Prediction of m6A Reader substrate sites using deep convolutional and recurrent neural network

Yuxuan Wu  
 Department of Biological Science   
 Xi’an Jiaotong and Liverpool University  
Suzhou Jiangsu China  
yuxuan.wu17@student.xjtlu.edu.cn

Yuxin Zhang  
Department of Biological Science  
Xi’an Jiaotong and Liverpool University  
 Suzhou Jiangsu China   
yuxin.zhang17@student.xjtlu.edu.cn

Ruoqi Wang  
Department of Biological Science  
 Xi’an Jiaotong and Liverpool University  
Suzhou Jiangsu China   
ruoqi.wang1802@student.xjtlu.edu.cn

Jia Meng  
 Department of Biological Science  
 Xi’an Jiaotong and Liverpool University  
 Suzhou Jiangsu China  
jia.meng@xjtlu.edu.cn

Kunqi Chen  
Department of Biological Science  
 Xi’an Jiaotong and Liverpool University  
 Suzhou Jiangsu China  
Kunqi.Chen@xjtlu.edu.cn

Yiyou Song  
Department of Biological Science  
 Xi’an Jiaotong and Liverpool University  
 Suzhou Jiangsu China  
yiyou.song15@alumni.xjtlu.edu.cn

Daiyun Huang\*  
Department of Biological Science  
 Xi’an Jiaotong and Liverpool University  
 Suzhou Jiangsu China  
daiyun.huang@liverpool.ac.uk

ABSTRACT

N6-methyladenosine (m6A) is the most prevalent post-transcriptional modification in mRNA since it could regulate some significant biological functions with the binding of some m6A reader proteins. Multiple readers exist in the human genome, however, the binding specificity was not clarified due to the limited wet experiments on this topic. Therefore, we devised a deep learning approach which incorporated CNN and RNN frameworks together to predict the epitranscriptome-wide targets of six m6A reader proteins (YTHDF1-3, YTHDC1-2, EIF3A). We also utilized layer-wise relevance calculation to obtain contribution of each input feature. Our model achieved state-of-the-art performance with the average AUROC of 0.942 in EIF3A full transcript, compared with 0.929 in CNN-only framework and 0.817 in Support Vector Machine (SVM) method under same condition. Besides, we identified the optimal sequence length (1001bp) in the m6A reader substrate prediction. The results provide new insight into epitranscriptome target prediction and functional characterization of m6A readers.

CONCEPTS

• Applied computing • Life and medical sciences • Bioinformatics

KEYWORDS

Deep learning, convolutional neural network, recurrent neural network, m6A, Readers.

**1 Literature Review**

Post-transcriptional and post-translational modifications are very important to organisms. Common post-transcriptional modifications include RNA post-transcriptional modifications that occur in all types of RNA, and common post-translational modifications include protein modifications such as Protein S-sulfenylation [1]. N6-methyladenosine (m6A) is a kind of post-transcriptional modification, and it is mediated by methyltransferase (writer), demethylases (eraser), and recognition proteins (reader) [2]. Considering the limit of the wet-lab experiment, computer methods were used to predict the mechanisms and functions of modification. At present, most computer-based predictions use deep learning or machine learning to extract sequence features from base-resolution transcriptome datasets to predict possible RNA modification sites. [3].

iRNA and iDNA6mA-PseKNC are two machine learning-based predictors to perform m6A site identification. iRNA was designed with pseudo dinucleotide composition. By constructing an efficient benchmark dataset, formulating RNA sequences into three different RNA physiochemical properties, and finally developing an algorithm to conduct prediction [4]. iDNA6mA-PseKNC is a bioinformatics tool to identify N6-methyladenine (m6A) sites in DNA sequences, which is established by incorporating nucleotide physicochemical properties into Pseudo K-tuple Nucleotide Composition (PseKNC). iDNA6mA-PseKNC has similar model building steps with iRNA, they all contain constructing benchmark dataset, formulating sequence sample, operating engine or algorithm, conducting cross-validation, and building a web server [5]. For post-transcriptional modifications, the DNA sequence is processed by testing the continuous bags of nucleobases, and the result is input as a feature into the support vector machine (SVM) algorithm for identification to improve the sensitivity and accuracy of identifying DNA N6-methyladenine sites [6].

CNN (Convolutional Neural Networks) and RNN (recurrent neural network, i.e. Bidirectional LSTM Network) are two deep learning models to perform predictions. Specifically, CNN can learn potential motifs from the RNA sequence since CNN is suitable for non-tabular data like text (i.e. RNA sequence). The input sequences were firstly encoded to generate features and then fed into models to predict the target site, for instance, one-hot encoding is widely used for sequence analysis [6]. RNN can learn non-linear sequential features and the interaction between each nucleotide from the RNA sequences, such as long short-term memory unit (LSTM) and gated recurrent unit (GRU)[3]. For post-translational modifications, RNN was used to predict the function of the protein through a scoring matrix that represents the control loop unit and a specific location, and at the same time improve the accuracy of the prediction [7].

Gene2vec [8], DeepPromise [9], iN6-Methyl (5-step) [10] and Deep-m6A [11] are CNN based model to predict m6A or m1A modifications. Gene2vec is a neural embedding predictor that extends the sliding window length to the thousand base-pair level, uses word embedding to represent mRNA subsequences, and performs classification with CNNs [8]. DeepPromise uses three encoding methods: enhanced nucleic acid composition, one-hot encoding, and RNA embedding, then inputs seven consecutive layers of convolutional neural networks to predict N1-methyladenosine and N6-methyladenosine [9]. iN6-Methyl is a deep learning model, it can characterize the human genome without any clear definition and input the learned features into the convolutional neural network model for classification [10]. Deep-m6A is a CNN model for a single base m6A predictor, which first combines mRNA sequence feature and MeRIP-Seq data [11]. Deep-m6A is also the first predictor that attempts to use only computational methods to make overall predictions on a large number of human MeRIP-Seq samples, thereby predicting m6A functions and related diseases [11].

2 Introduction

Although a series of computer technologies have been developed in this regard for m6A site prediction. So far there is less computational tool available focus on m6A reader binding site prediction. A deep learning model was designed to predict the binding sites in the m6A marked transcripts for 5 YTH family proteins m6A reader (YTHDC1, YTHDC2, YTHDF1, YTHDF2, YTHDF3, and eIF3).

mRNA methylation occurs in the N6-position of adenosine is called N6-methyladenosine (m6A) is the most prevalent internal modification on eukaryotic mRNA [12]. On average, within 1000 nucleotides, one or two m6A residues can be found [13, 14]. m6A is usually found in a stop codon, 3' untranslated region (UTR), and long internal exon [15, 16]. m6A has roles in altering gene regulation, gene stability maintenance, cell renewal and differentiation [17]. m6A also involves RNA metabolisms such as mRNA translation, degradation, splicing, nuclear export, and folding [18, 19].

There are three types of m6A regulators: methyltransferases (writer), demethylases (erasers), and readers [12]. RNA reader proteins can bind RNA, and each reader has a specific binding site to perform the corresponding functions [12]. In previous research, a prediction framework WHISTLE was developed, aiming to predict the whole-transcriptome m6A site [20]. Please note, we only focus on the YTH family and EIF3 reader in this project.

m6A is recognized by the YT521-B homology (YTH) domain-containing proteins. YTH domains exist in 174 different proteins [21]. Early studies have shown that they play an important role in RNA metabolism [18, 19]. YTH domain, namely YTHDF1-3 (i.e. YTHDC1, YTHDC2, YTHDF1) and YTHDC1, YTHDC2, which are stored in the mammalian genome [17]. From the sequence alignment, YTHDF1, YTHDF2, and YTHDF3 are similar to each other, whereas YTHDC1 and YTHDC2 are considered as different subtypes [17].

Although these five readers all came from the YTH family, they play different roles in the cell to modulate gene regulation, DNA repair, and cell fate determination [22]. YTHDC1 a nuclear protein, it involves exon alternative selection during gene splicing by binding to SRSF3 (serine/arginine-rich splicing factor 3) [23]. YTHDC2 is an RNA helicase that recognizes the YTH domain of m6A, regulates RNA levels during meiosis [24]. YTHDF1–3 are cytoplasmic m6A readers, by binding to m6A, YTHDF1 improves the translation efficiency of mRNA independent of m7G cap [25]. YTHDF2 uses the CCR4- NOT deadenylase complex to destroy and further decay the target mRNA in mRNA clearance [26].

For the RNA binding protein that lacks YTH domain-containing m6A readers, eILF3 (eukaryotic translation initiation factor 3) was considered as the most complex eukaryotic translation initiation factor [27]. It can specialize in translation initiation, in IRES-mediated translation initiation, eIF3 directly identifies the structure of the IRES (Internal ribosomal entry) [28]. eIF3 plays an important role in protein translation such as dissociating the post-termination 80S ribosome, participating in the 43S pre-initiation complex formation, and stimulating mRNA binding with 43S pre-initiation complex [29].

To date, there is a prediction model or comprehensive online platform that can be used to annotate the internal modification function of mRNA, such as m7GHub [30]. In this project, the combination of CNN (Convolutional Neural Networks) and RNN (Bidirectional LSTM Network) was used to predict binding sites. One hot encoding was used as an encoding method. After the model was built, the ROC and PR curves were drawn to evaluate and interpret the generalization ability of the deep learning model. The performance was compared with the SVM (support vector machine) machine learning algorithm and CNN framework only. Additionally, the product of inputs and its corresponding gradient was calculated to estimate the contribution of each input feature.

3 Material and Methods

3.1 Collection of m6A sites and the target sites of m6A readers

In this experiment, the base-resolution and high-resolution m6A sites (DRACH motif) were downloaded from m6A -Atlas [31]. We mainly focused on six types of m6A readers recognized by Par-CLIP or iCLP, including YTHDC1, YTHDC2, YTHDF1, YTHDF2, YTHDF3, EIF3A (Table 1). For YTHDC1 and YTHDC2, the total sites are 16,664, 1,234 respectively, the number of genes is 4722, 275, the cell line used both from HeLa to HEK293T. Similarly, for three YTHDF proteins (YTHDF1-3), the total sites are 25,597, 28,970 and 7,253 respectively, the number of genes is 6,714, 6,677 and 3,495, the cell lines used covered from HeLa, Huh7, HEK293T. The remaining EIF3A was used in HEK293T cell line, the total sites and the gene number are 756 and 470 respectively.

Table 1 Target sites of m6A readers identified by Par-CLIP or iCLIP

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Reader | Source | Site | Total | Gene | Cell line |
| D1 | YTHDC1 | GSE74397 | 482 [32] | 16,664 | 4,722 | HeLa |
| D2 | GSE58352 | 2,633 [33] |
| D3 | GSE71096 | 2,430 [34] |
| D4 | GSE78030 | 12,309 [35] | HEK293T |
| D5 | YTHDC2 | GSE98085 | 1,183 [36] | 1,234 | 275 | HeLa |
| D6 | GSE78030 | 131 [35] | HEK293T |
| D7 | YTHDF1 | GSE63591 | 4,541 [25] | 25,597 | 6,714 | HeLa |
| D8 | GSE83438 | 2,527 [37] | Huh7 |
| D9 | GSE78030 | 20,694 [35] | HEK293T |
| D10 | YTHDF2 | GSE49339 | 22,688 [26] | 28,970 | 6,677 | HeLa |
| D11 | GSE83438 | 5,147 [37] | Huh7 |
| D12 | GSE78030 | 6,280 [35] | HEK293T |
| D13 | YTHDF3 | GSE86214 | 2,608 [38] | 7,253 | 3,495 | HeLa |
| D14 | GSE83438 | 177 [37] | Huh7 |
| D15 | GSE78030 | 5,082 [35] | HEK293T |
| D16 | EIF3A | GSE65004 | 45 [39] | 756 | 470 | HEK293T |
| D17 | GSE73405 | 731 [40] |

3.2 Identification of m6A reader binding sites

For the benefits of modeling, we need to define the positive and negative samples of m6A reader binding site. As can be seen from Figure 1, there are three significant factors to consider when determining the m6A reader binding sites, which is DRACH motif, known m6A Sites and CLIP labeled Sites. Generally, m6A readers have a tendency to bind known m6A sites and DRACH motif, but these two are not robust enough, without CLIP labeled sites, they could still be considered as negative samples. Therefore, in this experiment, three factors have to be satisfied simultaneously. In addition, the position of gene model would not influence the samples. To minimize the bias in selecting the polyA RNAs, we prepared the full transcript data and mature RNA data. In detail, mature RNA data exclude the sites on the intron region while the full transcript data covered either the exon or intron region.

**A picture containing timeline

Description automatically generated**

**Figure 1: Criteria of determining the m6A reader binding sites**

3.3 Deep learning model construction

Previously, we conducted a traditional machine learning about m6A reader by different encoding methods, however, the performance of one-hot method was not ideal, there was still a gap to improve [41]. In addition, one-hot encoding method learned from the convolutional neural network (CNN) is suitable for learning potential motifs in the bioinformatics field, therefore, we opted for deep learning techniques in this research. Furthermore, recurrent neural network (RNN) was incorporated since it could capture the information in sequence, for instance, the potential relationship between each nucleotide.

To build the deep learning model, we used Keras v2.3.0 and R v4.0.2 to conduct the learning part and process the raw data for prediction. For the data preparation part, we used R to extract n nucleotides (bp) of flanking sequences centered on the target adenosine, ranging from 251 to 2001bp to explore a suitable length. The processed sequence data were then inputted to Python3 for encoding, in this case, we chose One-hot encoding method for better model interpretability, for instance, A (1,0,0,0), C (0,1,0,0), G (0,0,1,0), T (0,0,0,1). The overall framework can be seen in **Figure 2**. Each sequence was then transformed to an n4 matrix and fed into two combinations of 1D convolution (Conv1D) layer and max-pooling layer. For the first combination, we set 90 kernels with size equaled 5 and applied L2 regulation to prevent overfitting. The rectified linear unit (ReLU) was used as the activation function to provide our necessary non-linearity. The following max-pooling layer was set in size equaled 4 with strides 2 to reduce the dimension of output from the previous layer. The dropout rate was incorporated to 0.25 to further reduce the possibility of overfitting. A second 1D convolution (Conv1D) layer with 100 filters and size equaled 3 to extract the feature of the previous data. Similarly, the ReLU function and L2 regulation were applied. However, the max-pooling size was 10 with 1 stride, under which circumstance could the model achieve higher performance.

The recurrent neural networks long short-term memory (LSTM) layer was used to aggregate the outputs of CNNs for predicting the RBP binding [42], in this case, the m6A readers’ substrates sites. LSTM processed sequentially of the sequence element, hoping to capture the inter-dependencies between motifs. Moreover, the fully connected layer with 1000 neurons would receive the output from the LSTM layer, and the non-linear activation function n, sigmoid, would calculate the prediction probability in each training class. The overall tuning process was used the loss function, binary cross-entropy to conduct the weight-tuning, optimizing the learning process, additionally, we found that Adam is the most suitable for this task. Finally, the output would be the probability of being m6A reader substrate sites.

A screenshot of a cell phone

Description automatically generated**Figure 2: Overall architecture of the deep learning model**

The sequence data are encoded by One-hot method and fed into the convolution layer and followed by the pooling layer twice to extract the significant features. The LSTM layer learns the long-term dependencies between sequence data generated by convolution layers. The flatten layer combines the previous kernels into a vector and inputs to the fully connected layer to calculate the probability of being m6A reader substrate site

3.4 Training strategy and performance evaluation

We separated each gene data set into three categories, training, testing, validation dataset, the ratio was 8:1:1 respectively. Moreover, to reduce the bias caused by imbalanced data samples, we ensured the same number of positive and negative samples in each category. The early stopping method was included to reduce the unnecessary computation during the learning process and the patience was designed as 10. The loss plot was drawn to document the training procedure and monitor the potential overfitting.

To validate the model performance, four commonly used performance metrics, including area under the ROC curve (AUC) [43], area under the Precision-Recall curve (PR-AUC) [44], accuracy (ACC) [45] and Mathew’s correlation coefficient (MCC) [46]. The formula of ACC and MCC are demonstrated as follows:

(1)

where TP and TN are denoted as True Positive and True Negative, FN and FT are denoted as False Negative and False Positive. To sum up, the higher the performance metrics value, the more accurate the prediction. Additionally, we compared the performance with the previous research using machine learning method, the combination of CNN + RNN frameworks and the CNN framework only to determine the optimal choice [41].

We also exploited DeepExplain's epsilon-LRP method (gradient-based) [47] to calculate the contribution in each feature input. With the assistance of this approach, we could rank the nucleotides’ significance in identifying the m6A readers’ substrates. Moreover, we extend the sequence upstream/downstream length from 50bp to 250bp, hoping to cover more information in determining each nucleotide contribution. 4 Results and discussion

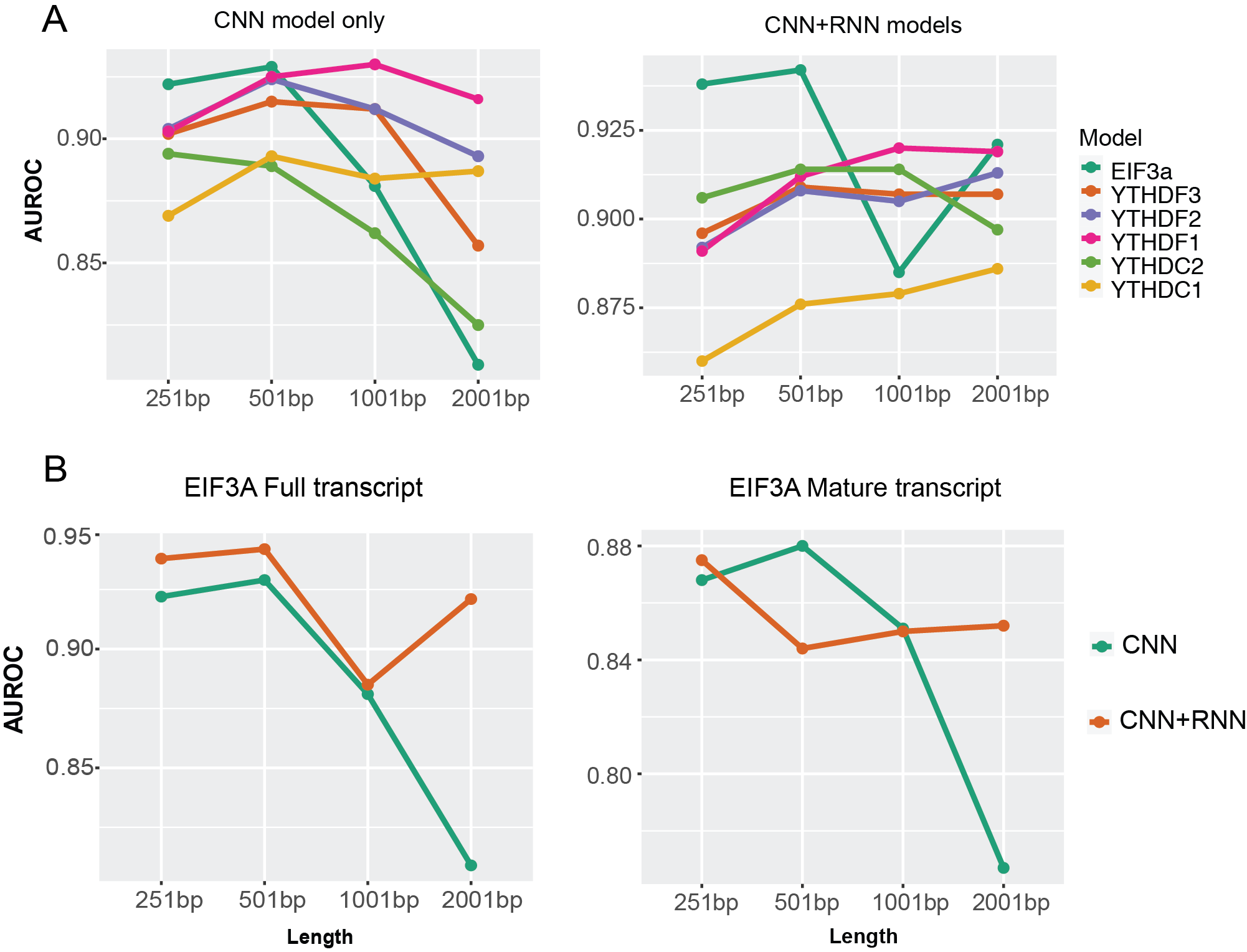
4. Results and discussion

4.1 Model comparison in different algorithms

To compare the performance of machine learning and deep learning models in predicting m6A reader, models with six different classifiers were trained: machine learning algorithm including SVM (Support Vector Machine), LR (Logistic Regression), RF (Random Forest), XGBoost and deep learning algorithm CNN (Convolutional Neural Network), the combination of CNN and RNN algorithm (Recurrent Neural Networks, specifically, Bidirectional LSTM Network), were built. Six mammalian m6A reader binding site datasets were used to train models, namely, EIF3a, YTHDF1-3, YTHDC1 and YTHDC2. For the traditional machine learning model, a commonly used transcript with size of 41bp was used and encoded by One-hot method. For the two deep learning models, full transcripts and mature transcripts with the length of 1001bp were chosen and encoded by one-hot as well. The performance metrics for EIF3A full transcripts by six models were summarized in **Table 2**. The model with CNN classifier achieved better prediction performance (AUROC = 0.929, PRAUC = 0.939) compared with the Random Forest (RF) algorithm, with AUROC of 0.818 and PRAUC of 0.825. The combination of CNN and RNN classifiers achieved the highest performance among the other models (AUROC = 0.942, PRAUC = 0.95). The AUROC for the leading machine learning model (RF) is respectively 11.94% and 13.16% lower than the CNN model and CNN + RNN model, which suggested that that deep learning algorithms are more powerful in characterizing m6A sites.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | AUROC | PRAUC | MCC | | ACC |
| LR | 0.789 | 0.785 | | 0.429 | 0.714 |
| XGBoost | 0.806 | 0.817 | | 0.517 | 0.758 |
| RF | 0.818 | 0.825 | | 0.504 | 0.751 |
| SVM | 0.811 | 0.816 | | 0.503 | 0.751 |
| CNN | 0.929 | 0.939 | | 0.736 | 0.868 |
| CNN+RNN | 0.942 | 0.95 | | 0.742 | 0.868 |

**Table 2:** Performance metrics of EIF3A full transcripts were chosen to demonstrate the AUROC, PRAUC, MCC and Accuracy for machine learning models, CNN and CNN+RNN models.

Classifiers might achieve varied performance on different datasets. To assess the fitness between the 6 reader binding site datasets and the two deep learning classifiers, models were built on full transcripts and their performance were analyzed. Similarly, different size of full transcripts was encoded with One-hot method. As shown in **Figure 3(A)**, models using CNN classifier achieved theoretically good performance with overall AUROC larger than 0.8. It seems that the CNN classifier fit the YTHDF1 binding datasets better than other reader binding sites, with overall AUROC exceeding 0.9 and highest AUROC of 0.93. In addition, CNN model achieves good performance with YTHDF2 binding datasets as well, with highest AUROC of 0.929. It is noticeable that the performance of CNN models with EIF3a varied dramatically along with the size of transcript, from 0.96 to 0.81, which suggests that the performance of CNN classifier is depend on the size of transcripts. Similar trends can be seen in the YTHDC2 datasets, the trained model with different input size achieves different AUROC score, with optimal input transcript size of 251bp (AUROC = 0.89).

Regarding the fitness of CNN+RNN classifier with the six reader datasets, models shows similar performance for YTHDF3, YTHDF2, YTHDF1, YTHDC2 datasets (with AUROC around 0.9). Moreover, the model trained with these four datasets as well as the YTHDC1 datasets (with AUROC around 0.875) seems transcript-size independent since lines are relatively stable. Interestingly, the performance of models trained with EIF3a datasets varied greatly from length to length (AUROC varied from 0.88 to 0.94). The structure variation between YTH family protein and EIF3a might contribute to the difference on model performance.

**Figure 3:** (A) Compared the performance of CNN model and CNN+RNN model in the prediction of six m6A reader substrates under different length in full transcripts. (B) Compared the AUROC value in either full transcript or mature transcript when predicting the EIF3A reader substrates.

To assess the feasibility of the two classifiers, namely CNN and RNN, performance of models was interpreted and compared. **Figure 3(B)** compares the performance of models with different size of EIF3a transcripts. As indicated, the combination of CNN and RNN classifier achieves overall better performance than the CNN classifier for both full transcript and mature transcript. Since the trend of line graph for CNN+RNN model is more stable than the line for CNN model, we can infer that the combination of CNN and RNN makes the model less dependent on the length of transcript used. Furthermore, the Receiver Operating Characteristic (ROC) curve and Precision-Recall curve for EIF3A datasets under the combination of CNN and RNN were visualized in **Figure 4.** As can be seen from the figure that, although the performance was different under different sequence size, the overall trend was stable and devoid of fluctuation. In addition, the overall performance in full transcripts could outperform the mature transcripts, probably the reason that full transcript data could cover either the exon or intron region.

Here, we mainly opted EIF3A dataset for easy demonstration in this paper. More details of the other five datasets could be achieved in the supplementary file.



**Figure 4:** (A) Compared the ROC curve and the regarding AUROC of mature and full transcript of EIF3A under CNN+RNN model in various sequence lengths. (B) Compared the PR curve and the regarding PRAUC of mature and full transcript of EIF3A under CNN+RNN model in various sequence lengths.

4.2 Layer-wise relevance calculation

Each input feature was calculated to obtain its contribution to the results by DeepExplain’s epsilon-LRP method. The feature importance plots were based on the EIF3a binding site datasets **(Figure 4).** The higher score that the position gets, the larger probability that the center nucleotide is an EIF3a reader binding site if this nucleotide present at that position. As shown in the graph, positions located around the predicted m6A sites got significantly higher scores than other positions, which means those positions are more important in determining whether the center nucleotide is m6A reader substrate site or not. Additionally, the prediction of modification site would benefit from taking sequence more than 50bp upstream or downstream the predicted site since they include positions with high importance score

Specifically, a site would be less likely to be m6A modification site if the adenosine represents in 100bp downstream since the majority of position within this sequence got importance scores smaller than 0. In comparison, the presence of cytosine in 50 upstream/downstream the predicted site tends to boost the chance of the center nucleotide being modified. No specific patterns were found for guanine and thymine as the importance plot present a shape like the sine function.

The results showed that if those positions 34bp, 59bp, 11bp, 58bp, 27bp, 49bp, 72bp upstream, 21bp, 27bp, 24bp, 25bp, 116bp downstream the modification site is cytosine, the site would more likely to be the EIF3a reader binding site. In addition, the probability of the modification site being EIF3a substrate site would decrease if guanosine was found on positions 21bp, 71bp, 33bp, 32bp, 31bp, 22bp upstream the center site or uridine was found on positions 54bp upstream or 53bp downstream the center site. The screened top 20 nucleotides that will decrease the change of the site being EIF3a modification site include: adenosines on positions 39bp, 27bp, 47bp, 61bp, 10bp, 12bp, 23bp, 170bp, 157bp, 51bp, 14bp, 226bp, 52bp upstream the center nucleotide, cytosine on positions 93bp upstream and 185bp downstream the center nucleotide, guanosine on positions 92, 97bp downstream the center site as well as uridines on positions 56bp, 63bp upstream the modification site.

A screenshot of a cell phone

Description automatically generated

**Figure 4** Feature importance scores in EIF3A full transcript prediction. We both extracted upstream/downstream 50bp and upstream/downstream 250 bp of the sequence to rank the contribution of each nucleotide in determining the binding site. In each position, the higher score it gains, the higher contribution towards the binding sites.

Conclusion

In this experiment, we developed a deep learning method combining both CNN and RNN frameworks to predict the m6A reader substrate sites from the sequence derived feature. Six m6A readers (YTHDC1-2, YTHDF1-3, EIF3A) were investigated and divided into mature and full transcripts parts to reduce the potential bias existing in polyA selection step during RNA library construction. We compared the performance with the traditional machine learning algorithm SVM and CNN framework only and identify the optimal sequence input length, which is 1001bp. Our model achieved outstanding performance, with average AUROC over 0.85, outperforming the other two frameworks. Feature importance scores were calculated and interpreted to demonstrate the impacts in determining the m6A reader substrates.

However, this study was restricted to the human genome only and could not extend to other important organisms, mainly owing to the scarce base-resolution epitranscriptome profiling data. In the future, with the existence of other species data, for instance, yeast, it would be promising to apply the model to predict the epitranscriptome target and functional characterization on other organisms. Additionally, more encoding methods could be incorporated to extract the important feature and combined features, such as genomic features, could improve the overall performance.

ACKNOWLEDGMENTS

This work has been supported by National Natural Science Foundation of China [31671373]; XJTLU Key Program Special Fund [KSF-T-01]

REFERENCES

[1] D. T. Do, T. Q. T. Le, and N. Q. K. Le. Using deep neural networks and biological subwords to detect protein S-sulfenylation sites. Brief Bioinform bbaa128 (Jul, 2020). DOI: https://doi.org/10.1093/bib/bbaa128

[2] S. Zaccara, R. J. Ries, and S. R. Jaffrey. Reading, writing and erasing mRNA methylation. Nature Reviews Molecular Cell Biology20, 10 (Sep, 2019), 608-624. DOI: https://doi.org/10.1038/s41580-019-0168-5

[3] L. Liu, B. Song, J. Ma, Y. Song, S.-Y. Zhang, Y. Tang, X. Wu, Z. Wei, K. Chen, J. Su, R. Rong, Z. Lu, J. P. de Magalhães, D. J. Rigden, L. Zhang, S.-W. Zhang, Y. Huang, X. Lei, H. Liu, and J. Meng. Bioinformatics approaches for deciphering the epitranscriptome: Recent progress and emerging topics. Computational and Structural Biotechnology Journal18 (Jun, 2020), 1587-1604. DOI: https://doi.org/10.1016/j.csbj.2020.06.010

[4] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Analytical Biochemistry 490 (Dec, 2015), 26-33. DOI: https://doi.org/10.1016/j.ab.2015.08.021

[5] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, and K.-C. Chou. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKN. Genomics 111, 1 (Jan, 2019), 96-102. DOI: https://doi.org/10.1016/j.ygeno.2018.01.005

[6] N. A.-O. Le. Fertility-GRU: Identifying Fertility-Related Proteins by Incorporating Deep-Gated Recurrent Units and Original Position-Specific Scoring Matrix Profiles. J Proteome Res 18, 9 (Sep, 2019), 3503-3511. DOI: https://pubs.acs.org/doi/10.1021/acs.jproteome.9b00411

[7] N. A.-O. Le. iN6-methylat (5-step): identifying DNA N(6)-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule. Mol Genet Genomics 294, 5 (Oct, 2019), 1173-1182. DOI: https://doi.org/10.1007/s00438-019-01570-y

[8] Q. A.-O. Zou, P. Xing, L. Wei, and B. Liu. Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. RNA 25,2 (Feb, 2019), 205-218. DOI: https://doi.org/10.1261/rna.069112.118

[9] Z. Chen, P. Zhao, F. Li, Y. Wang, A. I. Smith, G. I. Webb, T. Akutsu, A. Baggag, H. Bensmail, and J. Song. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief Bioinform 21, 5 (Sep, 2020), 1676-1696. DOI: https://doi.org/10.1093/bib/bbz112

[10] I. Nazari, M. Tahir, H. Tayara, and K. T. Chong. iN6-Methyl (5-step): Identifying RNA N6-methyladenosine sites using deep learning mode via Chou's 5-step rules and Chou's general PseKNC. Mol Genet Genomics 294, 5 (Oct, 2019), 1173-1182. DOI: https://doi.org/10.1007/s00438-019-01570-y

[11] S.-Y. Zhang, S.-W. Zhang, X.-N. Fan, J. Meng, Y. Chen, S.-J. Gao, and Y. Huang. Global analysis of N6-methyladenosine functions and its disease association using deep learning and network-based methods. PLoS Comput Biol 15, 1 (Jan, 2019), e1006663. DOI: https://doi.org/10.1371/journal.pcbi.1006663

[12] L. He, H. Li, A. Wu, Y. Peng, G. Shu, and G. Yin. 2019. Functions of N6- methyladenosine and its role in cancer. Molecular cancer 18, 1 (Dec, 2019), 176. DOI: https://doi.org/10.1186/s12943-019-1109-9

[13] R. M. Krug, M. A. Morgan, and A. J. Shatkin. 1976. Influenza viral mRNA contains internal N6-methyladenosine and 5'-terminal 7-methylguanosine in cap structures. Journal of virology 20, 1 (Oct, 1976), 45-53. Available at: https://jvi.asm.org/content/20/1/45.short

[14] K. Beemon, and J. Keith. 1977. Localization of N6-methyladenosine in the Rous sarcoma virus genome. Journal of molecular biology 113, 1 (Jun, 1977), 165-179. DOI: https://doi.org/10.1016/0022-2836(77)90047-X

[15] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, and M. Kupiec. 2012. Topology of the human and mouse m 6 A RNA methylomes revealed by m 6 A-seq. Nature 485, 7397 (May, 2012), 201-206. DOI: https://doi.org/10.1038/nature11112

[16] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey. 2012. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell 149, 7 (Jun, 2012), 1635-1646. DOI: https://doi.org/10.1016/j.cell.2012.05.003

[17] S. Liao, H. Sun, and C. Xu. 2018. YTH domain: A family of N6- methyladenosine (m6A) readers. Genomics, proteomics & bioinformatics 16, 2 (Apr, 2018), 99-107. DOI: https://doi.org/ 10.1016/j.gpb.2018.04.002

[18] Q. Liu, and R. I. Gregory. 2019. RNAmod: an integrated system for the annotation of mRNA modifications. Nucleic acids research 47, W1 (Jul, 2019), W548-W555. DOI: https://doi.org/ 10.1093/nar/gkz479

[19] N. Liu, K. I. Zhou, M. Parisien, Q. Dai, L. Diatchenko, and T. Pan. 2017. N6- methyladenosine alters RNA structure to regulate binding of a lowcomplexity protein. Nucleic acids research 45, 10 (Jun, 2017), 6051-6063. DOI: https://doi.org/10.1093/nar/gkx141

[20] S. Bowen, T. Yujiao, C. Kunqi, W. Zhen, R. Rong, L. Zhiliang, S. Jionglong, d. M. J. Pedro, D. J. Rigden, and M. Jia, m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human. Bioinformatics 36, 11 (Jun, 20020), 3528-3536 . DOI: https://doi.org/10.1093/bioinformatics/btaa178

[21] P. Stoilov, I. Rafalska, and S. Stamm. 2002. YTH: a new domain in nuclear proteins. Trends in biochemical sciences 27, 10 (Oct, 2002), 495-497. DOI: https://doi.org/10.1016/S0968-0004(02)02189-8

[22] B. S. Zhao, I. A. Roundtree, and C. He. 2017. Post-transcriptional gene regulation by mRNA modifications. Nature reviews Molecular cell biology 18, 1 (Jan, 2017), 31. DOI: https://doi.org/10.1038/nrm.2016.132

[23] Y. Xiang, B. Laurent, C.-H. Hsu, S. Nachtergaele, Z. Lu, W. Sheng, C. Xu, H. Chen, J. Ouyang, and S. Wang. 2017. RNA m 6 A methylation regulates the ultraviolet-induced DNA damage response. Nature 543, 7646 (Mar, 2017), 573-576. DOI: https://doi.org/10.1038/nature21671

[24] D. Jain, M. R. Puno, C. Meydan, N. Lailler, C. E. Mason, C. D. Lima, K. V. Anderson, and S. Keeney. ketu mutant mice uncover an essential meiotic function for the ancient RNA helicase YTHDC2. Elife 7 (Jan, 2018), e30919. DOI: https://doi.org/10.7554/eLife.30919

[25] X. Wang, B. S. Zhao, I. A. Roundtree, Z. Lu, D. Han, H. Ma, X. Weng, K. Chen, H. Shi, and C. He. N6-methyladenosine modulates messenger RNA translation efficiency. Cell 161, 6 (Jun, 2015), 1388-1399. DOI: https://doi.org/10.1016/j.cell.2015.05.014

[26] X. Wang, Z. Lu, A. Gomez, G. C. Hon, Y. Yue, D. Han, Y. Fu, M. Parisien, Q. Dai, and G. Jia. 2014. N 6-methyladenosine-dependent regulation of messenger RNA stability. Nature 505, 7481 (Jan, 2014), 117-120. DOI: https://doi.org/10.1038/nature12730

[27] S. D. Kasowitz, J. Ma, S. J. Anderson, N. A. Leu, Y. Xu, B. D. Gregory, R. M. Schultz, and P. J. Wang. 2018. Nuclear m6A reader YTHDC1 regulates alternative polyadenylation and splicing during mouse oocyte development. PLoS genetics 14, 5 (May, 2018), 1007412. DOI: https://doi.org/10.1371/journal.pgen.1007412

[28] M. J. Walker, M. D. Shortridge, D. D. Albin, L. Y. Cominsky, and G. Varani. 2020. Structure of the RNA Specialized Translation Initiation Element that Recruits eIF3 to the 5′-UTR of c-Jun. Journal of Molecular Biology (Jan, 2020). DOI: https://doi.org/10.1016/j.jmb.2020.01.001

[29] Y. Yin, J. Long, Y. Sun, H. Li, E. Jiang, C. Zeng, and W. Zhu. 2018. The function and clinical significance of eIF3 in cancer. Gene 673 (Oct, 2018), 130-133. DOI: https://doi.org/10.1016/j.gene.2018.06.034

[30] Y. Tang, K. Chen, X. Wu, Z. Wei, S. Y. Zhang, B. Song, S. W. Zhang, Y. Huang, and J. Meng, 2019. DRUM: Inference of Disease-Associated m6A RNA Methylation Sites From a Multi-Layer Heterogeneous Network. *Frontiers in Genetics* 10 (Apr, 2019). DOI: https://doi.org/10.3389/fgene.2019.00266

[31] Y. Tang, K. Chen, B. Song, J. Ma, X. Wu, Q. Xu, Z. Wei, J. Su, G. Liu, and R. Rong. 2020. m6A-Atlas: a comprehensive knowledgebase for unraveling the N6- methyladenosine (m6A) epitranscriptome. Nucleic Acids Research (Aug, 2020). DOI: https://doi.org/10.1093/nar/gkaa692

[32] I. A. Roundtree, G.-Z. Luo, Z. Zhang, X. Wang, T. Zhou, Y. Cui, J. Sha, X. Huang, L. Guerrero, and P. Xie. 2017. YTHDC1 mediates nuclear export of N6- methyladenosine methylated mRNAs. Elife 6 (Oct, 2017), e31311. DOI: https://doi.org/10.7554/eLife.31311

[33] C. Xu, X. Wang, K. Liu, I. A. Roundtree, W. Tempel, Y. Li, Z. Lu, C. He, and J. Min. 2014. Structural basis for selective binding of m 6 A RNA by the YTHDC1 YTH domain. Nature chemical biology 10, 11 (Nov, 2014), 927-929. DOI: https://doi.org/10.1038/nchembio.1654

[34] W. Xiao, S. Adhikari, U. Dahal, Y.-S. Chen, Y.-J. Hao, B.-F. Sun, H.-Y. Sun, A. Li, X.-L. Ping, and W.-Y. Lai. 2016. Nuclear m6A reader YTHDC1 regulates mRNA splicing. Molecular cell 61, 4 (Feb, 2016), 507-519. DOI: https://doi.org/10.1016/j.molcel.2016.01.012

[35] D. P. Patil, C.-K. Chen, B. F. Pickering, A. Chow, C. Jackson, M. Guttman, and S. R. Jaffrey, “m 6 A RNA methylation promotes XIST-mediated transcriptional repression,” Nature 537, 7620 (Sep, 2016), 369-373. DOI: https://doi.org/10.1038/nature19342

[36] P. J. Hsu, Y. Zhu, H. Ma, Y. Guo, X. Shi, Y. Liu, M. Qi, Z. Lu, H. Shi, and J. Wang. 2017. Ythdc2 is an N 6-methyladenosine binding protein that regulates mammalian spermatogenesis. Cell research 27, 9 (Sep, 2017), 1115-1127. DOI: https://doi.org/10.1038/cr.2017.99

[37] N. S. Gokhale, A. B. McIntyre, M. J. McFadden, A. E. Roder, E. M. Kennedy, J. A. Gandara, S. E. Hopcraft, K. M. Quicke, C. Vazquez, and J. Willer. 2016. N6- methyladenosine in Flaviviridae viral RNA genomes regulates infection. Cell host & microbe 20, 5 (Nov, 2016), 654-665. DOI: https://doi.org/10.1016/j.chom.2016.09.015

[38] H. Shi, X. Wang, Z. Lu, B. S. Zhao, H. Ma, P. J. Hsu, C. Liu, and C. He. 2017. YTHDF3 facilitates translation and decay of N 6-methyladenosine-modified RNA. Cell research 27, 3 (Mar, 2017), 315-328. DOI: https://doi.org/10.1038/cr.2017.15

[39] A. S. Lee, P. J. Kranzusch, and J. H. Cate. 2015. eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. Nature 522, 7554 (Jun, 2015), 111-114. DOI: https://doi.org/10.1038/nature14267

[40] K. D. Meyer, D. P. Patil, J. Zhou, A. Zinoviev, M. A. Skabkin, O. Elemento, T. V. Pestova, S.-B. Qian, and S. R. Jaffrey. 2015. 5′ UTR m6A promotes capindependent translation. Cell 163, 4 (Nov, 2015), 999-1010. DOI: https://doi.org/10.1016/j.cell.2015.10.012

[41] D. Zhen, Y. Wu, Y. Zhang, K. Chen, B. Song, H. Xu, Y. Tang, Z. Wei, and J. Meng. m6A Reader: Epitranscriptome Target Prediction and Functional Characterization of N6-Methyladenosine (m6A) Readers. Frontiers in Cell and Developmental Biology 8 (Aug, 2020), 741. DOI: https://doi.org/10.3389/fcell.2020.00741

[42] X. Pan, P. Rijnbeek, J. Yan, and H.-B. Shen. 2018. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC genomics 19, 1 (Dec, 2018), 511. DOI: https://doi.org/10.1186/s12864-018-4889-1

[43] A. P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition 30, 7 (1997), 1145-1159. DOI: https://doi.org/10.1016/S0031-3203(96)00142-2

[44] J. Keilwagen, I. Grosse, and J. Grau. 2014. Area under precision-recall curves for weighted and unweighted data. PloS one 9, 3 (Mar, 2014), 92209. DOI: https://doi.org/10.1371/journal.pone.0092209

[45] J. Huang, and C. X. Ling. 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering 17, 3 (Jan, 2005), 299-310. DOI: https://doi.org/10.1186/s12864-019- 6413-7

[46] D. Chicco, and G. Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics 21, 1 (Dec, 2020), 6. DOI: https://doi.org/10.1186/s12864-019-6413-7

[47] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, “Towards better understanding of gradient-based attribution methods for deep neural networks,” *arXiv preprint arXiv:1711.06104*, 2017.