A novel method for predicting DNA N4-methylcytosine sites based on deep forest algorithm

Yonglin Zhang[1†], Mei Hu[2†], Qi Mo[2], Wenli Gan[2], and Jiesi Luo[2*]

[1]Department of Pharmacy, The Affiliated Hospital of North Sichuan Medical College, Nanchong 637000, China

[2]Department of Pharmacology, School of Pharmacy, Southwest Medical University, Luzhou 646000, China

Corresponding authors:

Jiesi Luo, E-mail: ljs@swmu.edu.cn

Abstract

N^4-methyladenosine (4mC) methylation is an essential epigenetic modification of DNA that plays a key role in many biological processes, such as gene expression, gene replication and transcriptional regulation. The genome-wide identification and analysis of the 4mC sites can better reveal the epigenetic mechanisms that regulate various biological processes. Although some high-throughput genomic experimental methods can effectively facilitate the identification in a genome-wide scale, they are still too expensive and laborious for routine use. The computational methods can compensate for these disadvantages, but they still have plenty of room for performance improvement. In this study, we develop a non-NN-style deep learning-based approach for accurately predicting 4mC sites from genomic DNA sequence. We generate various informative features represented sequence fragments around 4mC sites, and subsequently implement them into a deep forest (DF) model. After training the deep model using tenfold cross-validation, the overall accuracies of 85.0%, 90.0% and 87.8% were achieved for three representative model organisms, *A. thaliana*, *C. elegans*, and *D. melanogaster*, respectively. In addition, extensive experiment results show that our proposed approach outperforms other existing state-of-the-art predictors in the 4mC identification. Our approach stands for the first deep forest-based algorithm for the prediction of 4mC sites, providing a novel idea in this field.

Keywords: DNA N4-methylcytosine; deep forest algorithms; feature importance analysis

1. Introduction

DNA methylation involves the covalent addition of a methyl group to the fifth position of cytosine in a CpG dinucleotide [1]. It is a major mechanism of epigenetic inheritance that can affect gene expression without altering the DNA sequence. Several different forms of DNA methylation were discovered to exist in many living organisms. For example, N^6-methyladenine (6mA), N^4-methylcytosine (4mC) and C^5-methylcytosine (5mC) are the most common types of methylation modification that play key roles in various important biological processes and have therefore attracted increasing attention in recent years [2]. N^6-methyladenine (6mA) is by far the most prevalent form of DNA methylation, widely distributed in prokaryotes and eukaryotes, which specifically regulates gene expression, genomic imprinting, and DNA mismatch repair and cell developments [3]. C^5-methylcytosine (5mC) is the dominant form in eukaryotic genomes, and it is actively involved in a broad range of biological processes, including gene expression, imprinting, regulation and transposon suppression [4]. N^4-methylcytosine (4mC), which is restricted to prokaryotes and archaea, plays a crucial role in the regulation of DNA replication, repair, expression, and differentiation, and can even prevent the enzymatic degradation of host DNA [5]. The past decades have witnessed the great progress in DNA methylation research, yet the presence and potential role of these epigenetic markers remain elusive and the exact mechanisms governing their states or patterns remain unclear. Compared with the studies for the other two types of methylation, many recent theoretical and experimental efforts have been devoted to the study of the 4mC due to the lack of effective detection methods.

With the rapid development of high-throughput sequencing technologies, several experiment-based have been developed to identify the 4mC modifications different DNA methylation modifications in the whole genome. For example, Lister *et al*. proposed a technique based on next-generation DNA sequencing to map genome-wide 5mC modification sites at single-base resolution using bisulfite treatment [6]. Although this method is a common technique for detecting DNA methylation sites on a genome-wide scale, it is only applicable to 5mC modifications in prokaryotes. Single molecule real time sequencing (SMRT) based on third generation sequencing

technology is emerging as the mainstream DNA methylation detection method because of its capability to sequence long reads and detect various forms of methylation, such as 4mC, 5mC and 6mA [7]. However, library preparation of SMRT is more expensive and time-consuming than next-generation sequencing (NGS)-based bisulfite treatment. To address these issues, Yu *et al*. proposed an NGS-based method, called 4mC-Tet-assisted bisulfite sequencing, for detecting genome-wide 4mC sites in bacterial species [8]. This is a 4mC-specific method that can effectively distinguish 4mC from other DNA methylation modification sites. Recently, Rathiet *et al*. proposed another 4mC-specific technique that applies a transcriptional activator-like effector approach to reveal 4mC sites in DNA sequences [9]. Although these experimental methods are accurate and reliable, they also suffer from several limitations, such as the high cost, time-consuming and labor-intensive. Therefore, computational methods emerge as a complementary strategy to improve the identification of 4mC sites and provide strong support for experimental studies.

To date, a number of computational methods have been developed for predicting 4mC sites. These methods can be broadly divided into two categories: traditional machine learning-based and deep learning-based approaches [10]. The traditional machine learning algorithms require a predefined feature vector to make predictions. For example, Wei *et al*. generated informative features proposed by an iterative feature representation learning scheme to predict the 4mC sites using SVM [11]. Wei *et al*. also adopt a two-step feature optimization strategy coupled with SVM to train the optimal predictive model for six species [12]. He *et al*. used position-specific trinucleotide propensity (PSTNP) and electron-ion interaction potential features to encode DNA sequences [13]. Chen *et al*. constructed feature vectors using nucleotide chemical properties (NCP) and nucleotide frequencies [14]. In recent years, deep learning, the most advanced branch of machine learning, has become a hot topic in various domains. Unlike those traditional machine learning methods, deep learning models can offer promising results without the need for manual feature engineering. For example, Zeng *et al*. developed a multi-layer deep learning model by integrating residual network and recurrent neural network [15]. Xu *et al*. used convolutional

neural networks with four representative features to predict 4mC sites [16]. Zeng *et al*. employed multi-layer convolutional neural networks with an inception module integrating with bidirectional long short-term memory for improved prediction of 4mC sites [17]. In our previous study, we reviewed the currently available deep learning-based 4mC prediction methods and provided a valuable reference for many issues for model development [18]. Despite remarkable progress in the identification of 4mC sites, computational methods are still inaccurate compared to experiment techniques, and there remains much room for improvement in prediction performance.

Thus, in this work, we have developed an a non-NN-style deep learning-based approach for predicting DNA 4mC sites from sequence alone, as portrayed by the schematic in Figure 1. Three datasets of experimentally determined 4mC and non-4mC modification sites from three species, including *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*, were used to construct the deep models capable of classifying a given DNA sequence as methylated or non-methylated. Starting from the +/- 20-bp flanking sequence of cytosine site, each sample was uniquely represented by a numeric vector of biological features capturing information on sequence orders, compositions, physicochemical properties and structures of DNA sequences. With these features as inputs, the classification models were constructed and trained using the deep forest algorithm, while their performance was evaluated through 10-fold cross-validation. Three different machine learning predictors for model classification (forest, lightgbm, and xbgoost) were tested, and those with the best performance were selected for future predictions. In addition, the Gini impurity-based feature importance analysis was performed to reduce noisy features and further improve performance. Experimental results indicated that our final models achieved the average accuracies of 85.0%, 90.0% and 87.8% on *A. thaliana*, *C. elegans*, and *D. melanogaster* benchmark datasets, which overall proved to be better than other existing approaches through comparison. To the best of our knowledge, this is the first attempt to predict 4mC sites using a deep forest algorithm.

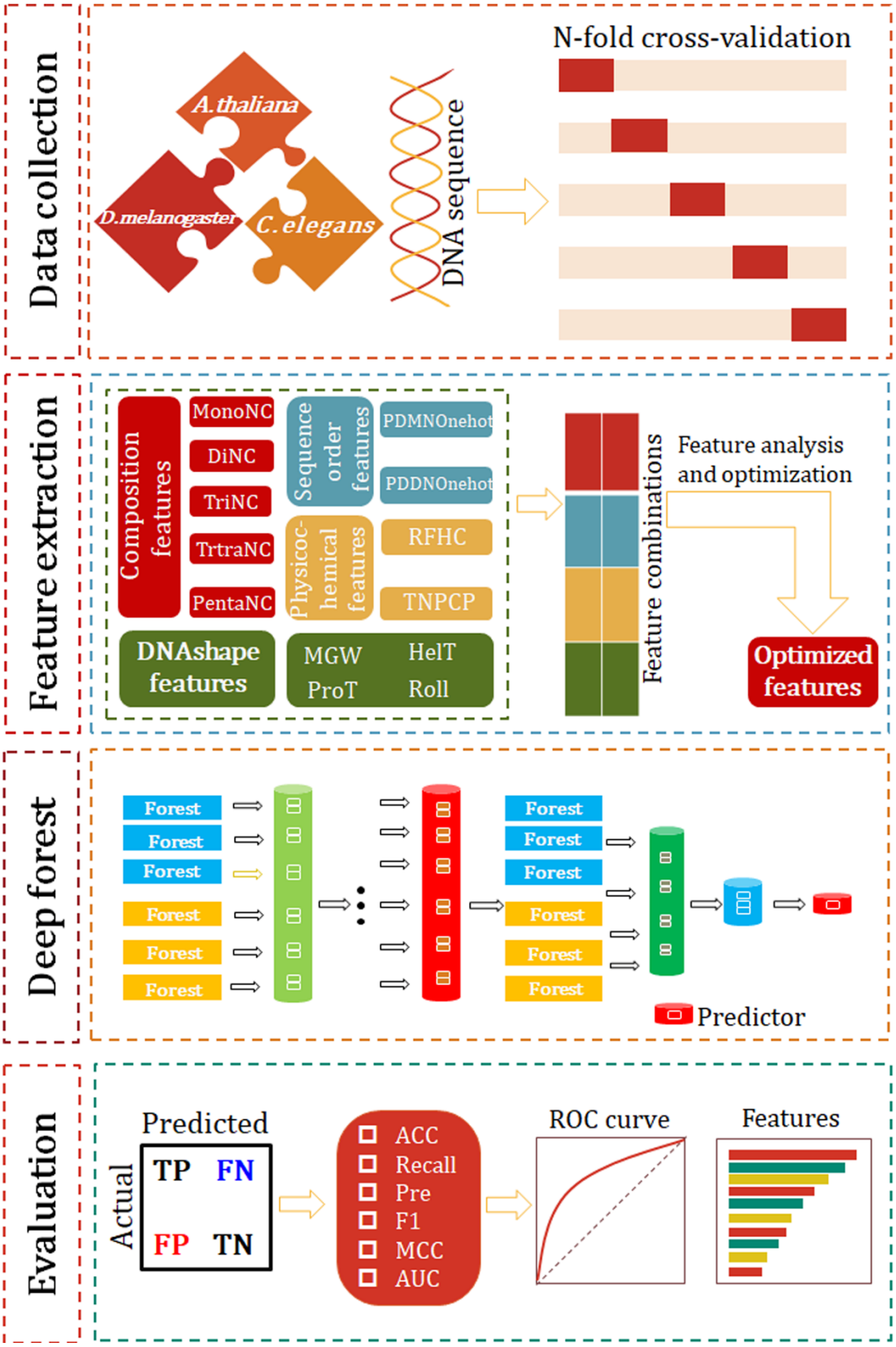

Figure 1. General overview of the algorithm design procedure. The deep forest model was trained on multiple sequential, compositional, physiochemical, and structural features for three species data using n-fold cross-validation to predict 4mC and non-

4mC sites. The trained models with best hyperparameters were evaluated using various performance metrics.

2. Materials and Methods

2.1 Data collection and processing

Rigorous and high quality benchmark datasets are crucial for training and evaluating the prediction models. Here, we used the benchmark dataset constructed by *Zeng et al*. to develop our computational model [19]. The total dataset contains 120, 000 samples from three species (*A. thaliana*, *C. elegans*, and *D. melanogaster*), each consisting of 20,000 positive samples (4mC) and 20,000 negative samples (non-4mC), respectively. All positive samples were extracted from the MethSMRT database [20] with a sequence length of 41 bp, where the methylated cytosine is located at the center position. Negative samples were collected from the whole-genome sequences by randomly sampled the same numbers of non-methylated cytosine sites. In addition, CD-HIT tool [21] was used to reduce the redundancy and homology bias with an 80% identity threshold.

2.2 Feature encoding methods

2.2.1 Composition-based features

Nucleotide composition is the basic feature of DNA sequences. It is calculated by dividing the number of Kmer of type i by the total number of Kmer of all types, where i=1,2,…,4^k. We generated five composition-based feature sets, namely mono-nucleotide composition (MonoNC), di-nucleotide composition (DiNC), tri-nucleotide composition (TriNC), Tetra-nucleotide composition (TetraNC) and pennta-nucleotide composition (PentaNC). A total of 1,364 nucleotide compositions were calculated, where the number of features for MonoNC, DiNC, TriNC, TetraNC and PentaNC are 4, 16, 64, 256 and 1,024 respectively.

2.2.2 Sequence-based features

We transformed nucleotide sequences into numerical vectors using the "one-hot" encoding to capture the position order information. A nucleotide *j* is converted to a 1 × 4 vector with a value of 0 at all positions except the *j*-th position; for example, Adenine (A) is represented as (1, 0, 0, 0). Each DNA sequence is thus represented by a binary vector of size $L \times 4$, where L is the sequence length and the "4" corresponds the number of nucleotide types (A, T, C, G). They are position-dependent mono-nucleotide onehot features (PDMNOnehot) [22]. We further considered all adjacent pairwise nucleotides as categorical variables, such as AA/AT/AC/etc. There are 4 × 4 such pairs, thus a single variable representing one such pair gests one-hot encoded into 16 binary vector, and the whole nucleotide sequence is represented as a (L-1) × 16-demensional vector (position-dependent di-nucleotide onehot, PDDNOnehot) [23].

2.2.3 Physicochemical property-based features

The calculation of tri-nucleotide physicochemical property features (TNPCP) is divided into two steps. We first transformed the retrieved 4mC and non-4mC sequences into numerical profiles where each tri-nucleotide is replaced by its corresponding physicochemical value. Seven physicochemical properties obtained from various bio-chemical experiments are used for property value assignment, including MW_kg, MW_Daltons, Nucleosome, Nucleosome_Rigid, Nucleosome_positioning, Dnase_I, and Dnase_I_Rigid [24]. Second, we used a sliding window approach to smooth the raw profiles with a window size of 5 bp and a step size of 1 bp. When the window slides along the sequence, we calculated the average value on each position.

According to the chemical properties of bases, such as ring structure, hydrogen bond and function group (RHFC), the nucleotides can also be classified into the following types:

$$\text{Ring structure} = \begin{cases} Purine, 1 \in \{A,G\} \\ Pyrimidinne, 0 \in \{C,T\} \end{cases} \quad (1)$$

$$\text{Hydrogen bond} = \begin{cases} Weak, 1 \in \{A,T\} \\ Strong, 0 \in \{C,G\} \end{cases} \quad (2)$$

$$\text{Function group} = \begin{cases} Amino, 1 \in \{A, C\} \\ Keto, 0 \in \{G, T\} \end{cases} \quad (3)$$

Thus, based on the chemical properties, A, C, G, and T can be encoded as (1, 1, 1), (0, 0, 1), (1, 0, 0) and (0, 1, 0), respectively [25].

2.2.4 Shape-based features

An increasing number of studies have shown that DNA structure plays an important role in determining the DNA binding preferences of transcription factors and other DNA binding proteins [26]. However, due to the limitations of experimental DNA structure determination, the amount of information about DNA structure lags far behind the amount of known DNA sequence data. To address this issue, Zhou *et al*. developed the DNAshape method for predicting DNA structural features for massive sequence data in a high-throughput manner [27]. Through the mining of atomic resolution data obtained from Monte Carlo simulations, they have characterized multiple important structural features of all 1,024 unique pentanucleotide fragments. DNAshape creates a sequence-structure dictionary to translate the input DNA sequences into their corresponding shape features using a sliding-window strategy. Here, we selected four shape features: the minor groove width (MGW), roll (Roll), propeller twist (ProT) and helix twist (HelT), to represent the DNA helical structures.

2.3 Deep forest and feature importance analysis

Deep learning models are generally built with neural network (NN)-based algorithms, such as convolutional neural network, recurrent neural network, deep belief network and generative neural network [28]. Currently, Zhou *et al*. proposed a non-NN-style deep learning concept that uses non-differentiable modules such as decision trees to build deep models instead of the parameterized differentiable non-linear neural networks [29]. Under the guidance of this concept, they constructed deep forest, a novel decision-tree ensemble, with a cascade structure that enables representation learning by forest. Deep forest shares the same crucial characteristics behind the success of deep neural networks, i.e. layer-by-layer processing, in-model feature

transformation and sufficient model complexity; on the other hand, it has several advantages over DNNs, such as it has much fewer hyper-parameters than DNNs and its model complexity can be automatically determined in a data-dependent way. Extensive experimental results demonstrate that the performance of deep forest is quite robust to hyper-parameter setting; and in most cases, it achieves achieve excellent performance even when using the default setting. In addition, deep forest algorithm supports the feature importance analysis of inner cascade layer. The importance of each feature in the cascade layers can be computed based on the criteria of Gini impurity or information gain. The deep forest algorithm was implemented by the DF21 package in Python.

2.4 Performance evaluation

We evaluated the performance of the models by 10-fold cross-validation. We partitioned the dataset into 10 non-overlapping equally sized sets, and trained the model on the union of 9 of these before testing on the remaining set. This process was repeated 10 times such that each of the 10 sets was used as the test set exactly once, and the average measures was recorded. We used the following five metrics to assess prediction performance: accuracy (*ACC*), precision (*Pre*), recall (*Recall*), F-value (*F1-value*) and Mathews correlation coefficient (MCC). They are defined as follows:

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

$$PRE = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1{-}value = 2 \times \frac{TP}{2TP + FP + FN} \tag{7}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{8}$$

where TP, TN, FP and FN are true positives (e.g., 4mC sites predicted as 4mC), true negatives, false positives and false negatives, respectively. The receiver operating

characteristic (ROC) curve is plotted with the true-positive rate against false-positive rate for various thresholds, and the area under the curve (AUC) was calculated to provide an aggregate measure of performance.

3. Results

3.1 Performance and robustness evaluation of different feature coding methods, predictors and datasets

To explore how two key factors of a deep forest model, input features and output predictors, determine the ability to effectively detect 4mC sites, we first developed thirty classifiers with different combinations of these factors. We incorporated sequential, compositional, physiochemical, and structural features in these models, some previously reported to be associated with N^4-methylcytosine. The features incorporated included position-dependent mono-nucleotide onehot (PDMNOnehot), position-dependent di-nucleotide onehot (PDDNOnehot), tri-nucleotide physicochemical property (TNPCP), ring structure, hydrogen bond and function group (RHFC), DNAshape, mono-nucleotide composition (MonoNC), di-nucleotide composition (DiNC), tri-nucleotide composition (TriNC), tetra-nucleotide composition (TetraNC) and pennta-nucleotide composition (PentaNC) (Materials and Methods). Additionally, we selected three predictors (forest, lightGBM and XGBoost), which are commonly used for deep forest classifiers, to test the discrimination power of these features. We used tenfold cross-validation on the benchmark data for each species to derive the corresponding predictive model based on binary classification (4mC and non-4mC). Supplementary Table 1, 2 and 3 summarized the performance of a total of 90 classification models with respect to different feature encoding methods, predictors and datasets. We found that the performance of different feature encoding methods varied greatly. PDDNOnehot achieved higher predictive performance in all test cases as measured by accuracy, recall, precision, *F*1-score and MCC compared to other features. The other nine features, ordered by above metrics, were RHFC, DNAshape, PDMNOnehot, TNPCP,

TetraNC, PentaNC, TriNC, DiNC and MonoNC, respectively. We also observed that all composition-based features consistently performed significantly worse than the other features, implying that they were less important in 4mC identification. When comparing the contribution of a single predictor for 4mC prediction, lightGBM was the predictor exerting the greatest effect, which is consistent with all independent features and datasets (Supplementary Table 1, 2 and 3). The effect of forest and XGBoost was similar or worse than that of lightGBM. Taken together, this benchmarking demonstrated that PDDNOnehot and lightGBM were the best features and predictors for predicting 4mC sites in the deep forest models.

The receiver operating characteristic curve of tenfold cross-validation was exploited to evaluate whether proposed feature encoding methods could be used to predict 4mC and non-4mC sites. To quantitatively assess the discriminative power of all features, we calculated the area under the ROC curves by using the classification probabilities computed by the lightGBM predictor. The PDDNOnehot achieved superior performance as indicated by the area under the curve (AUC) compared to other features in all species (Figure 2; AUC%: 91.7 for *A. thaliana*; 95.8 for *C. elegans*; 94.0 for *D. melanogaster*). RFHC showed a slightly worse performance, but the largest difference in the AUC scores is just 0.10. None of composition-based features alone could achieve the level of performance achieved by PDDNOnehot or RFHC, which is consistent with their classification results.

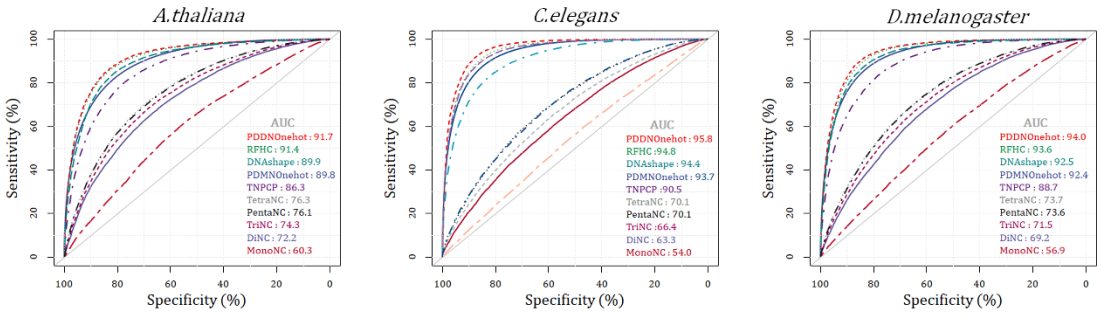

Figure 2. Receiver operating characteristic (ROC) curves and the corresponding AUC values of different feature encoding methods based on lightGBM predictor across three species: *A. thaliana*, *C. elegans* and *D. melanogaster*.

3.2 Model optimization and feature importance analysis

To evaluate the factors associated with 4mC methylation in a more systematic manner, we next performed feature importance analysis on the integrative 2,684 features that included all proposed types. For the deep forest model, we used its built-in layer feature importance method to assess the importance of each feature based on the Gini score [29]. Figure 3 showed the top 30 most important features for each species, ranked by the order of Gini score among all features in the full model. For simplicity, the name of each feature was represented by an abbreviated form and listed as the variable name, followed by '-' and the position of input sequence. We found that *A. thaliana* and *C. elegans* were ranked similarly in terms of feature importance, especially for the top seven features; whereas the most important features of *D. melanogaste* were completely different from the other two species. For *A. thaliana* and *C. elegans* species, the sequence-based PDMNOnehot and PDDNOnehot were the most important, and their top seven features were as follows: TA-26, CC-20, C-20, H-26, C-19, T-26 and G-28. Physicochemical property-based TNPCP feature had a greater impact on the *D. melanogaste*, with five of the top seven features belonging to this type, including MW_kg-18, MW_Daltons-18, Nucleosome-18, MW_Daltons-19 and MW_kg-19. In addition, some of the important features such as ProT-18 (DNAshape) were shared by all three species, implying that DNA structural features also have an important role in the identification of 4mC sites. To explore this more deeply, we showed ProT-18 value distributions for the 4mC and non-4mC sites in the entire data set (Supplementary figure 1). At the 18^{th} position of the input sequence, the positive samples (4mC sites) showed higher propeller twist degree or DNA flexibility than negative samples (non-4mC sites) in all species.

To build accurate and reliable predictive models, and to simplify the learning process, we eliminated irrelevant features by ranking their importance and pruning the least important ones. To investigate how many features are sufficient for making model predictions, we extracted several optimized feature subsets based on the number of top features and compared their performance using ACC and MCC as

evaluation criteria. Specifically, we trained a deep forest model for each species using an increasing number of features. Here, the number of selected input features were 10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1,000, 1,500, 2,000, 2,500 and 2,684, respectively. We observed a significant improvement in the performance of trained models as the number of feature increased, however, the inflection point of curve occurred at number 300, after which the improvement in model performance plateaus and additional features did not improve performance (Supplementary figure 2). Nevertheless, when the number of features was 1,500, the trained model obtained the best performance. The optimized model can predict 4mC sites with 85.0% accuracy, 86.3% recall, 84.1% precision, 85.2% F-value and 0.701 *MCC* when benchmarked against the *A. thaliana* dataset. For the other species, the values of *ACC*, *Recall*, *Pre*, *F1* and *MCC* were 90.0%, 92.0%, 88.4%, 90.2%, 0.801 for *C.elegans*, and 87.8%, 90.2%, 86.2%, 88.1%, 0.758 for *D.melanogaster*, respectively.

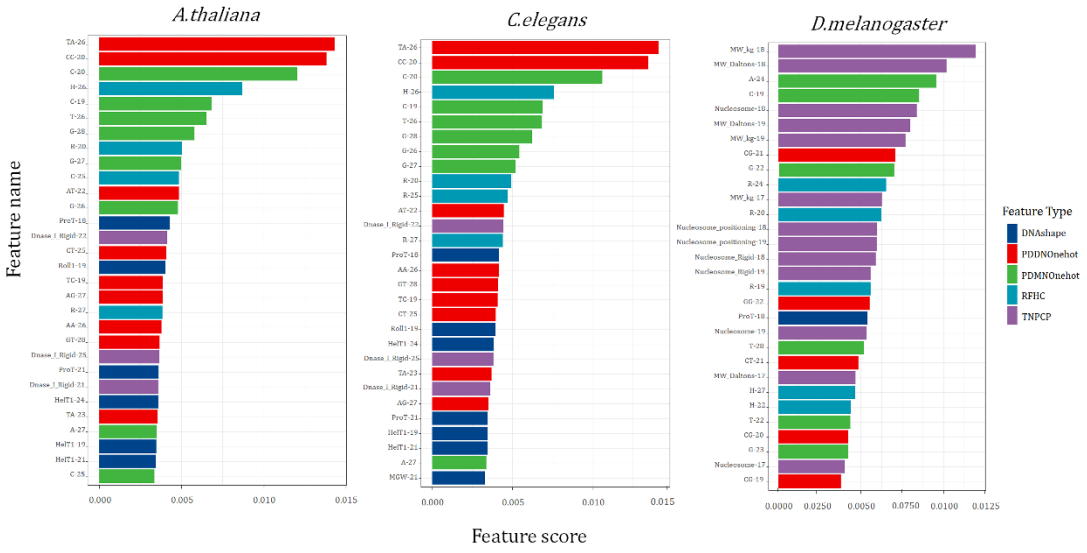

Figure 3. The Gini contribution scores of the top 30 most important features for each species.

3.3 Comparison with other methods

First, we compared the performance of deep forest (DF) with that of random forest (RF) [30] and deep neural network (DNN) [31] using the same training data and same model features for model calibration. The Scikit-learn [32] and Keras [33] packages

were used to implement the RF and DNN models, respectively. For the RF model, we mainly optimized the hyperparameters for the number of decision trees, ranging from 500 to 2,500 in steps of 500. The DNN model was constructed from three fully connected layers with 1,000, 500 and 100 hidden units, respectively. Each fully connected layer applied ReLU and dropout to prevent overfitting [34]. The last output layer was a sigmoid activation node. We trained the model with 50 epochs and a batch size of 75. In this comparison, we observed that deep forest produced the best performance (*A. thaliana* ACC: 85.0%, *C. elegans* ACC: 90.0%, *D. melanogaster* ACC: 87.8%) compared with RF (*A. thaliana* ACC: 82.7%, *C. elegans* ACC: 87.6%, *D. melanogaster* ACC: 84.8%) and DNN (*A. thaliana* ACC: 63.4%, *C. elegans* ACC: 70.2%, *D. melanogaster* ACC: 68.4%). We noted that the neural network-based DNN model failed to differentiate between these two samples. One possible reason for this might be that for the processing of proposed discrete features, the non-differentiable deep forests have an advantage over multilayer neural networks with parameterized differentiable non-linear modules.

We next compared the performance of our approach against several previously published predictors, including iDNA4mc [35], 4mcPred [36], 4mcPred_SVM [37], 4mcPred_IFL, Deep4mcPred [38] and DeepDNA4mC [39]. For comparison purposes, we referred to our as DF4mC. Figure 4 showed the accuracy and MCC evaluated per species for each competitor method. Overall, we observed that DF4mC consistently achieved higher predictive performance across all species compared to other methods.

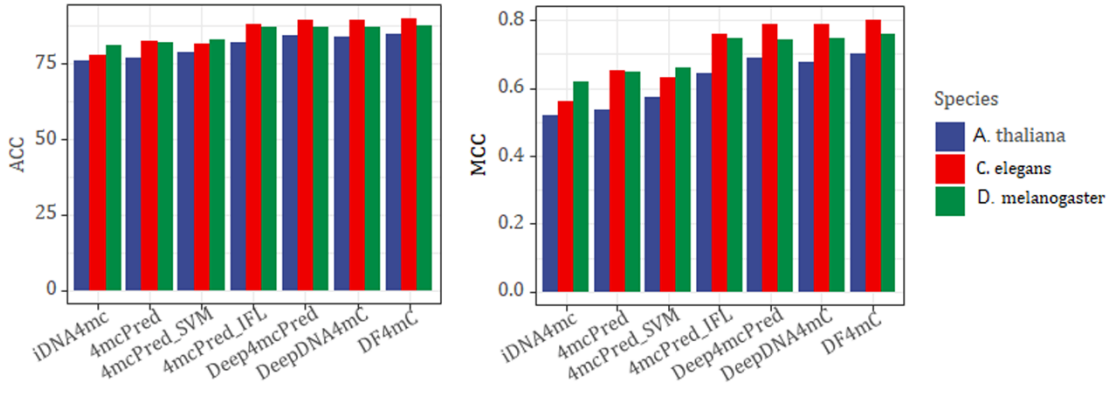

Figure 4. Performance comparison of DF4mC and other six existing predictors illustrated by accuracy and MCC

4. Conclusion

In this paper, we have presented DF4mC, a method that applies a non-NN-style deep model for elucidating DNA methylation on N^4-cytosine. Several simple and generalized classification models were learned using a dataset of experimentally determined 4mC sequences for three representative model organisms. These models provide a quick, inexpensive and interpretable estimate of 4mC and non-4mC pattern in the whole-genome of each individual species. To train the models, a comprehensive set DNA features was compiled to capture information about the sequences of upstream and downstream of the modification sites. These features were further refined and optimized using the permutation feature importance technique. The classification models were trained using the deep forest algorithm, while their performance was evaluated through 10-fold cross-validation. Various feature types and model predictors were assessed; specifically, we sought to answer the question of if and whether these factors influence performance. Our method has shown to be robust across different species and outperformed traditional machine learning algorithms and other existing predictors in terms of accuracy and MCC. The code for DF4mC is freely available at http://github.com/z513591251/DF4mC. The classification models are fast, robust and easy to set up and run.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (No. 21803045), and Joint project of Luzhou Municipal People's Government and Southwest Medical University (2020LZXNYDJ39).

Notes

The authors declare no competing financial interest.

Reference

1. Du K, et al. Epigenetic DNA modification N6-methyladenine inhibits DNA replication by Sulfolobus solfataricus Y-family DNA polymerase Dpo4. Arch Biochem Biophys 2019;675:108120. http://doi.org/10.1016/j.abb.2019.108120.
2. Luo GZ, et al. DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? Nat Rev Mol Cell Biol 2015;16(12):705-10. http://doi.org/ 10.1038/nrm4076.
3. Hong T, et al. Selective detection of N6-methyladenine in DNA via metal ion-mediated replication and rolling circle amplification. Chem Sci 2017 ;8(1):200-205. http://doi.org/10.1039/c6sc02271e.
4. Liu X, et al. N6-methyladenine is incorporated into mammalian genome by DNA polymerase. Cell Res 2021;31: 94-97. https://doi.org/10.1038/s41422-020-0317-6.
5. Lyko,F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nat Rev Genet 2018;19(2):81-92. http://doi.org/10.1038/nrg.2017.80.
6. Song CX, Yi C, He C. Mapping recently identified nucleotide variants in the genome and transcriptome. Nat Biotechnol 2012;30(11):1107-16. http://doi.org/10.1038/nbt.2398
7. Jin B, Li Y, Robertson KD. DNA methylation: superior or subordinate in the epigenetic hierarchy? Genes Cancer 2011;2(6):607-17. http://doi.org/10.1177/1947601910393957.
8. Ehrlich M, et al. N4-methylcytosine as a minor base in bacterial DNA. J Bacteriol 1987;169(3):939-43. http://doi.org/10.1128/jb.169.3.939-943.1987.
9. Flusberg BA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 2010;7(6):461-5. http://doi.org/10.1038/nmeth.1459.
10. Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res 2009; 19(6):959-66. http://doi.org/10.1101/gr.083451.108.
11. Rathi P, Maurer S, Summerer D. Selective recognition of N4-methylcytosine in DNA by engineered transcription-activator-like effectors. Philos Trans R Soc Lond B Biol Sci 2018; 373(1748):20170078. http://doi.org/10.1098/rstb.2017.0078.
12. Yu M, et al. Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite- sequencing. Nucleic Acids Res 2015;43(21):e148. http://doi.org/10.1093/nar/gkv738.
13. Wei L, et al. Iterative feature representations improve N4-methylcytosine site prediction. Bioinformatics 2019;35(23):4930-4937. http://doi.org/10.1093/bioinformatics/btz408.
14. Wei L, et al. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics 2019;35(8):1326-1333. http://doi.org/10.1093/bioinformatics/bty824.
15. He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics 2019;35(4):593-601. http://doi.org/10.1093/bioinformatics/bty668.
16. Chen W, et al. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide

chemical properties. Bioinformatics 2017;33(22):3518-3523. http://doi.org/10.1093/bioinformatics/btx479.
17. Zeng F, Fang G, Yao L. A Deep Neural Network for Identifying DNA N4-Methylcytosine Sites. Front Genet 2020;11:209. http://doi.org/10.3389/fgene.2020.00209.
18. Ye P, et al. MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Res 2017;45(D1):D85-D89. http://doi.org/10.1093/nar/gkw950.
19. Sood AJ, Viner C, Hoffman MM. DNAmod: the DNA modification database. J Cheminform 2019;11(1):30. http://doi.org/10.1186/s13321-019-0349-4.
20. Liu ZY, et al. MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. Hortic Res 2019;6:78. http://doi.org/10.1038/s41438-019-0160-4.
21. Zeng R, Liao M. Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications. Front Bioeng Biotechnol 2020;8:274. http://doi.org/10.3389/fbioe.2020.00274.
22. Fu L, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28(23):3150-2. http://doi.org/10.1093/bioinformatics/bts565.
23. Zhou Y, et al. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res 2016;44(10):e91. http://doi.org/10.1093/nar/gkw104.
24. Sehi P, et al. i6mA-DNC: Prediction of DNA N6-Methyladenosine sites in rice genome based on dinucleotide representation using deep learning. Chemometr Intell Lab Syst 2020;204,104102. http://doi.org/10.1016/j.chemolab.2020.104102
25. Chen W, et al. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 2013;41(6):e68. http://doi.org/10.1093/nar/gks1450.
26. Grabherr MG, et al. Exploiting nucleotide composition to engineer promoters. PLoS One 2011;6(5):e20136. http://doi.org/10.1371/journal.pone.0020136.
27. Panwar B, Raghava GP. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. Genomics 2015;105(4):197-203. http://doi.org/10.1016/j.ygeno.2015.01.005.
28. Iqbal M, Hayat M. "iSS-Hyb-mRMR": Identification of splicing sites using hybrid space of pseudo trinucleotide and pseudo tetranucleotide composition. Comput Methods Programs Biomed 2016;128:1-11. http://doi.org/10.1016/j.cmpb.2016.02.006.
29. Nagpal G, et al. VaccineDA: Prediction, design and genome-wide screening of oligodeoxynucleotide-based vaccine adjuvants. Sci Rep 2015;5:12478. http://doi.org/10.1038/srep12478.
30. Chen W, et al. iRNA-PseU: Identifying RNA pseudouridine sites. Mol Ther Nucleic Acids 2016;5(7):e332. http://doi.org/10.1038/mtna.2016.37.
31. Hasan MM, 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 2020;103(1-2):225-234. http://doi.org/10.1007/s11103-020-00988-y.
32. Zhou T, et al. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci U S A 2015;112(15):4654-9. http://doi.org/10.1073/pnas.1422023112.

33. Zhou ZH, Feng J. Deep forest. Natl Sci Rev 2019;6(1):74-86. http://doi.org/10.1093/nsr/nwy108.
34. Khanal J, et al. i6mA-stack: A stacking ensemble-based computational prediction of DNA N6-methyladenine (6mA) sites in the Rosaceae genome. Genomics 2021;113(1 Pt 2):582-592. http://doi.org/10.1016/j.ygeno.2020.09.054.
35. Xu H, Jia P, Zhao Z. Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. Brief Bioinform 2021;22(3):bbaa099. http://doi.org/
36. Li Z, et al. Deep6mA: A deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species. PLoS Comput Biol 2021;17(2):e1008767. http://doi.org/10.1093/bib/bbaa099.
37. Liu Q, et al. DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Brief Bioinform 2021;22(3):bbaa124. http://doi.org/10.1093/bib/bbaa124.
38. Yan J, et al. Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. Mol Ther Nucleic Acids 2020;20:882-894. http://doi.org/10.1016/j.omtn.2020.05.006.
39. Basith S, et al. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. Comput Struct Biotechnol J 2018;16:412-420. http://doi.org/10.1016/j.csbj.2018.
40. Lv H, et al. iDNA-MS: An Integrated Computational Tool for Detecting DNA Modification Sites in Multiple Genomes. iScience 2020;23(4):100991. http://doi.org/10.1016/j.isci.2020.100991.