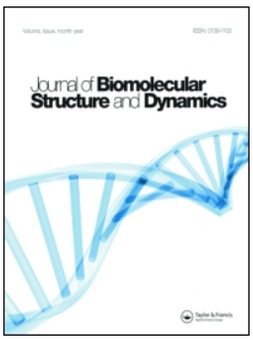

Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: 0739-1102 (Print) 1538-0254 (Online) Journal homepage: http://www.tandfonline.com/loi/tbsd20

MethyRNA: a web server for identification of N^6-methyladenosine sites

Wei Chen, Hua Tang & Hao Lin

To cite this article: Wei Chen, Hua Tang & Hao Lin (2017) MethyRNA: a web server for identification of N^6-methyladenosine sites, Journal of Biomolecular Structure and Dynamics, 35:3, 683-687, DOI: 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.
Published online: 04 May 2016.

Journal of Biomolecular Structure and Dynamics, 2017
Vol. 35, No. 3, 683–687, http://dx.doi.org/10.1080/07391102.2016.1157761

Taylor & Francis
Taylor & Francis Group

LETTER TO THE EDITOR

MethyRNA: a web server for identification of N^6-methyladenosine sites

Wei Chen[a]*, Hua Tang[b] and Hao Lin[c]*

[a]*Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China;* [b]*Department of Pathophysiology, Sichuan Medical University, Luzhou 646000, China;* [c]*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*

Communicated by Ramaswamy H. Sarma

(*Received 20 January 2016; accepted 20 February 2016*)

1. Introduction

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

Recently, using high-throughput sequencing techniques, m^6A data were 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^6A sites. However, to our best knowledge, the existing computational tools for the detection of m^6A sites are only applicable for *S. cerevisiae* (Chen, Feng, et al., 2015; Chen, Tran, et al., 2015). Therefore, there is an urgent need to develop new automated methods for m^6A 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^6A 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^6A 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. Data-set

Using MeRIP-Seq and m^6A-seq, m^6A sites have been identified in *S. cerevisiae*, *H. sapiens,* and *M. musculus* (Dominissini et al., 2012; Schwartz et al., 2013). These experimentally annotated m^6A sites have been checked and deposited in the RMBase (Sun et al., 2015). Therefore, from RMBase, we obtained 94,895 and 28,002 m^6A site containing sequences in *H. sapiens* and *M. musculus*, respectively. All of these sequences are 41-nt long with the m^6A 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, Niu, Zhu, Wu, & Li, 2012). After such a screening procedure, we obtained 1130 and 725 m^6A site containing sequences and deemed them as the positive samples for *H. sapiens* and *M. musculus*, respectively.

Considering the m^6A 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 data-set is dramatically greater than that of positive data-set. In machine-learning problems, imbalanced data-sets can affect the accuracy of learning models. To balance out the numbers between positive and negative samples in model training, 1130 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 data-sets as formulated by

*Corresponding authors. Emails: chenweiimu@gmail.com (W. Chen), hlin@uestc.edu.cn (H. Lin)

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

where the positive data-set $\mathbb{S}_1^+$ contains 1130 true m^6A site containing sequences, while the negative data-set $\mathbb{S}_1^-$ contains 1130 false m^6A site containing sequences; $\mathbb{S}_2^+$ contains 725 true m^6A site containing sequences, while the negative data-set $\mathbb{S}_2^-$ contains 725 false m^6A site containing sequences; and the symbol $\cup$ means the union in the set theory. All of the positive and negative samples in the benchmark data-set are available at http://lin.uestc.edu.cn/server/Methy/data.

2.2. Support vector machine

SVM is a machine learning algorithm and has been successfully used in the realm of bioinformatics (Chen, Feng, Lin, & Chou, 2013; Chen, Feng, Deng, Lin, & Chou, 2014b; Lin, Deng, Ding, Chen, & Chou, 2014; Liu, Fang, Liu, Wang, & Chou, 2016; Liu et al., 2014; Lin et al., 2015; Liu et al., 2015; Liu, Fang, Long, Lan, & Chou, 2016; Zou et al., 2015). 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 & 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 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 γ using a grid search approach in the range $2^{-5} \le C \le 2^{15}$ with step of 2 and $2^{-15} \le \gamma \le 2^{-5}$ 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, Rokeya Reaz, & Jeong, 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).

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

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), and 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.

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

where l is the sequence length, $|N_i|$ is the length of the i-th prefix string $\{n_1, n_2, \ldots, n_i\}$ in the sequence, $q \in \{A, C, G, U\}$. Suppose an example sequence 'GUACCUGAUG'. The density of 'A' is .33 (1/3), .25 (2/8) at positions 3 and 8, respectively. The density of 'C' is .25 (1/4) and .4 (2/5) at positions 4 and 5, respectively. The density of 'G' is 1 (1/1), .29 (2/7), and .30 (3/10) at positions 1, 7, and 10, respectively. The density of 'U' is .5 (1/2), .33 (2/6), and .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, .5), (1, 1, 1, .33), (0, 1, 0, .25), (0, 1, 0,.4), (0, 0, 1, .33), (1, 0, 0, .29), (1, 1, 1, .25), (0, 0, 1, .33), (1, 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 (Chen, Lin, Feng, & Wang, 2014c; Chen, Feng, et al., 2015; Chen, Tran, et al., 2015; Chen, Wang, & Liu, 2016; Lin, Chen, & Ding, 2013a; Lin, Chen, Yuan, Li, & Ding, 2013b; Wei et al., 2014), 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

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

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 (Chen, Feng, Lin, & Chou, 2014a; Feng, Chen, Lin, & Chou, 2013; Feng, Chen, & Lin, 2014; Feng, Lin, Chen, & Zuo, 2014; Guo et al., 2014; Liu, Fang, Chen, Liu, & Wang, 2015; Mohabatkar, Beigi, Abdolahi, & Mohsenzadeh, 2013). Hence, the jackknife test was used to examine the performance of the proposed model. In the jackknife test, each sample in the training data-set 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 data-set is 41 bp with the m^6A in the center, the corresponding predictive results were most promising. Therefore, each sample in the benchmark data-set was encoded by a 164 (4×41)-dimensional vector (see Section 2) and was used as the input vector of SVM. In the jackknife test, the proposed model accurately identified the m^6A 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, Feng, et al., 2015). Accordingly, we encoded the sequences in the benchmark data-set according to the rules of *iMethyl-RNA* and carried out the jackknife test on the benchmark data-set 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^6A 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^6A sites.

Table 1. Comparison of *MethyRNA* with the other method in identifying m^6A sites.

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

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 Figure 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 of 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 1st and 2nd are as following: The A at position 21 in the 1st query sequence is methylated; The A at position 21 in the 2nd 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.

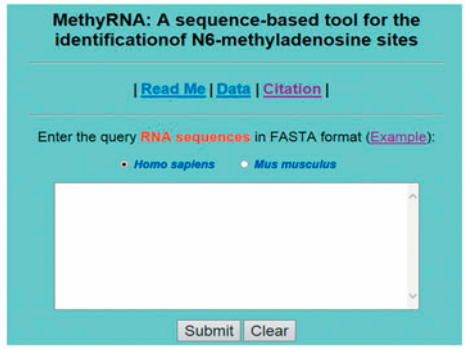

Figure 1. A semi-screenshot for the top page of the MethyRNA web server, which is available at http://lin.uestc.edu.cn/server/methyrna.

Step 4. Click on the *Data* button to download the data-sets 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*.

Note: While our paper was in proof, we were alerted to a study by Yuan Zhou and colleagues reporting similar researches on identifying m^6A sites (Zhou, Zeng, Li, Zhang, & Cui, 2016).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported the Nature Scientific Foundation of Hebei Province [No. C2013209105]; Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province [No. BJ2014028]; the Fundamental Research Funds for the Central Universities of China [grant number ZYGX2015J144], [grant number ZYGX2015Z006]; the Scientific Research Foundation of the Education Department of Sichuan Province [11ZB122]; the Applied Basic Research Program of Sichuan Province [No. 2015JY0100], [No. LZ-LY-45]; National Nature Scientific Foundation of China [61202256].

References

Cantara, W. A., Crain, P. F., Rozenski, J., McCloskey, J. A., Harris, K. A., Zhang, X., … Agris, P. F. (2011). The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Research, 39*, D195–D201.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 1–27.

Chen, W., Feng, P. M., Lin, H., & Chou, K. C. (2013). iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Research, 41*, e68.

Chen, W., Feng, P. M., Deng, E. Z., Lin, H., & Chou, K. C. (2014). iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical Biochemistry, 462*, 76–83.

Chen, W., Feng, P. M., Lin, H., & Chou, K. C. (2014). iSS-PseDNC: Identifying splicing sites using pseudo dinucleotide composition. *BioMed Research International, 2014*, 623149.

Chen, W., Lin, H., Feng, P. M., & Wang, J. P. (2014). Exon skipping event prediction based on histone modifications. *Interdisciplinary Sciences: Computational Life Sciences, 6*, 241–249.

Chen, W., Feng, P., Ding, H., Lin, H., & Chou, K. C. (2015). iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Analytical Biochemistry, 490*, 26–33.

Chen, T., Hao, Y. J., Zhang, Y., Li, M. M., Wang, M., Han, W., … Zhou, Q. (2015). m(6)A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell, 16*, 289–301.

Chen, W., Tran, H., Liang, Z., Lin, H., & Zhang, L. (2015). Identification and analysis of the N(6)-methyladenosine in the Saccharomyces cerevisiae transcriptome. *Scientific Reports, 5*, 13859.

Chen, J., Wang, X., & Liu, B. (2016). iMiRNA-SSF: Improving the identification of microRNA precursors by combining negative sets with different distributions. *Scientific Reports, 6*, 19062.

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology, 273*, 236–247.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., … Rechavi, G. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature, 485*, 201–206.

Feng, P. M., Chen, W., Lin, H., & Chou, K. C. (2013). iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Analytical Biochemistry, 442*, 118–125.

Feng, P. M., Chen, W., & Lin, H. (2014). Prediction of CpG island methylation status by integrating DNA physicochemical properties. *Genomics, 104*, 229–233.

Feng, P. M., Lin, H.,Chen, W., & Zuo, Y. C. (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, A. T. M., Rokeya Reaz, M., & Jeong, B. S. (2014). DNA encoding for splice site prediction in large DNA sequence. *MATCH Communications in Mathematical and in Computer Chemistry, 71*, 241–258.

Guo, S. H., Deng, E. Z., Xu, L. Q., Ding, H., Lin, H., Chen, W., & Chou, K. C. (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. (2013). AcalPred: A sequence-based tool for discriminating between acidic and alkaline enzymes. *PLoS One, 8*, e75726.

Lin, H., Chen, W., Yuan, L. F., Li, Z. Q., & Ding, H. (2013). Using over-represented tetrapeptides to predict protein submitochondria locations. *Acta Biotheoretica, 61*, 259–268.

Lin, H., Deng, E. Z., Ding, H., Chen, W., & Chou, K. C. (2014). iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Research, 42*, 12961–12972.

Lin, H., Liu, W. X., He, J., Liu, X. H., Ding, H., & Chen, W. (2015). Predicting cancerlectins by the optimal g-gap dipeptides. *Scientific Reports, 5*, 16964.

Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., … Chou, K. C. (2014). Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics, 30*, 472–479.

Liu, B., Fang, L., Chen, J., Liu, F., & Wang, X. (2015). miRNA-dis: MicroRNA precursor identification based on distance structure status pairs. *Molecular BioSystems, 11*, 1194–1204.

Liu, B., Fang, L., Liu, F., Wang, X., Chen, J., & Chou, K. C. (2015). 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, K. C. (2016). iMiRNA-PseDPC: MicroRNA precursor identification with a pseudo distance-pair composition approach. *Journal of Biomolecular Structure and Dynamics, 34*, 223–235.

Liu, B. , Fang, L. Y., Long, R., Lan, X., & Chou, K. C. (2016). iEnhancer-2L: A two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics, 32*, 362–369.

Meyer, K. D., & Jaffrey, S. R. (2014). The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nature Reviews Molecular Cell Biology, 15*, 313–326.

Mohabatkar, H., Beigi, M. M., 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. *Medicinal Chemistry, 9*, 133–137.

Nilsen, T. W. (2014). Internal mRNA methylation finally finds functions. *Science, 343*, 1207–1208.

Schwartz, S., Agarwala, S. D., Mumbach, M. R., Jovanovic, M., Mertins, P., Shishkin, A., … Regev, A. (2013). High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell, 155*, 1409–1421.

Sun, W. J., Li, J. H., Liu, S., Wu, J., Zhou, H., Qu, L. H., & Yang, J. H. (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.

Zhou, Y., Zeng, P., Li, Y. H., Zhang, Z. D., & Cui, Q. H. (2016). SRAMP: Prediction of mammalian N6-methyladenosine (m^6A) sites based on sequence-derived features. *Nucleic Acids Research*. doi: 10.1093/nar/gkw104.

Zou, Q., Guo, J. S., Ju, Y., Wu, M. H., Zeng, X. X., Hong, Z. L. (2015). Improving tRNAscan-SE annotation results via ensemble classifiers. *Molecular Informatics, 34*, 761–770.