Deep-2′-O-Me: Predicting 2′-O-methylation sites by Convolutional Neural Networks

Milad Mostavi, Sirajul Salekin and Yufei Huang

***Abstract*— 2′-O-methylation (2′-O-me) of ribose moiety is one of the significant and ubiquitous post-transcriptional RNA modifications which is vital for metabolism and functions of RNA. Although recent development of new technology (Nm-seq) enabled biologists to find precise location of 2′-O-me in RNA sequences, there is still a lack of computational tools that can also provide high resolution prediction of this RNA modification. In this paper, we propose a deep learning based method that takes advantage of an embedding method to learn complex feature representation of pre-mRNA sequences and employs a Convolutional Neural Network to fine-tune the features required for accurate prediction of such alteration. Specifically, we adopted dna2vec, a biological sequence embedding method originally inspired by the word2vec model of text analysis, to yield embedded representation of sequences that may or may not contain 2-O-me sites before feeding those features into CNN for classification. Our model was trained using the data collected from Nm-seq experiment. The proposed method achieved AUC and auPRC scores of 90% outperforming existing state-of-the-art algorithms by a significant margin in both balanced and unbalanced class testing scenarios.**

I. Introduction

Cellular RNA has more than 150 chemical modifications, known as epitranscriptomic modification that plays active roles in the functions of RNA and influences the mechanism of gene expression [1]. 2′-O-Me is one of the prevalent and ubiquitous post-transcriptional modifications, which was discovered in tRNA, rRNA, snRNAs, mRNAs, and miRNAs [2].

In past few years, several biological experiments are being developed to identify precise location of 2′-O-Me in different species [2]. Recently, Nm-seq experiment that exploits the differential reactivity of 2′-O-methylated and 2′-hydroxylated nucleosides toward periodate oxidation, revealed thousands of 2′-O-Me sites with base precision in mammalian mRNA [3]. The emergence of large scale genomic sequence datasets has enabled us to computationally predict a wide variety of post-transcriptional modifications. However, computational prediction of 2-O-me has largely been ignored by the research community primarily due to the lack of high precision training data. The widespread identification of 2-O-me sites by Nm-seq technology is likely to infuse an influx of research in this domain. To the best of our knowledge, [4] is the only paper that has released a computational method to solve 2′-O-Me site prediction problem. The method utilizes support vector machine (SVM) that takes nucleotide composition and chemical properties of 41 bp sequence surrounding the region of a 2′-O-Me site as input and outputs probability of whether the input sequence is 2′-O-Me site or not. The nucleotide and compositional properties of sequence were converted to one hot encoding producing an input of dimension 41*4 and make a binary prediction with proposed SVM. Although the reported accuracy of this method is 95%, the algorithm suffers from several pitfalls.

First of all, this method is highly biased towards the dataset that it has been trained. They filter out positive samples and reach to 147 positive 2′-O-Me sites to avoid overfitting problem in training of SVM while the real number of positive 2′-O-Me sites is around 4500 [3]. Furthermore, they consider the training and testing case that the dataset is balanced. In other words, the number of non-2′-O-Me sites or negative samples is 147 which yields the ratio of 1:1 between positive and negative samples; however, in reality such assumption is not true. Due to this fact, we observed a huge number of False Positives while testing the new dataset with their algorithm.

In this paper, we are motivated to facilitate the aforementioned disadvantages regarding previous method by providing a fully automated computational method to predict the exact location of 2′-O-Me based on the available surrounding sequence code context of 2′-O-Me.

The first section gives a brief overview of the dataset which is used. The second section explains how RNA sequences containing 2′-O-Me and non-2′-O-Me sites are fed into a recently developed embedding method to produce unique matrices for proposed CNN. The third section illustrates specifications of designed CNN followed by its implementation and results in the fourth section. Finally we will summarize our findings and give some directions for future works in the last section.

II. Dataset

2′-O-Me resides on the 2′ hydroxyl ribose moiety of all four ribonucleosides [5], namely 2′-O methylcytidine(Cm), 2′-O-methyladenosine(Am), 2′-O methylguanosine(Gm), 2′-O-methyluridine (Um) as it is shown in Figure *1*. Recently developed method in [3] revealed 2802 new 2′-O-Me sites. This dataset was downloaded and aligned with BedTools in hg38 in order to extract surrounding sequences around each

Milad Mostavi, Sirajul Salekin and Yufei Huang are with the Electrical and Computer Engineering Department, University of Texas at San Antonio, 1 UTSA Circle, San Antonio, TX 78249, USA. (e-mail: yufei.huang@utsa.edu)

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

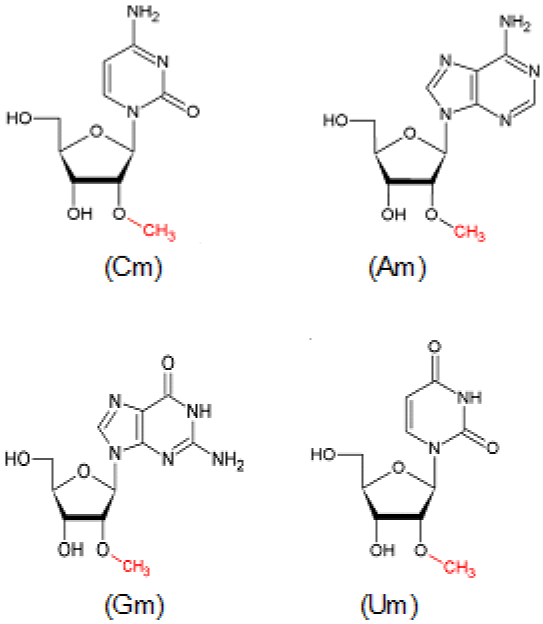


Figure 1: Structure of all four chemical modifications in 2′-O-methylation

2′-O-Me site. We picked 25 bp (it is a random number) nucleotide window size from left and right of 2′-O-Me sites as input sequences with positive label. In order to create non-2′-O-Me sites or negative samples, we randomly searched the surrounding area of positive sites. If there is not positive 2′-O-Me, we capture that site with its surrounding nuclides as negative sample. Since in reality the number of negative samples is bigger than positive samples, we collected two times more negative samples (i.e. 11200) corresponding to positive samples. As a matter of fact, the computational approach that we are going to apply needs to be able to cope with such unbalanced dataset.

III. RNA EMBEDDING PROCEDURE

Natural biological information in DNA, RNA and protein sequences needs to be fed into the computational methods where they take such representations as input. There are several different embedding approaches that calculate the distribution and relation of each individual nucleotides within its context and emit one vector of numbers as the representation of that sequence fragment. This paper utilizes recently developed dna2vec embedding method in [6] to encode 51 bp nucleotide sequences containing 2′-O-Me and non-2′-O-Me sites.

dna2vec method is developed based on word2vec [7] method which is a two layer neural network model. It is noteworthy to mention that dna2vec is able to omit the curse of dimensionality which exist in one-hot encoding method and gives arithmetic vectors that are similar to nucleotide concatenation[6].

We trained the proposed dna2vec model based on whole RNA and created our own pre-trained rna2vec model. By doing so, we were able to feed 2′-O-Me and non-2′-O-Me sequences into this pre-trained model and embed each sequence to specific vectors. We fed every three nucleotides in the sequences at a time to our rna2vec model and created vectors with 100 length for each three letters (i.e. every three letters produces one vector with length 100). Furthermore, we stacked all generated 17 yielded vectors vertically to produce matrices with 17*100 dimension for each sequence. Figure 2 depicts the architecture of model with graphical symbols. Finally, the generated matrices were normalized between -1 and 1 to enhance the power of prediction in the next step.

IV. CONVOLUTIONAL NEURAL NETWORK

Deep Learning (DL) is a set of computational models that take advantage of multiple processing layers to learn representations of data through each layer [8]. The intricate structure in data is learned with associated weights in each layer which are computed by backpropagation method [8]. These models have reached best performance results in several areas such as Natural Language Processing and Computer Vision.

DL was adopted to genomics by Deepbind and DeepSEA in early 2015 by improving the sate-of-art[9, 10].These pioneer works converted genomic sequences into 2D images and harnessed existing CNN models to accomplish either binary or multi-class classification tasks. Followed by the striking results yielded from CNN, there are several recent in-depth development and analysis of both computational DL models and applications in genomics [11-14].

CNN takes inputs as multiple arrays such as color 2D images and pass it through multiple modules to extract spatial connectivity between pixels. CNN slides different kernels, which are usually learnable multi-arrays, to find feature representation of data. In order to normalize these values, the calculated feature maps are passed through activation functions. Furthermore, the extracted feature maps are usually fed into fully connected layers to find combination of nodes in the output.

As it is shown in Figure 2, the embedded 17*100 matrices for each sequence from previous section are fed into our designed CNN. CNN operation is done by 32 kernels with 3*100 dimension which are initialized randomly and slide one cell at each time. Moreover, the values of each kernel are passed through batch normalization and LeakyRelu activation function modules for normalization purpose[15]. Finally, these feature maps are flattened and connected to two nodes in dense layer with tanh activation function for 2′-O-Me and non-2′-O-Me classification task.

V. RESULTS AND DISCUSSIONS

In this section, we will show and compare the results of Deep-2′-O-Me based on the existing standard evaluation metrics with the previous method.

A. Performance evaluation

The performance of Deep-2′-O-Me was evaluated based on several different metrics, namely accuracy (Acc), area under Precision-Recall Curve (auPRC) which is the correlation between Precision and Recall values with different thresholds, and area under Receiver Operating Characteristic (ROC) curve or Area Under Curve (AUC) scores.

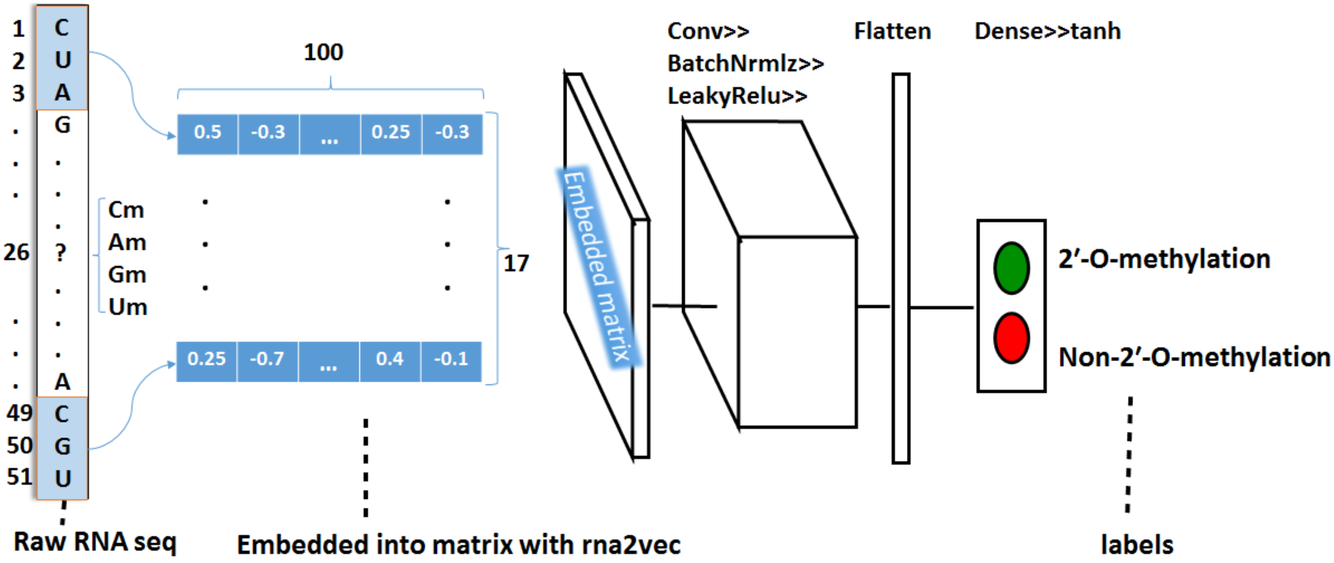


Figure 2: Architecture of Deep-2′-O-Me from raw RNA sequence to the predicted output

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}$$
$$Precision = \frac{TP}{TP + FP} \qquad (1)$$
$$Recall = \frac{TP}{TP + FN}$$

Where TP, TN, FP, and FN are variables for True Positive, True Negative, False Positive and False Negative, respectively [14].

B. Implementation

The entire implementation of Deep-2′-O-Me and its hyper parameters such as the number and size of kernels were chosen by automatic hyper parameter tuning provided in Keras DL platform namely hyperas. We opted Adam with its default parameters and mean square error as our optimization and loss functions respectively. In every iteration 64 samples were fed into model up to 20 epochs. We used early-stopping command inside the callback function provided in Keras platform with verbose equal to three on validation loss to avoid overfitting. The model stopped after 14 epochs by early-stopping command.

Deep-2′-O-Me was trained and tested in several different cases with balanced and unbalanced datasets. We held out 12 percent of total positive samples as fixed test samples and added same portion of negative samples in each iteration for unbalanced testing scenarios. We ran the code in four different cases, where the ratio of positive to negative samples varies from 1:1 to 1:4, respectively. After separating test samples, rest of the samples were fed into the CNN model for training purpose. During the training session, we used 10 fold cross validation, and kept 10 percent of the total samples for validation during training. Both training and testing sample pools were shuffled to avoid any bias decision making inside the model.

C. Results

The motive behind developing a new computational method in this paper was to reduce the number of False Positive rate in prediction evaluation of previous method[4]. We uploaded the sequences of recently published dataset in [3] containing new positive 2′-O-Me sites with total length of 41 bp (as it is needed in their algorithm) in the webserver that was provided by the previous method, and it resulted in high accuracy in predicting positive samples. On the other hand, we uploaded randomly created sequences as negative samples that don’t contain 2′-O-Me sites, and interestingly enough, we realized that the model is predicting negative samples as positive ones (i.e. producing False Positive prediction) with probability close to one (e.g. 0.99xx). Since the main reported performance evaluation criteria was accuracy, we use the same index for comparison. Finally we measure our method prediction strength based on auPRC and AUC scores which are more precise measurements while dealing with unbalanced dataset.

Table *1* shows the comparison results based on accuracy between Deep-2′-O-Me and previous method. As it is clear, the more negative samples were added to the testing pool, the more drop rate in accuracy performance index of previous method was achieved. Furthermore, Table 2 shows the performance of Deep-2′-O-Me method based on auPRC and AUC scores in four different testing cases. As it is clear, Deep-2′-O-Me produces all stable and robust prediction results across balanced and unbalanced cases.

Table 1: Accuracy of Deep-2′-O-Me and previous method with different ratios of positive and negative test samples

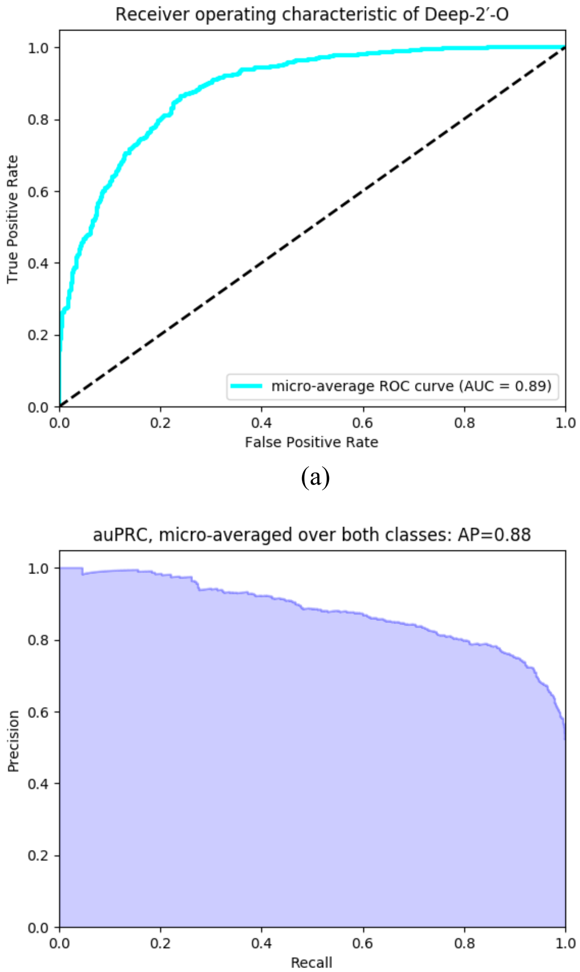
Method\Ratio	1:1	1:2	1:3	1:4
[3]	50.1	33.4	25.2	12.7
Deep-2′-O-Me	81.91	83.94	83.58	85.36

Table 2: AUC and auPRC scores of Deep-2′-O-Me with different ratios of positive and negative test samples

Index\Ratio	1:1	1:2	1:3	1:4
AUC	0.89	0.90	0.91	0.92
auPRC	0.88	0.90	0.91	0.93

As Table 2 shows the more negative samples were added to the test pool, the better performance results were achieved. This fact illustrates the fact that this method is able to predict negative samples better than positive samples which is in reality, it is a more realistic model.

Figure 3 (a) and (b) show ROC and Precision-Recall curve for balanced testing case with their auPRC and AUC scores, respectively.



(a)

(b)

Figure 3: Measuring prediction power of Deep-2′-O-Me in balanced data case with (a) AUC and (b) area under Precision-Recall Curve

VI. Conclusion

This paper presented a new and precise method for predicting 2′-O-Me sites from raw RNA sequences via a novel sequence embedding method and CNN model. This method was tested in balance and unbalance cases and showed stable and robust by both auPRC and AUC scores around 90 percent. The associated impact of 2′-O-Me sites with other RNA modifications can be considered as future work.

Acknowledgement

This work was supported by the National Institute of Health (R01GM113245).

References

1. Peer, E., G. Rechavi, and D. Dominissini, *Epitranscriptomics: regulation of mRNA metabolism through modifications.* Curr Opin Chem Biol, 2017. **41**: p. 93-98.
2. Incarnato, D., et al., *High-throughput single-base resolution mapping of RNA 2′-O-methylated residues.* Nucleic acids research, 2016. **45**(3): p. 1433-1441.
3. Dai, Q., et al., *Nm-seq maps 2'-O-methylation sites in human mRNA with base precision.* Nature methods, 2017. **14**(7): p. 695.
4. Chen, W., et al., *Identifying 2'-O-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions.* Genomics, 2016. **107**(6): p. 255-258.
5. Xuan, J.-J., et al., *RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data.* Nucleic Acids Research, 2018. **46**(D1): p. D327-D334.
6. Ng, P., *dna2vec: Consistent vector representations of variable-length k-mers.* arXiv preprint arXiv:1701.06279, 2017.
7. Mikolov, T., et al. *Distributed representations of words and phrases and their compositionality*. in *Advances in neural information processing systems*. 2013.
8. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436.
9. Alipanahi, B., et al., *Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning.* Nature biotechnology, 2015. **33**(8): p. 831.
10. Zhou, J. and O.G. Troyanskaya, *Predicting effects of noncoding variants with deep learning–based sequence model.* Nature methods, 2015. **12**(10): p. 931.
11. Salekin, S., et al. *Early disease correlated protein detection using early response index (eri)*. in *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*. 2016. IEEE.
12. Salekin, S., et al., *Early response index: a statistic to discover potential early stage disease biomarkers.* BMC bioinformatics, 2017. **18**(1): p. 313.
13. Salekin, S., J.M. Zhang, and Y. Huang. *A deep learning model for predicting transcription factor binding location at single nucleotide resolution*. in *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*. 2017. IEEE.
14. Salekin, S., J.M. Zhang, and Y. Huang, *Base-pair resolution detection of transcription factor binding site by deep deconvolutional network.* bioRxiv, 2018: p. 254508.
15. Ioffe, S. and C. Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. in *International conference on machine learning*. 2015.