MethyRNA: a web server for identification of N⁶-methyladenosine sites

Wei Chen, Hua Tang & Hao Lin


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.

Submit your article to this journal

Article views: 138

View related articles

View Crossmark data

Citing articles: 3 View citing articles
LETTER TO THE EDITOR

MethyRNA: a web server for identification of N6-methyladenosine sites

Wei Chena*, Hua Tangb and Hao Linc*

aDepartment of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China; bDepartment of Pathophysiology, Sichuan Medical University, Luzhou 646000, China; cKey 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

N6-methyladenosine (m6A) is the most abundant post-transcriptional modification and has been found in the three domains of life (Cantara et al., 2011). m6A plays fundamental regulatory roles in a series of biological processes, 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 m6A sites in RNA will help to expand our understanding of its potential roles.

Recently, using high-throughput sequencing techniques, m6A 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 m6A sites. However, to our best knowledge, the existing computational tools for the detection of m6A 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 m6A 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 m6A sites by encoding RNA sequence using nucleotide chemical property and frequency. Results from the jackknife test show that the proposed model could accurately identify m6A 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 m6A-seq, m6A sites have been identified in S. cerevisiae, H. sapiens, and M. musculus (Dominissini et al., 2012; Schwartz et al., 2013). These experimentally annotated m6A sites have been checked and deposited in the RMBase (Sun et al., 2015). Therefore, from RMBase, we obtained 94,895 and 28,002 m6A site containing sequences in H. sapiens and M. musculus, respectively. All of these sequences are 41-nt long with the m6A 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 m6A site containing sequences and deemed them as the positive samples for H. sapiens and M. musculus, respectively.

Considering the m6A site in H. sapiens and M. musculus harboring the consenus 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: chenweimu@gmail.com (W. Chen), hlin@uestc.edu.cn (H. Lin)

© 2016 Informa UK Limited, trading as Taylor & Francis Group
where the positive data-set \( S_1^+ \) contains 1130 true m6A site containing sequences, while the negative data-set \( S_1^- \) contains 1130 false m6A site containing sequences; \( S_2^+ \) contains 725 true m6A site containing sequences, while the negative data-set \( S_2^- \) contains 725 false m6A 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.

\[ s_k = s_k^+ \cup s_k^- \quad k = \begin{cases} 1 & \text{for } H. \text{sapiens} \\ 2 & \text{for } M. \text{musculus} \end{cases} \]

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 \( \gamma \) using a grid search approach in the range \( 2^{-5} \leq C \leq 2^{15} \) with step of 2 and \( 2^{-15} \leq \gamma \leq 2^{-3} \) 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).

\[
\begin{align*}
\text{A} & = \{1 \text{ if } s_i = \{A, G\} \\
\text{C} & = \{1 \text{ if } s_i = \{C, U\} \\
\text{G} & = \{1 \text{ if } s_i = \{A, C\} \\
\text{U} & = \{1 \text{ if } s_i = \{A, U\} \\
\end{align*}
\]

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 \((0, 0, 1)\), 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 m6A site, the density \( d_i \) of any nucleotide \( n_j \) at position \( i \) in RNA sequence was defined by the following formula.

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

where \( l \) is the sequence length, \( |N| \) is the length of the \( i \)-th prefix string \( \{n_1, n_2, ..., n_l\} \) 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 \( \{1, 0, 0, 1, 0, 0, 1, 0, .5, 1, 1, 1, 33, 0, 1, 0, .25, 0, 1, 0, 4, 0, 0, 1, .33, 1, 0, 0, .29, 1, 1, 1, .25, 0, 0, 0, 1, 33, 1, 0, 0, .30\} \). 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
sensitivity, specificity, and accuracy were also reported in Table 1, from which we found that the accuracies obtained by \textit{iMethyl-RNA} are approximately 23\% lower than our proposed method for identifying m\(^6\)A sites in \textit{H. sapiens} and \textit{M. musculus}. These results indicate that the model proposed in this work is quite promising and may become a useful tool in identifying m\(^6\)A sites.

### 3.3. Web server and guide for users

For the convenience of most experimental scientists, a publicly accessible web server for \textit{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 \url{http://lin.uestc.edu.cn/server/methyrna} and you will see the top page of the \textit{MethyRNA} predictor on your computer screen, as shown in Figure 1. Click on the \textit{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 \textit{Example} button right above the input box.

Step 3. Click on the \textit{Submit} button to see the predicted result. For example, if you use the query RNA sequences in the \textit{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 \textit{H. sapiens}, the ‘\textit{H. sapiens}’ button is checked; if from \textit{M. musculus}, the ‘\textit{M. musculus}’ button is checked.

---

**Table 1.** Comparison of \textit{MethyRNA} with the other method in identifying m\(^6\)A sites.

<table>
<thead>
<tr>
<th>Method</th>
<th>Species</th>
<th>Sn (%)</th>
<th>Sp (%)</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{iMethyl-RNA}</td>
<td>\textit{H. sapiens}</td>
<td>57.47</td>
<td>76.92</td>
<td>67.19</td>
</tr>
<tr>
<td></td>
<td>\textit{M. musculus}</td>
<td>62.80</td>
<td>66.25</td>
<td>64.53</td>
</tr>
<tr>
<td>Current method</td>
<td>\textit{H. sapiens}</td>
<td>81.68</td>
<td>99.11</td>
<td>90.38</td>
</tr>
<tr>
<td></td>
<td>\textit{M. musculus}</td>
<td>77.79</td>
<td>100.00</td>
<td>88.39</td>
</tr>
</tbody>
</table>

---

**Figure 1.** A semi-screenshot for the top page of the MethyRNA web server, which is available at \url{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 documented 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 m6A 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. B2014028]; 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]; the Science and Technology Department of Sichuan Province (No. 2015JY0100); the Fundamental Research Funds for the Central Universities of China [grant number ZYGX2015J144], [grant number ZYGX2015Z006]; the Sichuan University Young Innovative Talents of Higher Learning Institution 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]; the National Nature Scientific Foundation of China [61202256].

References


