Methods xxx (xxxx) xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Plant6mA: A predictor for predicting N6-methyladenine sites with lightweight structure in plant genomes

Hua Shi [a], Shuang Li [b,*], Xi Su [c,*]

[a] School of Opto-electronic and Communication Engineering, Xiamen University of Technology, Xiamen 361024, Fujian, China
[b] Beidahuang Industry Group General Hospital, Harbin 150001, China
[c] Foshan Maternal and Child Health Hospital, Foshan, Guangdong, China

ARTICLE INFO

Keywords:
DNA methylation
Deep learning
Bioinformatics

ABSTRACT

N6-methyladenine (6mA) in DNA, a type of DNA methylation in epigenetic modification, has attracted extensive attention in recent years. In order to improve our understanding of 6mA biological activities and mechanisms in plant genomes, we need to be able to accurately identify 6mA sites. Because traditional wet-lab experiments frequently necessitate a large amount of manpower and time, a plethora of computational methods, particularly machine learning, have emerged to achieve fast and accurate 6mA site prediction. Traditional machine learning methods, on the other hand, rely heavily on manual features and integrated learning to improve performance, resulting in a reliance on prior knowledge and a large model scale. Furthermore, many models are only trained and tested for one species, with no comparison of model generalization performance, resulting in models with limited practical usability. In order to increase the generalization capability of the model, we propose a lightweight structure predictor Plant6mA based on Transformer encoder. Based on results on independent test sets, our proposed Plant6mA has better generalization performance than the most advanced methods in predicting 6mA location in plant genomes. Plant6mA's MultiHead Attention mechanism effectively enhances its expressive ability by capturing potential biological information from multiple scales of the input sequence. Furthermore, we used a dimensionality reduction tool to visualize Plant6mA's training process and visually demonstrate the effectiveness of our model.

1. Introduction

Among DNA methylation, histone modifications, nucleosome positioning, genomic imprinting, gene silencing, and other epigenetic modifications, DNA methylation is the most important, and it can regulate gene expression with the same DNA sequences [1]. 5-methylcytosine (5mC), 4-methylcytosine (4mC) and N6-methyladenine (6mA) are three common DNA modifications [2]. In bacteria, eukaryotes, and archaea, 6mA denotes DNA methylation at the 6th position of the adenine purine ring [3]. Because of the widespread use and adaptability of 5mC, this sort of generic DNA alteration has received a lot of attention. However, 6mA is distributed sporadically in the genome. There have been few research on 6mA, and its significance in eukaryotes is mainly unclear [4]. In recent years, 6mA has been investigated more and more due to its link to germ cell differentiation, stress response, embryonic development, nervous system, and other processes in eukaryotes [5–9]. As a result, accurate prediction of 6mA location in various species is critical for future 6mA research.

Bisulfite-converted DNA, DNA methylation–specific immunoprecipitation techniques, and DNA partitions formed by methylation-specific restriction enzymes are all examples of wet-lab procedures that may be used to identify DNA methylation sites [9–13]. The use of next-generation sequencing technologies such as single-molecule real-time sequencing (SMRT) and nanopore sequencing for DNA methylation site prediction has greatly reduced the cost of traditional DNA methylation site identification experiments, but they are still time-consuming and labor-intensive [14–22].

A growing number of computational methods have emerged for more convenient and rapid prediction of DNA methylation sites. Machine learning techniques have recently become popular for identifying the location of the 6mA in plant genomes. SDM6A proposed a two-layer ensemble approach by combining a final prediction from a support vector machine (SVM) and an extremely randomized tree, both of which identified the 6mA location in the Rice genome based on an optimal

* Corresponding authors.
E-mail addresses: lishuang0312@yeah.net (S. Li), xisu_fsfy@163.com (X. Su).

https://doi.org/10.1016/j.ymeth.2022.02.009
Received 28 January 2022; Received in revised form 20 February 2022; Accepted 24 February 2022
Available online 26 February 2022

feature set [23]. Then, to obtain the final site prediction score, i6mA-Fuse used the linear regression model to integrate the prediction results of five Randomized Forest (RF) models based on different single encoding methods [24]. By constructing a Markov model, MM-6mAPred uses the transition probability between adjacent nucleotides to identify the 6mA site [25]. Most of the methods mentioned above, on the other hand, classify and evaluate single species without examining the models' ability to generalize to other data sets. After that, Meta-i6mA was trained on a single data set, and then the performance of multiple species was evaluated to see how well the model generalized across species [26]. However, Meta-i6mA model used meta-learning approach to obtain the final prediction results for 30 baseline models of 5 features from 6 machine learning models (RF, SVM, extremely randomized tree, logistic regression, naïve Bayes and AdaBoost). Many deep learning models can be generalized to different species, and Meta-i6mA are very redundant and complex. Convolutional neural network (CNN) and recurrent neural network (RNN) are two common deep learning models that have been used in methylation prediction, such as Mouse4mC-BGRU [27], GC6mA-Pred [28], 4mCi6mA-BGC [29], etc. However, when compared to CNN and RNN, Transformer Encoder's network structure can achieve better generalization performance after training on larger data sets, attracting a lot of attention.

We present Plant6mA, a lightweight structure predictor for identifying DNA N6-methyladenine sites in plant genomes without manual features. To begin, we test the model's generalization by training on the same plant species' training set and predicting on test sets from different plant species. When compared to other models, ours outperforms the existing predictors' generalization performance. Second, unlike traditional manual features, we allow the embedding that represents the features to change together through back propagation, resulting in the best performance. We can avoid a lot of the prior knowledge required for manual feature calculation and improve the model's efficiency without having to increase the number of feature calculation steps by using automatic representation as embedding. Finally, it can be seen that the model continuously separates positive and negative samples through iteration using the commonly used linear PCA and nonlinear t-SNE dimensionality reduction visualization training process, demonstrating the model's effectiveness. To conclude, Plant6mA is a lightweight structural model with high generalization and prediction performance that does not require a lot of previous knowledge.

2. Methods

2.1. Datasets

The data set supplied by META-I6mA is utilised in this research to see if our suggested method's generalisation performance satisfies state of the art (SOTA) in diverse plants. *Rosaceae* and Arabidopsis thaliana are among the three plant genes included in the dataset (*A. thaliana*). As shown in Table 1, the *Rosaceae* dataset was divided into 80 percent training sets, 20 percent test sets, and the remaining *A. thaliana* test set was used as independent test sets. Positive *Rosaceae* dataset samples were first extracted from the MDR database, which contained 26,516 FV 6mA sites and 14,666 RC 6mA sites, both belonging to the *Rosaceae* family [30]. Second, each sequence was adjusted to 41 base pairs (bp) with an adenine nucleotide ('A') at the 21st position or the canter for 40,574 treated sequences after similar sequences were removed. After removing the sequences with an identity of greater than 90%, a total of 36,537 sequences were obtained. Negative samples, as in previous studies, were those with adenine nucleotide ('A') at position 21 but were detected by SMRT as non-6mA sites [23,25,31,32]. GSE81597 is the accession number for the A. thaliana dataset collected from the National Center for Biotechnology Information GEO, and processed as a *Rosaceae* dataset, yielding 31 873 6mA sites and 31 873 non-6mA sites.

Table 1
Data set partitioning details.

Datasets	Positive sample	Negative sample
Rosaceae (Training)	29,237	29,237
Rosaceae (Independent)	7,300	7,300
Arabidopsis thaliana (Independent)	31,873	31,873

2.2. The proposed method's framework

The proposed Plant6mA predictive framework is depicted in Fig. 1. (A) Data Preprocess Module, (B) Adaptive Embedding Module, (C) Encoding Module, and (D) Classification Module make up this network architecture. The data preprocess module is designed to adjust the original data to the 41 bp sequence format with ('A') in the center or 12th position, which is convenient for the input of the subsequent model, as shown in the Dataset section. Then, to prevent interfering elements from being introduced into the final metrics, similarity is removed, and sequences with consistency greater than or equal to 90 are removed using CD-HIT [33]. The adaptive embedding query look-up table then generates the corresponding embedding for each base in DNA sequences, then adds it to the embedding of the location to get the total embedding, which is constantly updated with backpropagation. The multihead attention mechanism can then be used to learn multiple scales of previous embedding, followed by the Positionwise Feed Forward Network (FFN) [34]. Finally, the final representation obtained by the Encoding Module is fed into the linear transformation layer via the classification layer, and the probability is output to determine whether the input sequence contains a 6mA site.

2.2.1. Embedding module

We used an embedding module to map the four letters of the nucleotide to a random initialization vector according to a look-up table, rather than the traditional one-hot encoding of nucleotide directly. We use position embedding to introduce location information because there is a correlation between location and nucleotide position in the sequence. Finally, in model training, back propagation is used to make continuous adjustments.

2.2.2. Encoding module

The Encoding Module is developed from the Transformer's encoder which is popularized by the use of BERT and features a MultiHead Attention mechanism as well as Positionwise FFN [35–38]. The input of the query matrix, key matrix, and value matrix is the embedding of input sequences to capture different scales of internal correlations within the same sequence. Multi-head attention consists of several self-attention mechanisms in which the input of the query matrix, key matrix, and value matrix is the embedding of input sequences to capture different scales of internal correlations within the same sequence. The following is a mathematical description of MultiHead Attention:

$$\begin{cases} Q_i = X_{embedding} W_Q^i \\ K_i = X_{embedding} W_K^i \\ V_i = X_{embedding} W_V^i \end{cases}, i = 1, 2, \cdots, h$$

$$head_i = Attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}} V_i\right), i = 1, 2, \cdots, h$$

$$MultiHead\ Attention(Q, K, V) = Concat(head_1, head_2, \cdots, head_h) W_O$$

where $X_{embedding} \in R^{L \times d_m}$ is the output of the Encoding Module, in which L = 41 is the length of sequence, and d_m is the embedding dimension. W_Q^i, W_K^i, W_V^i respectively represents the query matrix, key matrix and value matrix in i_{th} head, where d_k is the dimension of previous three vectors. $W_Q^i, W_K^i, W_V^i \in R^{d_m \times d_k}$ and h means the number of heads. Both d_k, d_m and h are hyper-

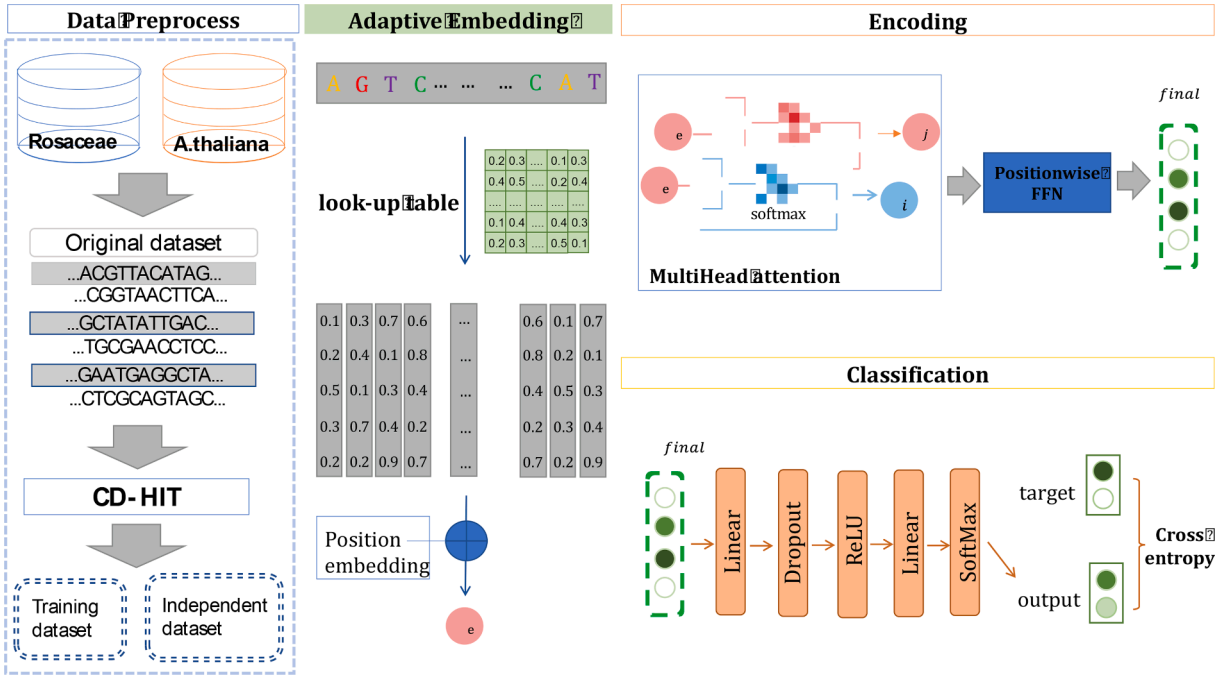

Fig. 1. The proposed method's framework. The Plant6mA consists of four modules: (A) Data Preprocess Module (B) Embedding Module that adjusts with back-propagation and first embeds the entire sequence with both tokenization and position embedding; (C) Encoding Module that extracts multi-scale representation of DNA sequences with multi-head attention. (D) Classification Module, which feeds the encoding module's output into the linear layer to predict whether the 6mA site is present or not, and the entire model is backward based on cross entropy loss.

parameters. Then, we merge h heads, and the result matrix is linearly transformed to the same dimension as the input matrix to facilitate multiple MultiHead Attention by using $W^O \in R^{d_m \times hd_k}$.

2.2.3. Classification module

After we get the representation from the Encoding Module, we input the vector to fully connected layers to calculate the probability of whether the sequence is true 6mA DNA methylation sites. The output of each fully linked layer with a nonlinear ReLU activation function is computed using the equation below:

$$X_i = ReLU(X_{i-1}W_d + b_i), i = 1, \cdots, m$$

where W_d is a matrix, b_i is bias in each layer *and m is* a hyper-parameter to decide the total layer of neural network. We use the cross-entropy loss, which is widely used in dichotomies to train the output module to improve the prediction performance, the function is at below:

$$\begin{cases} p_k = \dfrac{exp(y_{p,k})}{\sum_j exp(y_{p,k})}, k = 0, 1 \\ Loss_{CE}(p_1, y) = -ylogp_1 - (1-y)log(1-p_1) \end{cases}$$

where k = 0 or 1 donates 6mA or non-6mA, The ground truth result is y, and the probability that the model believes the sequence is k class is p_k.

2.3. Evaluation metrics

To evaluate the performance of our model and other existing models, We employ four common metrics: accuracy (ACC), specificity (SN), sensitivity (SP), and Matthew's correlation coefficient (MCC) [39–46]. The formulas for these metrics are as follows:

$$\begin{cases} ACC = \dfrac{TP + TN}{TP + TN + FP + FN} \\ SN = \dfrac{TP}{TP + FN} \\ SP = \dfrac{TN}{TN + FP} \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \end{cases}$$

where true positive, false negative, true negative, and false positive samples are represented as TP, FN, TN, and FP, respectively [47–61]. The number of correctly classified samples in an independent dataset is described by the ACC. Positive samples truly classified as positive in all classified truth and negative samples truly classified as negative in all classified false are shown in SN and SP, respectively. MCC is a comprehensive metirc that can assess accuracy even when datasets are unbalanced.

3. Results

3.1. Comparison of different numbers of Multihead Attention heads on Independent Rosaceae test dataset.

Although the number of heads in the above-mentioned Multihead Attention mechanism is a hyperparameter, many existing papers have demonstrated that the number of heads has a significant impact on the results. By referring to the number of heads in the transformer Encoder original paper [36], this paper selects 1, 2, 4, 6, 8, and 16 heads for testing on the Independent *Rosaceae* test dataset. As seen in Table 2, the Multihead Attention mechanism's eight heads outperformed the others. Because this paper is only a relatively simple binary classification task, it's possible that the performance in one head is good, but more than eight heads will introduce more noise and reduce performance [62]. As a result, throughout the rest of our research, we'll compare the results of eight heads of the Multihead Attention mechanism to encode with the

Table 2
The performance of different numbers of Multihead Attention heads on the independent *Rosaceae* test dataset.

The number of heads	ACC	SN	SP	MCC
1	0.9591	0.9573	0.9610	0.9182
2	0.9568	0.9558	0.9578	0.9136
4	0.9577	0.9630	0.9525	0.9155
8	**0.9633**	**0.9655**	**0.9611**	**0.9266**
16	0.9588	0.9599	0.9578	0.9177

results of previous approaches.

Because deep learning is a black box model, the actual learning process often piques people's interest. Principal Component Analysis (PCA) and t-SNE are used in this paper to reduce the dimension of the test results to a two-dimensional space so that the results can be visually divided [63–66]. In the *Rosaceae* data set, we visualized the process. As shown in Fig. 2, PCA and t-SNE were used to visualize Epoch 1, Epoch 30, and Epoch 60, respectively. D-F was a nonlinear classifier t-SNE, while A-C was a linear classifier PCA without curved structure. In Fig. 2, 0 and 1 represent negative samples and positive samples, which are yellow and blue respectively. From A to C, the number of misclassified blue points in the yellow points gradually decreases, and from D to F, the boundary between yellow and blue points gradually becomes clear, exhibiting the improved classification impact of Plant6mA (Fig. 2).

3.2. Performance comparison with previous models on independent dataset

We tested the model against three SOTA predictors, Meta-i6mA, SD6MA, and i6mA-Fuse, in terms of ACC, SN, SP, AUC, and MCC. The following are the experimental settings: We encode four transformer encoder layers in total, with a batch size of 32. There are eight heads, a feedforward dimension of 64, a learning rate of 0.002, a maximum epoch of 120, and a dropout rate of 0.2. It's worth noting that we uniformly trained the model on the *Rosaceae* training set and predicted test performance on the *Rosaceae* and *A. thaliana* independent test sets, respectively. In both datasets, Plant6mA outperformed the other predictors on all metrics, as shown in Fig. 3. Plant6mA was also 0.86%, 0.53 %, 0.55%, and 0.41 % higher than Meta-i6mA in MCC, ACC, SN, and SP, respectively, on Arabidopsis thaliana, and 2.99%, 1.74%, 2.39%, and 1.1% higher than Meta-i6mA in MCC, ACC, SN, and SP, respectively, on *Rosaceae*. The SP values of all classifiers are larger than SN on data set A, which is due to the data set's own distribution. Plant6mA's performance and generalization ability are clearly superior to those of other predictors.

4. Conclusion

We developed Plant6mA, a network based on Transformer encoder for identfying 6mA sites in the plant genome, in this study. We employ backpropagation to continually update and incorporate adaptive features, unlike standard machine learning approaches, which rely on a number of manual features. It's worth noting that, unlike the machine learning method that has become popular in recent years to improve performance through integrated learning, we use the advanced and simple structure of the Transformer encoder model directly for prediction, which outperforms existing machine learning-based methods on most metrics. MultiHead Attention can help us extract features from whole DNA sequences at various scales, and this paper looks at how the number of heads affects the final result in binary classification tasks. In addition, transformer encoder's model architecture can be better generalized to small data sets after training on big data. Finally, the model's ability to continuously divide test samples during training was visualized using linear and nonlinear dimensionality reduction methods, demonstrating that the model can effectively predict DNA methylation sites.

CRediT authorship contribution statement

Hua Shi: Conceptualization, Methodology, Data curation, Writing – original draft. **Shuang Li:** Visualization, Investigation, Supervision. **Xi Su:** Writing – review & editing.

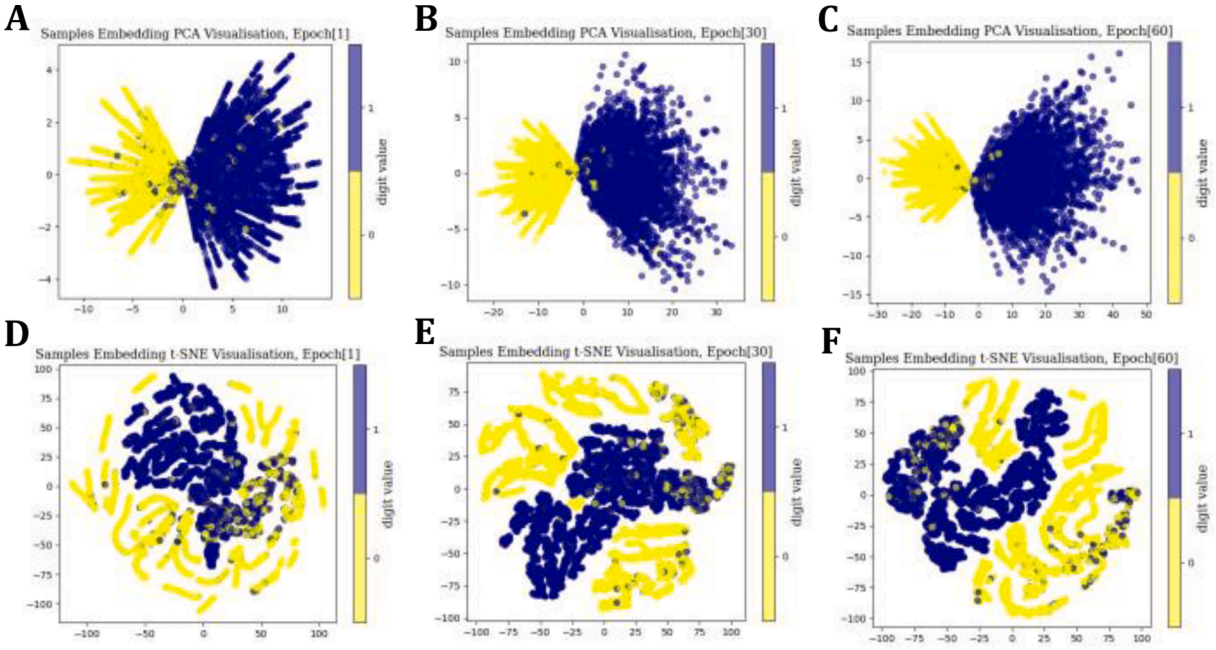

Fig. 2. Dimension reduction visualization of each sample on independent *Rosaceae* test dataset by PCA and t-SNE. Epoch[x] in the title of each subfigure denotes the model's response training epoch. Individually, 0 and 1 represent non-6mA and 6mA sites, respectively. PCA visual images at various epochs (A-C), and t-SNE visual images at various epochs (D-F).

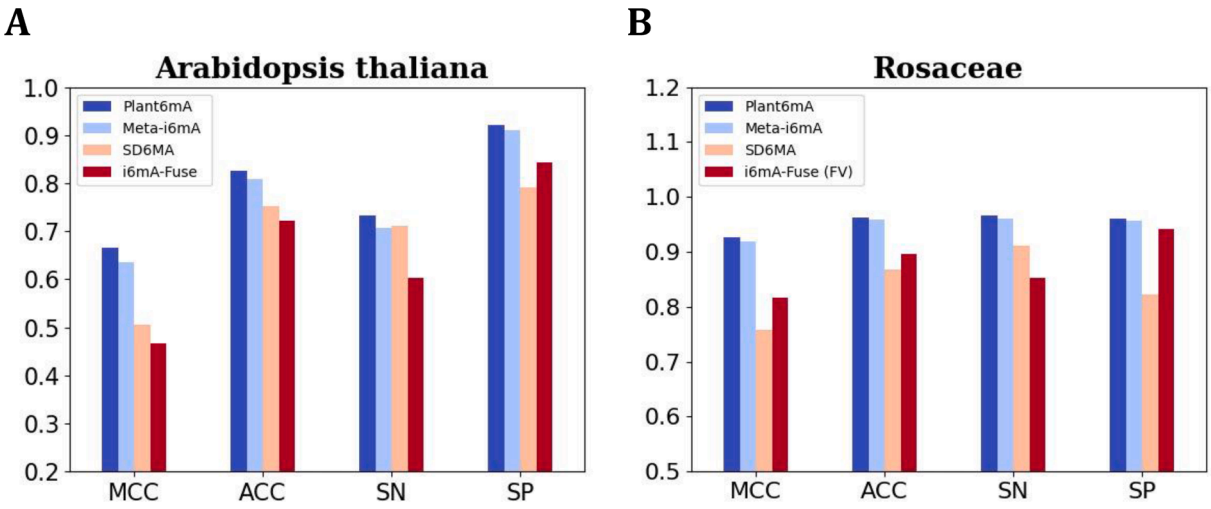

Fig. 3. Performance comparison between Plant6mA and previous machine learning-based methods Meta-i6mA, SD6MA and i6mA-Fuse on the independent test dataset *A. thaliana* and *Rosaceae*.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported by Scientific Research Project Foundation of the Educational Committee of Fujian Province for Middle and Young Teachers (JT180431).

References

[1] A. Portela, M. Esteller, Epigenetic modifications and human disease, Nat. Biotechnol. 28 (10) (2010) 1057–1068.
[2] H. Lv, et al., Advances in mapping the epigenetic modifications of 5-methylcytosine (5mC), N6-methyladenine (6mA), and N4-methylcytosine (4mC), Biotechnol. Bioeng. 118 (11) (2021) 4204–4216.
[3] Z. Liang, et al., DNA N(6)-adenine methylation in Arabidopsis thaliana, Dev. Cell 45 (3) (2018) 406–416 e3.
[4] S. Feng, et al., Conservation and divergence of methylation patterning in plants and animals, Proc. Natl. Acad. Sci. 107 (19) (2010) 8689.
[5] J. Liu, et al., Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig, Nat. Commun. 7 (1) (2016) 13052.
[6] B. Yao, et al., DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress, Nat. Commun. 8 (1) (2017) 1122.
[7] G. Zhang, et al., N6-methyladenine DNA modification in drosophila, Cell 161 (4) (2015) 893–906.
[8] Y. Hu, et al., rs1990622 variant associates with Alzheimer's disease and regulates TMEM106B expression in human brain tissues, BMC Med 19 (1) (2021) 11.
[9] D. Yalcin, H.H. Otu, An unbiased predictive model to detect DNA methylation propensity of CpG islands in the human genome, Curr. Bioinform. 16 (2) (2021) 179–196.
[10] R. Gupta, A. Nagarajan, N. Wajapeyee, Advances in genome-wide DNA methylation analysis, Biotechniques 49 (4) (2010) iii–xi.
[11] V. Ghafarpour, et al., DNA methylation association with stage progression of head and neck squamous cell carcinoma, Comput. Biol. Med. 134 (2021).
[12] G.-Z. Luo, et al., Characterization of eukaryotic DNA N6-methyladenine by a highly sensitive restriction enzyme-assisted sequencing, Nat. Commun. 7 (1) (2016) 11301.
[13] Y. Hu, S. Qiu, L. Cheng, Integration of multiple-omics data to analyze the population-specific differences for coronary artery disease, Comput. Math. Methods Med. 2021 (2021) 7036592.
[14] B.A. Flusberg, et al., Direct detection of DNA methylation during single-molecule, real-time sequencing, Nat. Methods 7 (6) (2010) 461–465.
[15] Y. Wang, et al., Nanopore sequencing technology, bioinformatics and applications, Nat. Biotechnol. 39 (11) (2021) 1348–1365.
[16] Y. Hu, et al., rs34331204 regulates TSPAN13 expression and contributes to Alzheimer's disease with sex differences, Brain 143 (11) (2020), e95.
[17] R. Garcia, et al., Identification of potential antiviral compounds against SARS-CoV-2 structural and non structural protein targets: a pharmacoinformatics study of the CAS COVID-19 dataset, Comput. Biol. Med. 133 (2021).
[18] Y. Zuo, et al., Analysis of the epigenetic signature of cell reprogramming by computational DNA methylation profiles, Curr. Bioinform. 15 (6) (2020) 589–599.
[19] M. Chagoyen, J.A.G. Ranea, F. Pazos, Applications of molecular networks in biomedicine, Biol. Methods Protoc. 4 (1) (2019).
[20] C. Chen, et al., A comprehensive comparison and overview of R packages for calculating sample entropy, Biol. Methods Protoc. 4 (1) (2019).
[21] S. Alguwaizani, et al., Predicting interactions between pathogen and human proteins based on the relation between sequence length and amino acid composition, Curr. Bioinform. 16 (6) (2021) 799–806.
[22] A. Dasti, et al., RNA-centric approaches to study RNA-protein interactions in vitro and in silico, Methods 178 (2020) 11–18.
[23] S. Basith, et al., SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome, Mol. Ther. Nucleic Acids 18 (2019) 131–141.
[24] M.M. Hasan, et al., i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation, Plant Mol. Biol. 103 (1–2) (2020) 225–234.
[25] C. Pian, et al., MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model, Bioinformatics 36 (2) (2020) 388–392.
[26] M.M. Hasan, et al., Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework, Brief Bioinform 22 (3) (2020) 1–16.
[27] J. Jin, Y. Yu, L. Wei, Mouse4mC-BGRU: deep learning for predicting DNA N4-methylcytosine sites in mouse genome, Methods (2022).
[28] J. Cai, G. Xiao, R. Su, GC6mA-Pred: a deep learning approach to identify DNA N6-methyladenine sites in the rice genome, Methods (2022).
[29] B. Yu, et al., Identification of DNA modification sites based on elastic net and bidirectional gated recurrent unit with convolutional neural network, Biomed. Signal Process. Control 75 (2022), 103566.
[30] Z.Y. Liu, et al., MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae, Hortic. Res. 6 (2019) 78.
[31] H. Xu, et al., 6mA-Finder: a novel online tool for predicting DNA N6-methyladenine sites in genomes, Bioinformatics 36 (10) (2020) 3257–3259.
[32] W. Chen, et al., i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome, Bioinformatics 35 (16) (2019) 2796–2800.
[33] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (13) (2006) 1658–1659.
[34] D. Wang, et al., DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism, Nucleic Acids Res. 49 (8) (2021), e46.
[35] J. Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.
[36] A. Vaswani, et al., Attention Is All You Need. arXiv, 2017.
[37] S. Ji, M. Holtta, P. Marttinen, Does the magic of BERT apply to medical code assignment? A quantitative study, Comput. Biol. Med. 139 (2021).
[38] S.M.A. Shah, Y.-Y. Ou, TRP-BERT: discrimination of transient receptor potential (TRP) channels using contextual representations from deep bidirectional transformer based on BERT, Comput. Biol. Med. 137 (2021).
[39] C. Ao, Q. Zou, L. Yu, NmRF: identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences, Briefings Bioinf. 23 (1) (1800).
[40] A. Alim, A. Rafay, I. Naseem, PoGB-pred: prediction of antifreeze proteins sequences using amino acid composition with feature selection followed by a sequential-based ensemble approach, Curr. Bioinform. 16 (3) (2021) 446–456.
[41] S. Akbar, et al., iAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model, Comput. Biol. Med. 137 (2021).
[42] H. Zulfiqar, et al., Deep-4mCW2V: a sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli, Methods (2021).

[43] S. Ayora, BIOMAP: a home for all biology methods, Biol. Methods Protocols 1 (1) (2016) bpv001.
[44] D. Lin, et al., PREDAIP: computational prediction and analysis for anti-inflammatory peptide via a hybrid feature selection technique, Curr. Bioinform. 16 (8) (2021) 1048–1059.
[45] I.T. Mariapushpam, S. Rajagopal, Improved algorithm for the detection of cancerous cells using discrete wavelet transformation of genomic sequences, Curr. Bioinform. 12 (6) (2017) 543–550.
[46] L. Yu, et al., Prediction of drug response in multilayer networks based on fusion of multiomics data, Methods 192 (2021) 85–92.
[47] R. Su, et al., Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools, Briefings Bioinf. 21 (2) (2020) 408–420.
[48] R. Su, X. Liu, L. Wei, MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy, Briefings Bioinf. 21 (2) (2020) 687–698.
[49] R. Su, et al., Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response, Methods 166 (2019) 91–102.
[50] R. Su, et al., Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction, Briefings Bioinf. 21 (3) (2020) 996–1005.
[51] Z. Hong, et al., Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism, Bioinformatics 36 (4) (2020) 1037–1043.
[52] Q. Jin, et al., DUNet: a deformable network for retinal vessel segmentation, Knowl.-Based Syst. 178 (2019) 149–162.
[53] B. Manavalan, et al., Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation, Mol. Ther.-Nucleic Acids 16 (2019) 733–744.
[54] B. Manayalan, et al., mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation, Bioinformatics 35 (16) (2019) 2757–2765.
[55] L. Wei, et al., A novel hierarchical selective ensemble classifier with bioinformatics application, Artif. Intell. Med. 83 (2017) 82–90.
[56] Q. Zou, et al., Gene2vec: gene subsequence embedding for prediction of mammalian N-6-methyladenosine sites from mRNA, RNA 25 (2) (2019) 205–218.
[57] L. Wei, et al., Fast prediction of protein methylation sites using a sequence-based feature selection technique, Ieee-Acm Trans. Comput. Biol. Bioinform. 16 (4) (2019) 1264–1273.
[58] L. Wei, et al., Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier, Artif. Intell. Med. 83 (2017) 67–74.
[59] L. Wei, et al., ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides, Bioinformatics 34 (23) (2018) 4007–4016.
[60] C. Ao, Q. Zou, L. Yu, RFhy-m2G: identification of RNA N2-methylguanosine modification sites based on random forest and hybrid features, Methods (San Diego, Calif.) (1800).
[61] Y. Fan, H. Xu, Prediction of off-target effects in CRISPR/Cas9 System by ensemble learning, Curr. Bioinform. 16 (9) (2021) 1169–1178.
[62] E. Voita, et al., Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. 2019.
[63] I.T. Jolliffe, Principal component analysis, J. Mark. Res. 87 (4) (2002) 513.
[64] V.D.M. Laurens, G. Hinton, Visualizing data using t-SNE, J. Machine Learn. Res. 9 (2605) (2008) 2579–2605.
[65] P. Wang, et al., Stochastic neighbor embedding algorithm and its application in molecular biological data, Curr. Bioinform. 15 (9) (2020) 963–970.
[66] K. Cao, et al., Machine learning with a reduced dimensionality representation of comprehensive Pentacam tomography parameters to identify subclinical keratoconus, Comput. Biol. Med. 138 (2021).