Bioinformatics, 36(11), 2020, 3336–3342
doi: 10.1093/bioinformatics/btaa155
Advance Access Publication Date: 5 March 2020
Original Paper

Sequence analysis

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

Kewei Liu[1] **and Wei Chen**[1,2,*]

[1]School of Life Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China and [2]Innovative 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 m^6A, m^5C, m^1A, 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 N^6-methyladenosine (m^6A), N^1-methyladenosine (m^1A), 5-methylcytosine (m^5C), 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, neurological disorders and cardiovascular diseases. For example, m^6A 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); m^1A 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 m^5C and disease has also been reported, such as autistic features (Hussain and Bashir, 2015), breast cancer (Yi *et al.*, 2017), intellectual disability syndromes (Abbasi-Moheb *et al.*, 2012; Khan *et al.*, 2012; Martinez *et al.*, 2012) and some others (Bohnsack *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 m^6A's

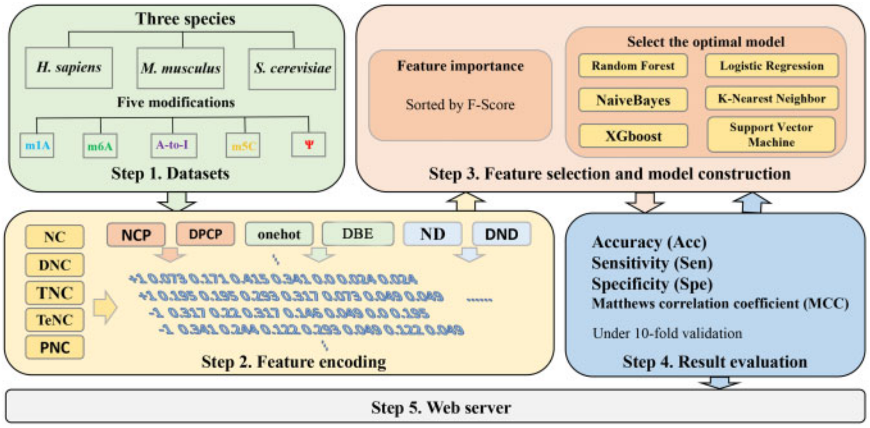


Fig. 1. The framework of developing iMRM. Step 1. Collecting experimentally confirmed m^6A, m^5C, m^1A, ψ and A-to-I data from *H.sapiens*, *M.musculus* and *S.cerevisiae*. Step 2. Encoding the RNA segments by using different kinds of sequence representation methods. Step 3. Selecting optimal features and the optimal machine learning algorithm to build the computational model. Step 4. Evaluation and comparison of the model. Step 5. Construction of the web server

miCLIP (Linder *et al.*, 2015), m6A-CLIP (Ke *et al.*, 2015), Oxford Nanopore Technologies (Liu *et al.*, 2019b), m^1A-seq (Dominissini *et al.*, 2016), m^1A-ID-seq (Li *et al.*, 2016), bisulfite sequencing (Squires *et al.*, 2012), m^5C-RIP (Edelheit *et al.*, 2013), Aza-IP (Edelheit *et al.*, 2013), miCLIP (Hussain *et al.*, 2013), ψ-seq (Lovejoy *et al.*, 2014), CeU-seq (Li *et al.*, 2015) and ICE-seq (Suzuki *et al.*, 2015). It should be noticed that all these methods are not able to simultaneously detect multiple modifications.

Reminiscent of the histone code, different kinds of RNA modifications might mediate the biological processes in a combinational pattern. For example, Xiang *et al.* found the interaction between m^6A 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 m^5C, m^1A and ψ 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 m^6A 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-Methyl (Chen *et al.*, 2015), SRAMP (Zhou *et al.*, 2016), iRNA-PseU (Chen *et al.*, 2016), PPUS (Li *et al.*, 2015), iRNA-2OM (Yang *et al.*, 2018), iRNAD (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 m^6A and m^1A (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 m^6A, m^5C, m^1A, ψ 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.

Table 1. The datasets used in this article

Species	a	b	c	d	e	f
H.sapiens	m^1A	6366	6366	41	RAMPred	80%
	m^5C	120	120	41	Feng *et al.*	70%
	m^6A	1130	1130	41	MethyRNA	60%
	ψ	495	495	21	XG-PseU	60%
	A-to-I	3000	3000	51	iRNA-AI	60%
S.cerevisiae	m^1A	483	483	41	RAMPred	80%
	m^5C	211	211	41	iRNA-m5C	—
	m^6A	1307	1307	51	iRNA-Methyl	—
	ψ	313	314	31	XG-PseU	60%
M.musculus	m^1A	1064	1064	41	RAMPred	80%
	m^5C	97	97	41	iRNA-m5C	—
	m^6A	725	725	41	MethyRNA	60%
	ψ	472	472	21	XG-PseU	60%

The columns from a to f represent the name of modification, number of positive samples, number of negative samples, sample length, source of data and sequence similarity, respectively.

2 Materials and methods

2.1 Benchmark dataset

In the present work, the benchmark datasets used to train and test the proposed methods were collected from previous works (Chen *et al.*, 2015, 2016, 2017a, b; Feng *et al.*, 2016; Liu *et al.*, 2020, Lv *et al.*, 2019). These datasets cover five distinct kinds of RNA modifications, namely m^1A, m^6A, m^5C, ψ and A-to-I from *H.sapiens*, *M.musculus* and *S.cerevisiae*, respectively. They can be downloaded from http://www.bioml.cn/XG_iRNA/download, and their detail information is provided in Table 1.

2.2 Feature extraction

Since the length of RNA samples in the dataset is different for distinct kinds of RNA modifications (m^6A, m^5C, m^1A, ψ or A-to-I), for the convenience of analysis, the RNA samples were represented as following,

$$\boldsymbol{R} = R_{-\xi}R_{(-\xi+1)}\cdots R_{-1}\mathbf{M}R_1\cdots R_{(\xi-1)}R_{\xi} \qquad (1)$$

where the bold-font **M** represents the center nucleotide (A, C or U). ξ is an integer. R_{ξ} is the ξth downstream nucleotide and $R_{-\xi}$ is the ξth upstream nucleotide from the center, respectively. The $(2\xi+1)$-long RNA sample could be a true or false modification site containing sequence according to whether the center nucleotide is one of the five types of RNA modifications or not.

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 method, 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

$$D_k = \left[f_1^k, f_2^k, \cdots f_i^k \cdots, f_{4^k}^k \right] \tag{2}$$

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, 1, 0, 0), C as (0, 0, 1, 0) and G as (0, 0, 0, 1). The RNA sample in the dataset will be converted to a $4 \times (2\xi + 1)$-dimensional vector.

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

Nucleotide density

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

$$d_j = \frac{1}{||S_i||} \sum_{j=1}^{p} f(N_j) \tag{3}$$

where d_j is the density of the nucleotide N_j at position i of RNA sample, the length of the sliding substring is $||S_i||$, p is the position of corresponding nucleotide.

$$f(N_j) = \begin{cases} 1, \text{ if } \; N_j \; \textit{is the corresponding nucleotide} \\ \qquad 0, \; \textit{other nucleotide} \end{cases} \tag{4}$$

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

$$x_i = \begin{cases} 1, if \;\; N_i \in \{A, G\} \\ 0, if \;\; N_i \in \{C, U\} \end{cases}; y_i = \begin{cases} 1, if \;\; N_i \in \{A, C\} \\ 0, if \;\; N_i \in \{G, U\} \end{cases}; \\ z_i = \begin{cases} 1, if \;\; N_i \in \{A, U\} \\ 0, if \;\; N_i \in \{C, G\} \end{cases} \tag{5}$$

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

Dinucleotide physicochemical properties

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

$$\text{DPCP}(i) = f \;\; (i) \times \;\; \text{dpcp}(i)_j \tag{6}$$

where f is frequency of the dinucleotide i. dpcp(i) is the value of the jth ($j = 1, 2, \ldots, 11$) DPCP for the ith dinucleotide and is also listed in Supplementary Table S1. Therefore, the dimension of DPCP is (16×11).

By integrating all these features, a RNA sample in the dataset will be converted into a $(1548 + 2\xi)$-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,

$$\begin{cases} \text{Sen} = 1 - \dfrac{N_-^+}{N^+} \\ \text{Spe} = 1 - \dfrac{N_+^-}{N^-} \\ \text{Acc} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ N^-} \\ \text{MCC} = \dfrac{1 - \dfrac{N_-^+ + N_+^-}{N^+ N^-}}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} \end{cases} \tag{7}$$

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 (Fushing 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.

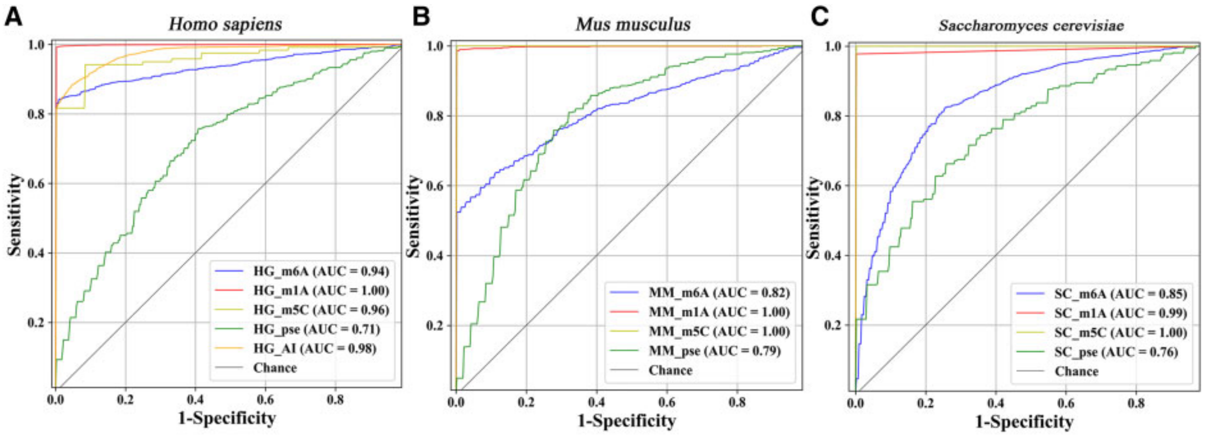


Fig. 2. The ROC curves for identifying distinct kinds of RNA modifications in (**a**) *H.sapiens*, (**b**) *M.musculus* and (**c**) *S.cerevisiae*. The AUC was also provided at the right corner of each figure

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

Species	Modification	Sn (%)	Sp (%)	Mcc	Acc (%) this article	Acc (%) existing
H.sapiens	m^1A	99.04	99.78	0.988	99.41	99.13[a]
	m^5C	90.83	93.33	0.842	92.08	92.90[b]
	m^6A	82.48	99.56	0.820	91.02	90.38[c]
	ψ	62.00	67.11	0.293	64.55	64.24[d]
	A-to-I	87.33	95.80	0.831	91.57	90.71[e]
S.cerevisiae	m^1A	97.72	100	0.977	98.86	97.83[a]
	m^5C	99.05	100	0.991	99.52	100.00[b]
	m^6A	77.04	78.50	0.555	77.77	76.51[f]
	ψ	68.69	73.48	0.422	71.08	65.13[d]
M.musculus	m^1A	98.49	99.90	0.984	99.20	98.73[a]
	m^5C	97.94	98.97	0.969	98.45	100.00[b]
	ψ	76.90	69.28	0.462	73.09	70.44[d]
	m^6A	78.34	99.57	0.779	88.97	88.39[c]

[a], [b], [c], [d], [e] and [f] appearance in the table represent RAMPred (Chen *et al.*, 2016), iRNA-M5C (Lv *et al.*, 2019), MethyRNA (Chen *et al.*, 2017), PseUI (He *et al.*, 2018), iRNA-AI (Chen *et al.*, 2017) and RNA-MethylPred (Jia *et al.*, 2016), respectively. M stands for modification site.

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.

To demonstrate the superiority of the proposed methods for identifying RNA modifications, we compared them with state-of-the-art methods under the jackknife test. The jackknife test results obtained by the proposed methods for identifying RNA modifications are listed in Table 2, where, for facilitating comparison, listed are also the corresponding results obtained by the best of the existing predictors for the same aim. As we can see from Table 2, our proposed methods outperforms the best existing predictor for identifying m^1A, m^6A, ψ, A-to-I modifications in *H.sapiens*, *M.musculus* and *S.cerevisiae*. Although the accuracies of the proposed methods are not the best for identifying m^5C, they remain comparable to that obtained by the best predictor iRNA-m5C (Lv *et al.*, 2019).

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), naïve 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 m^5C in *M.musculus*, and identifying m^1A and m^5C 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 m^6A, m^5C, m^1A, ψ 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

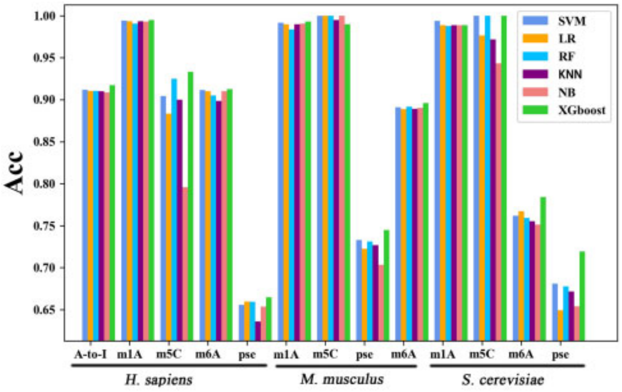


Fig. 3. Comparison of different algorithms for identifying RNA modification sites in *H.sapiens*, *M.musculus* and *S.cerevisiae*

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

	SVM	LR	RF	KNN	NB	XGboost
hg_AI	91.20	91.03	91.03	91.00	90.85	**91.73**
hg_m1A	99.43	99.35	99.04	99.35	99.29	**99.47**
hg_m5C	90.42	88.33	92.50	90.00	79.58	**93.33**
hg_m6A	91.15	91.02	90.49	89.87	91.02	**91.28**
hg_pse	65.56	65.96	65.94	63.60	65.36	**66.47**
mm_m1A	99.15	98.67	98.36	98.87	99.06	**99.29**
mm_m5C	**100.00**	**100.00**	**100.00**	99.50	**100.00**	99.00
mm_pse	73.30	72.25	73.11	72.67	70.35	**74.48**
mm_m6A	89.10	88.89	89.17	88.90	89.04	**89.59**
sc_m1A	**99.38**	98.86	98.76	98.87	98.87	98.87
sc_m5C	**100.00**	97.64	**100**	97.19	94.32	**100.00**
sc_m6A	76.20	76.74	75.93	75.55	75.13	**78.41**
sc_pse	68.10	64.91	67.78	67.17	65.39	71.91

Note: The best accuracies are in bold font.

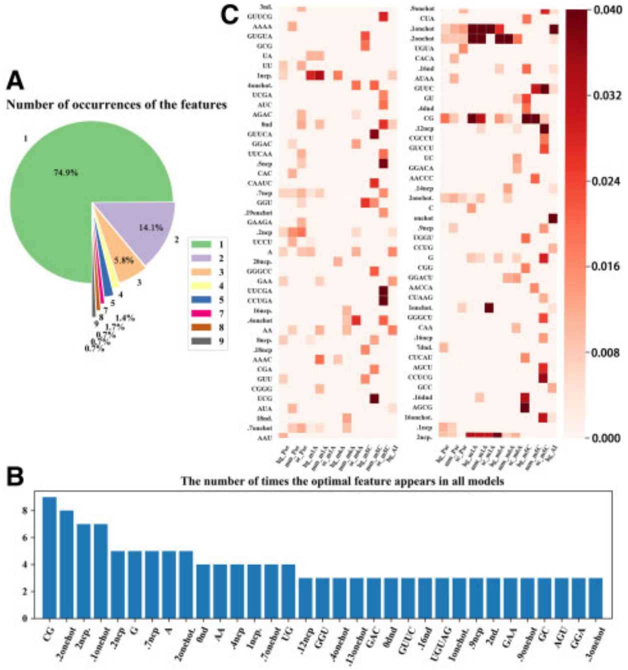


Fig. 4. Optimal features analysis. (**a**) From 1 to 9 is the number of occurrences of optimal features in the 13 models. (**b**) The optimal features with the occurrences greater than 2 in the 13 models. The ordinate is the occurrence, and the abscissa is the features. The '.' in the feature name represents RNA modification site, the number is the position relative to the modification site and the remaining characters indicate the sequence encoding method (see Section 2). For example, '.2onehot' stands for the scond nucleotide downstream of the modification site and encoded by using the one-hot encoding method. '2ncp.' stand for the second nucleotide upstream of the modification site and encoded by using the nucleotide chemical property method. (**c**) A heat map to describe the F-scores of the feature with the F-Score higher than 0.01. 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 type, and the ordinate is the features in the 13 models. The deeper the boxes color, the higher the feature's *F*-score

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 '.' 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 and so forth.

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 encoded using different colors according to their F-scores (Supplementary Fig. S14). It was found that the contributions of the features are different for identifying RNA modifications in different species.

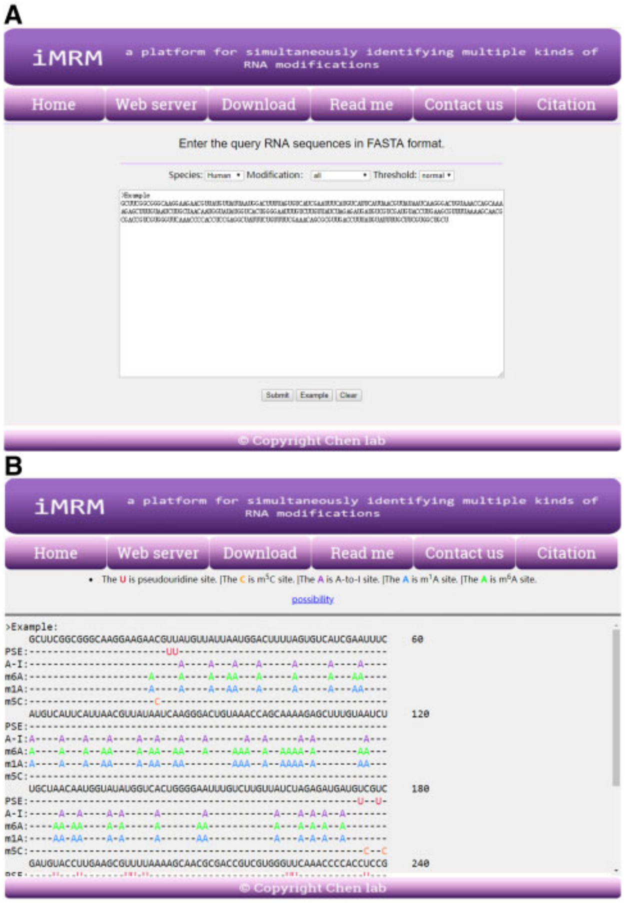
For more clarity, we picked out the features with F-scores greater than 0.01 for further analysis, Figure 4c. It was found that some features, such as CG, '2ncp.', '.2onehot' and '.1onehot', are used in more than eight models, indicating that they are universal features for identifying distinct kinds of RNA modifications in different species. In contrast, some other features, such as UCG, AUA and AUC, only appeared in less than three models or even appeared in one model, indicating that they are specific features for identifying a specific kind of RNA modifications in a specific species. These results suggest that it is reasonable to extract the sequence-based information to construct prediction models for identifying RNA modifications.

3.4 Web server

For the convenience of scientific community, an online web server called iMRM was built to simultaneously identify m^{1}A, m^{6}A, m^{5}C, ψ and A-to-I modifications in *H.sapiens*, *M.musculus* and *S.cerevisiae*, respectively. The iMRM can be accessed at http://www.bioml.cn/XG_iRNA/home.

The user guide of iMRM is as following. Open the home page at http://www.bioml.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 predictive results will appear in a new page as shown in Figure 5b. The identified distinct kinds of modifications are highlighted in different colors. User can click the 'possibility' link to download the predicted results.

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.



Enter the query RNA sequences in FASTA format.

Species: Human ▾ Modification: all ▾ Threshold: normal ▾

>Example
[illegible]

- The U is pseudouridine site. |The C is m^{5}C site. |The A is A-to-I site. |The A is m^{1}A site. |The A is m^{6}A site.

```
>Example:
    GCUUCGGCGGGCAAGGAAGAACGUUAUGUUAUUAAUGGACUUUUAGUGUCAUCGAAUUUC    60
PSE:-----------------------UU-----------------------------------
A-I:-------------------------A----A---A---A-----A-----A----A----
m6A:--------------------A----A----A--AA---A-----A-----A---AA----
m1A:--------------------A----A----A--AA---A-----A-----A---AA----
m5C:---------------------C--------------------------------------
    AUGUCAUUCAUUAACGUUAUAAUCAAGGGACUGUAAACCAGCAAAAGAGCUUUGUAAUCU    120
PSE:------------------------------------------------------------
A-I:A----A---A---A----A--A---A---A------A--A-----A-A--------A---
m6A:A----A---A--AA----A-AA--AA---A----AAA--A--AAAA-A-------AA---
m1A:A----A---A--AA----A-AA--AA--------AAA--A--AAAA-A-------AA---
m5C:------------------------------------------------------------
    UGCUAACAAUGGUAUAUGGUCACUGGGGAAUUUGUCUUGUUAUCUAGAGAUGAUGUCGUC    180
PSE:-------------------------------------------------------U--U-
A-I:-----A--A----A-A-----A-------A-----------A---A-A-A--A-------
m6A:----AA-AA----A-A-----A------AA-----------A---A-A-A--A-------
m1A:----AA-AA----A-A-----A------AA-----------A---A-A-A--A-------
m5C:--------------------------------------------------------C--C
    GAUGUACCUUGAAGCGUUUUAAAAGCAACGCGACCGUCGUGGGUUCAAACCCCACCUCCG    240
```

Fig. 5. A semi-screenshot for the web server of iMRM available at http://www.bioml.cn/XG_iRNA/home. (**a**) Web server interface of iMRM. (**B**) The predictive results returned obtained by iMRM. The identified distinct kinds of modifications are highlighted in different colors

4 Conclusion

In this work, we developed a XGboost-based prediction tool that can simultaneously identify multiple kinds of RNA modifications in *H.sapiens*, *M.musculus* and *S.cerevisiae*, respectively. In order to verify the validity of the proposed method, we not only compared its performance with that of existing methods, but also with that of the other six machine learning algorithms. It is excited to find that the performance of the proposed method is quite good for simultaneously identifying RNA modifications. For the convenience of scientific community, a freely online web server for the proposed method was provided at http://www.bioml.cn/XG_iRNA/home. We hope it will become a useful tool for identifying RNA modifications.

The iMRM is only based on sequence information and nucleotide physicochemical properties. In fact, RNA modification is a complicated process, in which multiple kinds of enzymes are included (Liu *et al.*, 2019a). As demonstrated by Chen *et al.* (2019), the genomic features are also important for the identification of RNA modifications, which should be considered in developing computational methods for identifying RNA modification sites.

Considering the availability of experimental data that can be used for model training, the current method is only able to identify m^{1}A, m^{6}A, m^{5}C, ψ and A-to-I modifications. With the appearance of smart sequencing methods, such as the Oxford Nanopore Technology (Garalde *et al.*, 2018; Smith *et al.*, 2019), the transcriptome-wide profiles of the other kinds of RNA modifications will be available. Therefore, in the future work, we will collect such data to 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

Abbasi-Moheb,L. *et al.* (2012) Mutations in NSUN2 cause autosomal-recessive intellectual disability. *Am. J. Hum. Genet.*, **90**, 847–855.

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

Bari,A.G. *et al.* (2013) DNA encoding for splice site prediction in large DNA sequence. In: *International Conference on Database Systems for Advanced Applications*. Springer, pp. 46–58.

Boccaletto,P. *et al.* (2018) MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.*, **46**, D303–D307.

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

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.*, **2**, 121–167.

Chen,K. *et al.* (2019) WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.*, **47**, e41.

Chen,T. and Guestrin,C. (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. *2016*. doi: 10.1145/2939672.2939785.

Chen,W. *et al.* (2015) iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*, **490**, 26–33.

Chen,W. *et al.* (2016) iRNA-PseU: identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids*, **5**, e332.

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

Chen,W. *et al.* (2017b) MethyRNA: a web server for identification of N(6)-methyladenosine sites. *J. Biomol. Struct. Dyn.*, **35**, 683–687.

Chen,W. *et al.* (2018) iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. *Mol. Ther. Nucleic Acids*, **11**, 468–474.

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.

Chen,Z. *et al.* (2019) Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. *Brief. Bioinform.* doi: 10.1093/bib/bbz112.

Cox,D.R. (1958) The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B*, **21**, 215–242.

Dominissini,D. *et al.* (2016) The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA. *Nature*, **530**, 441–446.

Du,T. *et al.* (2015) An association study of the m6A genes with major depressive disorder in Chinese Han population. *J. Affect. Disorders*, **183**, 279–286.

Edelheit,S. *et al.* (2013) Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. *PLoS Genetics*, **9**, e1003602.

Falk,M.J. *et al.* (2016) A novel HSD17B10 mutation impairing the activities of the mitochondrial RNase P complex causes X-linked intractable epilepsy and neurodevelopmental regression. *RNA Biol.*, **13**, 477–485.

Feng,P. *et al.* (2016) Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. *Mol. Biosyst.*, **12**, 3307–3311.

Fujiwara,T. and Harigae,H. (2013) Pathophysiology and genetic mutations in congenital sideroblastic anemia. *Pediatr. Int.*, **55**, 675–679.

Fushing,H. and Turnbull,B.W. (1996) Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann Statist.*, **24**, 25–40.

Fustin,J.M. *et al.* (2013) RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell*, **155**, 793–806.

Garalde,D.R. *et al.* (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods*, **15**, 201–206.

Guzzi,N. *et al.* (2018) Pseudouridylation of tRNA-derived fragments steers translational control in stem cells. *Cell*, **173**, 1204–1216, e26.

Han,L. *et al.* (2015) The genomic landscape and clinical relevance of A-to-I RNA editing in human cancers. *Cancer Cell*, **28**, 515–528.

He,J. *et al.* (2018) PseUI: Pseudouridinesites identification based on RNA sequence information. BMC Bioinformatics, **19**, 306.

Hussain,S. and Bashir,Z. (2015) The epitranscriptome in modulating spatiotemporal RNA translation in neuronal post-synaptic function. *Front. Cell. Neurosci.*, **9**, 420.

Hussain,S. *et al.* (2013) NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep.*, **4**, 255–261.

Jia,C.-Z. *et al.* (2016) RNA-MethylPred: A high-accuracy predictor to identify N6-methyladenosine in RNA. *Anal. Biochem.*, **510**, 72–75.

Jia,G. *et al.* (2011) N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.*, **7**, 885–887.

Jonkhout,N. *et al.* (2017) The RNA modification landscape in human disease. *RNA*, **23**, 1754–1769.

Ke,S. *et al.* (2015) A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev.*, **29**, 2037–2053.

Keller,J.M. *et al.* (1985) A fuzzy K-nearest neighbor algorithm. *IEEE Trans. Sys. Man Cybernetics*, **SMC-15**, 580–585.

Khan,M.A. *et al.* (2012) Mutation in NSUN2, which encodes an RNA methyltransferase, causes autosomal-recessive intellectual disability. *Am. J. Hum. Genet.* **90**, 856–863.

Khoddami,V. *et al.* (2019) Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc. Natl. Acad. Sci. USA*, **116**, 6784–6789.

Li,X. *et al.* (2015) Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat. Chem. Biol.*, **11**, 592–597.

Li,X. *et al.* (2016) Transcriptome-wide mapping reveals reversible and dynamic N(1)-methyladenosine methylome. *Nat. Chem. Biol.*, **12**, 311–316.

Li,Y.H. *et al.* (2015) PPUS: a web server to predict PUS-specific pseudouridine sites. *Bioinformatics*, **31**, 3362–3364.

Lichinchi,G. *et al.* (2016) Dynamics of human and viral RNA methylation during Zika virus infection. *Cell Host Microbe*, **20**, 666–673.

Linder,B. *et al.* (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods*, **12**, 767–772.

Liu,D. *et al.* (2019a) Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Brief. Bioinform.*, **20**, 1826–1835.

Liu,H. *et al.* (2019b) Accurate detection of m6A RNA modifications in native RNA sequences. *Nat. Commun.*, **10**, 4079.

Liu,K. *et al.* (2020) XG-PseU: an eXtreme gradient boosting based method for identifying pseudouridine sites. *Mol. Gen. Genom.*, **295**, 13–21.

Liu,N. *et al.* (2015) N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature*, **518**, 560–564.

Lovejoy,A.F. *et al.* (2014) Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. *PLoS One*, **9**, e110799.

Lv,H. *et al.* (2019) Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinform.*, doi: 10.1073/pnas.1817334116.

Manavalan,B. *et al.* (2019) Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucl. Acids*, **16**, 733–744.

Martinez,F.J. *et al.* (2012) Whole exome sequencing identifies a splicing mutation in NSUN2 as a cause of a Dubowitz-like syndrome. *J. Med. Genet.*, **49**, 380–385.

Metodiev,M.D. *et al.* (2016) Recessive mutations in TRMT10C cause defects in mitochondrial RNA processing and multiple respiratory chain deficiencies. *Am. J. Hum. Genet.*, **98**, 993–1000.

Morena,F. *et al.* (2018) Above the epitranscriptome: RNA modifications and stem cell identity. *Genes*, **9**, 329.

Paz,N. *et al.* (2007) Altered adenosine-to-inosine RNA editing in human cancer. *Genome Res.*, **17**, 1586–1595.

Qiang,X. *et al.* (2018) M6AMRFS: Robust prediction of N6-methyladenosine sites with sequence-based features in multiple species. *Front. Genet.*, **9**, 495.

Sasaki,S. *et al.* (2015) Autophagy in spinal motor neurons of conditional ADAR2-knockout mice: an implication for a role of calcium in increased autophagy flux in ALS. *Neurosci. Lett.*, **598**, 79–84.

Schwartz,S. *et al.* (2013) High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*, **155**, 1409–1421.

Smith,A.M. *et al.* (2019) Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One*, **14**, e0216709.

Squires,J.E. *et al.* (2012) Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.*, **40**, 5023–5033.

Suzuki,T. *et al.* (2015) Transcriptome-wide identification of adenosine-to-inosine editing using the ICE-seq method. *Nat. Protocol*, **10**, 715–732.

Tohru,F. and Hideo,H. (2018) Molecular pathophysiology and genetic mutations in congenital sideroblastic anemia. *Free Radic. Biol. Med.*, **133**, 179–185.

Wang,X. *et al.* (2014) N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, **505**, 117–120.

Wei,L. *et al.* (2019) Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics*, **35**, 1326–1333.

Xiang,J.F. *et al.* (2018) N(6)-methyladenosines modulate A-to-I RNA editing, *Mol. Cell*, **69**, 126–135, e126.

Xu,Z.C. *et al.* (2019) iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics*, **35**, 4922–4929.

Yang,H. *et al.* (2018) iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in homo sapiens. *J. Comput. Biol.*, **25**, 1266–1277.

Yang,X. *et al.* (2017) 5-methylcytosine promotes mRNA export- NSUN2 as the methyltransferase and ALYREF as an m5C reader. *Cell Res.*, **27**, 606–625.

Yi,J. *et al.* (2017) Overexpression of NSUN2 by DNA hypomethylation is associated with metastatic progression in human breast cancer. *Oncotarget*, **8**, 20751–20765.

Yu,B. *et al.* (2020) SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics*, **36**, 1074–1081.

Yu,J. *et al.* (2019) PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization. *Bioinformatics*, **35**, 2749–2756.

Zhang,H. *et al.* (2016) Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals, *Food Chem. Toxicol.*, **97**, 141–149.

Zhang,X. *et al.* (2012) The tRNA methyltransferase NSun2 stabilizes p16INK4 mRNA by methylating the 3′-untranslated region of p16. *Nat. Commun.*, **3**, 712.

Zhao,Z. *et al.* (2018) Imbalance learning for the prediction of N6-Methylation sites in mRNAs. *BMC Genomics*, **19**, 574.

Zhou,Y. *et al.* (2016) SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res.*, **44**, e91.

Zuo,Y. *et al.* (2017) PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics*, **33**, 122–124.