Sequence analysis

iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications

Kewei Liu¹ and Wei Chen¹,²,*

¹School of Life Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China and ²Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China

*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier

Received on January 8, 2020; revised on February 26, 2020; editorial decision on February 27, 2020; accepted on February 28, 2020

Abstract

Motivation: RNA modifications play critical roles in a series of cellular and developmental processes. Knowledge about the distributions of RNA modifications in the transcriptomes will provide clues to revealing their functions. Since experimental methods are time consuming and laborious for detecting RNA modifications, computational methods have been proposed for this aim in the past five years. However, there are some drawbacks for both experimental and computational methods in simultaneously identifying modifications occurred on different nucleotides.

Results: To address such a challenge, in this article, we developed a new predictor called iMRM, which is able to simultaneously identify m6A, m5C, m1A, ψ and A-to-I modifications in Homo sapiens, Mus musculus and Saccharomyces cerevisiae. In iMRM, the feature selection technique was used to pick out the optimal features. The results from both 10-fold cross-validation and jackknife test demonstrated that the performance of iMRM is superior to existing methods for identifying RNA modifications.

Availability and implementation: A user-friendly web server for iMRM was established at http://www.bioml.cn/XG_iRNA/home. The off-line command-line version is available at https://github.com/liukeweiaway/imrm.

Contact: greatchen@ncst.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Since the first kind of RNA modification was discovered in 1957, more than 150 distinct kinds of RNA modifications have been identified in the major classes of RNA (Boccaletto et al., 2018). Among these modifications, the N6-methyladenosine (m6A), N4-methyladenosine (m4A), 5-methylcytosine (m5C), pseudouridine (ψ) and adenosine to inosine (A-to-I) are the most common modifications in cellular RNA. Increasing evidences have demonstrated that, as a dynamic process, RNA modifications play critical roles in a series of cellular and developmental processes, such as RNA localization and degradation (Wang et al., 2014), dynamic changes in RNA structure (Chen et al., 1993), RNA localization and degradation (Yang et al., 2017; Zhang et al., 2012), RNA splicing (Guzzi et al., 2018; Liu et al., 2015), circadian rhythm (Fustin et al., 2013), etc.

Recent studies have also demonstrated that RNA modifications are associated with human diseases, such as metabolic diseases, cancer, neurologological disorders and cardiovascular diseases. For example, m6A is associated with cancer, obesity (Jia et al., 2011), acute myelogenous leukemia (Bansal et al., 2014), zika virus (Lichinchi et al., 2016) and depressive disorders (Du et al., 2015); m1A is linked to X-linked intractable epilepsy, multiple respiratory chain deficiencies (Metodiev et al., 2016) and neurodevelopmental regression (Falk et al., 2016). A-to-I is related to cancer (Han et al., 2015; Paz et al., 2007) and neurological disease (Sasaki et al., 2015). The relationship between m5C and disease has also been reported, such as autistic features (Hussain and Bashir, 2015), breast cancer (Yi et al., 2017), intellectual disability syndromes (Abbasi-Mohb et al., 2012; Khan et al., 2012; Martinez et al., 2012) and some others (Bohnack et al., 2019). Besides the modification itself, the enzymes that catalyze their formations are also linked to human disease. Mutations of the RNA pseudouridine synthase enzyme will cause mitochondrial myopathy and sideroblastic anemia (Fujiwara and Harigae, 2013; Tohru and Hideo, 2018). More details about associations between RNA modification and human diseases can refer to a recent review (Jonkhout et al., 2017). However, lack of efficient tools to detect RNA modifications precludes researches on the mechanisms that may lead to diseases. What we found is only the tip of the iceberg. To fill such a gap, it is urgent to develop effective methods to detect RNA modifications on the transcriptome wide.

With the development of high-throughput sequencing technology, a series of methods have been developed to identify distinct kinds of RNA modifications with high resolution, including m6A's
miCLIP (Linder et al., 2015), m6A-CLIP (Ke et al., 2015), Oxford Nanopore Technologies (Liu et al., 2019), m4A-SEQ (Dominissini et al., 2016), m5A-1D-SEQ (Li et al., 2016), bisulfite sequencing (Squires et al., 2012), m5C-RIP (Edelheit et al., 2013), Aza-IP (Edelheit et al., 2013), miCLIP (Hussain et al., 2013), jRNA-seq (Lovejoy et al., 2014), Cel-seq (Li et al., 2015) and ICE-seq (Su et al., 2015). It should be noticed that all these methods are not able to simultaneously detect multiple modifications.

Recent work from Chang et al. (2013) demonstrated that A-to-I modifications might mediate the biological processes in a combinational pattern. For example, Xiang et al. (2016) found the interaction between m5A and A-to-I (Xiang et al., 2018). Recently, Vahid Khoddami et al. developed a RNA bisulfite sequencing-based method that is able to simultaneously detect m5C, m5A and p modifications (Khoddami et al., 2019). However, the obtained results are greatly diverged from previous works. Moreover, all these experimental methods are still expensive and time consuming for transcriptome-wide detection of RNA modifications. Therefore, it is urgent to develop effective and low-cost approaches to automatically identify RNA modifications.

As excellent complements to experimental techniques, computational methods are in high demand to detect RNA modifications. In 2013, Schwartz et al. (2013) proposed the first computational model to identify m5A site in Saccharomyces cerevisiae. However, no public web server or software package was provided for this method. Inspired by Schwartz et al.’s work, a series of machine learning-based methods have been proposed to identify different kinds of RNA modifications over the past several years, such as iRNA-1Methyl (Chen et al., 2015), SRAMP (Zhou et al., 2016), iRNA-PseU (Chen et al., 2016), PPSU (Li et al., 2015), iRNA-m5SC (Yang et al., 2018), iRNA (Xu et al., 2019), iRNA-3typeA (Chen et al., 2018) and so on. More details about them and some other representative online computational tools for predicting RNA modifications were summarized by Morena et al. (2018). More recently, Chen et al. developed another computational tool that could predict m5A and m5C (Chen et al., 2019). However, all these computational methods are limited to adenosine, and could not identify modifications occurred on other nucleotides in the transcriptome.

Keeping this in mind, in the present work, we developed a new computational method, called iMRM, which is able to simultaneously identify m5A, m5C, m4A, p and A-to-I modification in Human sapiens, Mus musculus and S. cerevisiae, respectively. The framework of developing iMRM is shown in Figure 1. In order to demonstrate its better performance, we first compared iMRM with existing methods for identifying RNA modifications. Subsequently, we systematically compared and analyzed the optimal features that make great contributions for identifying RNA modifications in each species. Finally, a freely accessible user-friendly web server was provided for the proposed method.

Sequence representation is a key point for developing computational methods (Zuo et al., 2017). In order to transfer the RNA sequence into a discrete vector that can be recognized by machine learning methods, the following six kinds of sequence encoding methods were used to represent the RNA samples in the dataset.

**k-tuple nucleotide composition**
The k-tuple nucleotide composition is defined as

$$D_k = [f_1, f_2, \ldots, f_n]$$

(2)

In the present work, we set \( k = 1, 2, 3, 4 \) and 5 indicating single-nucleotide component (NC), di-nucleotide component (DNC), tri-nucleotide component (TNC), tetra-nucleotide component (TeNC) and penta-nucleotide component (PNC), respectively. Accordingly, the dimension of the feature vector obtained by using k-tuple nucleotide composition is \( 1364 (4 + 16 + 64 + 256 + 1024) \).

**Onehot**
Onehot (Qiang et al., 2018; Wei et al., 2019) is a simple and effective encoding method, where A is represented as \( (1, 0, 0, 0) \), U as \( (0, 0, 1, 0) \), C as \( (0, 0, 1, 0) \) and G as \( (0, 0, 0, 1) \). The RNA sample in the database will be converted to a \( (1548 + 2 \times 11) \)-dimensional vector.

**Di-nucleotide binary encoding**
DINUC (di-nucleotide binary encoding) (Qiang et al., 2018) is an extension of onehot, in which AA, AU, AC, AG, \( \ldots \) and GG is encoded as \( (0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 1, 0), (0, 0, 1, 1) \) and \( (1, 1, 1, 1) \), respectively. By using DINUC, the dimension of the obtained feature vector is \( 4 \times (2^2 + 1) \).

**Nucleotide density**
Nucleotide density (ND) (Bari et al., 2013) considers both nucleotide location information and frequency information, which is defined as:

$$d_i = \frac{1}{|S|} \sum_{j=1}^{p} f(N_i)$$

(3)

where \( d_i \) is the density of the nucleotide \( N_i \) at position \( i \) of RNA sample, the length of the sliding substring is \(|S|\), \( p \) is the position of corresponding nucleotide.

$$f(N_i) = \begin{cases} 1, & \text{if } N_i \text{ is the corresponding nucleotide} \\ 0, & \text{other nucleotide} \end{cases}$$

Accordingly, we can obtain a \( (2^2 + 1) \)-dimensional vector. If we consider dinucleotide density, we will obtain a \( 2^2 \)-dimensional vector.

**Nucleotide chemical property**
Nucleotide chemical property (NCP) is also used as a coding method in RNA modification site prediction (Chen et al., 2018). The four nucleotides can be divided into three groups according to the number of ring structures, hydrazine or pyrimidine and the number of hydrogen bonds that can be formed. The A, C, G and U will be encoded by 0 and 1 in a three-dimensional coordinate system as follows:

$$x_i = \begin{cases} 1, & \text{if } N_i \in \{A, G\} \\ 0, & \text{if } N_i \in \{C, U\} \end{cases}$$

$$y_i = \begin{cases} 1, & \text{if } N_i \in \{A, C\} \\ 0, & \text{if } N_i \in \{G, U\} \end{cases}$$

$$z_i = \begin{cases} 1, & \text{if } N_i \in \{A, U\} \\ 0, & \text{if } N_i \in \{C, G\} \end{cases}$$

According to NCP, a RNA sample will be encoded by a \( 3 \times (2^2 + 1) \)-dimensional vector.

**Dinucleotide physicochemical properties**
Dinucleotide physicochemical properties (DPCP) integrate 11 DPCP, namely Shift, Slide, Rise, Tilt, Roll, Twist, Stacking energy, Entalphy, Entropy, Free energy and Hydrophilicity. DPCP is defined as (Manavalan et al., 2019),

$$\text{DPCP}(i) = f(i) \times \text{dpcp}(i)$$

(4)

where \( f \) is frequency of the dinucleotide \( i \), \( \text{dpcp}(i) \) is the value of the \( j \)-th \( (i = 1, 2, \ldots, 11) \) DPCP for the \( j \)-th dinucleotide and is also listed in Supplementary Table S1. Therefore, the dimension of DPCP is \( (16 \times 11) \).

By integrating all these features, a RNA sample in the dataset will be converted into a \( (1548 + 2 \times 11) \)-dimensional vector.

### 2.3 Machine learning method

XGBoost (eXtreme gradient boosting) (Chen and Guestrin, 2016) is a boosting algorithm for classifying based on tree models. Since the regularization term was added to the loss function of XGBoost, the complexity of the algorithm is controlled. On the other hand, by adding the function of parallel computing into XGBoost, its computational speed was also improved. Moreover, XGBoost is highly flexible and allows users to define custom optimization goals and evaluation criteria. Therefore, XGBoost is widely used to deal with bioinformatics problems (Qiang et al., 2018; Yu et al., 2019, 2020; Zhao et al., 2018). In the present work, the python package called xgboost (vision 0.90) which is available at https://pypi.org/project/xgboost/ was employed to perform the classifications. The range of its parameter is provided in Supplementary Table S2.

### 2.4 Feature selection

In order to avoid noise features that will reduce the stability and performance of a model, the two-step feature optimization strategy was performed to select optimal features. We first sorted the features according to their F-score obtained from XGBoost package. Considering the fact that most features might be noise and only a few features will be effective for the prediction, for saving computational time, we chose the top 50 features to build the optimal feature subsets by using the incremental feature selection (IFS) strategy.

### 2.5 Performance evaluation

The sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthews correlation coefficient (MCC) were used to evaluate the performance of the model and defined as,

\[
\begin{align*}
\text{Sen} &= \frac{N^+}{N^+ + N^-} \\
\text{Spe} &= \frac{N^-}{N^+ + N^-} \\
\text{Acc} &= 1 - \frac{N^+ + N^-}{N^+ + N^- + N^+ + N^-} \\
\text{MCC} &= \frac{(N^+ N^- - N^- N^+)}{(N^+ N^- + N^- N^+)}
\end{align*}
\]

(7)

where \( N^+ \) is the total number of the RNA sequence containing modification (M) site, \( N^- \) is a false negative sample, \( N^- \) is the total number of the RNA sequence which did not contain any modification (M) site, \( N^- \) is a false positive sample.

In addition, the receiver operating characteristic (ROC) curve (Fishing and Turnbull, 1996), which can intuitively evaluate the performance of the model by the graphics, was also used to evaluate the proposed model. The abscissa of the ROC curve is 1-specificity, and the ordinate is sensitivity. The area under the ROC curve (AUC) is an index reflecting the performance of a model. The larger the AUC is, the better the model’s performance.
3 Results and discussion

3.1 Comparison with existing state-of-the-art models

By using the two-step feature optimization strategy (see Section 2), the optimal features that used to build models for identifying distinct kinds of RNA modifications in H. sapiens, M. musculus, and S. cerevisiae were determined. The number of optimal features and the corresponding accuracy obtained in the 10-fold cross-validation test were shown in Supplementary Figures S1–S13. The ROC curves and AUC of these models for identifying RNA modifications in the three species were shown in Figure 2. The corresponding sensitivity, specificity, accuracy, and Matthews correlation coefficient were listed in Supplementary Table S3. Figure 2. The ROC curves for identifying distinct kinds of RNA modifications in (a) H. sapiens, (b) M. musculus, and (c) S. cerevisiae. The AUC was also provided at the right corner of each figure.

Table 2. Comparative results of different methods for identifying RNA modifications in different species under jackknife test

<table>
<thead>
<tr>
<th>Species</th>
<th>Modification</th>
<th>Sn (%)</th>
<th>Sp (%)</th>
<th>Mcc</th>
<th>Acc (%) this article</th>
<th>Acc (%) existing</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens</td>
<td>m1A</td>
<td>99.04</td>
<td>99.78</td>
<td>0.988</td>
<td>99.41</td>
<td>99.13</td>
</tr>
<tr>
<td></td>
<td>m3C</td>
<td>90.83</td>
<td>93.33</td>
<td>0.842</td>
<td>92.08</td>
<td>92.90</td>
</tr>
<tr>
<td></td>
<td>m6A</td>
<td>82.48</td>
<td>99.56</td>
<td>0.820</td>
<td>91.02</td>
<td>90.38</td>
</tr>
<tr>
<td></td>
<td>A-to-I</td>
<td>62.00</td>
<td>67.11</td>
<td>0.293</td>
<td>64.55</td>
<td>64.24</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>m1A</td>
<td>97.72</td>
<td>100</td>
<td>0.977</td>
<td>98.86</td>
<td>97.83</td>
</tr>
<tr>
<td></td>
<td>m3C</td>
<td>99.05</td>
<td>100</td>
<td>0.991</td>
<td>99.52</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>m6A</td>
<td>77.04</td>
<td>78.50</td>
<td>0.555</td>
<td>77.77</td>
<td>76.51</td>
</tr>
<tr>
<td></td>
<td>m5C</td>
<td>68.69</td>
<td>73.48</td>
<td>0.422</td>
<td>71.08</td>
<td>65.13</td>
</tr>
<tr>
<td>M. musculus</td>
<td>m1A</td>
<td>98.49</td>
<td>99.90</td>
<td>0.984</td>
<td>99.20</td>
<td>98.73</td>
</tr>
<tr>
<td></td>
<td>m3C</td>
<td>97.94</td>
<td>98.97</td>
<td>0.969</td>
<td>98.45</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>m6A</td>
<td>76.90</td>
<td>69.28</td>
<td>0.462</td>
<td>73.09</td>
<td>70.44</td>
</tr>
<tr>
<td></td>
<td>m5C</td>
<td>78.34</td>
<td>99.57</td>
<td>0.779</td>
<td>88.97</td>
<td>88.39</td>
</tr>
</tbody>
</table>

3.2 Comparison with other algorithms

To further demonstrate the power of XGBoost for identifying RNA modifications, we compared it with support vector machine (SVM) (Burges, 1998), random forest (RF) (Breiman, 2001), Logistic regression (LR) (Cox, 1958), naive Bayes (NB) (Zhang et al., 2016) and K-nearest neighbor (KNN) (Keller et al., 1985) on the benchmark dataset by using 10-fold cross-validation test. These algorithms were all performed by using the python package of scikit-learn (version: 0.21.3). The range of parameters for these methods were provided in Supplementary Tables S4–S8. Their results for identifying different kinds of RNA modifications in H. sapiens, M. musculus, and S. cerevisiae were shown in Table 3 and Figure 3. It was found that the XGBoost-based methods obtained the best accuracies in most cases (10 out of 13) for identifying RNA modifications in different species. Although the accuracies are lower for the XGBoost-based methods for identifying m5C in M. musculus and identifying m3A in S. cerevisiae, the discrepancy is only less than 1%. These results demonstrated the stability and superiority of the proposed XGBoost-based methods for identifying RNA modifications. Therefore, based on these XGBoost methods, we developed iMRM, which is able to simultaneously identify m1A, m3C, m6A, m5C, and A-to-I modification in H. sapiens, M. musculus, and S. cerevisiae, respectively.

3.3 Feature analysis

In order to understand the contributions of the specific features for identifying RNA modifications in different species, we extracted and analyzed the optimal features of the 13 models. The occurrence frequency of these optimal features are shown in Figure 4a. A total of 291 optimal features were used in the 13 models, and 74.9% of these features appeared only once in these models which could be
found in Supplementary Table S9. These results indicate that there exists specific features for identifying distinct kinds of RNA modifications in different species. The names and occurrences of the features that appeared more than two times in these models were shown in Figure 4b. Among these features, four kinds of features appear in more than half of the models. They are ‘.2onehot’, ‘2ncp.’, ‘.1onehot’ and ‘.2ncp’. The ‘.’ in the feature name represents RNA modification site, ‘.2onehot’ stands for the second nucleotide downstream of the modification site and encoded by using the one-hot encoding method. ‘2ncp.’ stands for the second nucleotide upstream of the modification site and encoded by using the nucleotide chemical property method. In order to demonstrate the contributions of the optimal features for identifying RNA modifications, we sorted them according to their F-scores (Supplementary Table S10) and plotted them by a heat map in which the elements represent the features and are encoded using different colors according to their F-score. The abscissa is the type of the RNA modification site and encoded by using the nucleotide chemical property method and so forth.

For the convenience of scientific community, an online web server called iMRM was built to simultaneously identify m1A, m6A, m5C, 5mC and A-to-I modifications in H.sapiens, M.musculus and S.cerevisiae respectively. The iMRM can be accessed at http://www.biolml.cn/XG_iRNA/home.

The user guide of iMRM is as following. Open the home page at http://www.biolml.cn/XG_iRNA/home. First, clicking the ‘Web server’ button, the page shown in Figure 5a will be appeared. Second, selecting the ‘Species’ and ‘Modification’ successively. In order to control the false positive rate, a ‘Threshold’ option was provided, whose corresponding value can be found in Supplementary Table S11. Third, type or copy/paste the query RNA sequence with a FASTA format in the input box. Fourth, clicking the ‘Submit’ button, the page shown in Figure 5b will be appeared. Second, selecting the ‘Species’ and ‘Modification’ successively. In order to control the false positive rate, a ‘Threshold’ option was provided, whose corresponding value can be found in Supplementary Table S11. Third, type or copy/paste the query RNA sequence with a FASTA format in the input box. Fourth, clicking the ‘Submit’ button, the page shown in Figure 5b will be appeared. Second, selecting the ‘Species’ and ‘Modification’ successively.

The off-line command-line version of the tool can be obtained either under the ‘Download’ module of the web server, or from https://github.com/liukeweiaway/iMRM.

Table 3. Accuracies of different algorithms for identifying RNA modification sites in H.sapiens, M.musculus and S.cerevisiae.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>H.sapiens</th>
<th>M.musculus</th>
<th>S.cerevisiae</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>91.20</td>
<td>91.03</td>
<td>91.00</td>
</tr>
<tr>
<td>LR</td>
<td>92.43</td>
<td>91.35</td>
<td>91.04</td>
</tr>
<tr>
<td>RF</td>
<td>91.05</td>
<td>90.75</td>
<td>90.52</td>
</tr>
<tr>
<td>KNN</td>
<td>90.85</td>
<td>90.53</td>
<td>90.42</td>
</tr>
<tr>
<td>NB</td>
<td>99.47</td>
<td>99.35</td>
<td>99.15</td>
</tr>
<tr>
<td>XGboost</td>
<td>99.73</td>
<td>99.65</td>
<td>99.50</td>
</tr>
</tbody>
</table>

Note: The best accuracies are in bold font.
update the current method and make it able to simultaneously identify much more kinds RNA modifications.

Acknowledgements

The authors thank the anonymous reviewers for their constructive comments.

Funding

This work was supported by the National Nature Scientific Foundation of China [grant number 31771471] and the Natural Science Foundation for Distinguished Young Scholar of Hebei Province [grant number C2017209244].

Conflict of Interest: none declared.

References


Bansal,H et al. (2014) WTAP is a novel oncogenic protein in acute myeloid Leukemia. Leukemia, 28, 1171–1174.


Bohnack,K. et al. (2019) Eukaryotic 5-methylcytosine (m5C) RNA methyltransferases: mechanisms, cellular functions, and links to disease. Genes, 10, 102.


Chen,W. et al. (2017a) iRNA-AtE: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget, 8, 4208–4217.


Chen,Y. et al. (1993) 5-Methylcytidine is required for cooperative binding of magnesium(2+) and a conformational transition at the anticodon stem-loop of yeast phenylalanine tRNA. Biochemistry, 32, 10249–10253.


