frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 21 April 2020
doi: 10.3389/fbioe.2020.00274

Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications

*Rao Zeng and Minghong Liao**

Department of Software Engineering, School of Informatics, Xiamen University, Xiamen, China

DNA N4-methylcytosine modification (4mC) plays an essential role in a variety of biological processes. Therefore, accurate identification the 4mC distribution in genome-scale is important for systematically understanding its biological functions. In this study, we present Deep4mcPred, a multi-layer deep learning based predictive model to identify DNA N4-methylcytosine modifications. In this predictor, we for the first time integrate residual network and recurrent neural network to build a multi-layer deep learning predictive system. As compared to existing predictors using traditional machine learning, our proposed method has two advantages. First, our deep learning framework does not need to specify the features when training the predictive model. It can automatically learn the high-level features and capture the characteristic specificity of 4mC sites, benefiting to distinguish true 4mC sites from non-4mC sites. On the other hand, our deep learning method outperforms the traditional machine learning predictors in performance by benchmarking comparison, demonstrating that the proposed Deep4mcPred is more effective in the DNA 4mC site prediction. Moreover, via experimental comparison, we found that attention mechanism introduced into the deep learning framework is useful to capture the critical features. Additionally, we develop a webserver implementing the proposed method for the academic use of research community, which is now available at http://server.malab.cn/Deep4mcPred.

Keywords: DNA N4-methylcytosine, deep learning, site prediction, webserver, feature representation

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Balachandran Manavalan,
Ajou University, South Korea
Bin Liu,
Beijing Institute of Technology, China

****Correspondence:***
Minghong Liao
liao@xmu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *15 January 2020*
Accepted: *16 March 2020*
Published: *21 April 2020*

Citation:
Zeng R and Liao M (2020) Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications. Front. Bioeng. Biotechnol. 8:274. doi: 10.3389/fbioe.2020.00274

INTRODUCTION

Epigenetics refers to the heritable phenotype changes in the function of genes that do not involve alterations in DNA sequence. DNA methylation refers to the binding of a methyl group on the nucleotide of DNA (Liu et al., 2019a) under the action of DNA methyltransferases (Dnmt). As one of the earliest discovered and most in-depth epigenetic regulation mechanisms, it is associated with normal development and plays an essential role in key biological processes including regulating gene expression, regulating mammalian growth and development, mediating X chromosome inactivation, and participating in gene imprinting (Jin et al., 2011). It can be divided into three categories according to the position of methylation modification: N6-methyladenine (6mA), 5-Methylcytosine (5mC) and N4-methylcytosine (4mC) (Chen et al., 2017; Wei et al., 2019a). The most prevalent methylation modification in eukaryotes is 5mC (Luo et al., 2015; Xiao et al., 2018)

which consists of methylation at the fifth position of the cytosine pyrimidine ring and has focused on epigenetic markers in mammals and plants (Liu et al., 2019a), while 6mA (methylations on the sixth position of the adenine purine ring) (Liu et al., 2019a) is the most predominant DNA modification in prokaryote and has been found to be related to the regulation of restriction-modification (R-M) system, DNA mismatch repair, gene expression, and other aspects (Luo et al., 2015; Xiao et al., 2018). With the development of high-throughput techniques, the 4mC (methylations on the fourth position of the cytosine pyrimidine ring) was discovered in bacteria, and found to play an important role in protecting genome from invasion in restriction-modification (R-M) system. Developing methods to explore more biological functions of 4mC is of significance.

Single-molecule real time sequencing (SMRT) technology has been proposed to detect the 4mC and 6mA sites from the whole genome (Flusberg et al., 2010). However, using SMRT techniques to analyze the genome is costly inefficient. Therefore, Yu et al. (2015) proposed 4mC-Tet-assisted bisulfite-sequencing (4mC-TABseq) as a new generation of sequencing technology (Illumina sequencing systems) to identify the genome-wide locations of 4mC for bacterial species more rapidly and cost efficiently. Although the prediction of 4mC sites by this sequencing technique has been improved to some extent, recent studies focus more on the recognition of 4mC sites using machine learning, which is capable of predicting 4mC sites based on genome sequences, without any prior experimental knowledge. There are currently four methods available in literature to identify 4mC sites, including iDNA4mC (Chen et al., 2017), 4mCPred (Su et al., 2018), 4mcPred-SVM (Wei et al., 2018a), and 4mcPred-IFL (Wei et al., 2019a). iDNA4mC, as the first machine learning predictor, encodes sequences by nucleotide chemical properties and nucleotide frequency to features and trains support vector machine (SVM) models for prediction (Liang et al., 2018). Although this method has the ability to distinguish between 4mC and non-4mC sites, the prediction accuracy is relatively low overall. Afterwards, He et al. proposed 4mCPred, an SVM-based predictive model trained with position-specific trinucleotide propensity (PSTNP) and electron-ion interaction potential features. More recently, 4mcPred-SVM and 4mcPred-IFL, proposed by Wei et al., further improve the predictive performance on the same golden benchmark datasets. The former employs a two-step feature optimization strategy to improve the feature representation ability, while the latter uses an iterative feature representation algorithm to learn critical information from several sequential feature models. Even though the above methods have improved the performance for identifying 4mC sites, too few data sets have been adopted to fully reflect the whole genome and to build robust models. Consequently, it is eager and indispensable to develop a robust and strong model to more accurately identify 4mC sites.

In recent years, deep learning is not only developed as a new research direction in machine learning, but also has made a lot of achievements in data mining (Lan et al., 2018), speech recognition (Amodei et al., 2014), machine translation (Sutskever et al., 2014), natural language processing (Collobert and Weston, 2008; Young et al., 2018), and other related fields (Hong et al., 2019; Li and Liu, 2019; Liu et al., 2019b; Yang et al., 2019; Zeng et al., 2019a,b). In the field of computational biology, deep learning has been widely applied, especially in solving the problems of genome sequence-based by convolutional neural networks (CNN) (Nie et al., 2018; Peng et al., 2018; Lv et al., 2019a; Wang et al., 2019; Zhang et al., 2019a; Zou et al., 2019). In this paper, we proposed Deep4mcPred, a multi-layer deep learning based predictive model to identify DNA N4-methylcytosine modifications. In this predictor, we for the first time integrate residual network (He et al., 2016) and recurrent neural network, together with attention mechanism, to build a multi-layer deep learning predictive system. We evaluated and compared our predictor with existing predictors. The comparative results demonstrate that our proposed model can more accurately identify 4mC sites than the state-of-the-art predictors. In addition, the proposed method is implemented by the simple and easy-to-use webserver which is freely available on http://server.malab.cn/Deep4mcPred.

METHODS AND MATERIALS

Dataset Collection

Previous study has demonstrated that a stringent dataset is essential for building a robust predictive model (Zeng et al., 2016, 2017a; Liu et al., 2017; Wei et al., 2017a, 2018b,c; Jin et al., 2019; Liu, 2019; Su et al., 2019). In existing studies, there is one golden benchmark dataset proposed by Chen et al. for performance evaluation and comparison. However, the size of the dataset is too small to train a deep learning model. Accordingly, we constructed a larger dataset in this study. We strictly followed the data processing procedure as introduced in Chen's study. By doing so, we can guarantee our dataset the most representative.

Positive Samples Collection

Specifically, there are three main steps for collecting the positive samples. Firstly, we collected all 41bp long sequences centered with true 4mC sites from the MethSMRT database (Ye et al., 2016). Next, we removed the sequences with Modification QV (modQV) score not <30 as it is the default threshold for invoking the modification location according to the Methylome Analysis Technical Note. Next, we used CD-HIT software (with the threshold of 80%) (Fu et al., 2012) to reduce the identity of the positives, avoiding the potential of performance biased-estimation. Ultimately, following the procedure, we collected the positive samples from three species: *Arabidopsis thaliana* (*A. thaliana*), *Caenorhabditis elegans* (*C. elegans*), and *Drosophila melanogaster* (*D. melanogaster*). The details of the positive samples in the three species are presented in **Table 1**. Note that we randomly picked 20,000 positive samples for model training.

Negative Samples Collection

The negative samples were also cytosine-centered sequences with a length of 41bp but are not recognized by the SMRT sequencing technology. In this case, the number of negative samples per species are much larger than the corresponding positive samples. To avoid the data imbalance problem, we randomly selected

TABLE 1 | Summary of benchmark datasets in three species.

Species	Positives	Negatives	Total
A. thaliana	20,000	20,000	40,000
C. elegans	20,000	20,000	40,000
D. melanogaster	20,000	20,000	40,000

the same number of negative samples with that of the positive samples in corresponding species for model training.

The Framework of the Proposed Deep Learning Method

Figure 1 illustrates the overall predictive framework of the proposed multi-layer deep learning network. For given DNA sequences, neural network is composed of four layers: the input layer, the ResNet layer, the LSTM layer and the attention layer, as seen in **Figure 1**. The first layer is the input layer. The sequences of the dataset are encoded by one-hot method and the obtained features are fed into the subsequent ResNet layer. Through this residual network model, deeper networks can be built than plain CNN models for extracting effective global features. The output feature vectors are utilized as inputs of the LSTM layer. In the LSTM layer, the bidirectional LSTM model is utilized to gather feature information from two directions which has been proven to be more effective than the unidirectional LSTM model. In the last attention layer, the attention mechanism is introduced to integrate the output of the LSTM layer for more relevant feature information. Finally, a fully-connected neural network (FC) is attached after the attention model and the softmax activation function is performed to make predictions.

Sequence Representation Using One-Hot Encoding

Genomic sequences are consisting of four nucleotides: "A" (adenine), "G" (guanine), "C" (cytosine), and "T" (thymine). Undetermined bases are annotated as "N." The nucleotides are represented using one-hot encoding over four bits. For example, "A" is represented as the binary vector (1,0,0,0); "G" is encoded as (0,1,0,0); "C" is encoded as (0,0,1,0); "T" is encoded as (0,0,0,1); and "N" is (0,0,0,0).

Deep Learning Model Architecture

We developed a novel prediction method, namely Deep4mcPred, that integrates Long Short Term Memory (LSTM) recurrent neural network and the attention mechanism into the Residual Networks (ResNet). The overall architecture of our proposed model is shown in **Figure 1**.

Residual Networks (ResNet)

Studies have showed that the overall performance of the network is greatly affected by the number of network layers when it comes to convolutional neural network (CNN). To be specific, the accuracy of the network increases as the depth increases, but when the depth reaches a certain level, the accuracy begins to drop rapidly. This is called the degradation problem, making it difficult to generate very deep neural networks.

To address this, ResNet introduces a residual learning framework to improve the degradation, which has achieved great success in the areas of image classification and item identification in recent studies. The internal residual blocks of ResNet utilize jump connections, alleviating the problem of gradient disappearance caused by the increase of depth in convolutional neural networks.

For an input x, ResNet learns a specific residual function $F(x) = H(x) - x$, whereas $F(x) = H(x)$ for plain CNN. Supposing the residual $F(x) = 0$, then it occurs identity mapping "shortcut." The residual block is performed as follows:

$$y = F(X, \{W_i\}) + x$$

where

$$F = W_2\sigma(x,\ W_i)$$

where the function F denotes the learned residual mapping and σ represents relu. F and x are added element by element under the premise of shortcut connections.

But in fact, the residual $F(x)$ will not be zero, so the dimensions of F and x will be different. The output of the ResNet layer can be formulated as follows:

$$y = F(X, \{W_i\}) + W_s x$$

where W_s is introduced to perform a linear mapping to match the dimensions. Taking consider of ResNet, it allows the stacked layer to extract more distinct features of the input x, resulting in better performance.

Long Short Term Memory (LSTM)

Recurrent Neural Network (RNN) is a powerful neural network for processing sequential data. The parameter learning of the RNN is performed by the back-propagation algorithm over time. When the input sequence is long, a gradient disappearance or gradient explosion problem occurs, which is termed as long-term dependency problem.

LSTM is one type of RNN, which introduces the conception of self-loop to generate a path of continuous gradient flow for a long time and gating mechanism to control the information flow, solving the long-term dependency problem. It was firstly proposed by Hochreiter and Schmidhuber in 1997. From then on, LSTM has achieved considerable success and has been widely used in the fields of handwriting recognition, machine translation, and speech recognition, etc.

The stacked architecture of LSTM is shown in **Figure 1**. The output from the ResNet layers is fed into the subsequent LSTM layer as the input. Then, the LSTM components are updated by the following formulations:

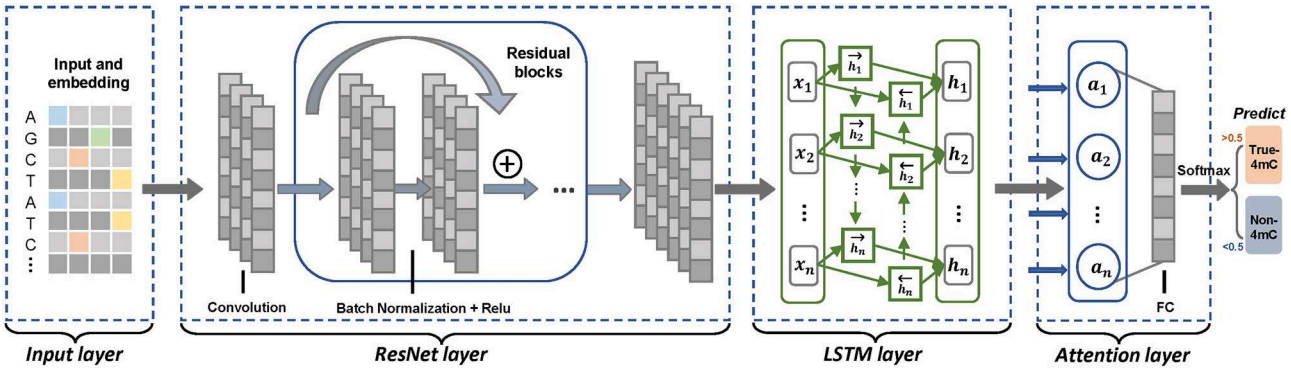

FIGURE 1 | The illustration of the deep neural network. It is a four-layer prediction system. Firstly, given sequences are fed to the input layer for feature representation using one-hot encoding, thus generating feature matrixes. Next, we feed the matrixes to the ResNet layer for extracting global features. After that, in LSTM layer, we use bidirectional LSTM model to gather feature information from two directions. In Attention layer can learn more relevant feature information. Ultimately, the features are connected with full connected layer and Softmax can make predictions. If the prediction score is higher than 0.5, the predicted sequences are 4mc sequences; otherwise, they are not.

$$\begin{pmatrix} i_t \\ f_t \\ C_t' \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ tanh \end{pmatrix} \left(\begin{pmatrix} W_x \\ W_h \\ W_c \end{pmatrix} [h_{t-1}, x_t, c_{t-1}] + \begin{pmatrix} b_i \\ b_f \\ b_g \end{pmatrix} \right)$$

$$\begin{aligned} C_t &= i_t C'_t + f_t C_{t-1} \\ o_t &= \sigma \left(W_x h_{t-1} + W_h x_t + W_c C_t + b_o \right) \\ h_t &= o_t \tanh \left(C_t \right) \end{aligned}$$

where i_t, f_t and o_t represent the input, forget and output gate, respectively; C'_t is an auxiliary value for calculating the cell memory C_t; t denotes the recurrent time step; W_x, W_h, W_c, and b are the corresponding weight values for each equation; and the current output of LSTM cell is h_t at time step t.

In consideration of bidirectional LSTM, the final LSTM network is composed of two LSTM networks with opposite directions. Hence, the i-th deoxynucleotide of the DNA sequence can be encoded as below:

$$h_i = [\overrightarrow{h_i} \bigoplus \overleftarrow{h_i}]$$

Attention Mechanism

Inspired by human attention, the attention mechanism is an idea for solving problems that focuses on the important factors while ignoring the unimportant. The attention mechanism can quickly filter out high-level information from noises, which has recently demonstrated great success in many relevant classification tasks. To take advantage of this, we applied the attention mechanism after the LSTM layer in the model to obtain the final distinctive feature representation. Let H be the output vectors $[h_1, h_2, \ldots, h_s]$ generated by LSTM layer, where s is the length of the DNA sequence. As shown in **Figure 1**, the following formulations are performed in the attention layer:

$$\begin{aligned} M &= \tanh \left(H \right) \\ \alpha &= softmax \left(W^T M \right) \\ r &= H \alpha^T \end{aligned}$$

where W^T is a transpose of the trained parameter vector W. Then the final representation of the attention layer can be encoded as below:

$$h^* = \tanh \left(r \right)$$

Softmax

The generated vectors h^* after the attention module are fed into a softmax layer for classification as input. The softmax score of class k will be calculated as follows:

$$\alpha_k = \frac{e^{h^*}}{\sum_{k=1}^{C} e^{h^*}}$$

where C denotes the total number of categories, and $C = 2$ when dealing with the binary classification tasks.

The softmax function maps and the output of neurons to numbers between (0–1) and normalizes the sum to 1. In other words, the output scores of each category can be converted into a relative probability by softmax. Therefore, the predicted label can be determined by comparing the predicted probability α_k for each class.

At last, we generated a multi-layered neural network integrating ResNet with a LSTM layer and an attention module, which incorporates the strengths behind ResNet, LSTM, and the attention mechanism. Through applying such a comprehensive network structure, feature extraction and learning are combined in an end-to-end manner, which can significantly improve the prediction performance.

Performance Indicators

In our experiment, we used the following four indicators to evaluate the predictive performance of our proposed model, including Accuracy (ACC), Sensitivity (SN), Specificity (SP), and Mathew's Correlation Coefficient (MCC). They are the four commonly used indicators for classifier performance evaluation in other Bioinformatics fields (Zhang et al., 2008, 2018a,b,c, 2019b,c,d; Wei et al., 2017b, 2019b; Zeng et al., 2017b, 2019c; Chen et al., 2018; Lu et al., 2018a,b; Fu et al., 2019; Gong et al., 2019; Jin et al., 2019; Liu and Li, 2019; Liu et al., 2019c,d; Manavalan et al., 2019a,b,c,d; Basith et al., 2020). Their calculation formulas are as follows:

$$\begin{cases} Sn = \frac{TP}{TP+FN} & 0 \le Sn \le 1 \\ Sp = \frac{TN}{TN+FP} & 0 \le Sp \le 1 \\ ACC = \frac{TP+TN}{TP+FP+TN+FN} & 0 \le ACC \le 1 \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN+FN) \times (TN+FP) \times (TP+FN) \times (TP+FP)}} & -1 \le MCC \le 1 \end{cases}$$

where TP (True Positive) represents the number of positive samples correctly predicted; TN (True Negative) represents the number of negative samples correctly predicted; FP (False Positive) represents the number of negative samples incorrectly predicted to be the positives; FN (False Negative) represents the number of positive samples incorrectly predicted to be the negatives.

Moreover, we also used the area under the ROC curve (AUC) is to quantitively measure the predictive performance of the model (Yang et al., 2018; Lv et al., 2019b; Niu et al., 2019). A higher AUC represents a better predictor (Hanley and McNeil, 1982; Liu et al., 2018; Feng et al., 2019; Lai et al., 2019).

RESULTS AND DISCUSSIONS

Comparison of the Proposed Method and Existing Predictors

To examine the predictive performance of our deep learning model, we compared several existing predictors with our model, including iDNA4mC (Chen et al., 2017), 4mCPred (Su et al., 2018), 4mcPred-SVM (Wei et al., 2018a), and 4mcPred-IFL (Wei et al., 2019a). It is worth noting that besides our predictor using deep learning, other compared predictors are all traditional machine learning algorithm -SVM and different handcrafted sequential features to train their respective models. For fair comparison, all the predictors are evaluated with 10-fold cross validation on the same dataset used in this study.

Table 2 lists the performances of the proposed method and four existing predictors. We can see that our proposed deep learning method achieves the highest performance in two out of three species (*C. elegans* and *A. thaliana*), with only one exception in *D. melanogaster*, in which our method is slightly worse than existing predictors. Specifically, for *C. elegans*, our predictor achieves 91.5%, 87.2%, 89.3%, and 0.787 in terms of SN, SP, ACC, and MCC, respectively. The overall performances (ACC and MCC) by our predictor are significantly better than the runner-up predictor—4mcPred-IFL (with the ACC of 88.0% and the MCC of 0.761). The more significant improvement is observed in *A. thaliana*, in which our predictor outperforms existing predictors in all metrics, leading by 5.7%, 2.2%, and 0.045 in terms of SN, ACC, and MCC, respectively. In addition, we found that our model remarkably improves the SN in all three species, demonstrating that our deep learning model can more accurately identify true 4mC sites. To better illustrate the difference between various models, we used Delong's test from the R package pROC to compare the ROC curves, confirming that the performance gain from fixed-length to full-length version is statistically significant ($p = 0.0005$). Generally, the comparative results demonstrate that our deep learning model is better than existing predictors using traditional machine learning algorithms in prediction of 4mC sites. More importantly, our deep learning model can automatically learn high-level feature representations to capture the characteristics of 4mC sites, rather than specify sequence-based features before model training as existing predictors did.

TABLE 2 | Performance comparison of the proposed Deep4mcPred and existing sequence-based predictors.

Species	Predictors	SN (%)	SP (%)	ACC (%)	MCC
C. elegans	iDNA4mc	79.0	77.0	78.0	0.560
	4mcPred	82.5	82.6	82.6	0.652
	4mcPred_SVM	82.4	80.7	81.5	0.631
	4mcPred_IFL	89.0	87.1	88.0	0.761
	Deep4mcPred	**91.5**	**87.2**	**89.3**	**0.787**
D. melanogaster	iDNA4mc	83.3	79.0	81.2	0.620
	4mcPred	82.4	82.1	82.2	0.646
	4mcPred_SVM	83.8	82.2	83.0	0.661
	4mcPred_IFL	86.5	88.0	87.3	0.745
	Deep4mcPred	**87.6**	**86.6**	**87.1**	**0.742**
A. thaliana	iDNA4mc	76.6	75.5	76.1	0.520
	4mcPred	75.5	78.0	76.8	0.536
	4mcPred_SVM	77.8	79.6	78.7	0.573
	4mcPred_IFL	80.3	84.0	82.2	0.644
	Deep4mcPred	**86.0**	**82.9**	**84.4**	**0.689**

The performances are evaluated with 10-fold cross validation. Note that the performances of the other methods are cited from existing studies, since the source codes of existing methods are not available. The value in bold indicates the optimal value of the indicator.

TABLE 3 | Performance comparison of the model using the attention mechanism and the model not using the attention mechanism.

Species	Models	SN (%)	SP (%)	ACC (%)	MCC
C. elegans	ResNet_LSTM_Attention	91.5	87.2	89.3	0.787
	ResNet_LSTM	90.9	87.2	89.0	0.781
D. melanogaster	ResNet_LSTM_Attention	87.6	86.6	87.1	0.742
	ResNet_LSTM	87.7	86.3	87.0	0.740
A. thaliana	ResNet_LSTM_Attention	86.0	82.9	84.4	0.689
	ResNet_LSTM	84.9	83.9	84.4	0.688

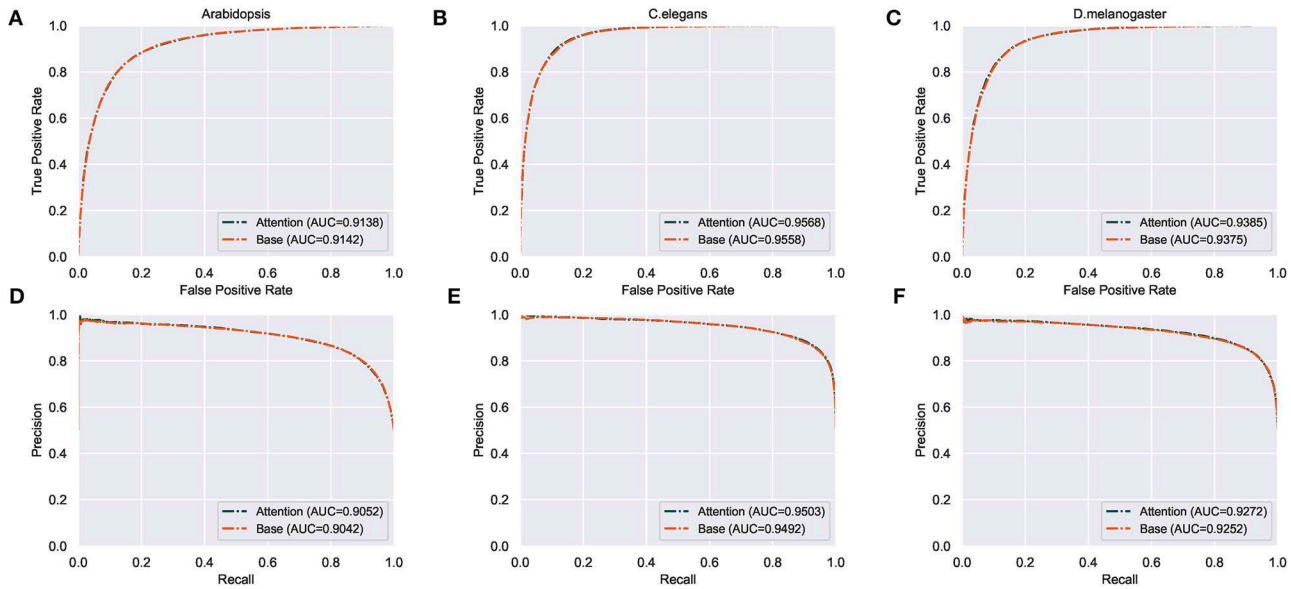

FIGURE 2 | Performance of the model using the attention mechanism and the model not using the attention mechanism. **(A–C)** represent the ROC curves of the two models in the three species. **(D–F)** represent the PR curves of the two models in the three species, respectively.

Performance Impact by Integrating Attention Mechanism

In this section, we evaluated whether or not the attention mechanism can improve the performance of 4mC site prediction. Subsequently, we compared the models taking into account attention mechanism and the model not taking into account attention mechanism for prediction. Both models were trained and evaluated with 10-fold cross validation on the dataset used in this study.

Results in **Table 3** show that training with the attention mechanism, the model achieves 89.3% in ACC and 0.787 in MCC for *C. elegans* dataset, achieves 87.1% in ACC and 0.742 in MCC for the *D. melanogaster* dataset, achieves 84.4% in ACC and 0.689 in MCC for the *A. thaliana* dataset, respectively. These results demonstrate that using the attention mechanism we can achieve good performances for 4mC sites prediction for different species. The comparison between the models using and not using the attention mechanism is shown in **Figure 2**. We can observe that the model using attention mechanism performs better than the model not using the attention mechanism in ROC and PR curves. The details of the performances for both models are listed in **Table 3**. Results show that using the attention mechanism, the deep learning model can achieve the average improvement of 0.1% roughly in three species as compared to the model not using the attention mechanism. This demonstrates that the attention mechanism indeed helps to capture discriminative feature representations.

CONCLUSIONS

In this study, we have proposed Deep4mcPred, a novel predictor for the prediction of DNA 4mC sites. Different from existing predictors using traditional machine learning algorithms (like SVM), Deep4mcPred is the first deep learning-based predictor, in which we integrate residual network and recurrent neural network–biLSTM to build a multi-layer deep learning predictive system. As compared to existing predictors, our proposed method has two advantages. First, our deep learning framework does not need to specify the features when training the predictive model. It can automatically learn the high-level features and capture the characteristic specificity of 4mC sites, benefiting to distinguish true 4mC sites from non-4mC sites. On the other hand, our deep learning method outperforms the traditional machine learning predictors in performance by benchmarking comparison. It demonstrates that the proposed Deep4mcPred is more effective in the DNA 4mC site prediction. Moreover, via experimental comparison, we found that attention mechanism introduced into the deep learning framework is useful to capture the critical features.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://server.malab.cn/Deep4mcPred.

AUTHOR CONTRIBUTIONS

RZ wrote the manuscript, designed experiments, and did the results analysis. ML provided the idea.

FUNDING

This work was supported by National Natural Science Foundation of China.

REFERENCES

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., et al. (2014). Deep speech 2: end-to-end speech recognition in english and mandarin. *International Conference on Machine Learning 2016* (New York, NY), 173–182.

Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020). Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med. Res. Rev*. 2020, 1–39. doi: 10.1002/med.21658

Chen, W., Feng, P., Liu, T., and Jin, D. (2018). Recent advances in machine learning methods for predicting heat shock proteins. *Curr Drug Metab*. 20, 224–228. doi: 10.2174/1389200219666181031105916

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Collobert, R., and Weston, J. (2008). "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning* (New York, NY: ACM), 160–167. doi: 10.1145/1390156.1390177

Feng, C. Q., Zhang, Z. Y., Zhu, X. J., Lin, Y., Chen, W., Tang, H., et al. (2019). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 35, 1469–1477. doi: 10.1093/bioinformatics/bty827

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*. 7:461. doi: 10.1038/nmeth.1459

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150. doi: 10.1093/bioinformatics/bts565

Fu, X., Ke, L., Cai, L., Chen, X., Ren, X., and Gao, M. (2019). Improved prediction of cell-penetrating peptides via effective orchestrating amino acid composition feature representation. *IEEE Access*. 7, 163547–163555. doi: 10.1109/ACCESS.2019.2952738

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

Hanley, J. A., and McNeil, B. J. J. R. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770–778. doi: 10.1109/CVPR.2016.90

Hong, Z., Zeng, X., Wei, L., and Liu, X. J. B. (2019). Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 36, 1037–1043. doi: 10.1093/bioinformatics/btz694

Jin, B., Li, Y., and Robertson, K. D. (2011). DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer*. 2, 607–617. doi: 10.1177/1947601910393957

Jin, Q., Meng, Z., Pham, T. D., Chen, Q., Wei, L., and Su, R. (2019). DUNet: a deformable network for retinal vessel segmentation. *Knowledge-Based Syst*. 178, 149–162. doi: 10.1016/j.knosys.2019.04.025

Lai, H. Y., Zhang, Z. Y., Su, Z. D., Su, W., Ding, H., Chen, W., et al. (2019). iProEP: a computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids*. 17, 337–346. doi: 10.1016/j.omtn.2019.05.028

Lan, K., Wang, D.-T, Fong, S., Liu, L.-S, Wong, K. K., and Dey, N. (2018). A survey of data mining and deep learning in bioinformatics. *J. Med. Syst*. 42:139. doi: 10.1007/s10916-018-1003-9

Li, C.-C., and Liu, B. (2019). MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. *Brief. Bioinf*. bbz133. doi: 10.1093/bib/bbz133

Liang, S., Ma, A., Yang, S., Wang, Y., and Ma, Q. (2018). A review of matched-pairs feature selection methods for gene expression data analysis. *Comput. Struct. Biotechnol. J*. 16, 88–97. doi: 10.1016/j.csbj.2018.02.005

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

Liu, B., Gao, X., and Zhang, H. (2019d). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res*. 47:e127. doi: 10.1093/nar/gkz740

Liu, B., Han, L., Liu, X., Wu, J., and Ma, Q. (2018). Computational prediction of sigma-54 promoters in bacterial genomes by integrating motif finding and machine learning strategies. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 99:1. doi: 10.1109/TCBB.2018.2816032

Liu, B., Li, C., and Yan, K. (2019b). DeepSVM-fold: protein fold recognition by combining Support Vector Machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinf*. bbz098. doi: 10.1093/bib/bbz098

Liu, B., and Li, K. (2019). iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol. Ther. Nucleic Acids*. 18, 80–87. doi: 10.1016/j.omtn.2019.08.008

Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., Zou, Q., et al. (2019c). Computational methods for identifying the critical nodes in biological networks. *Brief. Bioinf*. bbz011. doi: 10.1093/bib/bbz011

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2017). Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinf*. 14, 905–915. doi: 10.1109/TCBB.2016.2550432

Liu, Z.-Y., Xing, J.-F., Chen, W., Luan, M.-W., Xie, R., Huang, J., et al. (2019a). MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. *Horticul. Res*. 6:78. doi: 10.1038/s41438-019-0160-4

Lu, X., Li, X., Liu, P., Qian, X., Miao, Q., and Peng, S. (2018b). The integrative method based on the module-network for identifying driver genes in cancer subtypes. *Molecules* 23:183. doi: 10.3390/molecules23020183

Lu, X., Qian, X., Li, X., Miao, Q., and Peng, S. (2018a). DMCM: a data-adaptive mutation clustering method to identify cancer-related mutation clusters. *Bioinformatics* 35, 389–397. doi: 10.1093/bioinformatics/bty624

Luo, G.-Z., Blanco, M. A., Greer, E. L., He, C., and Shi, Y. (2015). DNA N 6-methyladenine: a new epigenetic mark in eukaryotes? *Nat. Rev. Mol. Cell Biol*. 16:705. doi: 10.1038/nrm4076

Lv, H., Zhang, Z. M., Li, S. H., Tan, J. X., Chen, W., and Lin, H. (2019b). Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinform*. bbz048. doi: 10.1093/bib/bbz048

Lv, Z. B., Ao, C. Y., and Zou, Q. (2019a). Protein function prediction: from traditional classifier to deep learning. *Proteomics* 19:2. doi: 10.1002/pmic.201900119

Manavalan, B., Basith, S., Shin, T. H., Lee, D. Y., Wei, L., and Lee, G. (2019c). 4mCpred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome. *Cells* 8:1332. doi: 10.3390/cells8111332

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019a). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35, 2757–2765. doi: 10.1093/bioinformatics/bty1047

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019b). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids*. 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019d). AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput. Struct. Biotechnol. J*. 17, 972–981. doi: 10.1016/j.csbj.2019.06.024

Nie, L. L., Deng, L., Fan, C., Zhan, W. H., and Tang, Y. J. (2018). Prediction of protein S-sulfenylation sites using a deep belief network. *Curr. Bioinform*. 13, 461–467. doi: 10.2174/1574893612666171122152208

Niu, S.-Y., Liu, B., Ma, Q., and Chou, W.-C. (2019). rSeqTU—a machine-learning based R package for prediction of bacterial transcription units. *Front. Genetics* 10:374. doi: 10.3389/fgene.2019.00374

Peng, L., Peng, M. M., Liao, B., Huang, G. H., Li, W. B., and Xie, D. F. (2018). The advances and challenges of deep learning application in biological big data processing. *Curr Bioinform*. 13, 352–359. doi: 10.2174/1574893612666170707095707

Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2019). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1231–1239. doi: 10.1109/TCBB.2018.2858756

Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* 34, 4196–4204. doi: 10.1093/bioinformatics/bty508

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Montreal, QC), 3104–3012.

Wang, Y., Zhang, S., Yang, L., Yang, S., Tian, Y., and Ma, Q. (2019). Measurement of conditional relatedness between genes using fully convolutional neural network. *Front. Genet.* 10:1009. doi: 10.3389/fgene.2019.01009

Wei, L., Chen, H., and Su, R. (2018b). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids.* 12, 635–644. doi: 10.1016/j.omtn.2018.07.004

Wei, L., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2018a). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* 35, 1326–1333. doi: 10.1093/bioinformatics/bty824

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019a). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*. 35, 4930–4937. doi: 10.1093/bioinformatics/btz408

Wei, L., Tang, J., and Zou, Q. (2017a). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026

Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019b). Fast prediction of methylation sites using sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 16, 1264–1273. doi: 10.1109/TCBB.2017.2670558

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017b). Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intelligence Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018c). ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Xiao, C.-L., Zhu, S., He, M., Chen, D., Zhang, Q., Chen, Y., et al. (2018). N6-methyladenine DNA modification in the human genome. *Molecular Cell*. 71, 306–18. e7. doi: 10.1016/j.molcel.2018.06.015

Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in homo sapiens. *J.Comput. Biol.* 25, 1266–1277. doi: 10.1089/cmb.2018.0004

Yang, J., Ma, A., Hoppe, A. D., Wang, C., Li, Y., Zhang, C., et al. (2019). Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic Acids Res*. 47, 7809–7824. doi: 10.1093/nar/gkz672

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 Res*. 2016:gkw950. doi: 10.1093/nar/gkw950

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Magaz.* 13, 55–75. doi: 10.1109/MCI.2018.2840738

Yu, M., Ji, L., Neumann, D. A., Chung, D.-H, Groom, J., Westpheling, J., et al. (2015). Base-resolution detection of N 4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite-sequencing. *Nucleic Acids Res*. 43:e148. doi: 10.1093/nar/gkv738

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017a). Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Transact. Comput. Biol. Bioinform.* 14, 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017b). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* 13:e1005420. doi: 10.1371/journal.pcbi.1005420

Zeng, X., Wang, W., Chen, C., and Yen G. G. (2019c). A consensus community-based particle swarm optimization for dynamic community detection. *IEEE Trans. Cyber*. 99:1. doi: 10.1109/TCYB.2019.2938895

Zeng, X., Zhang, X., and Zou, Q. (2016). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinform.* 17, 193–203. doi: 10.1093/bib/bbv033

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2019b). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief. Bioinform*. bbz080. doi: 10.1093/bib/bbz080

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019a). deepDR: a network-based deep learning approach to *in silico* drug repositioning. *Bioinformatics* 35, 5191–5198. doi: 10.1093/bioinformatics/btz418

Zhang, J., Chen, Q., and Liu, B. (2019a). DeepDRBP-2L: a new genome annotation predictor for identifying DNA-binding proteins and RNA-binding proteins using convolutional neural network and long short-term memory. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 1–1. doi: 10.1109/TCBB.2019.2952338

Zhang, W., Chen, Y., Li, D., and Yue, X. (2018b). Manifold regularized matrix factorization for drug-drug interaction prediction. *J. Biomed. Inform*. 88, 90–97. doi: 10.1016/j.jbi.2018.11.005

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019c). SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions. *Inform. Sci*. 497, 189–201. doi: 10.1016/j.ins.2019.05.017

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

Zhang, W., Liu, J., Niu, Y. Q., Wang, L., and Hu, X. (2008). A Bayesian regression approach to the prediction of MHC-II binding affinity. *Comp. Methods Programs Biomed.* 92, 1–7. doi: 10.1016/j.cmpb.2008.05.002

Zhang, W., Liu, X., Chen, Y., Wu, W., Wang, W., and Li, X. (2018c). Feature-derived graph regularized matrix factorization for predicting drug side effects. *Neurocomputing* 287, 154–162. doi: 10.1016/j.neucom.2018.01.085

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018a). SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PLoS Comput. Biol.* 14:e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, X., Zou, Q., Rodriguez-Paton, A., Zeng X. J. (2019b). Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 16, 283–91. doi: 10.1109/TCBB.2017.2776280

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

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.