Published online DD MM YYYY

Nucleic Acids Research, YYYY, Vol. xx, No. xx **1–7**
doi:10.1093/nar/gkn000

i6mA-CNN: a convolution based computational approach towards identification of DNA N^6-Methyladenine sites in rice genome

Ruhul Amin [1], Chowdhury Rafeed Rahman [1,*], Md. Sadrul Islam Toaha [1] and Swakkhar Shatabda [1,* *]

[1]Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

DNA N6-methylation (6mA) in Adenine nucleotide is a post replication modification and is responsible for many biological functions. Experimental methods for genome wide 6mA site detection is an expensive and manual labour intensive process. Automated and accurate computational methods can help to identify 6mA sites in long genomes saving significant time and money. Our study develops a convolutional neural network based tool i6mA-CNN capable of identifying 6mA sites in the rice genome. Our model coordinates among multiple types of features such as PseAAC inspired customized feature vector, multiple one hot representations and dinucleotide physicochemical properties. It achieves area under the receiver operating characteristic curve of 0.98 with an overall accuracy of 0.94 using 5 fold cross validation on benchmark dataset. Finally, we evaluate our model on two other plant genome 6mA site identification datasets besides rice. Results suggest that our proposed tool is able to generalize its ability of 6mA site identification on plant genomes irrespective of plant species. Supplementary data (benchmark dataset, independent test dataset, comparison purpose dataset, trained model, physicochemical property values, attention mechanism details for motif finding) are available at Nucleic Acids Research online. Web tool for this research can be found here.

INTRODUCTION

N4-methylcytosine, N6-methyladenine and 5-methylcytosine exist in DNA of different species (1). DNA N^6 Methylation is a part of epigenetics where methyl groups are transferred to DNA molecules (2). It is one kind of post replication modification. Such modification has been identified in three kingdoms of life such as Bacteria, Archaea, and Eukaryotes (3). N^6 Methylation occurs when methyl group is transferred to the 6th position of purine ring of Adenine nucleotide. Such modification is responsible for discrimination between original DNA strand and newly synthesized DNA strand, gene transcription regulation, replication and repair (4).

Laboratory based experimental methods such as sequencing (SMRT-seq, MeDIP-seq) (5), methylated DNA Immunoprecipitation (6), capillary electrophoresis and laser-induced fluorescence (7) have been proposed to identify N^6-Methyladenine (6mA) sites in the genome. 6mA profile of rice has been obtained recently using mass spectrometry analysis and 6mA Immunoprecipitation followed by sequencing (IP-seq) (8). Genome wide identification of 6mA sites is labor intensive and costly using such experimental methods. Computational experiment based pLogo plot (9) on benchmark dataset regarding rice 6mA site identification shows nucleotide distribution to be different near 6mA and non-6mA sites (10). So, accurate computational methods can be effective for automatic 6mA site identification in rice genome.

Tool iDNA6mA-PseKNC was established for N^6 Methyladenine site identification on Mus musculus genome dataset containing 1934 samples in each class (11). They used custom feature vector for each sample inspired by PseAAC (12). They used LibSVM package 3.18 for their classification task. This same package was used in i6mA-Pred tool for the same task in rice genome dataset containing 880 samples in each class using similar feature vector (13). They used Maximum Relevance Maximum Distance (MRMD) method (14) along with Incremental Feature Selection (IFS) for limiting feature space. This same dataset was used to train and build iDNA6mA tool which utilized one hot matrix as sample feature (15). They used sequential one dimensional convolutional neural network architecture for classification. Same benchmark rice genome dataset was used to build i6mA-DNCP tool as well (16). They used di-nucleotide composition frequency and dinucleotide based DNA properties as sample features. TreeBagging algorithm was used as their classifier. They implemented heuristic based DNA property selection

*To whom correspondence should be addressed. Email: swakkhar@cse.uiu.ac.bd, rafeed@cse.uiu.ac.bd

© YYYY The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

for selecting four dinucleotide based DNA properties for their classification. Tool iDNA6mA-Rice was developed using a rice genome dataset containing 1,54,000 samples in each of the two classes (10). They used one hot vector based linear feature space with Random Forest algorithm.

The main limitation of these methods is that they require single feature vector for training and validation purpose irrespective of the heterogeneity of the features they are using. Moreover, these proposed algorithms are unable to use two dimensional matrix oriented features which limits local and sequential pattern detection capability.

We propose i6mA-CNN, a one dimensional CNN based branched classifier which can identify 6mA sites in plant genome. We work with four different types of feature matrices used for DNA sequence representation - PseAAC inspired customized feature matrix, monomer one hot matrix, dimer one hot matrix and dimer physicochemical properties. We use correlation based dimer physicochemical property selection technique in order to remove unnecessary and/ or redundant property features. Our proposed model is able to coordinate between these heterogeneous feature matrices and learn distinguishing features between 6mA and non-6mA sites successfully. Accurate experimental results on a large benchmark dataset and success on independent test datasets show the effectiveness of our proposed tool.

MATERIALS AND METHODS

Benchmark Dataset Construction

Gene Expression Omnibus (GEO) database maintained by National Center for Biotechnology Information (NCBI) (17) has a total of 2,65,290 6mA site containing sequences each of 41 base pair (bp) length with 6mA site at the center. CD-HIT program (18) with 80% (following similar research in (10)) similarity threshold has been implemented on these sequences in order to avoid redundancy bias which resulted in 1,54,000 mA site containing sequences. This is our positive set of samples. Same number of negative samples (no 6mA site) have been obtained from NCBI to formulate a class balanced dataset following some specific criteria. Each 41 bp long negative sequence has Adenine at the center, where the center is not methylated (proven by experiment). 6mA sites have a tendency of occurring near GAGG rich sequences (19). In order to avoid such bias while classification, negative sequences rich with GAGG motifs were considered only to make negative data more objective.

Sequence Representation

A sample sequence of our dataset looks as follows:
$D = N_1, N_2, N_3, \cdots, N_L$
where, L is the length of the DNA sequence, which is 41 in our case and $N_i \in \{A, T, C, G\}$. With appropriate mathematical representation of these sequences, deep learning models can learn class distinguishing features from the sequence local and global patterns (20). It is clear from Figure 1 that our model is capable of coordinating between multiple types of feature vectors and matrices using branch like structure which has been explained in detail in Subsection **Proposed Model**. We use the following sequence representation features in our model:

(a) PseAAC inspired feature representation

PseAAC (Pseudo Amino Acid Composition) has shown success in dealing with protein sequences (21, 22). Inspired by such success, PseKNC (Pseudo K-tuple Nucleotide Composition) has been introduced for dealing with DNA/RNA sequences in computational biology (23, 24). The general form of PseKNC has been designed in such a way that their values will depend on user specified feature extraction techniques from DNA samples.

The four types of nucleotides of DNA can be classified into three categories of attributes. These attributes along with representation code have been provided in Table 1. The categories are as follows:
Ring number: A and G have two rings, whereas C and T have only one ring.
Chemical functionality: A and C are from amino group, whereas G and T are from keto group.
Hydrogen bonding: C and G are bonded to each other with 3 hydrogen bonds (strong), whereas A and T with only two (weak).

The above mentioned properties have significant impact on biological functions as well (25). The ith nucleotide N_i of a sequence can be represented by (x_i, y_i, z_i), where x_i, y_i and z_i represent ring structure, functional group and hydrogen bonding code (see Table 1) of nucleotide N_i, respectively. Besides these three properties, we consider lingering density of each nucleotide to preserve sequence oriented features. Lingering density of nucleotide N_i is defined below:

$$D_i = \frac{1}{L_i} \sum_{j=1}^{L_i} f(N_j)$$

Here, L_i is the length of the substring up to nucleotide N_i and

$$f(N_j) = \begin{cases} 1, & \text{if } N_i = N_j \\ 0, & \text{otherwise} \end{cases}$$

So, for each nucleotide, we have four properties to calculate. Since each sequence in our dataset is 41 bp long, we have a 41×4 dimensional feature matrix for each sequence. We illustrate this four property representation scheme on a toy example sequence "ACGGA" in Table

Table 1: Nucleotide categorical attribute-wise encoding

Category	Attribute	Nucleotides	Code
Ring Structure	Purine	A, G	1
	Pyrimidine	C, T	0
Functional Group	Amino	A, C	1
	Keto	G, T	0
Hydrogen Bonding	Strong	C, G	1
	Weak	A, T	0

2. The resultant matrix representation is: $[[1,1,0,1],[0,1,1,0.5],[1,0,1,0.33],[1,0,1,0.5],[1,1,0,0.4]]$

(b) Monomer one hot representation

Tool iDNA6mA was developed based on monomer one hot encoding (15). Their one dimensional convolutional neural network (1D CNN) was quite successful in class discrimination based on one hot encoding based mathematical representation. Each branch of our designed model is also 1D CNN oriented. So, we have used this particular feature. There are four types of nucleotides in DNA sequence and each of them is represented by a four size one hot vector. A, T, C and G are represented by (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1), respectively. So, a 41 bp long sequence is represented by a 41×4 dimensional matrix.

(c) Dimer one hot representation

There is statistically significant difference (considering p-value of 0.05) between 6mA and non-6mA site containing samples in terms of dinucleotide composition frequencies (16). There are sixteen possible unique dimers such as AA, AT, AC, ... in DNA sequences. We represent each dimer by a 16 size one hot vector. 1D CNN based feature extraction technique is able to learn from class distinguishing dimer patterns which include difference in frequency. In a 41 length sequence, there are 40 overlapping dimers such as N_1N_2, N_2N_3, N_3N_4, ... So, each sample of our dataset is represented by a 40×16 dimensional matrix.

(d) Dimer physicochemical property based representation

According to some studies, Adenine methylation may have impact on DNA structure and its stability (26). We have used the 90 dinucleotide physicochemical properties used by PseKNC tool (23). So, each dimer of our sample sequence is represented by a 90 size Z-score normalized vector. As a result, each 41 bp long sample has been represented by a 40×90 dimensional matrix.

This is certainly a large matrix. If the dinucleotide physicochemical property set contains features which include redundant or less important information in terms of our classification task, it will lead to overfitting problem (27). The steps of dinucleotide physicochemical property (DPP) selection are as follows:

Table 2: Four property representation scheme detail of "ACGGA" sequence

Pos	Nuc	X	Y	Z	D
1	A	1	1	0	1/1 = 1
2	C	0	1	1	1/2 = 0.5
3	G	1	0	1	1/3 = 0.33
4	G	1	0	1	2/4 = 0.5
5	A	1	1	0	2/5 = 0.4

- We use Logistic regression based model for 6mA site identification task using one DPP at a time.
- We note the 5 fold cross validation Mathew's correlation coefficient (MCC) score for each DPP and select the the top 20 DPPs according to obtained MCC score. The higher the MCC, the better is the DPP for our identification task. It is to note that MCC score is better than F1 score and accuracy for binary classification evaluation (28).
- We select the topmost DPP. Now we want to remove redundant information. Each DPP is represented by a 16 size vector as there are 16 possible dimers. DPPs which have a high positive or negative correlation coefficient represent the same type of information. We calculate Pearson's correlation coefficient for each of the 19 other selected DPPs with the topmost DPP.
- We leave out those DPPs which have correlation score greater than 0.9 (positive correlation) or less than -0.9 (negative correlation) with the topmost DPP.

Thus we select nine DPPs out of 90 which include *Roll rise, major Groove distance, twist stiffness, protein induced deformability, mobility to bend towards major groove, melting temperature, Hartman trans free energy, protein DNA twist and tip.*

Model Architecture

Our model architecture has been shown in Figure 1. There are four branches for four different feature matrices (described in detail in Subsection **Sequence Representation**). Each branch has three 1D convolution layers put sequentially. The kernel size and filter number of each convolution layer have been shown in the figure. 1D CNN plays important role in position independent local feature extraction (29, 30). Each branch of the architecture learns class distinguishing features regarding its corresponding input feature matrix independent of the other three branches. The monomer one hot representation of the leftmost branch represents raw representation of sample sequence. Model parameters of this branch learn from this raw representation free from any feature specific knowledge bias. This branch has the potential of learning those class specific hidden features which may not be apparent to humans. The second branch which has dimer one hot representation as input helps the model learn statistically significant differences between 6mA and non-6mA site containing samples in terms of dinucleotide pattern and frequency of occurrence. Such difference has been pointed out in (16). The third branch uses dimer physicochemical property based representation. These properties played an important role in DNA specific classification and identification tasks such as promoter classification (30), 6mA site identification (16), recombination spot identification (31) and DNase I hypersensitive sites identification (32). This third branch facilitates learning the interaction between these property values in a sequence order basis. The last branch is based on learning from ring structure, functional group, hydrogen bonding and lingering density based nucleotide features. These features have impact on DNA

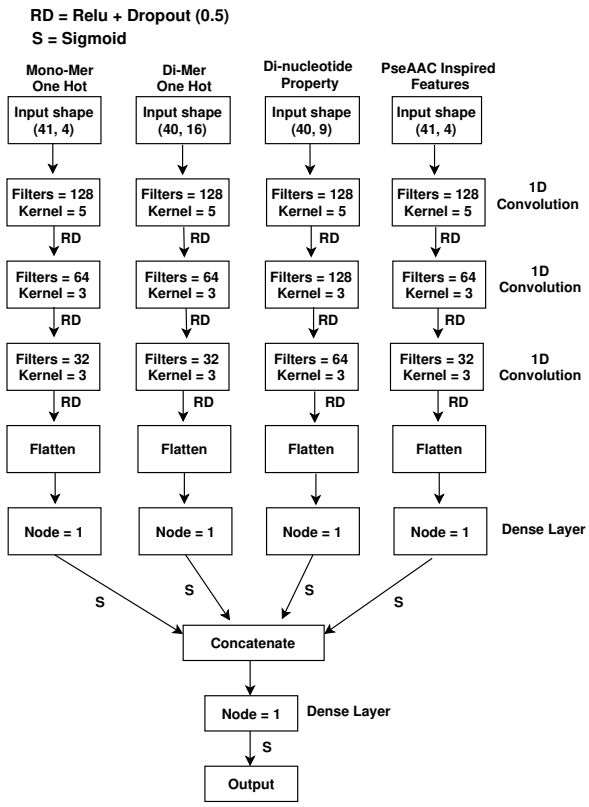

Figure 1. Proposed Model Architecture

biological functions (25, 33). This last branch parameters learn to identify patterns from these sequence coupled features.

We need to have a way to combine these branches for a mutual classification decision depending on the four heterogeneous feature matrices. The convolutional part of each branch i returns a matrix of dimension $m_i \times n_i$. We flatten and convert this matrix into a linear vector. We pass this vector through a dense node having Sigmoid activation function, which outputs a value in the range of 0 to 1. This single value obtained from the ith branch is its representative for 6mA identification task. We have total four values in a range of 0 to 1 from the four branches, which we concatenate and turn into a four size vector. This four size vector is the representative vector of all four branches. The final dense layer with Sigmoid activation function provides the output value depending on the four values of the concatenated vector. If the output value is greater than 0.5, then we consider the sample sequence to have 6mA site. Else, we decide that there is no 6 mA site in the input sequence.

Performance Evaluation and Model Selection

Accuracy (Acc), sensitivity (Sn), specificity (Sp) and Mathew's correlation coefficient (MCC) have been used as evaluation metrics which are described as follows:

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

$$Sn = \frac{TP}{TP+FN}$$

$$Sp = \frac{TN}{TN+FP}$$

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

Here, TP, FP, TN and FN represent true positive, false positive, true negative and false negative, respectively. We consider the 6mA site containing samples as positive class samples. Apart from the above four metrics, we have also used area under receiver operating characteristic (auROC) curve for performance evaluation. This metric provides classifier performance irrespective of provided threshold for class discrimination to the classifier. This score has an impact on classification confidence value for each sample as well. Acc, Sn, Sp and auROC lie in the range [0, 1] while for MCC score, the range is [-1, +1]. Higher value indicates better classification ability.

We have tuned and selected our model hyperparameters using 5 fold cross validation on our benchmark dataset. We took a local search oriented approach where we went in uphill direction in terms of obtained MCC value. The tuned model hyperparameters are as follows:

Convolution layer: Change in number of convolution layers in each branch, convolution filter number and kernel size in each convolution layer

Learning rate: Experimentation with different learning rates such as 0.01, 0.001, 0.0001, 0.00001 and adaptive learning rate scheme

Dropout rate: Testing out dropout rate of 0.1, 0.3 and 0.5.

Optimizer: Model training with adaptive moment estimation (ADAM), stochastic gradient descent (SGD) and Nesterov accelerated gradient (NAG) optimizer.

Activation function: Experimentation with Relu, Tanh and LeakyRelu activation function in initial convolution layers

As shown in Figure 1, Each branch has three convolution layers. These three consecutive layers contain kernel size of 5, 3 and 3, whereas the number of filters are 128, 64 and 32, respectively. Learning rate of 0.0001 has been used in order to avoid divergence while training. Dropout of 0.5 has been used for overfitting avoidance. Addition of dropout changes network connectivity on each training epoch and thus helps model to understand the training sample features instead of memorizing them (34). ADAM has been chosen as the optimizer for model weight update during training. Each intermediate convolution layer output goes through Relu activation function which is very popular for being effective and simple (35).

RESULTS AND DISCUSSION

i6mA-Pred (13), iDNA6mA-Rice (10), iDNA6mA (15) and i6mA-DNCP (16) are some of the state-of-art tools which can identify 6mA sites in rice genome. For comparison purpose of

our proposed tool with these state-of-the-art tools, consistency in dataset usage and evaluation method are necessary.

The benchmark dataset that we have used for this research has also been used for constructing the latest rice genome i6mA site identification tool **iDNA6mA-Rice** (10). Tools for the same task have been developed in (13, 15, 16), though they used a different dataset for benchmarking which contains 880 samples in each of the two classes. They used either Jackknife testing or 10 fold cross validation for algorithm and model selection. For comparison purpose, we also test our custom model on the same 880 sample per class dataset using similar validation methods. This comparison purpose dataset has been obtained from NCBI GEO under accession number GSE103145. All sequences with 6mA site in the center and with modification score greater than 30 (according to Methylome Analysis Technical Note, 30 is the minimum score to call a nucleotide as modified) have been considered as positive samples. Each sample sequence is 41 bp long. CD-HIT (18) with 60% similarity threshold has been applied on the chosen positive sequences in order to remove redundancy. Thus the 880 positive samples were obtained. The 880 negative samples of this dataset have been obtained in a manner similar to our negative benchmark dataset samples. Detailed performance comparison between the tools used for 6mA site identification has been provided in Table 3. All the tool results except for our proposed tool i6mA-CNN have been obtained from corresponding research articles (10, 13, 15, 16). Our proposed tool clearly outperforms all other tools for similar task by all five performance metrics. Tool i6mA-CNN achieves a high 5 fold cross validation auROC score of 0.98 on our benchmark dataset. ROC curve has been shown in Figure 2.

Independent test set performance of a tool can help in proving the generalization ability of that tool. Following procedures similar to our benchmark dataset construction, we have constructed two independent test datasets in order to evaluate the robustness of our proposed tool in terms of N^6-Methyladenine site identification in plant genome. Fragaria Vesca (Wild strawberry) and Arabidopsis thaliana (cress weed plant) 6mA site containing sequences have been obtained from the MDR database (36) and from NCBI GEO with accession number GSE81597 (37), respectively. Performance of i6mA-CNN on these datasets have been shown in detail in Table 4. Results show the robustness of the proposed tool for 6mA site identification in plant genome.

Our tool has also been tested on Mus musculus (house mouse) genome dataset. It was used for constructing iDNA6mA-PseKNC tool (11). The 6mA site containing positive samples were extracted from MethSMRT database (38). Our tool achieved less than 50% test accuracy on this particular dataset. This proves that 6mA site containing region pattern of plants are different from that of animals.

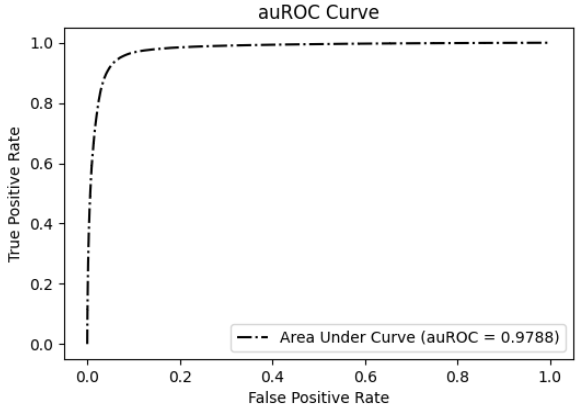

Figure 2. ROC curve of i6mA-CNN on benchmark dataset

Our model is a CNN based branched model with no time stamp oriented sequential information processing capability. There are reasons why we did not use recurrent layers which utilize time stamp information while working with sequence type data. First, it is not possible to provide multiple feature representations simultaneously in each time stamp input in recurrent neural networks. Second, each time stamp of such networks only accept linear vector representation which would limit our current ability of using matrices. Third, recurrent neural networks such as RNN, GRU or LSTM give importance to positional information (39). But classification of genome related sample sequence often depends on the presence of potential motif type patterns irrespective of their position in the sequence (19). Finally, training and validation using recurrent layer based networks are costly because of sequential type of computation. 1D CNN based model having multiple convolution layers have the capability of encoding and learning both short and long range patterns. Parallel processing is possible in such networks which allow fast computation. Our customized model architecture is able to work with multiple heterogeneous features simultaneously and is able to utilize the benefits of feature matrix

Table 3: i6mA-CNN performance comparison with state-of-the-art tools. Best performance results have been marked in bold. **Dataset 1** is the 880 sample per class dataset used for comparison purpose, while **Dataset 2** is our benchmark dataset with 1,54,000 samples in each class

Dataset	Evaluation Method	Tool	Sn	Sp	Acc	MCC	auROC
Dataset 1	Jackknife Testing	i6mA-Pred	83.41%	83.64%	83.52%	0.67	0.91
		i6mA-CNN	**90.61%**	**94.12%**	**93.15%**	**0.85**	**0.98**
	10 Fold Cross Validation	iDNA6mA	86.7%	86.59%	86.64%	0.73	0.93
		i6mA-DNCP	84.09%	88.07%	86.08%	0.72	0.93
		i6mA-CNN	**89.55%**	**93.18%**	**91.36%**	**0.83**	**0.97**
Dataset 2	5 Fold Cross Validation	iDNA6mA-Rice	93%	90.5%	91.7%	0.83	0.96
		i6mA-CNN	**95.13%**	**92.81%**	**93.97%**	**0.88**	**0.98**

representation. Because of such inherent characteristics, our proposed tool is fast, accurate and shows performance superior to the remaining state-of-the-art tools for 6mA site identification in rice genome and also generalizes well to other plant genomes for the same task.

We have found potential motifs which our deep learning based model actively searches for in order to perform 6mA site identification in rice genome. Motifs are class specific substrings found frequently in the sample sequences of that class. Following the procedure shown in (30), we have customized an attention model (details of this model have been provided as supplementary information) based on our proposed CNN model provided in Figure 1 with a view to identifying the potential motifs. A list of potential motifs found for 6mA site identification on our benchmark dataset have been provided in Table 5. The top motif that we have found is **ATAT**. It was found to be active 5.2% of the time during 6mA site identification. The pLogo plot constructed in (10) showed the significant difference of nucleotide distribution in postive and negative samples. This analysis has not been repeated in our current work.

CONCLUSION

Tool i6mA-CNN is the outcome of the current research aimed at 6mA site identification in rice genome. The combination of four different feature matrices and redundant feature reduction have facilitated our customized CNN based architecture to be robust and accurate. Experimental results on two rice genome datasets and two other independent plant genome datasets related to 6mA site identification show the effectiveness of our tool. Proposed i6mA-CNN is expected to be an effective tool for automation in epigenetics related research. Future research may aim at constructing a tool that can perform 6mA site identification in both plant and animal genome effectively.

Conflict of interest statement. None declared.

REFERENCES

1. Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.-C. (2019) iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics,* **111**(1), 96–102.
2. von Meyenn, F., Iurlaro, M., Habibi, E., Liu, N. Q., Salehzadeh-Yazdi, A., Santos, F., Petrini, E., Milagre, I., Yu, M., Xie, Z., et al. (2016) Impairment of DNA methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Molecular cell,* **62**(6), 848–861.
3. O'Brown, Z. K. and Greer, E. L. (2016) N6-methyladenine: a conserved and dynamic DNA mark. In *DNA Methyltransferases-Role and Function* pp. 213–246 Springer.
4. Wion, D. and Casadesús, J. (2006) N 6-methyl-adenine: an epigenetic signal for DNA–protein interactions. *Nature Reviews Microbiology,* **4**(3), 183–192.
5. Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., and Turner, S. W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods,* **7**(6), 461.
6. Pomraning, K. R., Smith, K. M., and Freitag, M. (2009) Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods,* **47**(3), 142–150.
7. Krais, A. M., Cornelius, M. G., and Schmeiser, H. H. (2010) Genomic N6-methyladenine determination by MEKC with LIF. *Electrophoresis,* **31**(21), 3548–3551.
8. Zhou, C., Wang, C., Liu, H., Zhou, Q., Liu, Q., Guo, Y., Peng, T., Song, J., Zhang, J., Chen, L., et al. (2018) Identification and analysis of adenine N 6-methylation sites in the rice genome. *Nature plants,* **4**(8), 554–563.
9. O'shea, J. P., Chou, M. F., Quader, S. A., Ryan, J. K., Church, G. M., and Schwartz, D. (2013) pLogo: a probabilistic approach to visualizing sequence motifs. *Nature methods,* **10**(12), 1211–1212.
10. Hao, L., Dao, F.-Y., Guan, Z.-X., Zhang, D., Tan, J.-X., Zhang, Y., Chen, W., and Lin, H. (2019) iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Frontiers in genetics,* **10**, 793.
11. Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.-C. (2019) iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics,* **111**(1), 96–102.
12. Shen, H.-B. and Chou, K.-C. (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical biochemistry,* **373**(2), 386–388.
13. Chen, W., Lv, H., Nie, F., and Lin, H. (2019) i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics,* **35**(16), 2796–2800.
14. Chen, W., Feng, P., Ding, H., and Lin, H. (2018) Classifying included and excluded exons in exon skipping event using histone modifications. *Frontiers in genetics,* **9**, 433.
15. Tahir, M., Tayara, H., and Chong, K. T. (2019) iDNA6mA (5-step rule): Identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. *Chemometrics and Intelligent Laboratory Systems,* **189**, 96–101.
16. Kong, L. and 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.
17. Long, C., Li, W., Liang, P., Liu, S., and Zuo, Y. (2018) Transcriptome comparisons of multi-species identify differential genome activation of mammals embryogenesis. *Ieee Access,* **7**, 7794–7802.
18. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* **22**(13), 1658–1659.
19. Zhou, C., Wang, C., Liu, H., Zhou, Q., Liu, Q., Guo, Y., Peng, T., Song, J., Zhang, J., Chen, L., et al. (2018) Identification and analysis of adenine N 6-methylation sites in the rice genome. *Nature plants,* **4**(8), 554–563.

Table 4: i6mA-CNN performance on independent test datasets

Independent Test Dataset	Total Sample No.	TP Sample No.	FN Sample No.	Success Rate
Fragaria Vesca	8979	8642	337	96.25%
Arabidopsis Thaliana	67650	63252	4398	93.5%

20. Umarov, R. K. and Solovyev, V. V. (2017) Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one,* **12**(2), e0171410.
21. Zhong, W.-Z. and Zhou, S.-F. Molecular science for drug development and biomedicine. (2014).
22. Zhou, G.-P. and Zhong, W.-Z. (2016) Perspectives in Medicinal Chemistry.. *Current topics in medicinal chemistry,* **16**(4), 381.
23. Chen, W., Lei, T.-Y., Jin, D.-C., Lin, H., and Chou, K.-C. (2014) PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Analytical biochemistry,* **456**, 53–60.
24. Chen, W., Lin, H., and Chou, K.-C. (2015) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Molecular BioSystems,* **11**(10), 2620–2634.
25. Chou, K.-C. (1988) Low-frequency collective motion in biomacromolecules and its biological functions. *Biophysical chemistry,* **30**(1), 3–48.
26. Cheng, S., Herman, G., and Modrich, P. (1985) Extent of equilibrium perturbation of the DNA helix upon enzymatic methylation of adenine residues.. *Journal of Biological Chemistry,* **260**(1), 191–194.
27. Hawkins, D. M. (2004) The problem of overfitting. *Journal of chemical information and computer sciences,* **44**(1), 1–12.
28. Chicco, D. and Jurman, G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics,* **21**(1), 6.
29. Chen, T., Xu, R., He, Y., and Wang, X. (2017) Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications,* **72**, 221–230.
30. Amin, R., Rahman, C. R., Ahmed, S., Sifat, M., Rahman, H., Liton, M. N. K., Rahman, M., Khan, M., Hossain, Z., Shatabda, S., et al. (2019) iPromoter-BnCNN: a Novel Branched CNN Based Predictor for Identifying and Classifying Sigma Promoters. *Bioinformatics,*.
31. Zhang, L. and Kong, L. (2019) iRSpot-PDI: Identification of recombination spots by incorporating dinucleotide property diversity information into Chou's pseudo components. *Genomics,* **111**(3), 457–464.
32. Zhang, S., Lin, J., Su, L., and Zhou, Z. (2019) pDHS-DSET: Prediction of DNase I hypersensitive sites in plant genome using DS evidence theory. *Analytical biochemistry,* **564**, 54–63.
33. Chou, K.-C. and Mao, B. (1988) Collective motion in DNA and its role in drug intercalation. *Biopolymers: Original Research on Biomolecules,* **27**(11), 1795–1815.
34. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research,* **15**(1), 1929–1958.
35. Agarap, A. F. (2018) Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375,*.
36. Liu, Z.-Y., Xing, J.-F., Chen, W., Luan, M.-W., Xie, R., Huang, J., Xie, S.-Q., and Xiao, C.-L. (2019) MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. *Horticulture research,* **6**(1), 1–7.
37. Liang, Z., Shen, L., Cui, X., Bao, S., Geng, Y., Yu, G., Liang, F., Xie, S., Lu, T., Gu, X., et al. (2018) DNA N6-adenine methylation in Arabidopsis thaliana. *Developmental cell,* **45**(3), 406–416.
38. Ye, P., Luan, Y., Chen, K., Liu, Y., Xiao, C., and Xie, Z. (2016) MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic acids research,* p. gkw950.
39. Yin, W., Kann, K., Yu, M., and Schütze, H. (2017) Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923,*.

Table 5: Potential motifs for N^6-Methyladenine site identification

Motif Sequence	Active Occurrence %
ATAT	5.2
AGGC	4.1
AAAA	4.0
AAAT	3.9
GAGG	3.5