Bioinformatics, 36(10), 2020, 3257–3259
doi: 10.1093/bioinformatics/btaa113
Advance Access Publication Date: 24 February 2020
Applications Note

Sequence analysis

6mA-Finder: a novel online tool for predicting DNA N6-methyladenine sites in genomes

Haodong Xu [1], Ruifeng Hu[1], Peilin Jia [1] and Zhongming Zhao[1,2,3,*]

[1]School of Biomedical Informatics, Center for Precision Health, [2]MD Anderson Cancer Center, UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA and [3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on November 23, 2019; revised on January 29, 2020; editorial decision on February 13, 2020; accepted on February 14, 2020

Abstract

Motivation: DNA N6-methyladenine (6 mA) has recently been found as an essential epigenetic modification, playing its roles in a variety of cellular processes. The abnormal status of DNA 6 mA modification has been reported in cancer and other disease. The annotation of 6 mA marks in genome is the first crucial step to explore the underlying molecular mechanisms including its regulatory roles.

Results: We present a novel online DNA 6 mA site tool, 6 mA-Finder, by incorporating seven sequence-derived information and three physicochemical-based features through recursive feature elimination strategy. Our multiple cross-validations indicate the promising accuracy and robustness of our model. 6 mA-Finder outperforms its peer tools in general and species-specific 6 mA site prediction, suggesting it can provide a useful resource for further experimental investigation of DNA 6 mA modification.

Availability and implementation: https://bioinfo.uth.edu/6mA_Finder.

Contact: zhongming.zhao@uth.tmc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

DNA N6-methyladenine (6 mA) represents an essential epigenetic modification in the genomes of diverse species (Greer *et al.*, 2015; Zhang *et al.*, 2015). DNA 6 mA refers to the modification of adding a methyl group to the 6th position of an adenine ring catalyzed by DNA methyltransferases. This process plays an important role in the regulation of various biological processes, including restriction-modification system, DNA repair and replication, and gene expression. Moreover, recent studies have shown that the abnormal status of DNA 6 mA modification is related to human cancer and other disease (Xiao *et al.*, 2018).

Due to the limitation of labor-intensive and expensive experiments, *in silico* prediction of DNA 6 mA sites in a genome has emerged to be an alternative approach. Here, we developed a novel predictor, 6 mA-Finder, by integrating seven types of sequence-derived information and three types of physicochemical-based features for both general and species-specific 6 mA site predictions. Recursive feature elimination (RFE) strategy (Granitto *et al.*, 2006) was used to choose the optimal feature group. We strictly evaluated the performance of 6 mA-Finder through cross-validation (CV). For the general prediction, the area under curve (AUC) values were 0.9193, 0.9197, 0.9203 and 0.9207 for 4-, 6-, 8- and 10-fold CVs, indicating the promising accuracy and robust of general model. In comparison, 6 mA-Finder outperforms other existing tools and achieved a considerable AUC value improvement using both benchmark and independent datasets. Furthermore, we found that 6 mA-Finder could achieve better performance than other existing tools in the species-specific manner. The web service of 6 mA-Finder is freely available at https://bioinfo.uth.edu/6mA_Finder.

2 Materials and methods

2.1 Data collection and processing

In 2018, Feng *et al.* (2019) built a benchmark DNA 6 mA site dataset in the mouse genome, containing 1934 positive data (1934 sequences that contain 6 mA site) and 1934 negative data. The 6 mA sites in the mouse genomes were taken from the MethSMRT database (Ye *et al.*, 2016), and its accession number of sequencing data in the Gene Expression Omnibus (GEO) was GSE71866. The data were generated from the mouse embryonic stem cells (Wu *et al.*, 2016). Another frequently used 6 mA dataset was built using the rice (*Oryza sativa* spp. *gengdao*) genome, including 880 positive data and 880 negative data (Chen *et al.*, 2019). These data were obtained from GEO with accession number GSE103145 (Zhou *et al.*, 2018). All the sequences are 41-bp long with the 6 mA site in the center. In addition to the two species-specific datasets, we

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/36/10/3257/5753942 by guest on 10 April 2024

compiled a cross-species dataset by merging these two datasets, which comprised 2768 positive data and 2716 negative data after eliminating sequence redundancy using CD-HIT software (Fu *et al.*, 2012). The compiled dataset was divided into benchmark (2500 positive data versus 2500 negative data) and independent dataset (268 positive data versus 216 negative data). All of the benchmark and independent dataset can be downloaded at: https://bioinfo.uth.edu/6mA_Finder/Download.php.

2.2 Feature encoding and classification algorithms

Seven sequence-based coding schemes were implemented, including Accumulated Nucleotide Frequency (ANF), Binary, Composition of K-spaced Nucleic Acid Pairs (CKSNAP), Dinucleotide Composition (DNC), Enhanced Nucleic Acid Composition (ENAC), Nucleic Acid Composition (NAC) and Trinucleotide Composition (TNC). Three types of physicochemical features were encoded, including electron–ion interaction pseudopotentials of trinucleotide (EIIP), Nucleotide Chemical Property (NCP) and Pseudo Dinucleotide Composition (PseDNC). Seven conventional classification algorithms were implemented and assessed by 10-fold CV in a pairwise way. RFE strategy was used to select the optimal feature group. The details of these methods were described in the Supplementary Material.

3 Results

Motif-based analysis for 6 mA modification sequences was performed. The results indicated recognition specificity for 6 mA-containing DNA fragments in the general data and different species (Supplementary Figs. S1 and S2). To extract the motif information, 10 types of features were encoded. The average AUC values of individual feature upon each classification algorithm were calculated and illustrated based on 10-fold CV (Fig. 1A, Supplementary Figs. S3 and S4). For the general prediction, the results showed that four features, i.e. Binary, NCP, EIIP and ENAC, performed well, obtaining the average AUC values as 0.8479, 0.8450, 0.8274 and 0.8260, respectively. The performances of TNC, CKSNAP, PseDNC, DNC and NAC were relatively moderate with the average AUC values ranging from 0.6856 (NAC) to 0.7509 (TNC), whereas ANF had the lowest average AUC value of 0.5849. Similarly, Binary, NCP, EIIP and ENAC could achieve high AUC values in species-specific prediction (Supplementary Figs. S3 and S4). Among them, EIIP and ENAC were novel features and not used in the previous 6 mA predictors. Our results demonstrated that all sequences and physicochemical features were efficient and informative. In addition, we found the most powerful classifier was random forest (RF), receiving an average AUC value of 0.7669, 0.8646 and 0.7479 across 10 types of features in general, mouse-specific and rice-specific prediction (Fig. 1B). Taking the general data as an example, the importance of each dimension feature was evaluated by the chi-square test (Fig. 1C), suggesting different features contributed to model performance unequally. Thus, the RFE strategy was used to further select the optimal feature group (Fig. 1D). Finally, the RF algorithms with 65-, 55- and 155-dimension representative features selected by RFE method were implemented as the predictive model for the general data, mouse-specific and rice-specific data in the 6 mA-Finder. Such integration of optimal features performed much better than the individual features (Fig. 1E). Moreover, we have attempted to build an ensemble method for the prediction of 6 mA sites. Specifically, the predicted probability from seven classical machine-learning algorithms were considered as the second feature vector and was input into the Logistic Regression classifier to build an ensemble model for prediction. However, we found the ensemble model could not improve the performance when compared to the original model.

We performed 4-, 6-, 8- and 10-fold CVs on the benchmark dataset. The average AUC values were 0.9201, 0.9938 and 0.9367 in general, mouse-specific and rice-specific prediction (Fig. 1F, Supplementary Fig. S5), suggesting the promising accuracy and the robustness of the model. Moreover, we compared 6 mA-Finder with five previously reported predictors (Fig. 1G): iDNA6mA-PseKNC (Feng *et al.*, 2019), iDNA6mA-Rice (Lv *et al.*, 2019), i6mA-Pred

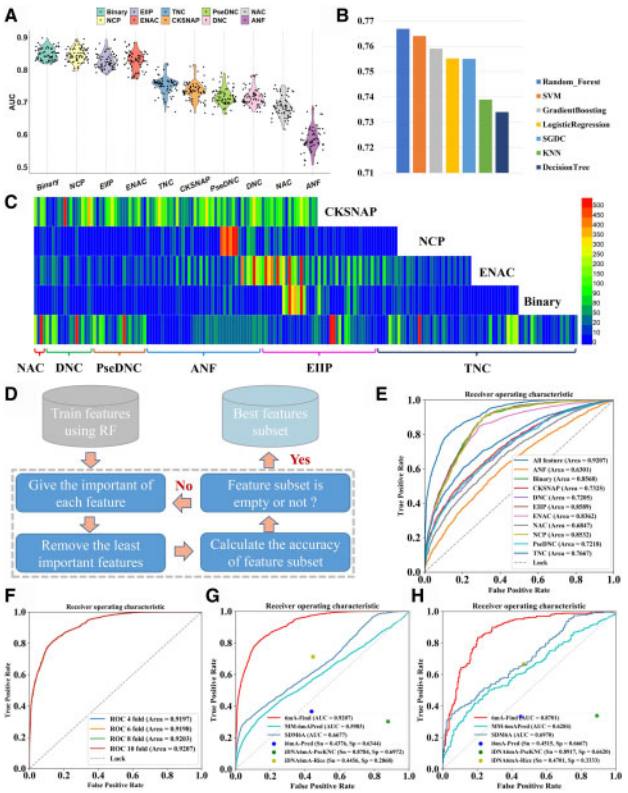

Fig. 1. (**A**) Average AUC values of individual feature upon different classification algorithms and (**B**) *vice versa*. (**C**) Feature was calculated by the chi-square test. (**D**) The workflow of RFE strategy. (**E**) Performance comparison between the integration of optimal features and the individual feature. (**F**) The ROC curves and AUC values of 6 mA-Finder in the benchmark dataset. (**G**, **H**) The comparison of 6 mA-Finder with other existing predictors using benchmark dataset (G) and independent dataset (H)

(Chen *et al.*, 2019), MM-6mAPred (Pian *et al.*, 2020) and SDM6A (Basith *et al.*, 2019). Of note, only the predictors with available web service were selected and compared. These tools do not have the option for customizing computational models by users. Thus, the benchmarks or independent datasets were directly submitted into each tool to calculate the performance. The ROC curves of MM-6mAPred and SDM6A were illustrated and the AUC values were computed as 0.5983 and 0.6677. Moreover, for those predictors that have pre-defined scoring cutoffs, the sensitivity and specificity values were calculated and labeled. We found that 6 mA-Finder achieved better sensitivity, with an improvement of 37.89% AUC value compared to the results by iDNA6mA-Rice and SDM6A in the benchmark data. In the independent dataset, the performance of 6 mA-Finder was also superior to the existing tools, as demonstrated by better sensitivity and AUC value (Fig. 1H). Furthermore, we compared 6 mA-Finder with previous tools in the species-specific manner. We found that 6 mA-Finder_Mouse achieved better performance than iDNA6mA-PseKNC and 6 mA-Finder_Rice had a higher AUC value than other existing tools in the rice genome (Supplementary Fig. S6). The web service of 6 mA-Finder is freely available at: https://bioinfo.uth.edu/6mA_Finder. Currently, only limited benchmark datasets are available for testing, but this algorithm and tool can be applied to other new data in future.

Funding

This work was supported by NIH [R01LM012806]; and Cancer Prevention & Research Institute of Texas (CPRIT) [RP180734].

Conflict of Interest: none declared.

Downloaded from https://academic.oup.com/bioinformatics/article/36/10/3257/5753942 by guest on 10 April 2024

References

Basith,S. *et al*. (2019) SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids*, **18**, 131–141.

Chen,W. *et al*. (2019) i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics*, **35**, 2796–2800.

Feng,P. *et al*. (2019) iDNA6mA-PseKNC: identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics*, **111**, 96–102.

Fu,L. *et al*. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Granitto,P.M. *et al*. (2006) Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chem. Intel. Lab. Syst*., **83**, 83–90.

Greer,E.L. *et al*. (2015) DNA methylation on N6-adenine in *C. elegans*. *Cell*, **161**, 868–878.

Lv,H. *et al*. (2019) iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Front. Genet*., **10**, 793.

Pian,C. *et al*. (2020) MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics*, **36**, 388–392.

Wu,T.P. *et al*. (2016) DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature*, **532**, 329–333.

Xiao,C.L. *et al*. (2018) N(6)-methyladenine DNA modification in the human genome. *Mol. Cell*, **71**, 306–318.e7.

Ye,P. *et al*. (2016) MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res*., **45**, gkw950.

Zhang,G. *et al*. (2015) N6-methyladenine DNA modification in Drosophila. *Cell*, **161**, 893–906.

Zhou,C. *et al*. (2018) Identification and analysis of adenine N 6-methylation sites in the rice genome. *Nat. Plants*, **4**, 554–563.

Downloaded from https://academic.oup.com/bioinformatics/article/36/10/3257/5753942 by guest on 10 April 2024