m6ABRP: Predicting m^{6}A-YTHDF2 Binding Regions via Sequence-based Properties

Ze Liu*
College of Water Resources and Architectural Engineering, Northwest A&F University, 3 Taicheng Road, Yangling
zeliu@nwafu.edu.cn

Jinghui Xu
College of Water Resources and Architectural Engineering, Northwest A&F University, 3 Taicheng Road, Yangling
x36936@163.com

Xiuli Mao
College of Water Resources and Architectural Engineering, Northwest A&F University, 3 Taicheng Road, Yangling
maoxl@nwafu.edu.cn

Jianzhao Qi
Shaanxi Key Laboratory of Natural Products & Chemical Biology, College of Chemistry & Pharmacy, Northwest A&F University, 3 Taicheng Road, Yangling
qjz@nwafu.edu.cn

Quanwu Li
College of Water Resources and Architectural Engineering, Northwest A&F University, 3 Taicheng Road, Yangling
liquanwu@nwafu.edu.cn

ABSTRACT

m^{6}A plays important roles in cell differentiation and tissue development via selectively binding with the YTH-containing proteins. However, the mechanism of the selectively binding events is largely unknown. The precise prediction of m^{6}A-YTH binding regions across the transcriptome will be helpful to learn the molecular basis of m^{6}A selectively binding with the YTH-containing proteins. In this study, we developed a machine learning-based model, named m6ABRP, for predicting m^{6}A-YTHDF2 binding regions. Sequence-based features were extracted and five different algorithms, Support vector machine (SVM), RandomForest, ExtraTrees, GradientBoosting and AdaBoost, were implemented for model training. Among them, the SVM-based model received the best AUC of 0.920 on the training dataset using five-fold cross validation and received the best AUC of 0.910 on the independent test dataset. The results suggest that m6ABRP is a powerful tool for predicting m^{6}A-YTHDF2 binding regions.

CCS CONCEPTS

• **Computing methodologies** → Machine learning.

KEYWORDS

YTHDF2, Selectively binding, Machine learning, Cross-validation

*Ze Liu received the Ph. D. degree in school of electronics and information technology from Xi'an Jiaotong University, Xi'an, China, in 2017. He is currently a lecture of the College of Water Resources and Architectural Engineering, Northwest A&F University, Shanxi, China. His research interests include bioinformatics and machine learning.

ICBBS '20, October 16–18, 2020, Xiamen, China

ACM ISBN 978-1-4503-8865-8/20/10…$15.00
https://doi.org/10.1145/3431943.3431950

ACM Reference Format:
Ze Liu, Jinghui Xu, Xiuli Mao, Jianzhao Qi, and Quanwu Li. 2020. m6ABRP: Predicting m^{6}A-YTHDF2 Binding Regions via Sequence-based Properties. In *2020 9th International Conference on Bioinformatics and Biomedical Science (ICBBS '20), October 16–18, 2020, Xiamen, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3431943.3431950

1 INTRODUCTION

More than 150 types of post-transcriptional modifications have been found on cellular RNA so far, e.g. N6-methyladenosine (m^{6}A) [1], 5-methylcytidine (m^{5}C) [2], pseudouridine [3] and N^1-methyladenosine (m^{1}A) [4]. Among them, m^{6}A occupied more than 50% of total methylated ribonucleotides, accounting for 0.1%~0.4% of all adenosines in total cellular RNA [5]. m^{6}A is the most abundant modification in eukaryotic mRNA and was first discovered in 1974 using chromatography [6]. However, it had been considered to be an inherently static modification in the past few decades. In 2011, Jia et al. found that the fat mass and obesity-associated protein (FTO) exhibits efficient m^{6}A demethylase activity both in vivo and in vitro, which suggests that m^{6}A is dynamically reversible [7]. Since then, some studies have shown that m^{6}A is always highly enriched within long exons and 3' UTR, especially near stop codons [8]. It is worth mentioning that there are almost no m^{6}A modifications in the poly(A) tails of mRNAs [1].

Subsequent studies show that m^{6}A plays important roles in alternative splicing [9], viral replication [10] and tumor proliferation [11], via the recognition by selective binding with different proteins. Many proteins are involved in the modification of m^{6}A (Figure 1). The m^{6}A methyltransferases, called the "writers", are responsible for the methylation of RNA and mainly consist of METTL3, METTL14, and WTAP proteins. The m^{6}A detransferases, named the "erasers", are in charge of m^{6}A removal and mainly including the ALKBH5 and FTO proteins. The RNA-binding proteins, called the "readers", are responsible for recognizing m^{6}A sites and can be divided into the direct acting-proteins and indirect-acting proteins. The direct-acting proteins mainly include five YT521-B homologous

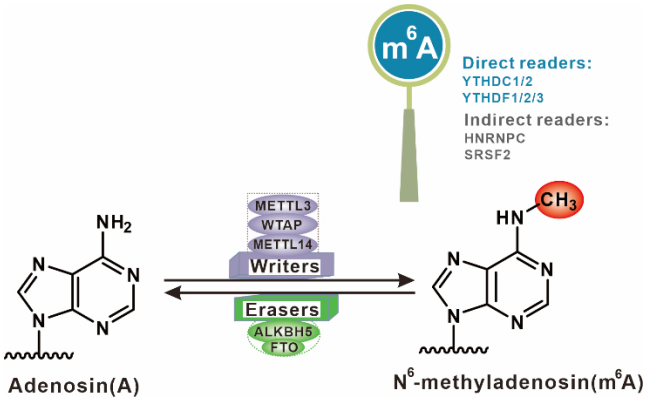

Figure 1: The dynamic regulation of m^{6}A using different proteins

(YTH) family of proteins, i.e. YTHDF1, YTHDF2, YTHDF3, YTHDC1 and YTHDC2, which can recognize m^{6}A via the YTH domain located at the C-terminus [8]. However, despite these progresses the underlying mechanisms of m^{6}A functions remain to be elucidated.

Accumulating evidences show that YTH-containing proteins play various functions at different stages of cell development. YTHDF1 facilitates the translation of m^{6}A-containing mRNAs in protein-synthesis [12]; YTHDF2 accelerates RNA degradation of target mRNAs [13]; YTHDF3 plays important roles in the initial stage of m^{6}A-driven translation from circular RNAs [14]; YTHDC1/2 is essential for the germline development in mouse [15]. By using photoactivatable ribonucleoside cross-linking and immunoprecipitation (PAR-CLIP) and Methylated RNA Immunoprecipitation Sequencing (MeRIP-Seq), Wang et al. found that 59% of the YTHDF2 binding regions overlapped with the m^{6}A peaks [13], and Shi et al. discovered that only 28% of the YTHDF3 binding regions overlapped with the m^{6}A peaks [12]. These results suggest that m^{6}A readers selectively recognized m^{6}A sites. However, the mechanism by which the m^{6}A readers selectively binding with m^{6}A sites is poorly understood.

In light of m^{6}A's increasingly recognized roles in cell development, plenty of efforts have been made toward predicting m^{6}A sites based on the machine learning methods. Such as MethyRNA [16], SRAMP [17], DeepM6ASeq [18], and WHISTLE [19] in mammalian genome, M6ATH [20], AthMethPre [21], and RFAthM6A [22] in plant genome. However, no effort has been made for predicting m^{6}A-YTH binding regions using a machine learning-based method. To further investigate the mechanism of the selectively binding events, it is a critical step to locate the m^{6}A-YTH binding regions precisely and find the sequence preferences in the m^{6}A-containing mRNAs. As YTHDF2 is the first identified reader that specifically recognizes and binds to m^{6}A-containing mRNAs, we take YTHDF2 as a case and attempt to find out the sequence and context preferences of m^{6}A-YTH binding regions. What's more, we also developed a machine learning-based predictor, m6ABRP, to predict m^{6}A-YTHDF2 binding regions in mRNAs. The m6ABRP classifier received good performance both on the training dataset and the independent test dataset. The results suggest that our model can be used to predict m^{6}A-YTHDF2 binding regions in the transcriptome.

2 METHODS

2.1 Original Data Acquisition

Combining MeRIP-seq, miCLIP-seq (m^{6}A-specific methylated RNA immunoprecipitation combined with high-throughput sequencing) and YTHDF2 RIP-seq (RNA-immunoprecipitation sequencing for YTHDF2), Zhang et al. investigated the function of m^{6}A during embryogenesis of zebrafish [23]. Accordingly, 55,886 and 43,371 YTHDF2 binding regions were extracted using the MCS2 peak-calling software [24]. To generate a high-quality benchmark dataset, we selected those located on the chromosomes and then extracted 20,404 YTHDF2 binding regions in the intersection as the positive samples. To obtain the negative samples, the R package developed by Ghandi et al. [25] was used to generate the negative dataset. To avoid sequence redundancy, the CD-HIT-EST software [26] was used with the most rigorous threshold at 0.8 and 19,796 positive samples and 34,208 negative samples were obtained in the original dataset. To keep a balanced rate of the positive samples and negative samples, we random select 19,796 negative samples to create the negative dataset. Finally, three datasets were launched in this study, i.e. the DatasetFP, consists of 1,583 positive samples and 1,583 negative samples, was used for feature selection and parameter optimization; the DatasetCV, includes 14,253 positive samples and 14,253 negative samples, was used for model training; and the DatasetIT, contains 3,960 positive samples and 3,960 negative samples, was used as the independent testing dataset.

2.2 Feature Extraction

Extract key features of the biological sequences is important to improve the performance of the constructed classifier. The composition of sequences and the order in which the various components in the sequences arranged are the key features for a biological sequence. In this study, K monomeric units (K-mer) and K-spaced nucleotide pair frequencies (KSNPF) were used for feature encoding. K-mer divides the sequence of length L into a string of k residues and can obtain n-k+1 components for each sequence. For example, an RNA sequence of length 100 is simple frequencies of k consecutive residues and have been widely used in the prediction of DNA/RNA modifications, such as 4mC [27], m^{5}C [28] and m^{6}A [29]. In this study, we set k as 2, 3, 4, and obtained $4^2+4^3+4^4$=336 kinds of K-mer components for each sequence. KSNPF are frequencies of nucleotide pairs gapped by k arbitrary residues [22], [30] and can be defined as follows:

$$f\left(n1Gap\left(k\right)n2\right) = \frac{S\left(n1Gap\left(k\right)n2\right)}{L-k-1} \quad (1)$$

Where n1Gap(k)n2 stands for a nucleotide pair (n1 and n2) separated by k arbitrary residues (Gap(k)). S(n1Gap(k)n2) represents the number of appearances of the nucleotide pair in a sequence of length L. In this study, we set k as 1, 2, 3, 4, and generated $4^2\times4$=64 kinds of KSNPF components for each sequence. Thus, the original feature space consists of 400 features for each sample.

2.3 Feature Selection

Feature selection refers to the selection of the most effective feature subset from the original feature space. It can reduce the dimension

of the original feature space and improve the generalization performance of the training model. The RandomForest algorithm is a tree-based ensemble learning model and can be used to assess feature importance [31]. For feature selection, the DatasetFP was used and the original feature space was ranked using the RandomForest algorithm. To further elucidate the impact of multi-feature integration, the principal component analysis (PCA) method [32] was also used. The PCA algorithm can transform a set of possibly linear correlation feature vectors into a set of linear uncorrelated vectors (Principal components). In this study, the ranked feature subset was processed using the PCA method, and each principal component was ranked according to the magnitude of its contribution. To find the best combination of the ranked features and principal components, the grid research method was used and the model performance with different combinations was compared.

2.4 Implementation of Five Different Classifiers

Five different classifiers, SVM, RandomForest, ExtraTrees, GradientBoosting, and AdaBoost, were implemented using sklearn module [33]. The SVM-based classifier intends to find the optimal classification hyperplane and maximize the classification interval. Selecting a proper kernel function has a great impact on the performance of the SVM classifier. For 'radial basis' kernel function, grid search was used to find the best combination of 'C' and 'gamma'. For the other four ensemble-based classifiers, RandomForest, ExtraTrees, GradientBoosting and AdaBoost, different numbers of estimators were selected for each of them.

2.5 Validation of the Proposed Prediction Model

The cross-validation method was first proposed by Stone et al. in 1974 [34]. Its main purpose is to overcome the optimism of training error estimates generated by training and testing on the same dataset. In this study, five-fold cross validation was used to evaluate the performance of the training model. The DatasetCV was roughly divided into 5 folds at first, and then in each iteration, 4 folds were taken in turn for model training and the left one was used as the test set. The mean results were used as estimates of the accuracy of the model.

Standard measures, Sensitivity (Sn), Specificity (Sp), Accuracy (ACC) and Matthews correlation coefficient (MCC) [35], [36], were used to evaluate the performance of our model. Their definitions are defined as follows:

$$\begin{cases} Sn = 1 - \frac{N_-^+}{N^+} \\ Sp = 1 - \frac{N_+^-}{N^-} \\ ACC = 1 - \left(\frac{N_-^+ + N_+^-}{N^+ + N^-}\right) \\ MCC = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \end{cases} \quad (2)$$

where N^+ and N^- refer to the number of positive samples or negative samples, respectively. N_-^+ stands for the number of positive samples that were predicted to be negatives, N_+^- refers to the number of negative samples that were predicted to be positives. To compare the performance of different classifiers, the receiver operating characteristic curves (ROC) corresponding to different classifiers were also drawn in the same coordinate system. The closer the ROC curve is to the upper left corner, the performance of the classifier is better. Besides, the AUC (Area under the curve) values were also used to compare the performance of different classifiers. Generally, the classifier performed better with larger AUC values.

3 RESULTS

3.1 YTHDF2 is Selectively Binding with m^{6}A Sites

Previous studies have demonstrated that YTHDF2 binding with m^{6}A sites through its YTH domain. However, YTHDF2 only selectively binding with a portion of m^{6}A sites. As shown in Figure 2(b) and Figure 2(c), only 6,272 and 5,486 YTHDF2 binding regions were overlapped with m^{6}A sites for each replication. We also investigated the distribution of m^{6}A sites and YTHDF2 binding regions and found that m^{6}A sites and YTHDF2 binding regions were enriched in promoter, 3' UTR and exon regions (Figure 2(d)). What's more, we also found that YTHDF2 binding domains were enriched near transcription initiation sites, which suggests that YTHDF2 binding domains play important roles in the transcription process.

3.2 Feature Ranking and Fusion

As irrelevant and redundant features tend to degrade the performance of the training model, the RandomForest algorithm was used to select the optimal feature subset from the original feature space. Five-fold cross validation was used on the DatasetFP and 100 estimators were selected into the RandomForest classifier for performance evaluation. The top ten ranked features were shown in Figure 3(a). Among them, the frequencies of 'ATA', with feature importance more than 0.015, is the most striking feature of all. Using all original features to train classifiers can result in slow training and poor generalization. The PCA algorithm was then used for feature fusion while maintaining most of the information contained in the ranked feature set. As shown in Figure 3(b), it is very hard to classify positive samples from negative samples using two components. Thus, more principal components were used to train the classify model, e.g. 10 principal components are shown in Figure 3(c), and a heat map was used to represent the contribution of each feature to each principal component. The grid research method was used to find the best combination of the ranked features and principal components. We set the number of the ranked features from 1 to 400 and set the number of principal components from 1 to 100. Finally, our model received the best AUC of 0.822 when the top 215 ranked features were used to generate 58 principal components.

3.3 Performance Evaluation of m6ABRP with Different Strategies

3.3.1 Five-fold Cross Validation on the Training Dataset. The DatasetCV was used for model training and five different algorithms, SVM, RandomForest, GradientBoosting, Adaboost, and ExtraTrees, were implemented in this study. For SVM, grid research

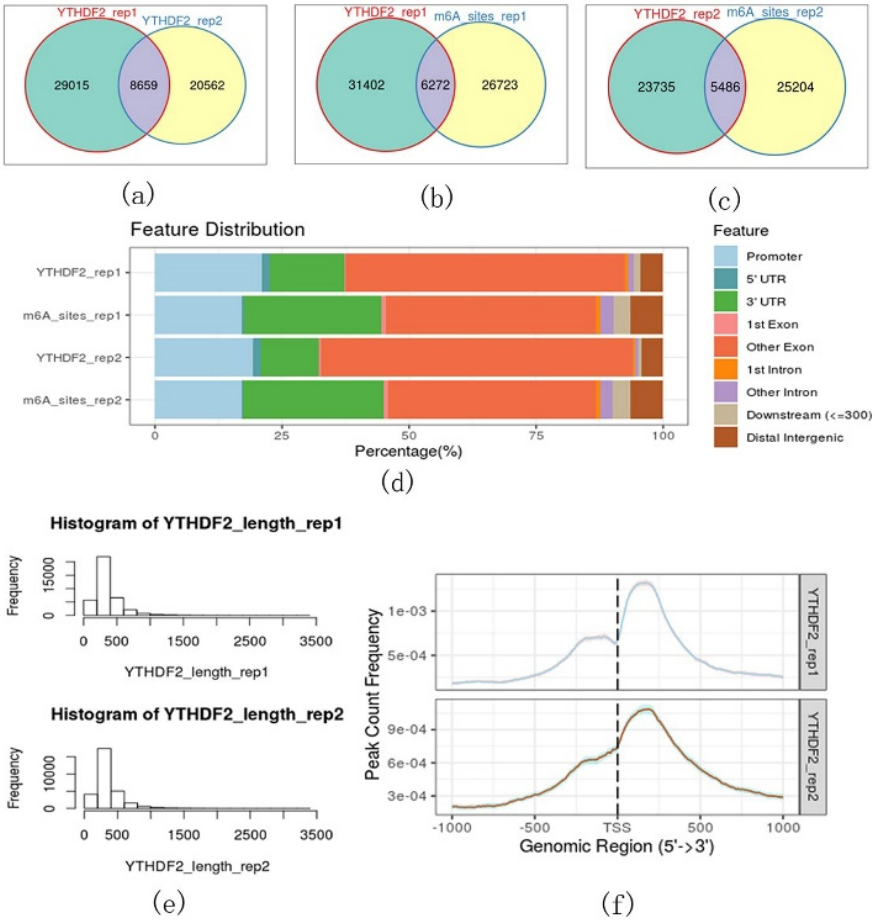

Figure 2: Distribution of YTHDF2 binding regions. (a) The reproducible regions detected in the comparison of replicates 1 and 2; (b) YTHDF2 were selectively binding with m^{6}A sites in replicate 1; (c) YTHDF2 were selectively binding with m^{6}A sites in replicate 2; (d) Genome-wide distribution of m^{6}A sites and YTHDF2 binding regions in each replicates; (e) Length distribution of YTHDF2 binding regions of replicates 1 and 2; (f) YTHDF2 binding domains were enriched near transcription initiation sites of replicates 1 and 2

was used and the classifier received the best performance with 'C' of 6.75 and 'gamma' of 0.02. For the other four ensemble-based classifiers, RandomForest, ExtraTrees, GradientBoosting and AdaBoost, 1000 estimators were selected for each of them. Five-fold cross validation was used to evaluate the performance of different models. The mean values of standard measures, such as ACC, AUC, Sn, Sp, and MCC, were used as estimates of the accuracy of the model. As shown in Table 1, the SVM classifier received the best mean ACC of 0.838, Sp of 0.879, and MCC of 0.680, while the AdaBoost classifier received the best Sn of 0.818. To obtain a more comprehensive view of the model performance, the AUC value was also calculated in each iteration. And the SVM classifier received the best mean AUC of 0.920. Thus, the SVM algorithm was used to train the optimized model.

3.3.2 Model Performance on the Independent Testing Dataset. To further evaluate the generalization performance of our model, all samples in the DatasetCV were used for model training and all samples in the DatasetIT were used for prediction. As shown in Table 2, the SVM classifier also received the best ACC of 0.832, AUC of 0.910, Sn of 0.783 and MCC of 0.667, while the RandomForest classifier received the best Sp of 0.903 and the AdaBoost classifier also achieved the best Sn of 0.783. The ROC and PRC curves are shown in Figure 4(a) and Figure 4(b), the SVM classifier also received the best AUC of 0.910 and the best AP of 0.927. Although the SVM classifier achieved the best performance, there were only a few differences between the five classifiers and each classifier can receive an acceptable performance for prediction of YTHDF2 binding regions.

4 DISCUSSION

As the experimental methods are expensive and time-consuming, it is very necessary to predict m^{6}A-YTH binding regions based on the machine learning method. In this study, an SVM-based classifier,

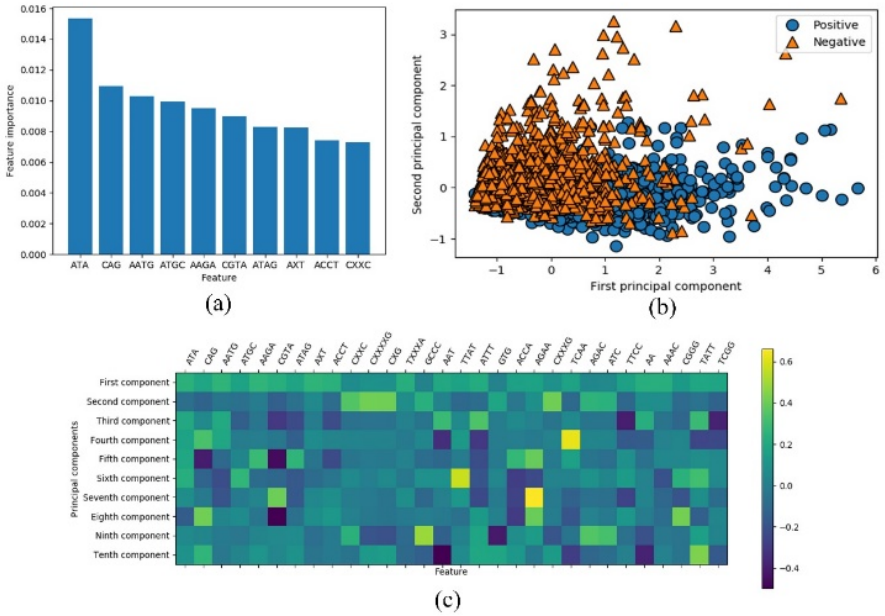

Figure 3: Feature ranking and fusion. (a) The original feature set was ranked using the RandomForest algorithm; (b) The ranked feature set was fused into two principal components using the PCA algorithm; (c) The contribution of each feature to the top ten principal components

Table 1: Model comparison on the training dataset

Classifier	ACC	AUC	Sn	Sp	MCC
RandomForest	0.820	0.895	0.767	0.873	0.643
GradientBoosting	0.825	0.910	0.812	0.839	0.651
AdaBoost	0.822	0.907	**0.818**	0.826	0.644
ExtraTrees	0.819	0.897	0.763	0.874	0.641
SVM	**0.838**	**0.920**	0.798	**0.879**	**0.680**

The highest value of each column is marked in bold.

Table 2: Model comparison on the independent testing dataset

Classifier	ACC	AUC	Sn	Sp	MCC
RandomForest	0.812	0.888	0.721	**0.903**	0.635
GradientBoosting	0.814	0.898	0.778	0.852	0.631
AdaBoost	0.815	0.894	**0.783**	0.847	0.632
ExtraTrees	0.817	0.897	0.755	0.880	0.639
SVM	**0.832**	**0.910**	**0.783**	0.880	**0.667**

The highest value of each column is marked in bold.

m6ABRP was developed for predicting YTHDF2 binding regions. Firstly, several differential enrichment motifs were discovered and sequence-based features, Kmer and KSNPF, were extracted to establish the original feature space; and then the RandomForest algorithm was used for feature ranking and the PCA method was used for feature fusion; Finally, five different algorithms, SVM, RandomForest, ExtraTrees, GradientBoosting, and AdaBoost, were implemented to construct the classifier. The five-fold cross validation results on the DatasetCV and the independent test results on the DatasetIT show the SVM-based classifier received the best performance on both datasets. These results suggest that m6ABRP can be further used for predicting m^{6}A-YTHDF2 binding regions in the transcriptome.

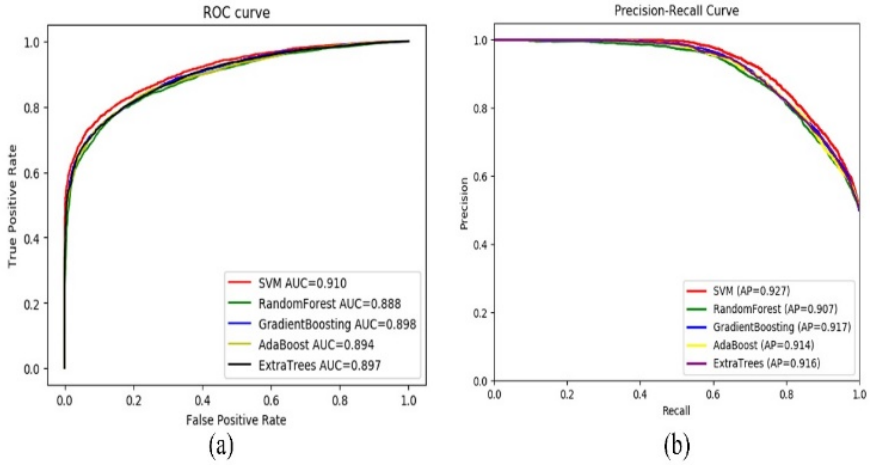

Figure 4: Performance comparison using ROC and PRC curves. Five different algorithms, SVM, RandomForest, GradientBoosting, AdaBoost, and ExtraTrees, were trained on the DatasetCV and independent testing on the DatasetIT. (a) The ROC curves of different classifiers. (b) The PRC curves of different classifiers

m6ABRP is the first machine learning-based model for predicting YTHDF2 binding regions up to now. Evidences show that other m^6A readers, YTHDF1, 3, YTHDC1 and YTHDC2, can also selectively binding with m^6A sites. Thus, the prediction framework can be also extended to predict binding regions of other m^6A readers, which makes it possible to reduce time-consuming and labor-intensive experimental methods. Although the prediction of YTHDF2 binding regions via sequence-based features achieved satisfactory results, there is still room for improvement. In the future, more feature encoding schemes, such as conformational and thermodynamic properties, will be used to improve the performance of m6ABRP. And a more complicated method, such as a Deep Learning algorithm will be adopted to train the classifier.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (61902323, 51909222, and 31800031). Z.L. participated in conceiving and performing the experiments. J.X., X.M. participated in analyzing the data. All authors contributed to the writing of the manuscript. The authors declare no competing interests.

REFERENCES

[1] Dominissini D, Moshitch-Moshkovitz S, Schwartz S, *et al.* 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature, 485(7397):201-206. https://doi.org/10.1038/nature11112.

[2] Khoddami V, Cairns BR. 2013. Identification of direct targets and modified bases of RNA cytosine methyltransferases. Nature Biotechnology, 31(5):458-464. https://doi.org/10.1038/nbt.2566.

[3] Schwartz S, Bernstein DA, Mumbach MR, *et al.* 2014. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell, 159(1):148-162. https://doi.org/10.1016/j.cell.2014.08.028.

[4] Li X, Xiong X, Wang K, *et al.* 2016. Transcriptome-wide mapping reveals reversible and dynamic N(1)-methyladenosine methylome. Nature Chemical Biology, 12(5):311. https://doi.org/10.1038/nchembio.2040.

[5] Yue H, Nie X, Yan Z, *et al.* 2019. N6-methyladenosine regulatory machinery in plants: composition, function and evolution. Plant Biotechnology Journal, 17(7). https://doi.org/10.1111/pbi.13149.

[6] Desrosiers R, Friderici K, Rottman F. 1974. Identification of Methylated Nucleosides in Messenger RNA from Novikoff Hepatoma Cells. Proceedings of the National Academy of Sciences, 71(10):3971-3975. https://doi.org/10.1073/pnas.71.10.3971.

[7] Jia G, Fu Y, Zhao X, *et al.* 2011. N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nature Chemical Biology, 7(12):885-887. https://doi.org/10.1038/nchembio.687.

[8] Shi H, Wei J, and He C. 2019. Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. Molecular Cell, 74(4):640-650. https://doi.org/10.1016/j.molcel.2019.04.025.

[9] Robinson M, Shah P, Cui Y, *et al.* 2018. The Role of Dynamic m6A RNA Methylation in Photobiology. Photochemistry and Photobiology, 95:374-387. https://doi.org/10.1111/php.12930.

[10] Kennedy E, Bogerd H, Kornepati AR, *et al.* 2016. Posttranscriptional m6A Editing of HIV-1 mRNAs Enhances Viral Gene Expression. Cell Host & Microbe, S1931312816301081. https://doi.org/10.1016/j.chom.2016.04.002.

[11] Deng X, Su R, Weng H, *et al.* 2018. RNA N6-methyladenosine modification in cancers: current status and perspectives. Cell Research, 28:507-517. https://doi.org/10.1038/s41422-018-0034-6.

[12] Shi H, Wang X, Lu Z, *et al.* 2017. YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA. Cell Research, 27(3):315-328. https://doi.org/10.1038/cr.2017.15.

[13] Wang X, Lu Z, Gomez A, *et al.* 2014. N6-methyladenosine-dependent regulation of messenger RNA stability. Nature, 505:117-120. https://doi.org/10.1038/nature12730.

[14] Yang Y, Fan X, Mao M, *et al.* 2017. Extensive translation of circular RNAs driven by N6-methyladenosine. Cell Research, 27:626-641. https://doi.org/10.1038/cr.2017.31.

[15] Hsu P, Zhu Y, Ma H, *et al.* 2017. YTHDC2 is an N6 -methyladenosine binding protein that regulates mammalian spermatogenesis. Cell Research, 27(9):1115-1127. https://doi.org/10.1038/cr.2017.99.

[16] Chen W, Tang H, Lin H. 2016. MethyRNA: A web-server for identification of N(6)-methyladenosine sites. Journal of biomolecular Structure & Dynamics, 35(3):1-11. https://doi.org/10.1080/07391102.2016.1157761.

[17] Zhou Y, Zeng P, Li YH, *et al.* 2016. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Research, 44(10): gkw104. https://doi.org/10.1093/nar/gkw104.

[18] Zhang Y, Hamada M. 2018. DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. BMC Bioinformatics, 19(S19). https://doi.org/10.1186/s12859-018-2516-4.

[19] Chen K, Wei K, Zhang Q, *et al.* 2019. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Research, 47(7):e41. https://doi.org/10.1093/nar/gkz074.

[20] Chen W, Feng P, Ding H, *et al.* 2016. Identifying N6-methyladenosine sites in the Arabidopsis thaliana transcriptome. Molecular Genetics and Genomics, 291(6):2225-2229. https://doi.org/10.1007/s00438-016-1243-7.

[21] Xiang S, Yan Z, Liu K, *et al.* 2016. AthMethPre: a web server for the prediction and query of mRNA m6A sites in Arabidopsis Thaliana. Mol. BioSyst. 12:3333-3337.

https://doi.org/10.1371/journal.pone.0162707.

[22] Wang X, Yan R. 2018. RFAthM6A: a new tool for predicting m6A sites in Arabidopsis thaliana. Plant Molecular Biology, 96(3):327-337. https://doi.org/10.1007/s11103-018-0698-9.

[23] Zhang C, Chen Y, Sun B, *et al.* 2017. m6A modulates haematopoietic stem and progenitor cell specification. Nature, 549:273-276. https://doi.org/10.1038/nature23883.

[24] Zhang Y, Liu T, Meyer CA, *et al.* 2008. Model-based Analysis of ChIP-Seq (MACS). Genome biology, 9: R137. https://doi.org/10.1186/gb-2008-9-9-r137.

[25] Ghandi M, Mohammad-Noori M, Ghareghani N, *et al.* 2016. gkmSVM: an R package for gapped-kmer SVM. Bioinformatics, btw203. https://doi.org/10.1093/bioinformatics/btw203.

[26] Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13):1658. https://doi.org/10.1093/bioinformatics/btl158.

[27] Wei L, Luan S, Zou Q, *et al.* 2019. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics, 35(8):1326-1333. https://doi.org/10.1093/bioinformatics/bty824.

[28] Song J, Zhai J, Bian, E, *et al.* 2018. Transcriptome-Wide Annotation of m5C RNA Modifications Using Machine Learning. Frontiers in Plant Science, 9:519. https://doi.org/10.3389/fpls.2018.00519.

[29] Zhao Z, Peng H, Lan C, *et al.* 2018. Imbalance learning for the prediction of N6-Methylation sites in mRNAs. BMC Genomics, 19:574. https://doi.org/10.1186/s12864-018-4928-y.

[30] Liu, Z, Dong W, Jiang W, *et al.* 2019. csDMA: an improved bioinformatics tool for identifying DNA 6mA modifications via Chou's 5-step rule. Scientific Reports, 9:13109. https://doi.org/10.1038/s41598-019-49430-4.

[31] Ganz M., *et al.* 2015. Relevant feature set estimation with a knock-out strategy and random forests. NeuroImage, 122:131-148. https://doi.org/10.1016/j.neuroimage.2015.08.006.

[32] Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. British Journal of Educational Psychology, 24(6):417-520. https://doi.org/10.1037/h0070888.

[33] Pedregosa F, Gramfort A, Michel V, *et al.* 2013. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(10):2825-2830. https://doi.org/10.1524/auto.2011.0951.

[34] Stone MA. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion). Journal of the Royal Statistical Society, 36(2): 111-47. https://doi.org/10.2307/2344741.

[35] Chou KC. 2001. Using subsite coupling to predict signal peptides. Protein Engineering Design and Selection, 14(2):75-79. https://doi.org/10.1093/protein/14.2.75.

[36] Chou KC. 2002. Prediction of signal peptides using scaled window. Peptides, 22(12):1973-1979. https://doi.org/10.1016/S0196-9781(01)00540-X.