JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

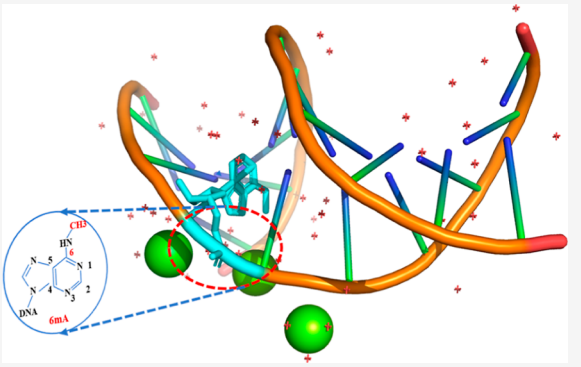
I-DNAN6mA: Accurate Identification of DNA N^6-Methyladenine Sites Using the Base-Pairing Map and Deep Learning

Xue-Qiang Fan, Bing Lin, Jun Hu,* and Zhong-Yi Guo*

Cite This: *J. Chem. Inf. Model.* 2023, 63, 1076−1086 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The recent discovery of numerous DNA N^6-methyladenine (6mA) sites has transformed our perception about the roles of 6mA in living organisms. However, our ability to understand them is hampered by our inability to identify 6mA sites rapidly and cost-efficiently by existing experimental methods. Developing a novel method to quickly and accurately identify 6mA sites is critical for speeding up the progress of its function detection and understanding. In this study, we propose a novel computational method, called I-DNAN6mA, to identify 6mA sites and complement experimental methods well, by leveraging the base-pairing rules and a well-designed three-stage deep learning model with pairwise inputs. The performance of our proposed method is benchmarked and evaluated on four species, i.e., *Arabidopsis thaliana*, *Drosophila melanogaster*, *Rice*, and *Rosaceae*. The experimental results demonstrate that I-DNAN6mA achieves area under the receiver operating characteristic curve values of 0.967, 0.963, 0.947, 0.976, and 0.990, accuracies of 91.5, 92.7, 88.2, 0.938, and 96.2%, and Mathew's correlation coefficient values of 0.855, 0.831, 0.763, 0.877, and 0.924 on five benchmark data sets, respectively, and outperforms several existing state-of-the-art methods. To our knowledge, I-DNAN6mA is the first approach to identify 6mA sites using a novel image-like representation of DNA sequences and a deep learning model with pairwise inputs. I-DNAN6mA is expected to be useful for locating functional regions of DNA.



INTRODUCTION

Prevalent DNA N^6-methyladenine (the sixth position of the purine ring in adenines, 6mA) modifications in prokaryotes have been demonstrated.[1−3] 6mA plays an important role in a series of biological processes, e.g., regulating bacterial DNA repair, DNA replication, gene expression regulation, and gene transcription.[4−6] In the past several years, 6mA function in prokaryote genomes has been extensively studied, and great progress has been made. On the contrary, since it has been a grand challenge to carry out an accurate detection for 6mA modifications sites in eukaryotes in computational biophysics for decades, the role of 6mA modifications in eukaryotes is just starting to be elucidated and remains to be extensively characterized.[1,7] Importantly, recent research results show that accurate 6mA knowledge can serve as a priori for cancer suppression studies in eukaryotes.[8,9] Furthermore, the 6mA can provide essential clues for the DNA structure prediction and cell differentiation process exploring.[10] Therefore, accurate determination of the 6mA sites of the DNA molecule in eukaryotes is critical for speeding up the progress of DNA function detection and understanding, especially in the post-genome era where a large volume of DNA sequences are rapidly being accumulated. Nevertheless, the biological experimental methods, e.g., methylated DNA immunoprecipitation sequencing (MeDIP-seq),[11] capillary electrophoresis and laser-induced fluorescence (CE-LIF),[12] and single-molecule real-time sequencing (SMRT-seq),[13] for detecting 6mA sites of DNA are expensive and time-consuming. In view of this situation, designing fast and cost-effective computational methods to directly predict 6mA sites from the DNA sequence in eukaryotes is highly desired and can be well complementary to experimental methods.

During the recent years, several computation-based methods have been proposed to address this challenge. Generally speaking, these existing methods can be broadly classified into two categories: traditional machine learning (ML)-based methods and deep learning (DL)-based methods. Early-stage methods of predicting 6mA sites are developed mainly based on traditional ML algorithms, such as i6mA-Pred,[14] i6mA-DNCP,[15] iDNA6mA-Rice,[16] SDM6A,[17] 6mA-Finder,[18] and i6ma-stack.[19] These ML-based methods generally use several sequence-based hand-crafted features and an ensemble of

Received: November 19, 2022
Published: February 1, 2023

https://doi.org/10.1021/acs.jcim.2c01465
J. Chem. Inf. Model. 2023, 63, 1076−1086

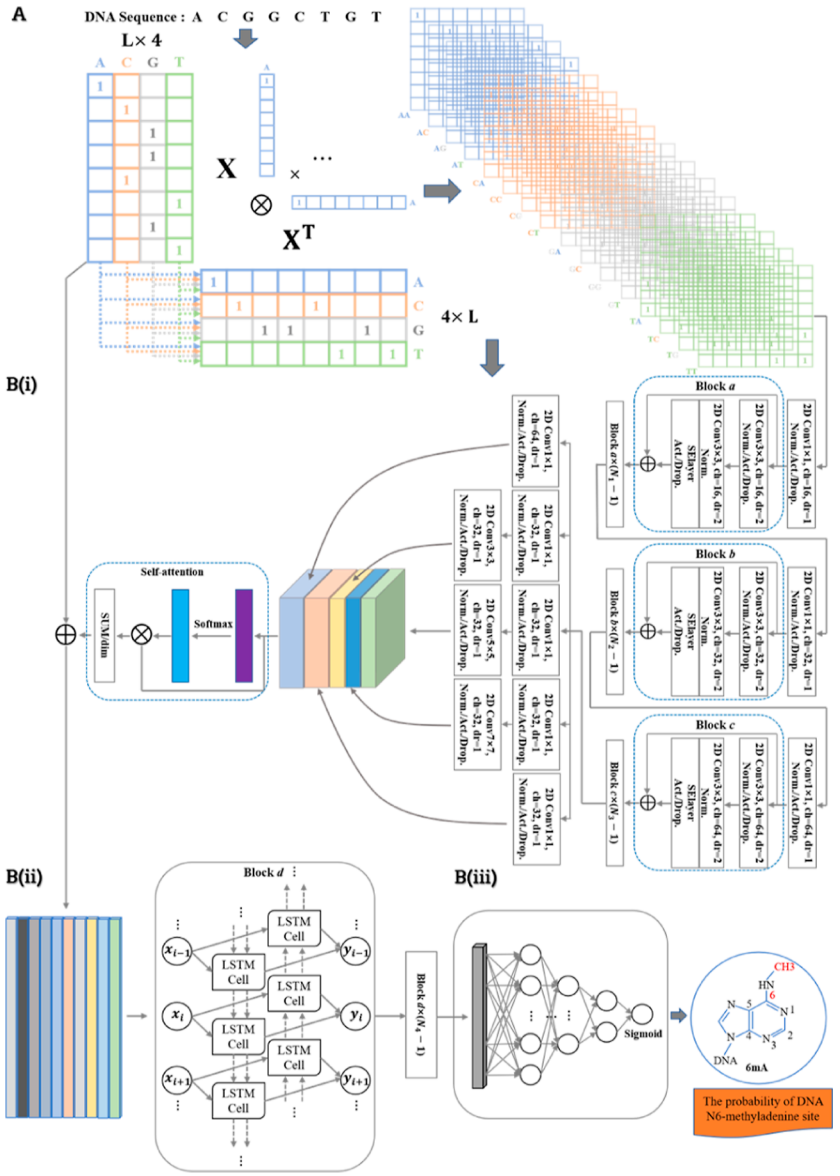


Figure 1. Flow chart of I-DNAN6mA. The workflow of I-DNAN6mA is composed of the base-paring "image" feature encoding (A) and DL-based prediction model (B). The network layout of the I-DNAN6mA consists of base-paring "image" feature mining [B(i)], capturing knowledge of time series dependencies between nucleotides [B(ii)], and calculating the probability score of 6mA sites [B(iii)].

multiple ML algorithms (i.e., support vector machine,[20] random forest,[21] Gaussian Naive Bayes,[22] bagging,[23] and second-order Markov model[24,25]) to predict 6mA sites. For example, i6mA-Pred[14] is the first computational method for 6mA site identification, which combined one support vector machine classifier with chemical features of nucleotides and position-specific nucleotide frequencies to learn a 6mA site prediction model. However, applying the ensemble of sequence-based hand-crafted features and traditional ML classifiers widely to process the DNA biology knowledge recognition inevitably suffers from certain disadvantages. For instance, with the continuous increase and rapid accumulation of DNA sequence data, traditional ML algorithms cannot effectively mine the hidden information in a multitude of sequences, making it difficult for the ML-based methods to be considered the optimal option for training classifiers. Furthermore, the prediction accuracy of ML-based methods heavily depends on the quality of hand-crafted features.

To overcome the drawback of ML-based methods, a few DL techniques that utilize multi-layer artificial neural networks to learn tasks have been successfully applied to solve computational biology problems, including 6mA site prediction. A few of these have been named below: SpineNet-6mA,[26] GC6mA-Pred,[27] iRicem6A-CNN,[28] Deep6mA,[29] i6mA-VC,[30] SNNRice6mA,[31] DeepM6A,[32] iDNA6mA (five-step rule),[33] i6mA-DNC,[34] LA6mA,[35] AL6mA,[35] 6mA-RicePred,[36] and Deep6mAPred.[37] These DL-based methods often utilize only one-hot encoding of the DNA sequence to recognize 6mA sites with one or more DL algorithms, such as convolutional neural networks[38] and fully connected hidden layers.[39] For instance, in Deep6mA, convolutional neural networks combined with bidirectional long short-term memory recurrent neural networks (BiLSTM)[40] are applied to extract discriminative information from solely one-hot encoding and accurately predict the 6mA sites. In LA6mA, one-hot encoding also is fed into BiLSTM and a self-attention mechanism for predicting

https://doi.org/10.1021/acs.jcim.2c01465
J. Chem. Inf. Model. 2023, 63, 1076−1086

6mA sites. Nevertheless, despite the efficiency and accuracy achieved, the existing methods still have several following critical deficiencies: (i) to develop powerful computational models for 6mA site prediction, a critical step is to extract sufficient discriminative features. By revisiting existing ML- and DL-based 6mA site identification methods, it was found that all of them employ a fused feature generated in series with multiple features and a limited-scale one-hot encoding feature to capture discriminative information. Despite digging a certain degree of discriminative information, they still lose the remote low-level feature information that is likely to aid in predicting 6mA sites. (ii) The problem of identification of 6mA sites should take into account all possible base pairing structure patterns in the DNA sequence.

To address the important issues mentioned above, in this study, we propose a novel DL-based method, called I-DNAN6mA, to further improve the performance of 6mA site prediction. Unlike existing 6mA site prediction methods, the input of our model contains all possible base-pairing maps, which are treated as one 2D "image" with 16 channels. Each channel, represented by a square matrix of the same dimension as the length of the input DNA sequence, indicates the occurrence of one of the 16 possible base pairs between the input nucleotides.[41] Benefitting from this novel feature representation, the first stage B(i) (see Figure 1) of our proposed DL-based model, which is an ensemble of deep hybrid networks of ResNet[42,43] coupled with the "inception" module[44] and self-attention networks,[45] can explicitly grab all short- and long-range interactions and all possible base pairings. Our model includes one additional stage B(ii) to store the output of B(i) and concatenate it with one-hot encoding as input features and uses BiLSTM capture knowledge of time series dependencies between nucleotides in the DNA sequence. The output of our model is the probability score of 6mA sites calculated by the third stage B(iii), which integrates a multilayer perceptron (MLP)[46] and a sigmoidal activation function.[47] We conduct experiments on four eukaryotes with known 6mA sites to compare the performance of I-DNAN6mA against that of the recent 6mA site prediction methods. Benchmarking results demonstrate that I-DNAN6mA yields substantially superior performance over previous methods, highlighting its promising potential in solving the 6mA site prediction problem.

MATERIALS AND METHODS

Benchmark Data Sets. High-quality benchmark data sets are important to establish reliable prediction models. To fairly compare the performance of I-DNAN6mA with that of other existing methods, in this study, the benchmark data sets containing the DNA 6mA site data for four species, i.e., *Arabidopsis thaliana*, *Drosophila melanogaster*, *Rice*, and *Rosaceae*, are derived from the recently published studies.[14,16,48−50] The above three species, i.e., *A. thaliana*, *D. melanogaster*, and *Rosaceae*, are further randomly divided into a training set and an independent testing set by the researchers, respectively. The *Rice* genome containing both 6mA-rice-Chen and *6mA-rice-Lv* data sets, which are widely used to examine the performance of the previous methods[29,31,37,51] over cross-validation tests, are also employed to evaluate our method by performing cross-validation tests. For more detailed information on the data set construction, please refer to refs 1416, and 48−50. Table 1 shows the details of the statistical composition of these data sets. We use the five benchmark data sets mentioned above since they are the latest public benchmark data set for 6mA site prediction. This enables us to directly compare our results with those of other methods. The complete benchmark data sets can be downloaded from https://github.com/XueQiangFan/I-DNAN6mA/.

Table 1. Statistical Summary of the Benchmark Data Sets in This Study

data set	number of positive samples	number of negative samples	total number
A. thaliana	19,616	19,616	39,232
D. melanogaster	10,653	10,653	21,306
6mA-rice-Chen	880	880	1,760
6mA-rice-Lv	154,000	154,000	308,000
Rosaceae	36,535	36,733	73,268

Feature Representation. Deep neural networks can automatically learn high-level representation from sequence-based features, such as nucleotide composition and nucleotide chemical property. Nevertheless, only the observed concatenated feature information of limited spatial dimensions related to target bases is not sufficient for training a high-performance model since the recognition of 6mA sites should be affected by strict base-pairing rules in the DNA sequence. Meanwhile, one of the most important but also the most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector and yet still retain the considerable sequence order information or key pattern characteristics. Recognizing this, in this study, instead of using commonly hand-crafted DNA sequence features, a novel base-pairing structure feature map is designed to improve the prediction performance of 6mA sites.

Unlike the existing 6mA site prediction methods, given an input DNA sequence, I-DNAN6mA first generates a binary one-hot vector representing DNA sequence order information. This one-hot vector is then transformed into one 16 channel "image". In detail, according to the one-hot encoding method, A, T, C, G, and N are encoded as [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1], and [0, 0, 0, 0], respectively. Let $S = (s_1, s_2, s_3, \cdots, s_L)$ be an input DNA sequence of length L, where $s_i \in \{A, C, G, T, N\}$. S is encoded as a matrix $X = [x_1, x_2, x_3, \cdots, x_L]$ using the one-hot encoding rule. Note that since the letter "N" indicates a non-sequenced nucleotide, the input DNA sequence is represented as a 4 by 41 encoding matrix.[29] X is then converted into a $16 \times L \times L$ matrix through a Kronecker product[52] between X and X^T as shown in Figure 1A

$$\Gamma = X \otimes X^T \tag{1}$$

Γ can be understood as an "image" of size $L \times L$ with 16 color channels. Each channel indicates one of the 16 possible base-pairing rules. For instance, the second channel of Γ represents a matrix where A and C are paired. The novel feature representation method makes it more efficient and convenient to mine both long-distance and local intra-sequence dependencies using dilated convolutional neural networks, while also taking into consideration all possible base pairing patterns.[41]

Deep Neural Networks. Figure 1B illustrates the overview of our proposed deep neural network architecture including three stages, i.e., B(i), B(ii), and B(iii), which is utilized to mine short- and long−range interactions and all possible base pairing information from the base-paring map [B(i)], capture knowledge of time series dependencies between nucleotides

https://doi.org/10.1021/acs.jcim.2c01465
J. Chem. Inf. Model. 2023, 63, 1076−1086

Table 2. Comparison of I-DNAN6mA with and without the Base-Pairing Map on the Training Data Sets over Five-Fold Cross-Validation Tests

training data set	method	Sen[a]	Spe[a]	ACC[a]	MCC[a]	auROC[a]	Sen[b]	Sen[c]
A. thaliana	I-DNAN6mA[d]	0.836	0.913	0.874	0.750	0.938	0.845	0.907
	I-DNAN6mA[e]	**0.882**	**0.942**	**0.912**	**0.825**	**0.962**	**0.917**	**0.947**
D. melanogaster	I-DNAN6mA[d]	0.881	0.905	0.893	0.786	0.949	0.885	0.930
	I-DNAN6mA[e]	**0.886**	**0.945**	**0.915**	**0.831**	**0.965**	**0.923**	**0.953**

[a]Results computed with the prediction cutoff threshold value set as 0.5. [b]Results computed with the fixed FPR at 0.1. [c]Results computed with the fixed FPR at 0.2. [d]Results computed by using I-DNAN6mA without the base-pairing map. [e]Results computed by using I-DNAN6mA with the base-pairing map.

[B(ii)], and calculate the probability scores of 6mA sites [B(iii)], for a given DNA sequence, where L is the sequence length of the target DNA, ch denotes the channel number, dr denotes the dilation rate of the convolutional neural network,[53] Norm. denotes the batch normalization layer,[54] Act. denotes the exponential linear unit (elu) activation function,[47] and Drop. denotes the dropout ratio.[55]

The input of I-DNAN6mA is one DNA sequence of length L. Each base, i.e., A, C, G, and T, is first encoded into four-dimensional embeddings. The $L \times 4$-dimensional sequence embeddings are then converted into an "image" of size $16 \times L \times L$ using a trainable embedding function.

The $16 \times L \times L$ matrix is entered into the network B(i) to dig out discriminative information of the base-paring channels map. The network architecture B(i) includes initialization convolution layers, block *a* repeated N_1 times, block *b* repeated N_2 times, block *c* repeated N_3 times, an inception module,[44] and one self-attention layer.[45] The initialization convolution layer, which is used before blocks *a*, *b*, and *c*, transforms the input feature maps into a spatial vector with a larger signal channel. Each block (*a*, *b*, and *c*) consists of two dilated convolutional layers with kernel sizes of 3×3 and 3×3, respectively, and the dilation rate (dr) of 2. After each traditional or dilated convolutional layer, batch normalization and elu activation functions are employed. A dropout strategy with a ratio of 25% is utilized to avoid overfitting. Besides, a module of SENet[56] is embedded in the blocks *a*, *b*, and *c* to capture key position knowledge. The output of block *c* is a feature map with a channel number of 64, which is followed by the inception module. The self-attention block converts the feature map with $192 \times L \times L$ into the feature map with $L \times 192$. The $L \times 196$-dimensional matrix generated by concatenating the output of B(i) and sequence-based one-hot encoding is entered to block *d* repeated N_4 times. The block *d* contains N_5 layers of BiLSTM. A dropout rate of 25% is again utilized in BiLSTM layers.

Finally, the probability scores of 6mA sites are calculated by N_6-layer MLP and a sigmoidal activation function. The layer normalization and elu function with a dropout rate of 25% are also employed.

Model Implementation. The models of I-DNAN6mA are implemented in PyTorch (version 1.7.1) library[57] and trained by using the Adam optimization algorithm[58] with the default learning rate of 0.001. To speed up training, the models are trained on one Nvidia GTX TITAN X graphics processing unit with a batch size of 380. In the model training process, we use the mean squared error function (MSE) to calculate the loss. We use the strategy of a grid search and adjust the network's hyperparameters, i.e., N_1, N_2, N_3, N_4, N_5, and N_6, by observing the model performance on the training data set over five-fold cross-validation tests. Finally, according to the best performance of the I-DNAN6mA model, we use the following values for the above hyperparameters: $N_1 = N_2 = N_3 = 8$, $N_4 = 3$, $N_5 = 2$, and $N_6 = 5$.

Evaluation Indexes. DNA 6mA site prediction is a binary classification problem. In this study, four evaluation metrics, i.e., sensitivity (Sen), specificity (Spe), accuracy (ACC), and Matthew's correlation coefficient (MCC), are used to evaluate the predictive method performances. They are, respectively, defined as follows

$$\text{Sen} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Spe} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

$$\text{MCC} = \frac{\text{TP}\cdot\text{TN} - \text{FP}\cdot\text{FN}}{\sqrt{(\text{TP} + \text{FP})\cdot(\text{TP} + \text{FN})\cdot(\text{TN} + \text{FP})\cdot(\text{TN} + \text{FN})}} \tag{5}$$

where TP (true positive) means the number of correctly predicted 6mA sites; FP (false positive) means the number of not correctly predicted 6mA sites; TN (true negative) means the number of correctly predicted non-6mA sites; and FN (false negative) means the number of not correctly predicted non-6mA sites. The MCC measures the correlation between the expected class and the predicted class. Furthermore, we also use the area under the receiver operating characteristic (ROC) curve (auROC), which is calculated by using the scikit-learn[59] tool, to comprehensively evaluate the performance of the prediction method.

EXPERIMENTAL RESULTS AND ANALYSIS

Base-Pairing Map Significantly Improves the Performance of I-DNAN6mA. This section examines to what extent the base-pairing map can help predict 6mA sites. The results of our proposed methods with (I-DNAN6mA*) and without (I-DNAN6mA#) the base-pairing map are obtained using five-fold cross-validation tests on the training data sets. Specifically, (i) the training data sets are randomly divided into five non-overlapping subsets; (ii) in each testing stage, the proposed method trains the prediction model on four subsets, and meanwhile, the remaining subset is employed to evaluate the performance of the trained model; and (iii) the averages of five-fold cross-validation tests are calculated as the final results. The five-fold cross-validation tests results of I-DNAN6mA without and with the base-pairing map are summarized in Table 2.

https://doi.org/10.1021/acs.jcim.2c01465
J. Chem. Inf. Model. 2023, 63, 1076−1086

Table 3. Comparison of I-DNAN6mA with Conventional ML-Based Methods on the Independent Validation Data Sets

testing data set	method	Sen[a]	Spe[a]	ACC[a]	MCC[a]	auROC[a]	Sen[b]	Sen[c]
A. thaliana	LR	0.859	0.853	0.856	0.712	0.925	0.826	0.892
	KNN	**0.910**	0.730	0.819	0.650	0.905	0.729	0.857
	DT	0.845	0.812	0.828	0.657	0.828	0.447	0.847
	NB	0.821	0.862	0.842	0.684	0.910	0.778	0.863
	Bagging	0.835	0.923	0.880	0.763	0.939	0.854	0.907
	AB	0.862	0.840	0.851	0.702	0.921	0.810	0.882
	GB	0.858	0.897	0.878	0.756	0.940	0.856	0.901
	LDA	0.859	0.849	0.854	0.709	0.923	0.824	0.885
	I-DNAN6mA	0.896	**0.935**	**0.915**	**0.831**	**0.967**	**0.917**	**0.955**
D. melanogaster	LR	0.882	0.884	0.882	0.766	0.941	0.864	0.916
	KNN	**0.929**	0.641	0.788	0.606	0.909	0.779	0.878
	DT	0.841	0.817	0.828	0.658	0.829	0.459	0.843
	NB	0.831	0.888	0.860	0.721	0.928	0.822	0.885
	Bagging	0.856	0.914	0.885	0.771	0.936	0.864	0.909
	AB	0.888	0.858	0.873	0.747	0.938	0.842	0.907
	GB	0.880	0.900	0.891	0.782	0.951	0.880	**0.924**
	LDA	0.883	0.872	0.878	0.756	0.939	0.854	0.911
	I-DNAN6mA	0.903	**0.952**	**0.927**	**0.855**	**0.963**	**0.885**	0.919

[a]Results computed with the prediction cutoff threshold value set as 0.5. [b]Results computed with the fixed FPR at 0.1. [c]Results computed with the fixed FPR at 0.2.

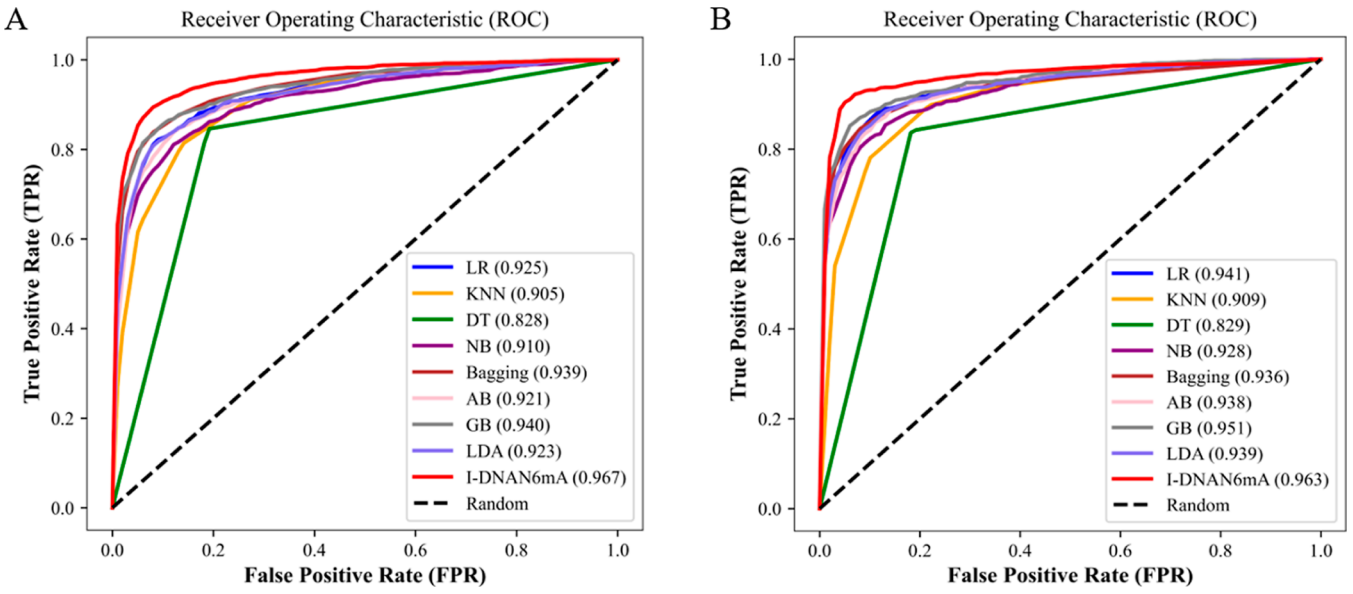


Figure 2. ROC curves of I-DNAN6mA and conventional ML-based methods on both independent testing data sets. (A) *A. thaliana* and (B) *D. melanogaster*.

From Table 2, we can observe that I-DNAN6mA* is consistently superior to I-DNAN6mA[#] concerning the five evaluation indexes, i.e., Sen, Spe, ACC, MCC, and auROC. The Sen, Spe, ACC, MCC, and auROC of I-DNAN6mA* are 0.882, 0.942, 0.912, 0.825, and 0.962, which are 5.50, 3.18, 4.35, 10.00, and 2.56% higher than those of I-DNAN6mA[#], respectively, on the training data set *A. thaliana*. Similar effects are well-established in the training data set *D. melanogaster*. Compared with I-DNAN6mA[#], I-DNAN6mA* achieves 0.57, 4.42, 2.46, 5.73, and 1.69% improvements in Sen, Spe, ACC, MCC, and auROC, respectively. In addition, Table 2 also lists the performance comparison of I-DNAN6mA* and I-DNAN6mA[#] concerning Sen under the false positive rate (FPR) = 0.1 and 0.2 (FPR = 1-Spe). For both species, it can be found that I-DNAN6mA* consistently performs the best under the fixed FPR. The comparison results shown in Table 2 suggest that the prediction performance is indeed improved after using the base-pairing map.

Performance Comparison with Conventional ML-Based Methods. To empirically examine the predictive performance of our method, we compare our method with other methods based on conventional ML algorithms, including logistic regression (LR), K-nearest neighbor (KNN), decision tree (DT), Gaussian NB (NB), bagging, AdaBoost (AB), gradient boosting (GB), and linear discriminant analysis (LDA). Among them, the number of LR iterations is 500 with an L2 penalty, the neighbors of the KNN method are set as 7, and DT, NB, bagging, AB, GB, and LDA use default parameters. We also try to use the support vector machine with the RBF kernel function to implement 6mA site prediction. However, it takes too long to produce the results. Noted that these eight different supervised algorithms take the

https://doi.org/10.1021/acs.jcim.2c01465
J. Chem. Inf. Model. 2023, 63, 1076−1086

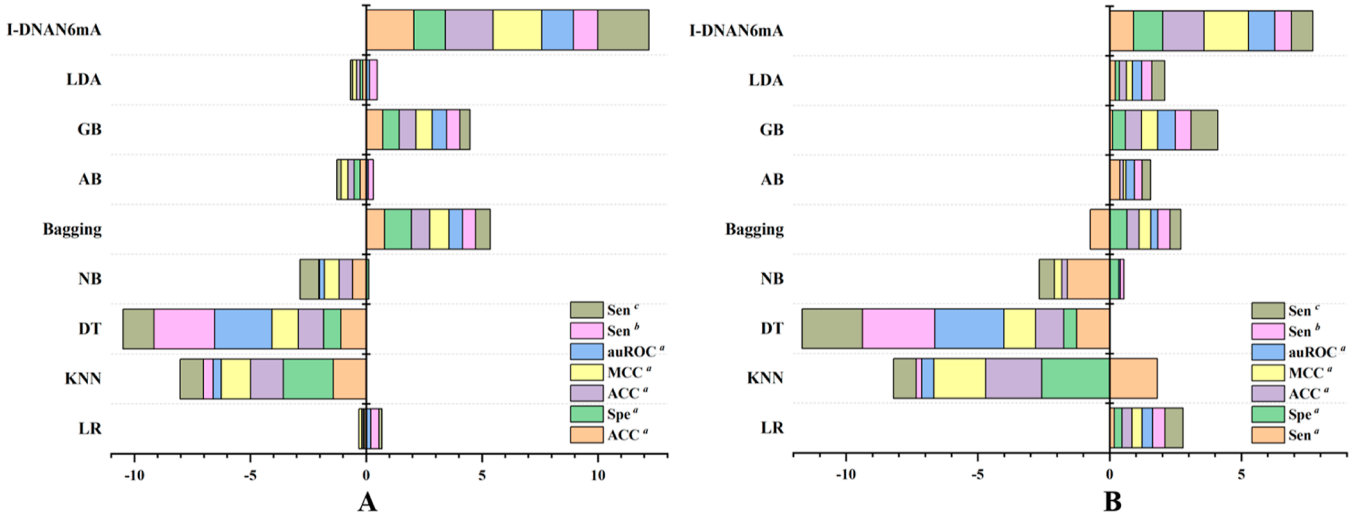


Figure 3. Ranking of the methods in the global performance evaluation. (A,B) are ranked according to the sum of the Z-scores of all the evaluation metrics on the *A. thaliana* and *D. melanogaster* data sets, respectively.

Table 4. Performance Comparison between the Proposed I-DNAN6mA Method and Other Existing Methods for Identifying 6mA Sites on the Independent Testing Data Sets in the *A. thaliana* and *D. melanogaster* Genomes

testing data set	method	Sen[a]	Spe[a]	ACC[a]	MCC[a]	auROC[a]	Sen[b]	Sen[c]
A. thaliana	DeepM6A[d]	0.894	0.931	0.913	0.826	0.966	**0.920**	**0.956**
	i6mA-DNC[d]	0.846	0.909	0.878	0.757	0.944	0.853	0.912
	iDNA6mA[d,e]	0.843	0.889	0.866	0.733	0.932	0.833	0.902
	3-mer-LR[d]	0.669	0.728	0.699	0.397	0.773	0.411	0.577
	LA6mA[d]	**0.899**	0.917	0.909	0.817	0.962	0.912	0.948
	AL6mA[d]	0.862	0.905	0.884	0.768	0.945	0.867	0.927
	I-DNAN6mA	0.896	**0.935**	**0.915**	**0.831**	**0.967**	0.917	0.955
D. melanogaster	DeepM6A[d]	0.901	0.939	0.920	0.841	**0.969**	**0.930**	**0.959**
	i6mA-DNC[d]	0.869	0.917	0.893	0.787	0.947	0.878	0.916
	iDNA6mA[d,e]	0.883	0.843	0.863	0.727	0.937	0.846	0.904
	3-mer-LR[d]	0.680	0.702	0.691	0.383	0.753	0.347	0.558
	LA6mA[d]	**0.909**	0.915	0.912	0.824	0.966	0.921	0.955
	AL6mA[d]	0.840	0.916	0.878	0.758	0.941	0.848	0.920
	I-DNAN6mA	0.903	**0.952**	**0.927**	**0.855**	0.963	0.885	0.919

[a]Results computed with the prediction cutoff threshold value set as 0.5. [b]Results computed with the fixed FPR at 0.1. [c]Results computed with the fixed FPR at 0.2. [d]Results excerpted from ref 48. [e]iDNA6mA stands for iDNA6mA (five-step rule).

one-hot encoding as their inputs and train the models of 6mA site identification on the training data sets. The Scikit-learn Python library[59] is adopted to implement and tune the prediction model of the aforementioned nine classifiers on Windows Server 10 (Version 21H1) Inter(R) Core (TM) i7-9750H CPU @2.60 Hz 2.59 GHz, and 16.0 GB of RAM. The independent validation results of I-DNAN6mA and these methods are displayed in Table 3.

As can be seen from Table 3, it is apparent that the performance of I-DNAN6mA is superior to that of other eight classifiers. Specifically, I-DNAN6mA achieves 3.52, 5.01, 4.13, and 9.63% average improvements in Sen, Spe, ACC, and MCC, respectively, on the two independent testing data sets, compared with the second best performer of GB. Taking results on *D. melanogaster* as an example, the Sen, Spe, ACC, and MCC of I-DNAN6mA are 0.903, 0.952, 0.927, and 0.855, respectively, which are 2.61, 5.78, 4.04, and 9.34% higher than the corresponding values measured for GB. To compare the nine methods more systematically, the ROC curves are plotted in Figure 2. The auROC values of I-DNAN6mA are 0.967 and 0.963 on both independent testing data sets, respectively. We also rank the methods by using the sum of the *Z*-scores of global metrics to analyze the comprehensive performance of various 6mA site prediction methods. It can be found that I-DNAN6mA has the best comprehensive performance in both *A. thaliana* (Figure 3A) and *D. melanogaster* (Figure 3B) genomes. The results show that our proposed method consistently outperforms other methods.

Performance Comparison in *A. thaliana* and *D. melanogaster* Genomes. The purpose of this section is to experimentally demonstrate the efficacy of the proposed I-DNAN6mA by comparing it with other recently state-of-the-art 6mA site prediction methods on both independent test data sets, including DeepM6A,[32] i6mA-DNC,[34] iDNA6mA (five-step rule),[33] 3-mer-LR,[35] LA6mA,[35] and AL6mA.[35] For an objective and fair comparison, all the methods use the same training data sets and independent testing data sets. Table 4 summarizes the compared results.

As described in Table 4, the novel method I-DNAN6mA proposed in this study achieves Sen, Spe, ACC, MCC, and auROC of 0.896 and 0.903, 0.935 and 0.952, 0.915 and 0.927, 0.831 and 0.855, and 0.967 and 0.885, respectively, on both

https://doi.org/10.1021/acs.jcim.2c01465
J. Chem. Inf. Model. 2023, 63, 1076−1086

Table 5. Performance Comparison between the Proposed I-DNAN6mA Method and Other Existing Methods for Identifying 6mA Sites over the Jackknife Tests in the *6mA-rice-Chen* Genome

testing data set	method	Sen	Spe	ACC	MCC	auROC	Sen[a]	Sen[b]
6mA-rice-Chen	i6mA-Pred[c]	0.830	0.833	0.831	0.660	0.886	-	-
	PseDNC[c]	0.635	0.656	0.646	0.290	0.636	-	-
	iDNA6mA (5-step rule)[d]	0.867	0.866	0.866	0.730	0.931	-	-
	iDNA6mA-Rice[d]	0.839	0.834	0.836	0.670	0.910	-	-
	i6mA-DNCP[e]	0.844	0.889	0.867	0.734	0.926	-	-
	SVM-based method[f]	0.863	0.839	0.851	0.701	0.915	-	-
	I-DNAN6mA	**0.850**	**0.917**	**0.882**	**0.763**	**0.947**	0.846	0.923

[a]Results computed with the fixed FPR at 0.1. [b]Results computed with the fixed FPR at 0.2. [c]Results excerpted from ref 14. [d]Results excerpted from ref 33. [e]Results excerpted from ref 16. [f]Results excerpted from ref 15; "-" means that the value is not given.

Table 6. Performance Comparison between the Proposed I-DNAN6mA Method and Other Existing Methods for Identifying 6mA Sites over Five-Fold Cross-Validation Tests in the *6mA-rice-Lv* Genome

testing data set	method	Sen	Spe	ACC	MCC	auROC	Sen[a]	Sen[b]
6mA-rice-Lv	Deep6mA[c]	**0.951**	0.929	**0.940**	**0.880**	**0.980**	-	-
	SNNRice6mA-large[c]	0.943	0.897	0.920	0.840	0.970	-	-
	MM-6mAPred[c]	0.934	0.895	0.914	0.830	0.960	-	-
	Deep6mAPred[d]	0.954	0.925	0.939	0.879	0.979	-	-
	I-DNAN6mA	0.947	**0.930**	0.938	0.877	0.976	0.963	0.979

[a]Results computed with the fixed FPR at 0.1. [b]Results computed with the fixed FPR at 0.2. [c]Results excerpted from ref 29. [d]Results excerpted from ref 37 "-" means that the value is not given.

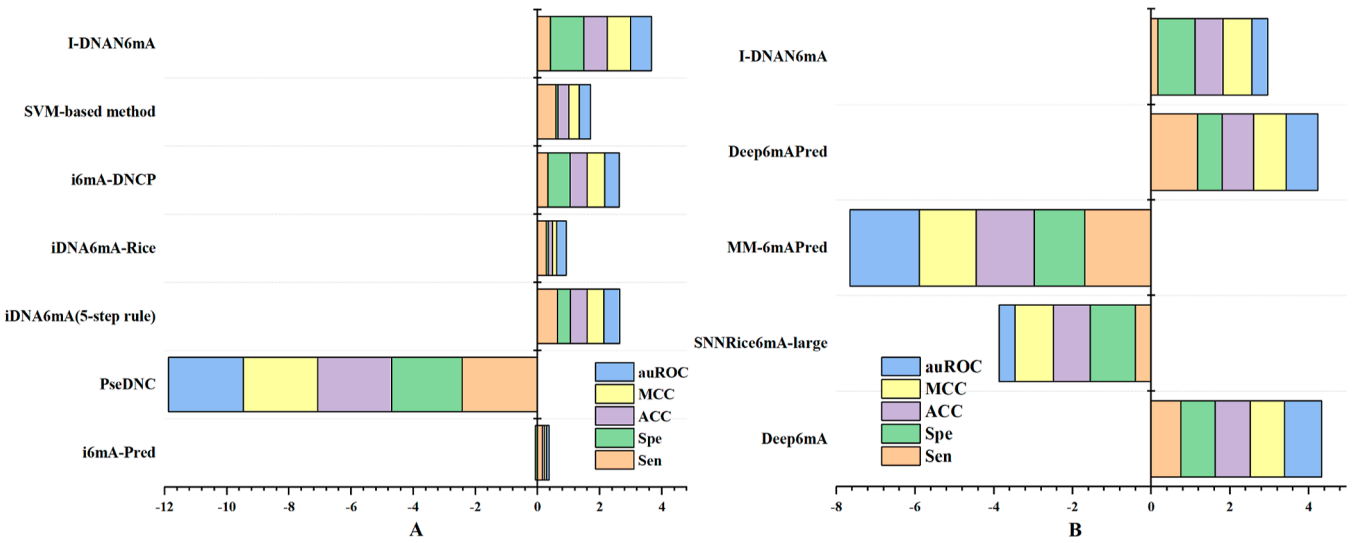


Figure 4. Ranking of the methods in the global performance evaluation. I-DNAN6mA and other existing methods are ranked according to the sum of the Z-scores of all the evaluation indexes in the *Rice* genome. (A) Performance on the *6mA-rice-Chen* data set. (B) Performance on the *6mA-rice-Lv* data set.

independent testing data sets. Although the existing methods gain a reasonable MCC of 0.383−0.841, the improvement of I-DNAN6mA is significant. Specifically, the ACC and MCC of I-DNAN6mA are 0.22−34.15% and 0.61−123.24% higher than those of the existing methods, respectively. As expected, 3-mer-LR, which is developed based on the individual classifier LR algorithm, gains the lowest prediction performance in terms of five evaluation indexes. Table 4 also presents the performance comparison of different methods concerning Sen for FPR = 0.1 and 0.2. For two independent testing data sets, it is easy to find that DeepM6A performs the best under the fixed FPR followed by I-DNAN6mA. By revisiting Table 4, it is noteworthy that although the five DL-based methods, i.e., DeepM6A, i6mA-DNC, iDNA6mA (five-step rule), LA6mA, and AL6mA, exhibit good performance, the proposed I-DNAN6mA is the sole approach that achieves Sen > 0.896, Spe > 0.935, ACC > 0.915, and MCC > 0.831 on both model organisms. In addition, it is clear that I-DNAN6mA has the largest improvement in the MCC, which is an overall index for evaluating the quality of binary identification, among the five evaluation indexes on two independent testing data sets, suggesting that I-DNAN6mA shows the best overall performance.

Performance Comparison in the *Rice* Genome. In this section, the effectiveness of our proposed I-DNAN6mA is evaluated on the data set *6mA-rice-Chen* over jackknife tests, compared with that of other 6mA site prediction methods, including i6mA-Pred,[14] PseDNC,[14,60] iDNA6mA (five-step rule),[33] iDNA6mA-Rice,[16] i6mA-DNCP,[15] and the SVM-based method.[15] The jackknife test results of these methods and I-

https://doi.org/10.1021/acs.jcim.2c01465
J. Chem. Inf. Model. 2023, 63, 1076−1086

Table 7. Performance Comparison between the Proposed I-DNAN6mA Method and Other Existing Methods for Identifying 6mA Sites on the Independent Testing Data Set in the *Rosaceae* Genome

tfesting data set	Method	Sen	Spe	ACC	MCC	auROC	Sen[a]	Sen[b]
Rosaceae	Meta-i6mA[c]	0.954	0.951	0.953	0.905	-	-	-
	i6mA-Fuse_FV[c]	0.924	**0.962**	0.943	0.887	-	-	-
	i6mA-Fuse_RC[c]	0.890	0.895	0.893	0.786	-	-	-
	i6mA-stack_FV[c]	0.928	0.927	0.928	0.856	-	-	-
	i6mA-stack_RC[c]	0.920	0.877	0.899	0.798	-	-	-
	i6mA-Pred[c]	0.897	0.782	0.840	0.684	-	-	-
	iDNA6mA-Rice[c]	0.951	0.805	0.878	0.764	-	-	-
	MM-6mAPred[c]	0.961	0.785	0.873	0.758	-	-	-
	6mA-Finder[c]	0.928	0.764	0.846	0.701	-	-	-
	i6mA-vote[c]	0.955	0.954	0.955	0.909	-	-	-
	I-DNAN6mA	**0.962**	0.961	**0.962**	**0.924**	0.990	0.980	0.991

[a]Results computed with the fixed FPR at 0.1. [b]Results computed with the fixed FPR at 0.2. [c]Results excerpted from ref 49; "-" means that the value is not given.

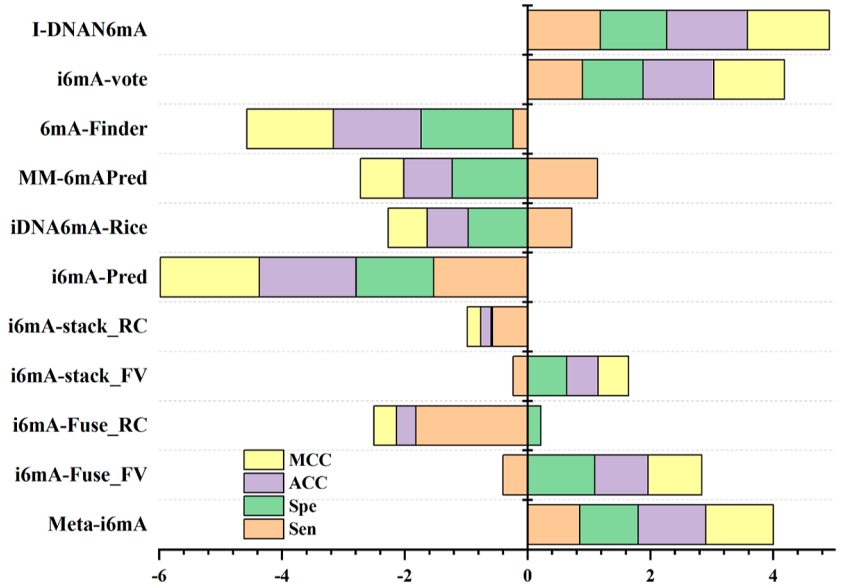


Figure 5. Ranking of the methods in the global performance evaluation. I-DNAN6mA and other existing methods are ranked according to the sum of the *Z*-scores of all the evaluation metrics on the *Rosaceae* genome.

DNAN6mA are summarized in Table 5. From Table 5, according to the evaluation indexes, we can find that I-DNAN6mA acts as the best performer followed by i6mA-DNCP, iDNA6mA (five-step rule), the SVM-based method, iDNA6mA-Rice, i6mA-Pred, and PseDNC. Compared with the second best method i6mA-DNCP, I-DNAN6mA achieves the improvements of 3.15, 1.73, 3.95, and 2.27% on Spe, ACC, MCC, and auROC, respectively. Although the SVM-based method obtains the highest Sen (0.863), its MCC and auROC are 8.13 and 3.38% lower than those of I-DNAN6mA.

To further verify the performance of I-DNAN6mA, we also use five-fold cross-validation to compare the prediction performance of Deep6mA,[29] SNNRice6mA-large,[31] MM-6mAPred,[51] and Deep6mAPred[37] on the *6mA-rice-Lv* data set. It is straightforward to find from Table 6 that the performance achieved by I-DNAN6mA is consistently equal to or better than that of the compared methods. Also, we rank these methods using the sum of Z-scores of all evaluation indexes to analyze the combined performance of I-DNAN6mA and other comparison methods. Figure 4 shows the comprehensive performance of all methods in the *Rice* genome. It can be found that the comprehensive performance of I-DNAN6mA is the best among all methods on the *6mA-rice-Chen* data set; meanwhile, I-DNAN6mA gives the third best performance on the *6mA-rice-Lv* data set.

Performance Comparison in the *Rosaceae* Genome. To further evaluate the power of I-DNAN6mA, its performance is also assessed on the independent testing data set of the *Rosaceae* genome, compared to eight control methods, i.e., Meta-i6mA,[50] i6mA-Fuse,[61] i6mA-stack,[19] i6mA-Pred,[14] iDNA6mA-Rice,[16] MM-6mAPred,[51] 6mA-Finder,[18] and i6mA-vote.[49] Among them, i6mA-Fuse contains both prediction models, which are renamed i6mA-Fuse_FV and i6mA-Fuse_RC, respectively, for the sake of description. The same is true for the i6mA-stack. Note that in this section, the prediction model of I-DNAN6mA is built on the training data set of the *Rosaceae* genome. The comparative results are demonstrated in Table 7. From Table 7, it is easily found that I-DNAN6mA consistently outperforms other state-of-the-art methods concerning all four evaluation indexes. More specifically, the Sen, Spe, ACC, and MCC of I-DNAN6mA are 0.962, 0.961, 0.962, and 0.924, respectively, which are 0.73, 0.74, 0.73, and 1.65% higher than those of the second best method i6mA-vote, respectively, where the prediction model is

https://doi.org/10.1021/acs.jcim.2c01465
J. Chem. Inf. Model. 2023, 63, 1076−1086

also trained on the same training data set. Furthermore, although i6mA-Fuse_FV achieves the highest Spe value (0.962), its MCC value is 4.00% lower than that of I-DNAN6mA.

In addition, to further investigate the overall performance of I-DNAN6mA and other comparative methods, we rank these methods using the sum of *Z*-scores of all evaluation metrics. The sum of *Z*-scores of all evaluation metrics given in Figure 5 shows that the comprehensive performance of I-DNAN6mA significantly outperforms other comparison methods. By revisiting Tables 3, 4, and 7 on the results of the independent testing experiments, we can find an interesting fact, which is that I-DNAN6mA gives the best performance in the *Rosaceae* genome with larger data, followed by the *A. thaliana* genome and the *D. melanogaster* genome. This phenomenon reveals that the 6mA site predictive performance of I-DNAN6mA should gradually increase as the training data grow. Thus, to improve the performance further, constructing one large high-quality data set will be an effective strategy.

CONCLUSIONS

Accurate identification of 6mA sites in DNA is crucial to elucidate the function of 6mA epigenetic modification. In this study, we have developed and implemented a novel method, termed I-DNAN6mA, to identify 6mA sites from DNA sequence information. Experimental results have demonstrated the efficacy of the proposed I-DNAN6mA via comparison to several state-of-the-art 6mA site prediction methods on five benchmark data sets. The excellent performance of I-DNAN6mA is mainly attributed to the following reasons: (i) several high-quality benchmark data sets are used; (ii) to the best of our knowledge, a novel image-like representation of DNA sequences is employed for the first time; and (iii) a well-designed three-stage DL model with pairwise input can effectively distinguish knowledge embedded between positive and negative samples. Finally, based on the proposed I-DNAN6mA, we implement a new standalone version predictor for predicting 6mA sites, which is freely available at https://github.com/XueQiangFan/I-DNAN6mA/for academic use.

Despite its good performance, the proposed I-DNAN6mA still has potential disadvantages. For instance, I-DNAN6mA may lack the ability to remove noisy information from feature sources. Furthermore, the performance of I-DNAN6mA for DNA 4mA site identification is not yet known. Our further research work comprises the following five directions to further enhance the prediction efficacy of 6mA sites: (i) designing a high discriminative feature source; (ii) developing an excellent feature optimization tool to get rid of the noise in the feature; (iii) designing a more accurate method by combining I-DNAN6mA and other state-of-the-art 6mA site prediction methods; (iv) investigating the applicability of I-DNAN6mA to identify DNA 4mA sites; and (v) based on the proposed I-DNAN6mA, establishing a user-friendly web server to help potential researchers. Finally, although the proposed I-DNAN6mA still has room for optimization, we believe that it will be exploited as a useful tool to speed up the progress of DNA function detection and understanding.

ASSOCIATED CONTENT

Data Availability Statement

The data and I-DNAN6mA tool are freely accessible at https://github.com/XueQiangFan/I-DNAN6mA/.

AUTHOR INFORMATION

Corresponding Authors

Jun Hu − *College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China*; Email: hujunum@zjut.edu.cn

Zhong-Yi Guo − *School of Computer and Information, Hefei University of Technology, Hefei 230009, China*; orcid.org/0000-0001-7282-2503; Email: guozhongyi@hfut.edu.cn

Authors

Xue-Qiang Fan − *School of Computer and Information, Hefei University of Technology, Hefei 230009, China*

Bing Lin − *School of Computer and Information, Hefei University of Technology, Hefei 230009, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c01465

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (no. 61775050) and the Fundamental Research Funds for the Central Universities (no. PA2019GDZC0098).

ABBREVIATIONS

6mA, N^6-methyladenine; MeDIP-seq, methylated DNA immunoprecipitation sequencing; CE-LIF, capillary electrophoresis and laser-induced fluorescence; SMRT-seq, single-molecule real-time sequencing; ML, machine learning; DL, deep learning; BiLSTM, bidirectional long short-term memory recurrent neural networks; elu, exponential linear unit activation function; MLP, multilayer perceptron; MCC, Matthew's correlation coefficient

REFERENCES

(1) Luo, G.-Z.; Blanco, M. A.; Greer, E. L.; He, C.; Shi, Y. DNA N6-methyladenine: a new epigenetic mark in eukaryotes? *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 705−710.

(2) Wu, K.-J. The epigenetic roles of DNA N6-methyladenine (6mA) modification in eukaryotes. *Cancer Lett.* **2020**, *494*, 40−46.

(3) Luo, G.-Z.; Wang, F.; Weng, X.; Chen, K.; Hao, Z.; Yu, M.; Deng, X.; Liu, J.; He, C. Characterization of eukaryotic DNA N6-methyladenine by a highly sensitive restriction enzyme-assisted sequencing. *Nat. Commun.* **2016**, *7*, 11301.

(4) Liu, J.; Zhu, Y.; Luo, G.-Z.; Wang, X.; Yue, Y.; Wang, X.; Zong, X.; Chen, K.; Yin, H.; Fu, Y. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nat. Commun.* **2016**, *7*, 13052.

(5) Yao, B.; Cheng, Y.; Wang, Z.; Li, Y.; Chen, L.; Huang, L.; Zhang, W.; Chen, D.; Wu, H.; Tang, B. DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. *Nat. Commun.* **2017**, *8*, 1122.

(6) Zhang, G.; Huang, H.; Liu, D.; Cheng, Y.; Liu, X.; Zhang, W.; Yin, R.; Zhang, D.; Zhang, P.; Liu, J.; Li, C.; Liu, B.; Luo, Y.; Zhu, Y.; Zhang, N.; He, S.; He, C.; Wang, H.; Chen, D. N6-methyladenine DNA modification in Drosophila. *Cell* **2015**, *161*, 893−906.

(7) Douvlataniotis, K.; Bensberg, M.; Lentini, A.; Gylemo, B.; Nestor, C. E. No evidence for DNA N 6-methyladenine in mammals. *Sci. Adv.* **2020**, *6*, No. eaay3335.

(8) Xie, Q.; Wu, T. P.; Gimple, R. C.; Li, Z.; Prager, B. C.; Wu, Q.; Yu, Y.; Wang, P.; Wang, Y.; Gorkin, D. U.; Zhang, C.; Dowiak, A. V.; Lin, K.; Zeng, C.; Sui, Y.; Kim, L. J. Y.; Miller, T. E.; Jiang, L.; Lee, C.

H.; Huang, Z.; Fang, X.; Zhai, K.; Mack, S. C.; Sander, M.; Bao, S.; Kerstetter-Fogle, A. E.; Sloan, A. E.; Xiao, A. Z.; Rich, J. N. N6-methyladenine DNA modification in glioblastoma. *Cell* **2018**, *175*, 1228−1243.
(9) Liu, L.; Wang, Y.; Wu, J.; Liu, J.; Qin, Z.; Fan, H. N6-methyladenosine: a potential breakthrough for human cancer. *Mol. Ther.–Nucleic Acids* **2020**, *19*, 804−813.
(10) Wu, T. P.; Wang, T.; Seetin, M. G.; Lai, Y.; Zhu, S.; Lin, K.; Liu, Y.; Byrum, S. D.; Mackintosh, S. G.; Zhong, M.; Tackett, A.; Wang, G.; Hon, L. S.; Fang, G.; Swenberg, J. A.; Xiao, A. Z. DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature* **2016**, *532*, 329−333.
(11) Pomraning, K. R.; Smith, K. M.; Freitag, M. Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods* **2009**, *47*, 142−150.
(12) Krais, A. M.; Cornelius, M. G.; Schmeiser, H. H. Genomic N6-methyladenine determination by MEKC with LIF. *Electrophoresis* **2010**, *31*, 3548−3551.
(13) Flusberg, B. A.; Webster, D. R.; Lee, J. H.; Travers, K. J.; Olivares, E. C.; Clark, T. A.; Korlach, J.; Turner, S. W. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **2010**, *7*, 461−465.
(14) Chen, W.; Lv, H.; Nie, F.; Lin, H. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* **2019**, *35*, 2796−2800.
(15) Kong, L.; Zhang, L. i6mA-DNCP: computational identification of DNA N6-methyladenine sites in the rice genome using optimized dinucleotide-based features. *Genes* **2019**, *10*, 828.
(16) Lv, H.; Dao, F.-Y.; Guan, Z.-X.; Zhang, D.; Tan, J.-X.; Zhang, Y.; Chen, W.; Lin, H. iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Front. Genet.* **2019**, *10*, 793.
(17) 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. *Mol. Ther.–Nucleic Acids* **2019**, *18*, 131−141.
(18) Xu, H.; Hu, R.; Jia, P.; Zhao, Z. 6mA-Finder: a novel online tool for predicting DNA N6-methyladenine sites in genomes. *Bioinformatics* **2020**, *36*, 3257−3259.
(19) Khanal, J.; Lim, D. Y.; Tayara, H.; Chong, K. T. i6ma-stack: a stacking ensemble-based computational prediction of dna n6-methyladenine (6ma) sites in the rosaceae genome. *Genomics* **2021**, *113*, 582−592.
(20) Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18−28.
(21) Qi, Y.Random forest for bioinformatics. In *Ensemble Machine Learning*; Springer, 2012; pp 307−323. DOI: 10.1007/978-1-4419-9326-7_11
(22) Lou, W.; Wang, X.; Chen, F.; Chen, Y.; Jiang, B.; Zhang, H. Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes. *PLoS One* **2014**, *9*, No. e86703.
(23) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(24) Yang, J.; Lang, K.; Zhang, G.; Fan, X.; Chen, Y.; Pian, C. SOMM4mC: a second-order Markov model for DNA N4-methylcytosine site prediction in six species. *Bioinformatics* **2020**, *36*, 4103−4105.
(25) Pian, C.; Yang, Z.; Yang, Y.; Zhang, L.; Chen, Y. Identifying RNA N6-Methyladenine sites in three species based on a markov model. *Front. Genet.* **2021**, *12*, 650803.
(26) Abbas, Z.; Tayara, H.; Chong, K. Spinenet-6ma: A novel deep learning tool for predicting dna n6-methyladenine sites in genomes. *IEEE Access* **2020**, *8*, 201450−201457.
(27) Cai, J.; Xiao, G.; Su, R. GC6mA-Pred: A deep learning approach to identify DNA N6-methyladenine sites in the rice genome. *Methods* **2022**, *204*, 14.
(28) Lv, Z.; Ding, H.; Wang, L.; Zou, Q. A convolutional neural network using dinucleotide one-hot encoder for identifying DNA N6-methyladenine sites in the rice genome. *Neurocomputing* **2021**, *422*, 214−221.
(29) Li, Z.; Jiang, H.; Kong, L.; Chen, Y.; Lang, K.; Fan, X.; Zhang, L.; Pian, C. Deep6mA: A deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species. *PLoS Comput. Biol.* **2021**, *17*, No. e1008767.
(30) Xue, T.; Zhang, S.; Qiao, H. i6mA-VC: A multi-classifier voting method for the computational identification of DNA N6-methyladenine sites. *Interdiscip. Sci.: Comput. Life Sci.* **2021**, *13*, 413−425.
(31) Yu, H.; Dai, Z. SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Front. Genet.* **2019**, *10*, 1071.
(32) Tan, F.; Tian, T.; Hou, X.; Yu, X.; Gu, L.; Mafra, F.; Gregory, B. D.; Wei, Z.; Hakonarson, H. Elucidation of DNA methylation on N6-adenine with deep learning. *Nat. Mach. Intell.* **2020**, *2*, 466−475.
(33) Tahir, M.; Tayara, H.; Chong, K. T. iDNA6mA (5-step rule): identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. *Chemom. Intell. Lab. Syst.* **2019**, *189*, 96−101.
(34) Park, S.; Wahab, A.; Nazari, I.; Ryu, J. H.; Chong, K. T. i6mA-DNC: Prediction of DNA N6-Methyladenosine sites in rice genome based on dinucleotide representation using deep learning. *Chemom. Intell. Lab. Syst.* **2020**, *204*, 104102.
(35) Zhang, Y.; Liu, Y.; Xu, J.; Wang, X.; Peng, X.; Song, J.; Yu, D. J. Leveraging the attention mechanism to improve the identification of DNA N6-methyladenine sites. *Briefings Bioinf.* **2021**, *22*, bbab351.
(36) Huang, Q.; Zhang, J.; Wei, L.; Guo, F.; Zou, Q. 6mA-RicePred: a method for identifying DNA N 6-methyladenine sites in the rice genome based on feature fusion. *Front. Plant Sci.* **2020**, *11*, 4.
(37) Tang, X.; Zheng, P.; Li, X.; Wu, H.; Wei, D.-Q.; Liu, Y.; Huang, G. Deep6mAPred: A CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species. *Methods* **2022**, *204*, 142.
(38) Albawi, S.; Mohammed, T. A.; Al-Zawi, S.Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET), 2017*; IEEE, 2017; pp 1−6. DOI: 10.1109/icengtechnol.2017.8308186
(39) Iliadis, M.; Spinoulas, L.; Katsaggelos, A. K. Deep fully-connected networks for video compressive sensing. *Digit. Signal Process.* **2018**, *72*, 9−18.
(40) Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computations* **2019**, *31*, 1235−1270.
(41) Fu, L.; Cao, Y.; Wu, J.; Peng, Q.; Nie, Q.; Xie, X. Ufold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **2022**, *50*, No. e14.
(42) Li, Y.; Hu, J.; Zhang, C.; Yu, D. J.; Zhang, Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* **2019**, *35*, 4647−4655.
(43) Sun, S.; Wang, W.; Peng, Z.; Yang, J. RNA inter-nucleotide 3D closeness prediction by deep residual neural networks. *Bioinformatics* **2021**, *37*, 1093−1098.
(44) Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015*, 2015; pp 1−9. DOI: 10.1109/cvpr.2015.7298594
(45) Gao, J.; Zheng, S.; Yao, M.; Wu, P. Precise estimation of residue relative solvent accessible area from Cα atom distance matrix using a deep learning method. *Bioinformatics* **2021**, *38*, 94−98.
(46) Ramchoun, H.; Ghanou, Y.; Ettaouil, M.; Janati Idrissi, M. A.*Multilayer Perceptron: Architecture Optimization and Training*, 2016.
(47) Nguyen, A.; Pham, K.; Ngo, D.; Ngo, T.; Pham, L.An Analysis of State-of-the-art Activation Functions For Supervised Deep Neural Network. In *2021 International Conference on System Science and Engineering (ICSSE)*; IEEE, 2021; pp 215−220. DOI: 10.1109/icsse52999.2021.9538437

(48) Zhang, Y.; Liu, Y.; Xu, J.; Wang, X.; Peng, X.; Song, J.; Yu, D.-J. Leveraging the attention mechanism to improve the identification of DNA N6-methyladenine sites. *Briefings Bioinf.* **2021**, *22*, bbab351.
(49) Teng, Z.; Zhao, Z.; Li, Y.; Tian, Z.; Guo, M.; Lu, Q.; Wang, G. i6mA-Vote: Cross-Species Identification of DNA N6-Methyladenine Sites in Plant Genomes Based on Ensemble Learning With Voting. *Front. Plant Sci.* **2022**, *13*, 845835.
(50) Hasan, M. M.; Basith, S.; Khatun, M. S.; Lee, G.; Manavalan, B.; Kurata, H. Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Briefings Bioinf.* **2021**, *22*, bbaa202.
(51) Pian, C.; Zhang, G.; Li, F.; Fan, X. MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics* **2020**, *36*, 388−392.
(52) Graham, A.*Kronecker Products and Matrix Calculus with Applications*; Courier Dover Publications, 2018.
(53) Yu, F.; Koltun, V.; Funkhouser, T.Dilated residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017; pp 472−480. DOI: 10.1109/cvpr.2017.75
(54) Ioffe, S.; Szegedy, C.Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*; PMLR, 2015; pp 448−456.
(55) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.
(56) Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011−2023.
(57) Pajankar, A.; Joshi, A.Neural Network and PyTorch Basics. In *Hands-on Machine Learning with Python*; Springer, 2022, pp 215−226. DOI: 10.1007/978-1-4842-7921-2_12
(58) Kingma, D.; Ba, J., Adam: A Method for Stochastic Optimization. **2014**, arXiv:1412.6980.
(59) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(60) Chen, W.; Zhang, X.; Brooker, J.; Lin, H.; Zhang, L.; Chou, K.-C. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* **2015**, *31*, 119−120.
(61) Hasan, M.; Manavalan, B.; Shoombuatong, W.; Khatun, M.; Kurata, H. i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. *Plant Mol. Biol.* **2020**, *103*, 225−234.