved the prediction accuracy and precision.

These methods were compared and evaluated using the performance metrics in Rice Nip dataset. As shown in **Table 2**, SICD6mA was the optimal method, with ***Acc*** of 95.99%, ***Sn*** of 96.97%, ***Sp*** of 95.00%, ***Pre*** of 95.1%, ***Re c*** of 96.97%, ***F1-score*** of 96.02% and ***MCC*** of 91.99% at the high threshold; ***Acc*** of 94.47%, ***Sn*** of 98.94%, ***Sp*** of 90.01 %, ***Pre*** of 90.83%, ***Rec*** of 98.94%, ***F1-score*** of 94.71% and ***MCC*** of 89.31% at the medium threshold; ***Acc*** of 89.8%, ***Sn*** of 99.58%, ***Sp*** of 80.02%, ***Pre*** of 83.29%, ***Rec*** of 99.58%, ***F1-score*** of 90.71% and ***MCC*** of 81.17% at the low threshold. Besides, the metrics of SICD6mA were much greater than those of other methods, with the only exception of ***Pre*** at the low threshold. Such results demonstrated that, SICD6mAachieved higher accuracy and lower false positive rate than the existing methods with regard to Rice Nip dataset. Consequently, SICD6mAwas associated with high prediction quality and stability.

i6mADNCP, csDMA, MM-6mAPred and other similar methods have combined the chemistry and frequency of nucleotides, and they classify sequences through traditional machine learning. However, SICD6mA is a deep learning network, with no manual design of feature extraction method. In this study, the AUC values, AP values and several performance metrics of SICD6mA were superior to those of other methods, which indicated that it was advantageous to use SICD6mA for identifying the 6mA site.

Comparison of Performance in Identifying 6mA sites across species

To prove the ability of SICD6mA in sensitively identifying 6mA sites in other species, the successful prediction rates of several methods were tested using the cross-species positive datasets. To simulate the processing scene of experimental data, the CD-HIT tool was not adopted to remove the redundant positive sequences. The above models that had been trained on training datasets were employed in every method to predict the positive sites. Typically, the sequence that predicted a higher score or probability than the threshold was deemed as the6mA site sequence of successful identification; otherwise it was regarded as mis-classification. Notably, the threshold was the predicted score (for SICD6mA) or probability (for other methods) of the above independent test set when ***Sp*** was 95%.

Based on the Human dataset, the successful prediction rates of ***Arabidopsis thaliana***, ***Drosophila***, ***Fragaria vesca***, ***Rice indica***, ***Rosa chinensis***, ***Tolypocladium*** and ***yeast*** were as high as 82.02%, 82.13%, 84.50%, 84.32%, 84.64%, 49.20% and 57.06%, by SICD6mA, respectively (**Table S3**). Apparently, the results obtained for ***Drosophila***, ***Tolypocladium***, ***yeast***, ***Fragaria vesca*** and ***Rosa chinensis*** were notably higher than those obtained by i6mA-DNCP, MM-6mAPred, SNNRice6mA and csDMA. Meanwhile, results for ***Arabidopsis thaliana*** and ***Rice indica*** were close to those obtained by SNNRice6mA and csDMA respectively. The mean successful prediction rate achieved 74.84%, which was higher than those of the other methods. Using the Rice Nip dataset, the successful prediction rates for ***Arabidopsis thaliana***, ***Drosophila***, ***Fragaria vesca***, ***Rice indica***, ***Rosa chinensis***, ***Tolypocladium*** and ***yeast*** reached 75.62%, 68.18%, 94.72%, 97.02%, 92.48%, 41.56% and 39.40%, respectively, by SICD6mA (**Table S4**). All results and mean successful prediction rate (72.71%) were notably higher than those obtained by i6mA-DNCP, MM-6mAPred, SNNRice6mA and csDMA. It was figured out after integrating tables S3 and S4 that, each method attained the highest successful prediction rate (**Table 3**). The successful prediction rates achieved by SICD6mA were 82.13%, 94.72%, 97.02%, 92.48%, 49.20% and 57.06% for ***Drosophila***, ***Fragaria vesca***, ***Rice indica***, ***Rosa chinensis***, ***Tolypocladium*** and ***yeast***, separately, which were higher than those attained by other methods. The successful prediction rate (82.02%) of ***Arabidopsis thaliana*** attained by SICD6mA was close to that acquired by SNNRice6mA, and the mean successful prediction rate (79.26%) of SICD6mA was the highest compared with those of i6mA-DNCP, MM-6mAPred, SNNRice6mA and csDMA.

The successful prediction rates for ***Fragaria vesca***, ***Rice indica*** and ***Rosa chinensis*** attained by SICD6mA were much higher than those by other methods using the Rice Nip dataset. In the meantime, the successful prediction rates of ***Drosophila***, ***Tolypocladium*** and ***yeast*** achieved by SICD6mA are significantly higher than those obtained by other methods using the Human dataset, indicating that SICm6A was more accurate in some genomes. Therefore, SICm6A will be a very useful tool to analyze 6mA across the genome-wide.

Application of the program

The computer memory used to predict the 6mA sites is quite large, which consumes lots of calculation time. For the convenience of most experimental scientists, a 6mA site prediction tool, SICD6mA (**Figure 4**), has been developed in this study, which provides two prediction modes of Human and Rice Nip. It allows to predicting the 6mA sites for human and rice species, respectively. If the user inputs the sequences and presses the button “submit”, the tool will filter the sites with "A" at the center and calculate the corresponding scores. Written using python, this tool can rapidly identify 6mA sites based on the CPU or GPU platforms.

CONCLUSIONS

N6-methyladenine (6mA) is closely linked to a series of biological processes, but the relationship of the uneven distribution of these sites with its biological functions remains largely unknown so far. To better understand the biological function of 6mA, the 6mA sites in the genome-wide have been detected experimentally. Nevertheless, the experimental techniques are time-consuming and expensive, making it necessary to develop a fast calculation method. In this study, a deep learning model SICD6mA is proposed to identify the 6mA sites in eukaryotic genomes, such as human and rice. SICD6mA encodes the sequence using the order numbers of three consecutive nucleic acids (3-*mer* ID), then extracts the information of interest and identifies the 6mA sites by the Gated Recurrent Units (GRU) networks based on the memory mechanism. In addition, the large benchmark datasets of human and rice are constructed from the initially reported 6mA sites, and the SICD6mA model is subsequently trained based on these datasets. According to a series of performance evaluations, SICD6mA attains stable performance on various species datasets, and obtains high accuracy and sensitivity. Besides, the AUC values of human and rice independent test sets reach 0.9824 and 0.9903, respectively, and the cross-species successful prediction rates are superior to those of the existing methods, thus confirming the effectiveness of the 3-*mer* ID features and the deep memory networks. More importantly, SICD6mA is also an open source tool that can quickly identify the 6mA sites across species by supporting GPU-accelerated operations. Therefore, SICD6mA will be a useful computational tool to identify the 6mA sites in eukaryotic genomes like human and rice.

MATERIALS AND METHODS

DataSet Collection and Generation

In this study, the benchmark datasets (**Table S1**) of human and rice were adopted for model training and performance evaluation, and the cross-species prediction dataset (**Table S2**) was utilized for comparing the identification ability of 6mA sites. In the datasets, the sequence length was 41nt, and A was located in the middle of the sequence. If the left side of A was shorter than 20, some gap symbols "-"were added to the front, so that the sequence length was equal to 41; similarly, gap symbols were also added to the right if its right length was shorter. Thereafter, every benchmark dataset was randomly split into a training set and an independent test set at a ratio of 8:2, and such ratios between positive and negative sequences were maintained consistent on the training and independent sets, respectively.

The single species benchmark datasets (**Table S1**) included two types, namely, Human (***Homo sapiens)*** and Rice Nip (***Nipponbare japonica***), which were created in this work. The 6mA positive sequences of Human were derived from 6mA-IP-sequencing[10] that was deposited into the Gene Expression Omnibus (accession number GEO: GSE104475), and then the redundancy was removed by CD-HIT-EST[26] using the threshold of 0.8. Finally, altogether 614,856 sequences were remained. At the same time, the human negative sequences were extracted from the GRCh38/hg38 genome (http://hgdownload.soe.ucsc.edu/). There are over 100 million bases A in the human genome. Nonetheless, human negative sequences were generated due to the insufficient computing resources, as shown below, the short sequences without duplication were only searched from the forward and reverse strands in this genome based on the base "A" in the middle and 41nt length. In addition, short sequences identical to positive sequences were deleted, and a total of 770,690 sequences were randomly selected. CD-HIT-EST was also used to remove the redundancy, and 614,856 sequences were retained randomly. On the other hand, the positive sequences of Rice Nip were downloaded from the eRice[27] websites (http://www.elabcaas.cn/rice/downloads.html). CD-HIT-EST was applied in removing the redundancy of positive

sequences, and eventually 380,650 positive sequences were remained. In the meantime, the negative sequences were extracted from the (https://rgp.dna.affrc.go.jp/E/IRGSP/Build5/build5.html) of ***Nipponbare japonica*** genome[28]. Altogether 761,300 short sequences distinct from positive sequences were randomly selected from the forward and reverse strands of the genome according to the base "A" in the middle and the length of 41nt. Thereafter, CD-HIT-EST was adopted for removing the redundancy, 662,710 balanced sequences were obtained, and 380,650 negative sequences were randomly remained.

The cross-species prediction dataset was established based on 6mApositive sequences of several eukaryotic genomes (**Table S2**), including ***Arabidopsis thaliana, Fragaria vesca, Rosa chinensis, Rice indica, Drosophila, Tolypocladium*** and ***yeast.*** Among them, the positive sequences of ***Drosophila, Tolypocladium*** and ***yeast*** were downloaded from the MethSMRT[29] database (http://sysbio.sysu.edu.cn/methsmrt/), and the sequences described as "6mA"were deemed as the 6mA positive sequences. The positive sequences of ***Rice indica*** were derived from the eRice websites (http://www.elabcaas.cn/rice/downloads.html), while those of ***Arabidopsis thaliana***[9] were collected from the NCBI Gene Expression Omnibus (GEO) with accession number GSE81596 (GSM2157793), and those of ***Fragaria vesca***[30]**,** and ***Rosa chinensis***[31] were obtained from the MDR database (http://mdr.xieslab.org/).

Sequence Representation

Nucleic Acid Encoding. The different chemical properties of four nucleic acids (A,G) and (T,C) are that, they have one and two rings, respectively; besides, each (A,C)and(T,G)contain the amino and ketogroups; whereas (A,T)and(C,G) possess two and three hydrogen bonds, respectively. These properties can be computed as follow[20]:

$$x=\begin{cases}1, n_i \in \{A,G\}\\ 0, n_i \in \{T,C\}\end{cases}, y=\begin{cases}1, n_i \in \{A,C\}\\ 0, n_i \in \{T,G\}\end{cases}, z=\begin{cases}1, n_i \in \{A,T\}\\ 0, n_i \in \{C,G\}\end{cases} \quad (1)$$

Thus, a nucleic acid is represented as a one-dimensional vector (*x, y, z*); in other words, A, G, C and T to (1,1,1), (1,0,0) , (0,1,0)and (0,0,1), respectively. This suggests that, chars A, G, C and T have the sort encoding of 1, 2, 3, and 4,respectively, if their vectors are converted the numbers of 111, 100, 10 and1, separately. In addition, the encoding of gap symbol "N" or "-"is 0in a sequence.

***3-Mer* ID Feature**. In this study, *3-Mer* ID represented the three chars from nucleic acids or gap symbol, which was encoded as an integer by combining the encoding sequentially; for example, TAG -> 412, CGA -> 321, and GA- ->210 (**Figure 1.A**). A sequence of *n* bits might be converted to *n* integers through*3-Mer* ID encoding. To make the converted length consistent with *n*, two gap symbols "-"are added after the last nucleic acid at first. Next, every three consecutive chars are extracted from left to right in turn in the sequence, after that, all *3-Mer* chars are encoded as *3-mer* ID one by one. Therefore, the sequence changes to an integer vector, namely, the *3-Mer* ID feature.

Construction of the SICD6mA Model

Memory network is the model that describes the long-term dependencies of sequential data, and the memory components are used to save scene information, achieve long-term memory[32] and extract important features. The Recurrent Neural Network (RNN), long short-term memory (LSTM) and Gated Recurrent Units (GRU) layers of deep learning are associated with certain memory mechanisms. Combined with other deep learning layers, a variety of deep memory networks can be constructed to deal with the sequential data such as the biological sequences.

Word-Vector Embedding Layer. It is found that the relationship between variables can be expressed quantitatively using a numerical matrix. Typically, LSTM is a recursive neural network for learning the long-term dependency, namely, a numerical description matrix between nucleic acids in a sequence. GRU is the simplified version of LSTM, which is sometimes superior and faster. Nonetheless, GRU is unable to directly handle the site

coordinates between nucleic acids, and it is only capable of reading numerical matrices that describe the original features. Thus, a layer called the word vector embedding[33] is required to convert the site information into the original feature matrix. In this study, after the nucleic acid sequence was transformed into a 3-*Mer* ID feature with a length of *n* nt, each site corresponded to a 3-*Mer* ID, which was an integer (**Figure 1.B)**. Then, the word vector embedding layer functioned to map the integers at each site into a *m*-dimensional numerical vector. In this paper, *m* was equivalent to 125. After converting anucleic acid sequence with *n* nt into a numerical matrix of *nxm*, important features were discovered from the GRU layerby learning the numerical matrix.

GRU Layer. It is more difficult for the recurrent neural networks to capture the dependency of nucleic acids with huge separation distances in the sequences, which is ascribed to the problem of gradient attenuation during training[34]. In this study, the proposed GRU was able to better capture the sequence dependency in the long distance (**Figure 1.B)**. The learning process of GRU conformed to the decay memory mechanism of human, to be specific, the fuzzy memory of past things and clear memory of recent things. This attenuation effect was determined by the product of the attenuation coefficient and the memory value. In addition, the sigmoid function was employed to generate the attenuation coefficient, and the memory value was obtained via the tanh function. This learning mechanism was exhibited in equation (2) - (5).

$$z_t = \sigma\left(W_z \bullet [h_{t-1,} x_t]\right) \tag{2}$$

$$r_t = \sigma\left(W_r \bullet [h_{t-1,} x_t]\right) \tag{3}$$

$$\widetilde{h}_t = \tanh\left(W \bullet [r_t * h_{t-1,} x_t]\right) \tag{4}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h}_t \tag{5}$$

where x_t represents the current features for learning, h_{t-1} stands for the features remembered after previous learning, $\widetilde{h}_t$ is equivalent to the first impression generated before learning, z_t and r_t indicate the attenuation coefficients, h_t suggests the features to be remembered after learning, $\bullet$ and $*$ represent the product and dot operations of matrix, whereas W_z, W_r, W are the model parameters representing the network ability to learn features, and these parameters are automatically fitted through model training. GRU achieves two attenuation effects, namely, when the first impression and final memory are generated, respectively. Of them, the first impression is affected by previous memory, while the final memory is composed of the sum of decayed values from the previous memory and the first impression memory.

The Bi-directional Gated Recurrent Units (BGRU) network suggests that, each training sequence is made up of two recurrent neural networks (forward and backward), which are connected to an output layer. UGRU is short for the non Bi-directional Gated Recurrent Units, which is the GRU.

SGRU Layer. As found from the above discussion, the decay memory mechanism of GRU is substantially disturbed by the previous memory, which is not conducive to discovering the new features that do not change significantly. Consequently, a memory layer called the Simple Gated Recurrent Units (SGRU) layer, which possesses two activation functions of tanh and sigmoid, was specifically designed in this study (**Figure 1.B).** Among them, the tanh function generates candidate memories, while the sigmoid function builds attenuation coefficients, and these two functions produce the effect of remembering important information and forgetting the unimportant one. SGRU synthesizes the valid memory using the results of these two functions; in this way, it extracts the abstract features by forgetting the unimportant features.

$$z_t = \sigma\ (W_z \bullet x_t) \tag{6}$$

$$\widetilde{h}_t = \tanh(W \bullet x_t) \tag{7}$$

$$h_t = z_t * \widetilde{h}_t \tag{8}$$

where x_t represents the current features for learning, $\widetilde{h}_t$ is equivalent to the first impression generated before learning, h_t stands for the features to be remembered after learning, while $\bullet$ and $*$ suggest the product and dot operations of matrix, respectively. Notably, SGRU has the one-time attenuation effect when the final memory is generated, its first impression remains unaffected by the previous memory, and the final memory stands for the attenuation value of the first impression memory. This memory network is able to discover the subtle differences between sequence features.

Other Layers. The LeakyReLU activation function divides all negative values by a constant of >1. Among them, the dropout layer means that all values less than a certain threshold are all set to 0, and this layer prevents overfitting by dropping operation when too few sample data are used for model training. During the process of training, BatchNorm1d normalizes data by batch, which also prevents overfitting. Moreover, the fully connected layer is equivalent to a "classifier" in the network. The log_softmax function generates the probability value from the numerical features learned, and later converts them into log values, which are equivalent to the scoring values.

Parameters of SICD6mA. SICD6mAwas able to accelerate training on GPU platform using Adam optimizer and learning rate of 0.0001. It adopted the loss function of Cross-Entropy Loss, and the num_embeddings and embedding_dim of the word-vector embedding layer were 1024 and 125, separately. At the same time, the number of BGRU layers was 2, input_size was 125, hidden_size was 125, dropout was 0.5, batch_first was “True”, and the bidirectional was equal to “True”. On the other hand, the number of UGRU layers was 2, input_size was 51, hidden_size was 128, dropout was 0.5, batch_first was “True”, and the bidirectional was equal to “False”. As for SGRU, the parameters were 125 and 250, respectively, and the fully connected layers were FC1: Output=16 and FC2: Output=2. Besides, the three layers of LeakyReLU (negative_slope=0.2), BatchNorm1d (momentum=0.5) and Dropout(p=0.25) were in the middle of FC1 and FC2.

Performance Evaluation

The receiver operating characteristic (ROC) curve was adopted to evaluate the model performance. Typically, the area under the ROC curve (AUC) is an important indicator for binary classification, which is a float value between 0 and 1, and a AUC value closer to 1 suggests the better model performance. In addition, the PR curve is the precision *vs* recall curve, in which recall serves as the abscissa axis, while precision as the ordinate axis, and a larger area under the PR curve (namely, the AP value) is indicative of the higher accuracy.

In this study, the prediction performance of classifiers was evaluated by five metrics: accuracy (***Acc***), sensitivity (***Sn***), specificity (***Sp***), precision (***Pre***), recall (***Rec***), ***F1-score*** and Matthews Correlation Coefficient (***MCC***), as defined below:

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

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{14}$$

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{15}$$

Where TP (true positive) is the correctly classified number of positive samples, TN (true negative) represents the correctly predicted number of negative samples, FP (false positive) stands for the incorrectly classified number of negative samples, and FN (false negative) is the error count in predicting the positive samples.

DATA AVAILABILITY

The training, independent and prediction datasets were obtained from https://github.com/lwzyb/SICD6mA.

CODE AVAILABILITY

All codes used to generate results that are reported in this paper are available upon request.

REFERENCES

1. Luo, G.-Z., Blanco, M.A., Greer, E.L., He, C. & Shi, Y. DNA N 6-methyladenine: a new epigenetic mark in eukaryotes? *Nature reviews Molecular cell biology* **16**, 705-710 (2015).
2. Ratel, D., Ravanat, J.L., Berger, F. & Wion, D. N6-methyladenine: the other methylated base of DNA. *Bioessays* **28**, 309-315 (2006).
3. Casadesús, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70**, 830-856 (2006).
4. Cheng, L., Hu, Y., Sun, J., Zhou, M. & Jiang, Q. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* **34**, 1953-1956 (2018).
5. von Meyenn, F. *et al.* Impairment of DNA methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Molecular cell* **62**, 848-861 (2016).
6. Janulaitis, A., Klimašauskas, S., Petrušyte, M. & Butkus, V. Cytosine modification in DNA by BcnI methylase yields N 4-methylcytosine. *FEBS letters* **161**, 131-134 (1983).
7. Zhang, G. *et al.* N6-methyladenine DNA modification in Drosophila. *Cell* **161**, 893-906 (2015).
8. Zhou, C. *et al.* Identification and analysis of adenine N 6-methylation sites in the rice genome. *Nature plants* **4**, 554-563 (2018).
9. Liang, Z. *et al.* DNA N6-adenine methylation in Arabidopsis thaliana. *Developmental cell* **45**, 406-416. e3 (2018).
10. Xiao, C.-L. *et al.* N6-methyladenine DNA modification in the human genome. *Molecular cell* **71**, 306-318. e7 (2018).
11. Tang, W., Wan, S., Yang, Z., Teschendorff, A.E. & Zou, Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* **34**, 398-406 (2018).
12. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**, 484-492 (2012).
13. Frelon, S. *et al.* High-performance liquid chromatography− tandem mass spectrometry measurement of

radiation-induced base damage to isolated and cellular DNA. *Chemical research in toxicology* **13**, 1002-1010 (2000).

14. Pomraning, K.R., Smith, K.M. & Freitag, M. Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods* **47**, 142-150 (2009).
15. Smith, Z.D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Reviews Genetics* **14**, 204-220 (2013).
16. Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nature biotechnology* **30**, 1232 (2012).
17. Ardui, S., Ameur, A., Vermeesch, J.R. & Hestand, M.S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic acids research* **46**, 2159-2168 (2018).
18. Kong, L. & Zhang, L. i6mA-DNCP: Computational Identification of DNA N6-Methyladenine Sites in the Rice Genome Using Optimized Dinucleotide-Based Features. *Genes* **10**, 828 (2019).
19. Liu, Z., Dong, W., Jiang, W. & He, Z. csDMA: an improved bioinformatics tool for identifying DNA 6 mA modifications via Chou's 5-step rule. *Scientific reports* **9**, 1-9 (2019).
20. Feng, P. *et al.* iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* **111**, 96-102 (2019).
21. Chen, W., Lv, H., Nie, F. & Lin, H. i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* **35**, 2796-2800 (2019).
22. Pian, C., Zhang, G., Li, F. & Fan, X. MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics* **36**, 388-392 (2020).
23. Basith, S., Manavalan, B., Shin, T.H. & Lee, G. SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Molecular Therapy-Nucleic Acids* **18**, 131-141 (2019).
24. Hao, L. *et al.* iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Frontiers in genetics* **10**, 793 (2019).
25. Yu, H. & Dai, Z. SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Frontiers in genetics* **10**, 1071 (2019).
26. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
27. Zhang, P., Wang, Y., Chachar, S., Tian, J. & Gu, X. eRice: a refined epigenomic platform for japonica and indica rice. *Plant Biotechnology Journal* (2020).
28. International, R.G.S.P. The map-based sequence of the rice genome. *Nature* **436**, 793 (2005).
29. Ye, P. *et al.* MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic acids research*, gkw950 (2016).
30. Edger, P.P. *et al.* Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. *Gigascience* **7**, gix124 (2018).
31. Raymond, O. *et al.* The Rosa genome provides new insights into the domestication of modern roses. *Nature genetics* **50**, 772-777 (2018).
32. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735-1780 (1997).
33. Pennington, J., Socher, R. & Manning, C. Stanford glove: Global vectors for word representation. (2017).
34. Zhou, G.-B., Wu, J., Zhang, C.-L. & Zhou, Z.-H. Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing* **13**, 226-234 (2016).

ACKNOWLEDGEMENTS

This work was partially supported by the Natural Science Foundation for Talent Introduction Project of Sichuan University of Science & Engineering (No. 2018RCL20).

LEGEND

Figure:

Figure 1. Overall Framework of SICD6mA. The two steps include: **A**. sequence encoding using *3-mer* ID feature. **B**. feature extracting and classification scoring by GRU networks based on memory mechanism.

Figure 2. Performance evaluation based on training sets. **A.** Receiver Operating Characteristic (ROC) curves between Human and Rice Nip data sets. **B.** Precision-Recall (PR) curves between Human and Rice Nip data sets.

Figure 3. Comparisons of performance evaluation based on independent test sets. **A.** Receiver Operating Characteristic (ROC) curves of independent test evaluation based on Human data set. **B.** ROC curves of independent test evaluation based on Rice Nip data set. **C.** Precision-Recall (PR) curves of independent test evaluation based on Human data set. **D.** PR curves of independent test evaluation based on Rice Nip data set.

Figure 4. Main interface of SICD6mA prediction tool.

Table:

Table 1. Performance comparison between SICD6mA and several methods on the independent test set of human genome.

Table 2. Performance comparison between SICD6mA and several methods on the independent test set of Rice Nip genome

Table 3. Performance comparison for identifying 6mA sites based on human and Rice Nip data sets.

SUPPLEMENTARY MATERIAL

Table S1. Summary of the training and independent test sets

Table S2. Summary of the cross species prediction set

Table S3. Performance comparison for identifying 6mA sites based on human data set

Table S4. Performance comparison for identifying 6mA sites based on Rice Nip data set

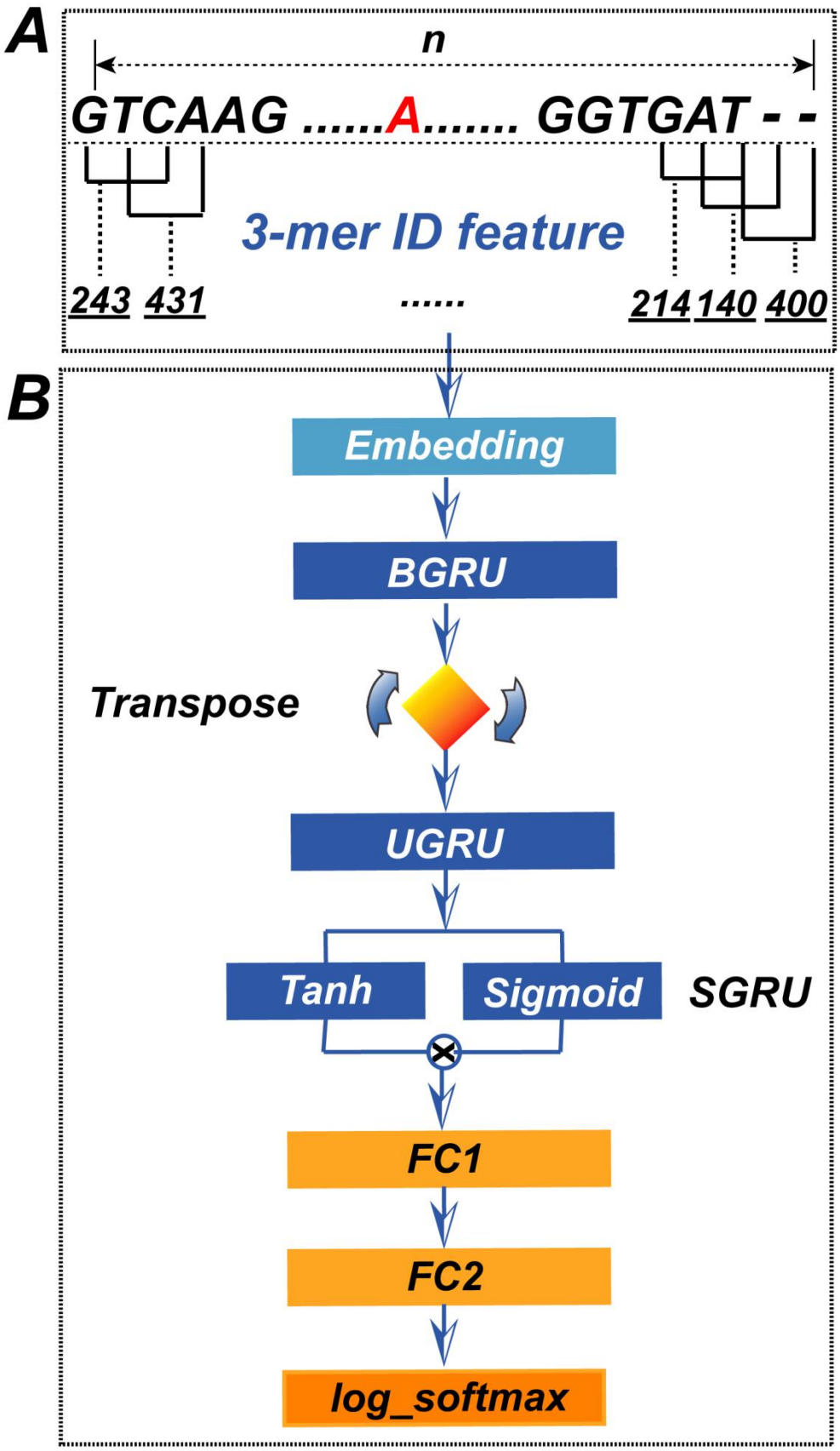
A
n
GTCAAG .....A..... GGTGAT - -
3-mer ID feature
243 431
.....
214 140 400
B
Embedding
BGRU
Transpose
UGRU
Tanh
Sigmoid
SGRU
FC1
FC2
log_softmax

Figure 1

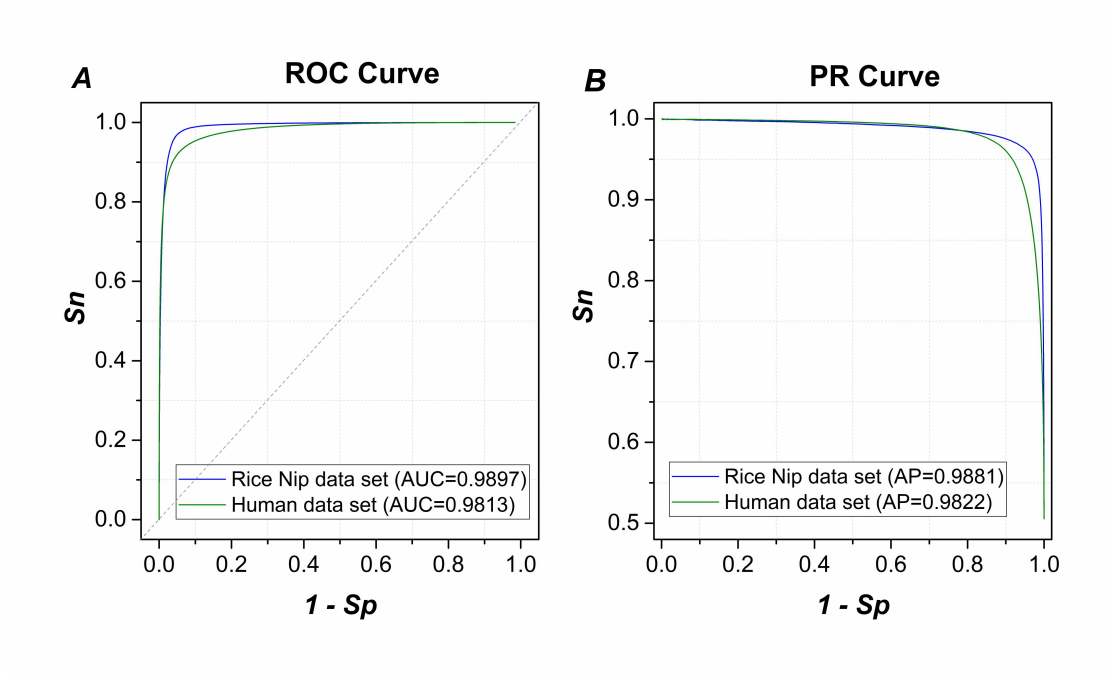
A
ROC Curve
Sn
1 - Sp
Rice Nip data set (AUC=0.9897)
Human data set (AUC=0.9813)
B
PR Curve
Sn
1 - Sp
Rice Nip data set (AP=0.9881)
Human data set (AP=0.9822)

Figure 2

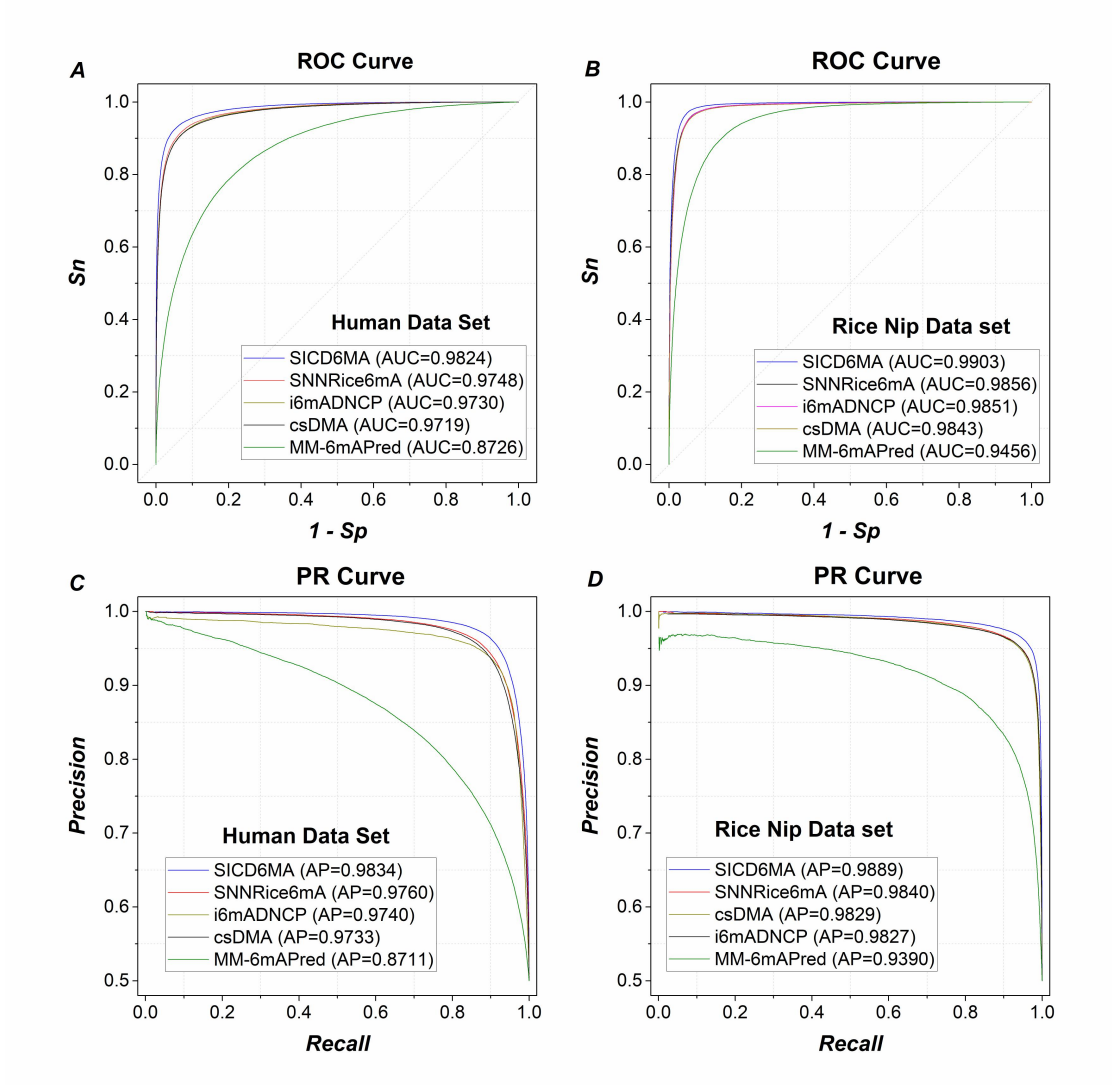
A
ROC Curve
Sn
1 - Sp
Human Data Set
SICD6MA (AUC=0.9824)
SNNRice6mA (AUC=0.9748)
i6mADNCP (AUC=0.9730)
csDMA (AUC=0.9719)
MM-6mAPred (AUC=0.8726)
B
ROC Curve
Sn
1 - Sp
Rice Nip Data set
SICD6MA (AUC=0.9903)
SNNRice6mA (AUC=0.9856)
i6mADNCP (AUC=0.9851)
csDMA (AUC=0.9843)
MM-6mAPred (AUC=0.9456)
C
PR Curve
Precision
Recall
Human Data Set
SICD6MA (AP=0.9834)
SNNRice6mA (AP=0.9760)
i6mADNCP (AP=0.9740)
csDMA (AP=0.9733)
MM-6mAPred (AP=0.8711)
D
PR Curve
Precision
Recall
Rice Nip Data set
SICD6MA (AP=0.9889)
SNNRice6mA (AP=0.9840)
csDMA (AP=0.9829)
i6mADNCP (AP=0.9827)
MM-6mAPred (AP=0.9390)

Figure 3

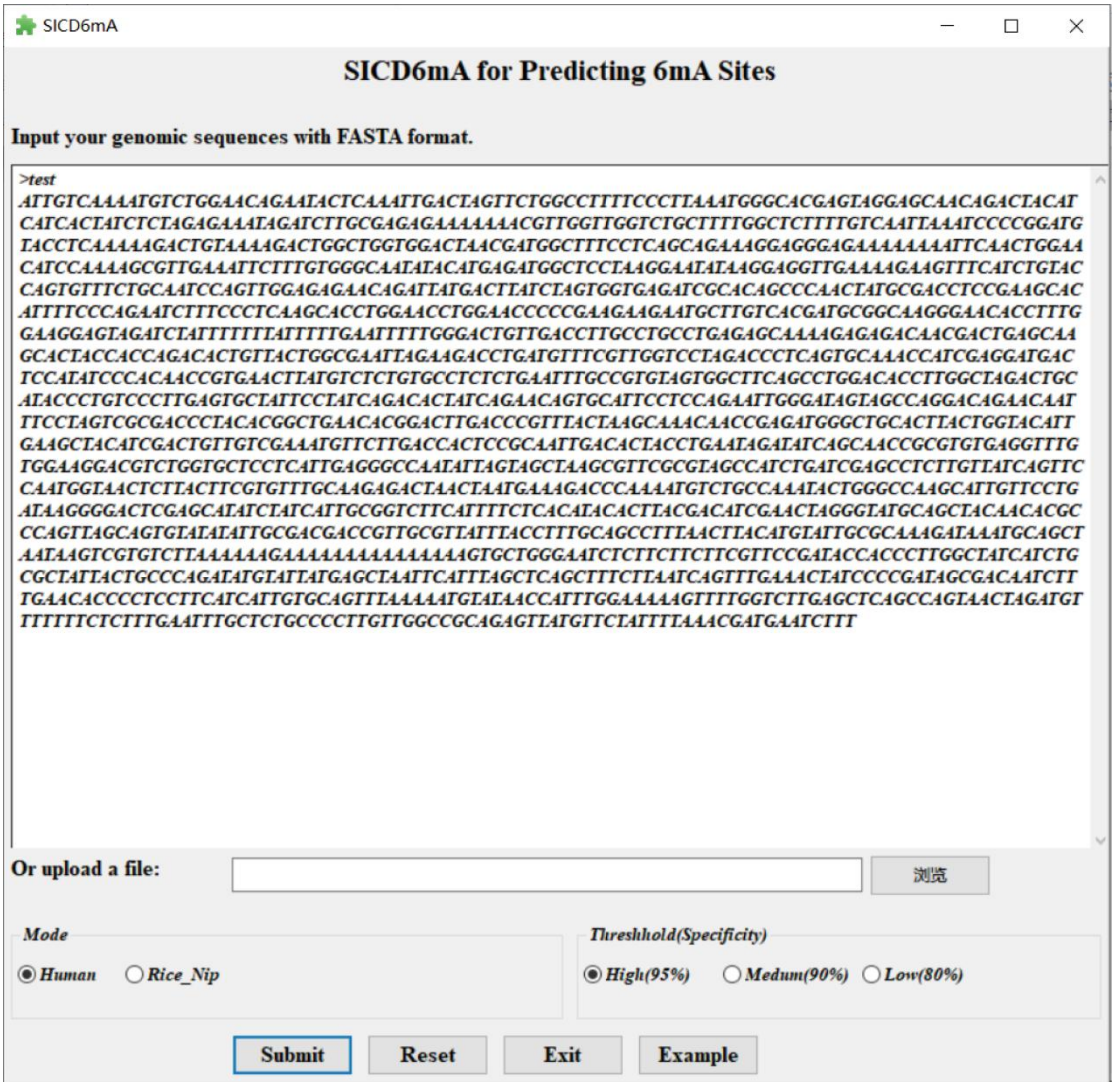
SICD6mA
SICD6mA for Predicting 6mA Sites
Input your genomic sequences with FASTA format.
>test
ATTGTCAAAATGTCTGGAACAGAATACTCAAATTGACTAGTTCTGGCCTTTTCCCTTAAATGGGCACGAGTAGGAGCAACAGACTACAT
CATCACTATCTCTAGAGAAATAGATCTTGCGAGAGAAAAAAACGTTGGTTGGTCTGCTTTTGGCTCTTTTGTCAATTAAATCCCCGGATG
TACCTCAAAAAGACTGTAAAAGACTGGCTGGTGGACTAACGATGGCTTTCCTCAGCAGAAAGGAGGGAGAAAAAAATTCAACTGGAA
CATCCAAAAGCGTTGAAATTCTTTGTGGGCAATATACATGAGATGGCTCCTAAGGAATATAAGGAGGTTGAAAAGAAGTTTCATCTGTAC
CAGTGTTTCTGCAATCCAGTTGGAGAGAACAGATTATGACTTATCTAGTGGTGAGATCGCACAGCCCAACTATGCGACCTCCGAAGCAC
ATTTTCCCAGAATCTTTCCCTCAAGCACCTGGAACCTGGAACCCCCGAAGAAGAATGCTTGTCACGATGCGGCAAGGGAACACCTTTG
GAAGGAGTAGATCTATTTTTTTATTTTTGAATTTTTGGGACTGTTGACCTTGCCTGCCTGAGAGCAAAAGAGAGACAACGACTGAGCAA
GCACTACCACCAGACACTGTTACTGGCGAATTAGAAGACCTGATGTTTCGTTGGTCCTAGACCCTCAGTGCAAACCATCGAGGATGAC
TCCATATCCCACAACCGTGAACTTATGTCTCTGTGCCTCTCTGAATTTGCCGTGTAGTGGCTTCAGCCTGGACACCTTGGCTAGACTGC
ATACCCTGTCCCTTGAGTGCTATTCCTATCAGACACTATCAGAACAGTGCATTCCTCCAGAATTGGGATAGTAGCCAGGACAGAACAAT
TTCCTAGTCGCGACCCTACACGGCTGAACACGGACTTGACCCGTTTACTAAGCAAACAACCGAGATGGGCTGCACTTACTGGTACATT
GAAGCTACATCGACTGTTGTCGAAATGTTCTTGACCACTCCGCAATTGACACTACCTGAATAGATATCAGCAACCGCGTGTGAGGTTTG
TGGAAGGACGTCTGGTGCTCCTCATTGAGGGCCAATATTAGTAGCTAAGCGTTCGCGTAGCCATCTGATCGAGCCTCTTGTTATCAGTTC
CAATGGTAACTCTTACTTCGTGTTTGCAAGAGACTAACTAATGAAAGACCCAAAATGTCTGCCAAATACTGGGCCAAGCATTGTTCCTG
ATAAGGGGACTCGAGCATATCTATCATTGCGGTCTTCATTTTCTCACATACACTTACGACATCGAACTAGGGTATGCAGCTACAACACGC
CCAGTTAGCAGTGTATATATTGCGACGACCGTTGCGTTATTTACCTTTGCAGCCTTTAACTTACATGTATTGCGCAAAGATAAATGCAGCT
AATAAGTCGTGTCTTAAAAAAGAAAAAAAAAAAAAAAGTGCTGGGAATCTCTTCTTCTTCGTTCCGATACCACCCTTGGCTATCATCTG
CGCTATTACTGCCCAGATATGTATTATGAGCTAATTCATTTAGCTCAGCTTTCTTAATCAGTTTGAAACTATCCCCGATAGCGACAATCTT
TGAACACCCCTCCTTCATCATTGTGCAGTTTAAAAATGTATAACCATTTGGAAAAAGTTTTGGTCTTGAGCTCAGCCAGTAACTAGATGT
TTTTTTCTCTTTGAATTTGCTCTGCCCCTTGTTGGCCGCAGAGTTATGTTCTATTTTAAACGATGAATCTTT
Or upload a file:
浏览
Mode
Human
Rice_Nip
Threshhold(Specificity)
High(95%)
Medum(90%)
Low(80%)
Submit
Reset
Exit
Example

Figure 4

Table 1. Performance comparison between SICD6mA and several methods on the independent test set of human genome

Methods	Threshold	***Acc***	***Sn***	***Sp***	***Pre***	***Rec***	***F1-score***	***MCC***	*p-value*
SIC6mA	High	**93.66**	**92.33**	95.00	**94.86**	**92.33**	**93.58**	**87.36**	-
	Medium	**92.81**	**95.61**	90.00	**90.53**	**95.61**	**93.00**	**85.75**	
	Low	**88.97**	**97.94**	80.00	**83.04**	**97.94**	**89.88**	**79.23**	
i6mA-DNCP	High	91.77	88.52	95.03	94.68	88.52	91.49	83.72	<0.1%
	Medium	91.70	93.28	90.12	90.42	93.28	91.83	83.44	
	Low	88.35	96.64	80.07	82.90	96.64	89.24	77.78	
MM-6mAPred	High	71.89	48.78	95.00	90.70	48.78	63.44	49.37	<0.1%
	Medium	76.74	63.47	90.00	86.39	63.47	73.18	55.46	
	Low	79.27	78.53	80.00	79.70	78.53	79.11	58.54	
SNNRice6mA	High	92.15	89.31	95.00	94.70	89.31	91.92	84.45	<0.1%
	Medium	91.97	93.94	90.00	90.38	93.94	92.12	84.00	
	Low	88.47	96.93	80.00	82.90	96.93	89.37	78.06	
csDMA	High	91.74	88.49	95.00	94.65	88.49	91.47	83.67	<0.1%
	Medium	91.57	93.10	90.03	90.33	93.10	91.69	83.17	
	Low	88.25	96.43	80.07	82.87	96.43	89.13	77.54	

Note:

1. *p-value* is the probability of the two-sample T-test between the current method and SICD6mA.
2. The threshold is the prediction score(for SICD6mA) or probability(for other methods) on independent test sets when ***Sp*** is equal to 95%.

Table 2. Performance comparison between SICD6mA and several methods on the independent test set of Rice Nip genome

Methods	Threshold	***Acc***	***Sn***	***Sp***	***Pre***	***Rec***	***F1-score***	***MCC***	*p-value*
SIC6mA	High	**95.99**	**96.97**	95.00	**95.10**	**96.97**	**96.02**	**91.99**	-
	Medium	**94.47**	**98.94**	90.01	**90.83**	**98.94**	**94.71**	**89.31**	
	Low	**89.80**	**99.58**	80.02	83.29	**99.58**	**90.71**	**81.17**	
i6mA-DNCP	High	94.81	94.58	95.03	95.01	94.58	94.80	89.62	<0.1%
	Medium	94.00	97.95	90.05	90.78	97.95	94.23	88.28	
	Low	89.71	99.16	80.25	**83.39**	99.16	90.60	80.88	
MM-6mAPred	High	82.55	70.10	95.00	93.34	70.10	80.07	67.22	<0.1%
	Medium	87.03	84.05	90.00	89.37	84.05	86.63	74.18	
	Low	87.02	94.05	80.00	82.46	94.05	87.88	74.79	
SNNRice6mA	High	94.83	94.66	95.00	94.98	94.66	94.82	89.66	<0.1%
	Medium	94.00	97.99	90.00	90.74	97.99	94.23	88.27	
	Low	89.58	99.15	80.01	83.22	99.15	90.49	80.65	
csDMA	High	94.64	94.25	95.02	94.98	94.25	94.61	89.27	<0.1%
	Medium	93.89	97.74	90.03	90.75	97.74	94.11	88.04	
	Low	89.55	99.01	80.09	83.26	99.01	90.45	80.56	

Note:

1. *p-value* is the probability of the two-sample T-test between the current method and SICD6mA.

2. The threshold is the prediction score(for SICD6mA) or probability(for other methods) on independent test sets when ***Sp*** is equal to 95%.

Table 3. Performance comparison for identifying 6mA sites based on human and Rice Nip data sets

Genome	SIC6mA	i6mA-DNCP	MM-6mAPred	SNNRice6mA	csDMA
Arabidopsis thaliana	82.02	82.70	54.95	**82.71**	81.37
Drosophila	**82.13**	75.64	49.43	78.04	77.28
Fragaria vesca	**94.72**	92.58	75.71	93.15	92.10
Rice indica	**97.02**	95.74	71.27	94.67	95.56
Rosa chinensis	**92.48**	90.34	71.03	90.60	89.81
Tolypocladium	**49.20**	43.82	23.20	46.97	47.18
yeast	**57.06**	50.39	32.78	54.27	51.10
Mean of Success Rates	**79.26**	75.89	54.05	77.20	76.34

Table S1. Summary of the training and independent test sets

Name of data sets	Content	Number of Training sets	Number of Independent test sets	Total Number
Human	Sum.	**983,770**	**245,942**	**1,229,712**
includes:	Positive	491,885	122,971	614,856
	Negative	491,885	122,971	614,856
Rice Nip	Sum.	**609,040**	**152,260**	**761,300**
includes:	Positive	304,520	76,130	380,650
	Negative	304,520	76,130	380,650

Table S2. Summary of the cross species prediction set

Genome	Positive Number
Arabidopsis thaliana	38,052
Drosophila	163,493
Fragaria vesca	26,514
Rice indica	776,858
Rosa chinensis	14,677
Tolypocladium	3,453
yeast	4,094

Table S3. Performance comparison for identifying 6mA sites based on human data set

Genome	SIC6mA	i6mA-DNCP	MM-6mAPred	SNNRice6mA	csDMA
Arabidopsis thaliana	82.02	82.70	54.95	**82.71**	81.37
Drosophila	**82.13**	75.64	49.43	78.04	77.28
Fragaria vesca	**84.50**	81.44	75.71	83.39	83.15
Rice indica	84.32	82.21	71.27	83.39	**84.54**
Rosa chinensis	**84.64**	82.77	71.03	83.68	84.49
Tolypocladium	**49.20**	43.82	23.20	46.97	47.18
yeast	**57.06**	50.39	32.78	54.27	51.10
Mean of Prediction Success Rates	**74.84**	71.28	54.05	73.21	72.73

Table S4. Performance comparison for identifying 6mA sites based on Rice Nip data set

Genome	SIC6mA	i6mA-DNCP	MM-6mAPred	SNNRice6mA	csDMA
Arabidopsis thaliana	**75.62**	70.57	52.29	72.22	68.77
Drosophila	**68.18**	57.31	47.10	63.99	57.50
Fragaria vesca	**94.72**	92.58	73.32	93.15	92.10
Rice indica	**97.02**	95.74	68.55	94.67	95.56
Rosa chinensis	**92.48**	90.34	68.58	90.60	89.81
Tolypocladium	**41.56**	36.81	21.52	39.53	37.59
yeast	**39.40**	29.73	30.90	35.69	29.60
Mean of Prediction Success Rates	**72.71**	67.58	51.75	69.98	67.28