



## MethyRNA: A web-server for identification of N<sup>6</sup>-methyladenosine sites

Wei Chen, Hua Tang & Hao Lin

To cite this article: Wei Chen, Hua Tang & Hao Lin (2016): MethyRNA: A web-server for identification of N<sup>6</sup>-methyladenosine sites, Journal of Biomolecular Structure and Dynamics, DOI: [10.1080/07391102.2016.1157761](https://doi.org/10.1080/07391102.2016.1157761)

To link to this article: <http://dx.doi.org/10.1080/07391102.2016.1157761>



Accepted author version posted online: 25 Feb 2016.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

**Publisher:** Taylor & Francis

**Journal:** *Journal of Biomolecular Structure and Dynamics*

**DOI:** <http://dx.doi.org/10.1080/07391102.2016.1157761>

**Letter to the Editor: A web-server**

**MethyRNA: A web-server for identification of N<sup>6</sup>-methyladenosine sites**

**Wei Chen<sup>1\*</sup>, Hua Tang<sup>2</sup>, Hao Lin<sup>3\*</sup>**

<sup>1</sup> Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China;

<sup>2</sup> Department of Pathophysiology, Sichuan Medical University, Luzhou 646000, China;

<sup>3</sup> Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China;

**\*Corresponding authors**

WC: [chenweimu@gmail.com](mailto:chenweimu@gmail.com)

HL: [hlin@uestc.edu.cn](mailto:hlin@uestc.edu.cn)

Tel: +86-315-3725715; Fax: +86-315-3725715

## 1. Introduction

N<sup>6</sup>-methyladenosine (m<sup>6</sup>A) is the most abundant post transcriptional modification and has been found in the three domains of life (Cantara et al. 2011). m<sup>6</sup>A plays fundamental regulatory roles in a series of biological process, such as protein translation and localization (Meyer and Jaffrey 2014), mRNA splicing and stability (Nilsen 2014), and stem cell pluripotency (Chen et al. 2015a). Therefore, accurately identifying m<sup>6</sup>A site in RNA will help expand our understanding of its potential roles.

Recently, using high-throughput sequencing techniques, m<sup>6</sup>A data is available for *Saccharomyces cerevisiae* (Schwartz et al. 2013), *Homo sapiens* (*H. sapiens*) and *Mus musculus* (*M. musculus*) (Dominissini et al. 2012). Since these methods are costly and time consuming in performing genome-wide analysis, with the increasing number of sequenced genomes, it is necessary to develop computational methods for timely identifying m<sup>6</sup>A sites. However, to our best knowledge, the existing computational tools for the detection of m<sup>6</sup>A sites are only applicable for *S. cerevisiae* (Chen et al. 2015b; Chen et al. 2015c). Therefore, there is an urgent need to develop new automated methods for m<sup>6</sup>A site identification.

Based on the high-resolution experimental data of *H. sapiens* and *M. musculus*, in the present study, a support vector machine (SVM) based model was proposed to identify m<sup>6</sup>A sites by encoding RNA sequence using nucleotide chemical property and frequency. Results from the jackknife test show that the proposed model could accurately identify m<sup>6</sup>A sites in *H. sapiens* and *M. musculus*. A web-server for the proposed model, called **MethyRNA**, is provided, which is freely available at <http://lin.uestc.edu.cn/server/methyrna>.

## 2. Materials and Methods

### 2.1 Dataset

Using MeRIP-Seq and m<sup>6</sup>A-seq, m<sup>6</sup>A sites have been identified in *S. cerevisiae*, *H. sapiens* and *M. musculus* (Dominissini et al. 2012; Schwartz et al. 2013). These experimentally annotated m<sup>6</sup>A sites have been checked and deposited in the RMBase (Sun et al. 2015). Therefore, from RMBase, we obtained 94,895 and 28,002 m<sup>6</sup>A site containing sequences in *H. sapiens* and *M. musculus*, respectively. All of these sequences are 41-nt long with the m<sup>6</sup>A site in the center. To overcome redundancy and reduce the homology bias, sequences with more than 60% sequence similarity were removed by using the CD-HIT program (Fu et al. 2012). After such a screening

procedure, we obtained 1,130 and 725 m<sup>6</sup>A site containing sequences and deemed them as the positive samples for *H. sapiens* and *M. musculus*, respectively.

Considering the m<sup>6</sup>A site in *H. sapiens* and *M. musculus* harboring the consensus motif RRACU (Dominissini et al. 2012), the negative samples were obtained by choosing adenines from the 41-nt long segments which are centered around the RRACU consensus motif in both *H. sapiens* and *M. musculus*, respectively. By doing so, we harvested a great number of negative samples. Therefore, the size of negative dataset is dramatically greater than that of positive dataset. In machine-learning problems, imbalanced datasets can affect the accuracy of learning models. To balance out the numbers between positive and negative samples in model training, 1,130 and 725 adenines containing sequences were randomly picked out to form the negative samples for *H. sapiens* and *M. musculus*, respectively. These negative samples were also 41-nt long and with the sequence similarity less than 60%. Finally, we obtained two benchmark datasets as formulated by

$$\mathbb{S}_k = \mathbb{S}_k^+ \cup \mathbb{S}_k^-, k = \begin{cases} 1 & \text{for } H. sapiens \\ 2 & \text{for } M. musculus \end{cases} \quad (1)$$

where the positive dataset  $\mathbb{S}_1^+$  contains 1,130 true m<sup>6</sup>A site containing sequences while the negative dataset  $\mathbb{S}_1^-$  contains 1,130 false m<sup>6</sup>A site containing sequences;  $\mathbb{S}_2^+$  contains 725 true m<sup>6</sup>A site containing sequences while the negative dataset  $\mathbb{S}_2^-$  contains 725 false m<sup>6</sup>A site containing sequences; and the symbol  $\cup$  means the union in the set theory. All of the positive and negative samples in the benchmark dataset are available at <http://lin.uestc.edu.cn/server/Methy/data>.

## 2.2 Support vector machine

Support vector machine (SVM) is a machine learning algorithm and has been successfully used in the realm of bioinformatics (Chen et al. 2013; Chen et al. 2014b; Lin et al. 2014; Liu et al. 2014; Lin et al. 2015; Liu et al. 2015b; Zou Q et al. 2015; Liu et al. 2016; Liu B 2016). The basic idea of SVM is to transform the input data into a high dimensional feature space and then determine the optimal separating hyperplane. In the current study, the LibSVM package 3.18 (Chang and Lin 2011) was used to implement SVM, which can be freely downloaded from <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>. Because of its effectiveness and speed in training process, the radial basis kernel function (RBF) was used to obtain the best classification hyperplane in the current study. In the SVM operation engine, the grid

search method was applied to optimize the regularization parameter  $C$  and kernel parameter  $\gamma$  using a grid search approach in the range  $2^{-5} \leq C \leq 2^{15}$  with step of 2 and  $2^{-15} \leq \gamma \leq 2^{-5}$  with step of  $2^{-1}$ , respectively.

### 2.3 Chemical property

There are four kinds of nucleotides found in RNA, namely, adenine (A), guanine (G), cytosine (C) and uracil (U). Since each nucleotide has different chemical structures and chemical binding, they can be classified into three different groups in terms of the chemical properties (Golam Bari et al. 2014). Adenine and guanine have two rings, while cytosine and uracil have only one ring. When forming secondary structures, guanine and cytosine have strong hydrogen bonds, whereas adenine and uracil have weak hydrogen bonds. In terms of chemical functionality, adenine and cytosine can be classified into the same group, called amino group, while guanine and uracil into the keto group. Hence, each nucleotide  $s_i=(x_i, y_i, z_i)$  in the sequence can be encoded by the following formula (Golam Bari et al. 2014).

$$x_i = \begin{cases} 1 & \text{if } s_i \in \{A, G\} \\ 0 & \text{if } s_i \in \{C, U\} \end{cases}, y_i = \begin{cases} 1 & \text{if } s_i \in \{A, C\} \\ 0 & \text{if } s_i \in \{G, U\} \end{cases}, z_i = \begin{cases} 1 & \text{if } s_i \in \{A, U\} \\ 0 & \text{if } s_i \in \{C, G\} \end{cases} \quad (2)$$

Thus, A can be represented by coordinates (1, 1, 1), C can be represented by coordinates (0, 1, 0), G can be represented by coordinates (1, 0, 0), U can be represented by coordinates (0, 0, 1).

### 2.4 Nucleotide frequency

In order to include the nucleotide frequency and nucleotide distribution around  $m^6A$  site, the density  $d_i$  of any nucleotide  $n_j$  at position  $i$  in RNA sequence was defined by the following formula.

$$d_i = \frac{1}{|N_i|} \sum_{j=1}^l f(n_j), f(n_j) = \begin{cases} 1 & \text{if } n_j = q \\ 0 & \text{other cases} \end{cases} \quad (3)$$

where  $l$  is the sequence length,  $|N_i|$  is the length of the  $i$ -th prefix string  $\{n_1, n_2, \dots, n_i\}$  in the sequence,  $q \in \{A, C, G, U\}$ . Suppose an example sequence “GUACCUGAUG”.

The density of ‘A’ is 0.33 (1/3), 0.25 (2/8) at positions 3 and 8, respectively. The density of ‘C’ is 0.25 (1/4), 0.4 (2/5) at positions 4 and 5, respectively. The density of ‘G’ is 1 (1/1), 0.29 (2/7), 0.30 (3/10) at positions 1, 7 and 10, respectively. The density of ‘U’ is 0.5 (1/2), 0.33 (2/6), 0.33 (3/9) at positions 2, 6 and 9, respectively.

By integrating both the nucleotide chemical property and accumulated nucleotide information, the sample sequence “GUACCUGAUG” can be encoded by the

following vector  $\{(1, 0, 0,1), (0, 0, 1, 0.5), (1, 1, 1, 0.33), (0, 1, 0, 0.25), (0, 1, 0,0.4), (0, 0, 1, 0.33), (1, 0, 0, 0.29) , (1, 1, 1, 0.25), (0, 0, 1, 0.33), (1, 0, 0, 0.30)\}$ . Both the chemical property and the long range sequence order information were incorporated in the vector.

## 2.5 Performance evaluation

As done in previous works (Lin et al. 2013a; Lin et al. 2013b; Chen et al. 2014c; Feng PM 2014; Wei et al. 2014; Chen et al. 2015b; Chen et al. 2015c; Chen et al. 2016), the performance of **MethyRNA** was also evaluated by using the following three metrics, namely sensitivity (Sn), specificity (Sp) and Accuracy (Acc), which are expressed as

$$\begin{cases} Sn = \frac{TP}{TP + FN} \times 100\% \\ Sp = \frac{TN}{TN + FP} \times 100\% \\ Acc = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \end{cases} \quad (4)$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

## 3. Results and Discussions

### 3.1 Cross-validation

Since the jackknife test is deemed as the least arbitrary and most objective cross-validation methods (Chou 2011), it has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors (Feng et al. 2013; Mohabatkar et al. 2013; Chen et al. 2014a; Feng et al. 2014; Guo et al. 2014; Liu et al. 2015a). Hence, the jackknife test was used to examine the performance of the proposed model. In the jackknife test, each sample in the training dataset is in turn singled out as an independent test sample and all the properties are calculated without including the one being identified.

Preliminary trials indicated that when the length of the sequences in the benchmark dataset is 41 bp with the m<sup>6</sup>A in the center, the corresponding predictive results were most promising. Therefore, each sample in the benchmark dataset was encoded by a 164 (4×41)-dimensional vector (see **Materials and Methods**) and was used as the input vector of SVM. In the jackknife test, the proposed model accurately

identified the m<sup>6</sup>A sites in *H. sapiens* and *M. musculus* with the accuracies of 90.38% and 88.39%, respectively (**Table 1**).

### 3.2 Comparison with other methods

To further demonstrate the power of the proposed method, we also compared its predictive accuracies with that of **iMethyl-RNA** (Chen et al. 2015b). Accordingly, we encoded the sequences in the benchmark dataset according to the rules of **iMethyl-RNA** and carried out the jackknife test on the benchmark dataset used in the current work. The predictive results, namely sensitivity, specificity and accuracy were also reported in **Table 1**, from which we found that the accuracies obtained by **iMethyl-RNA** are approximately 23% lower than our proposed method for identifying m<sup>6</sup>A sites in *H. sapiens* and *M. musculus*. These results indicate that the model proposed in this work is quite promising and may become a useful tool in identifying m<sup>6</sup>A sites.

### 3.3 Web-Server and Guide for Users

For the convenience of most experimental scientists, a publicly accessible web-server for **MethyRNA** has been established. Moreover, a step-by-step guide on how to use it to get the desired results is given below.

**Step 1.** Open the web server at <http://lin.uestc.edu.cn/server/methyrna> and you will see the top page of the **MethyRNA** predictor on your computer screen, as shown in **Fig. 1**. Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it.

**Step 2.** By clicking the open circle, the organism concerned will be selected. Either type or copy/paste the query RNA sequences into the input box at the center of **Figure 1**. The input sequence should be in FASTA format. Examples RNA sequences can be seen by clicking the Example button right above the input box.

**Step 3.** Click on the Submit button to see the predicted result. For example, if you use the query RNA sequences in the Example window as the input, the outcomes for the 1<sup>st</sup> and 2<sup>nd</sup> are as following: The A at position 21 in the 1<sup>st</sup> query sequence is methylated; The A at position 21 in the 2<sup>nd</sup> query sequence is unmethylated. All these results are fully consistent with the experimental observations. To get the anticipated prediction accuracy, the species button consistent with the source of query sequences should always be checked: if the query sequences are from *H. sapiens*, the '*H. sapiens*' button is checked; if from *M. musculus*, the '*M. musculus*' button is checked.

**Step 4.** Click on the Data button to download the datasets used to train and test



the model.

**Step 5.** Click on the Citation button to find the relevant paper that document the detailed development and algorithm of **MethyRNA**.

### Funding

This work was supported by the Fundamental Research Funds for the Central Universities of China (grant number ZYGX2015J144 and ZYGX2015Z006), the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122), the Nature Scientific Foundation of Hebei Province (No.C2013209105), the Applied Basic Research Program of Sichuan Province (No. 2015JY0100 and LZ-LY-45) and Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No.BJ2014028) .

### Reference

- Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, Zhang X, Vendeix FA, Fabris D, Agris PF (2011) The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res* 39:D195-201
- Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:1-27
- Chen J, Wang X, Liu B (2016) iMiRNA-SSF: Improving the Identification of MicroRNA Precursors by Combining Negative Sets with Different Distributions. *Sci Rep* 6:19062
- Chen T, Hao YJ, Zhang Y, Li MM, Wang M, Han W, Wu Y, Lv Y, Hao J, Wang L, Li A, Yang Y, Jin KX, Zhao X, Li Y, Ping XL, Lai WY, Wu LG, Jiang G, Wang HL, Sang L, Wang XJ, Yang YG, Zhou Q (2015a) m(6)A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell* 16:289-301
- Chen W, Feng P, Ding H, Lin H, Chou KC (2015b) iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal Biochem* 490:26-33
- Chen W, Feng PM, Deng EZ, Lin H, Chou KC (2014a) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal Biochem* 462:76-83
- Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res* 41:e68
- Chen W, Feng PM, Lin H, Chou KC (2014b) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed Res Int* 2014:623149
- Chen W, Lin H, Feng PM, Wang JP (2014c) Exon skipping event prediction based on histone modifications. *Interdisciplinary Sciences: Computational Life Sciences* 6:241-249
- Chen W, Tran H, Liang Z, Lin H, Zhang L (2015c) Identification and analysis of the



N(6)-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci Rep* 5:13859

Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. *J Theor Biol* 273:236-247

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, Sorek R, Rechavi G (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485:201-206

Feng PM, Chen W, Lin H (2014) Prediction of CpG island methylation status by integrating DNA physicochemical properties. *Genomics* 104:229-233

Feng PM, Chen W, Lin H, Chou KC (2013) iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal Biochem* 442:118-125

Feng PM LH, Chen W, Zuo YC. (2014) Predicting the types of J-proteins using clustered amino acids. *BioMed Research International* 2014:935719

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150-3152

Golam Bari ATM, Rokeya Reaz M, Jeong BS (2014) DNA Encoding for Splice Site Prediction in Large DNA Sequence. *MATCH Communications in Mathematical and in Computer Chemistry* 71:241-258

Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* 30:1522-1529

Lin H, Chen W, Ding H (2013a) AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. *PLoS One* 8:e75726

Lin H, Chen W, Yuan LF, Li ZQ, Ding H (2013b) Using over-represented tetrapeptides to predict protein submitochondria locations. *Acta Biotheor* 61:259-268

Lin H, Deng EZ, Ding H, Chen W, Chou KC (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res* 42:12961-12972

Lin H, Liu WX, He J, Liu XH, Ding H, Chen W (2015) Predicting cancerlectins by the optimal g-gap dipeptides. *Sci Rep* 5:16964

Liu B, Fang L, Chen J, Liu F, Wang X (2015a) miRNA-dis: microRNA precursor identification based on distance structure status pairs. *Mol Biosyst* 11:1194-1204

Liu B, Fang L, Liu F, Wang X, Chen J, Chou KC (2015b) Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS One* 10:e0121501

Liu B, Fang L, Liu F, Wang X, Chou KC (2016) iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. *J Biomol Struct Dyn* 34:223-235

Liu B FL, Long R, Lan X, Chou KC (2016) iEnhancer-2L: a two-layer predictor for

- identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* 32:362-369
- Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, Dong Q, Chou KC (2014) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* 30:472-479
- Meyer KD, Jaffrey SR (2014) The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nat Rev Mol Cell Biol* 15:313-326
- Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S (2013) Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med Chem* 9:133-137
- Nilsen TW (2014) Molecular biology. Internal mRNA methylation finally finds functions. *Science* 343:1207-1208
- Schwartz S, Agarwala SD, Mumbach MR, Jovanovic M, Mertins P, Shishkin A, Tabach Y, Mikkelsen TS, Satija R, Ruvkun G, Carr SA, Lander ES, Fink GR, Regev A (2013) High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* 155:1409-1421
- Sun WJ, Li JH, Liu S, Wu J, Zhou H, Qu LH, Yang JH (2015) RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res* 44:D259-D265
- Wei L, Liao M, Gao Y, Ji R, He Z, Zou Q (2014) Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. *IEEE/ACM Trans Comput Biol Bioinform* 11:192-201
- Zou Q, Guo JS, Ju Y, WU MH, Zeng XX, Hong ZL (2015) Improving tRNAscan-SE Annotation Results via Ensemble Classifiers *Molecular Informatics* 34:761-770

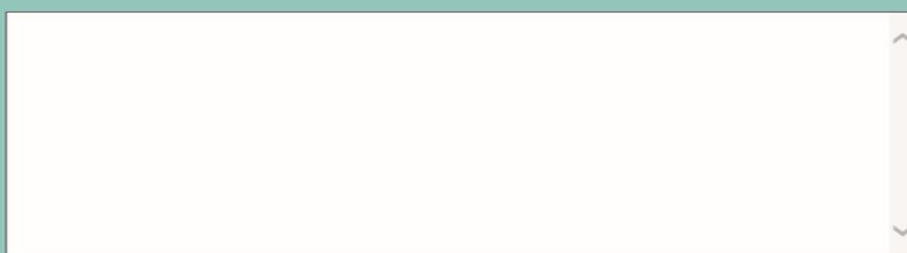
Method	Species	Sn (%)	Sp (%)	Acc (%)
iMethyl-RNA	<i>H. sapiens</i>	57.47	76.92	67.19
	<i>M. musculus</i>	62.80	66.25	64.53
Current method	<i>H. sapiens</i>	81.68	99.11	90.38
	<i>M. musculus</i>	77.79	100.00	88.39

## MethyRNA: A sequence-based tool for the identification of N6-methyladenosine sites

| [Read Me](#) | [Data](#) | [Citation](#) |

Enter the query **RNA sequences** in FASTA format ([Example](#)):

☒ *Homo sapiens*    ☐ *Mus musculus*



Submit

Clear