Briefings in Functional Genomics, 00(00), 2019, 1–10

doi: 10.1093/bfgp/elz018
Review paper

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019

A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of *Saccharomyces cerevisiae*

Xiaolei Zhu[†], Jingjing He[†], Shihao Zhao[†], Wei Tao, Yi Xiong and Shoudong Bi

Corresponding authors: Shoudong Bi, School of Sciences, Anhui Agricultural University, Hefei, Anhui 230036, China. Fax and Tel:+8655165786472; Email: bishoudong@163.com; Yi Xiong, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China. Fax and Tel:+862134204573; Email: xiongyi@sjtu.edu.cn
[†]The joint first authors

Abstract

N^6-methyladenosine (m6A) modification, as one of the commonest post-transcription modifications in RNAs, has been reported to be highly related to many biological processes. Over the past decade, several tools for m6A sites prediction of *Saccharomyces cerevisiae* have been developed and are freely available online. However, the quality of predictions by these tools is difficult to quantify and compare. In this study, an independent dataset M6Atest6540 was compiled to systematically evaluate nine publicly available m6A prediction tools for *S. cerevisiae*. The experimental results indicate that RAM-ESVM achieved the best performance on M6Atest6540; however, most models performed substantially worse than their performances reported in the original papers. The benchmark dataset Met2614, which was used as the training dataset for the nine methods, were further analyzed by using a position bias index. The results demonstrated the significantly different bias of dataset Met2614 compared with the RNA segments around m6A sites recorded in RMBase. Moreover, newMet2614 was collected by randomly selecting RNA segments from non-redundant data recorded in RMBase, and three different kinds of features were extracted. The performances of the models built on Met2614 and newMet2614 with the features were compared, which shows the better generalization of models built on newMet2614. Our results also indicate the position-specific propensity-based features outperform other features, although they are also easily over-fitted on a biased dataset.

Key words: N6-methyladenosine sites; computational predictor; dataset bias; position-specific propensity; web servers

Xiaolei Zhu is an associate professor at the School of Sciences, Anhui Agricultural University, China. His research interests include structural bioinformatics, machine learning and analysis of next-generation sequencing.
Jingjing He is currently an undergraduate student at the School of Life Sciences, Anhui University, China. Her research interests are biostatistics, machine learning and analysis of next-generation sequencing.
Shihao Zhao is currently a master student at the School of Sciences, Anhui Agricultural University, China. His research interests are computational biology, machine learning and data mining.
Wei Tao is currently an undergraduate student at the School of Sciences, Anhui Agricultural University, China. His research interests are applied statistics and mathematical modeling.
Yi Xiong is a research associate professor at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, China. His research interests are bioinformatics, machine learning and computational and systems biomedicine.
Shoudong Bi is the executive president and a professor at the School of Sciences, Anhui Agricultural University, China. He is also the director of the Institute of Applied Mathematics of Anhui Agricultural University, China. His research applies mathematical theories to study the ecosystem and biological systems.
© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
This article is published and distributed under the terms of the Oxford University Press, Standard Journals Publication Model (https://academic.oup.com/journals/pages/open_access/funder_policies/chorus/standard_publication_model)

Introduction

N^6-methyladenosine (m6A) modification is the first internal mRNA modification discovered [1, 2]. It is considered as one of the commonest post-transcription modifications in RNAs and has been founded in viruses [3, 4] and many different eukaryotes such as yeast [5, 6], plants [7, 8], mammals [1, 9–11] and insects [12]. m6A modification participates in a wide range of biological functional processes [13–17] and almost affects the entire mRNA metabolism [18–20], including splicing, transport, translation efficiency, secondary structure, etc. It was reported that m6A modification was associated with lots of diseases such as thyroid tumor [21], prostate cancer [22], breast cancer [23–25], pancreatic cancer [26, 27], leukemia [28], etc. Obviously, the identification of m6A sites would be of great benefit for cell biology and disease mechanism research.

Various types of methods have been developed to identify m6A sites. Experimental approaches such as two-dimensional thin layer chromatography [29], high performance liquid chromatography [30] and next-generation sequencing techniques (e.g. m6A-seq [31] and MeRIP-Seq [32]) have been used to identify m6A sites in mRNAs. However, these experimental methods are too costly and time-consuming to perform genome-wide analysis. These limitations of experimental methods may be overcome by developing computational methods for m6A sites identification.

Over the past few years, a variety of computational methods [33–56] have been developed for predicting m6A sites in different species of RNAs, such as *Saccharomyces cerevisiae*, *Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana*. *Saccharomyces cerevisiae* is one of the most widely utilized model organisms in biotechnology worldwide, and 13 computational methods [33, 34, 36, 38, 39, 42, 44, 45, 47, 48, 50–52] have been developed to predict its m6A sites. In this study, we focused on these methods. These models were summarized in Table 1; we noticed all the models were built on a training dataset Met2614 [33] which was based on the pioneering work of Schwartz *et al.* [57] and contains only 1307 positive examples. However, the number of m6A sites of *S. cerevisiae* collected in a RNA modification database (RMBase [58]) is over 60 000, of which 23 581 m6A sites are centered with GAC pattern.

Our original idea is to select a representative subset to validate the generalization of the available methods built on a small dataset Met2614. We first carefully collected an independent test set and carried out an unbiased evaluation of these predictors that have been publicly available as open access web

Table 1. A comprehensive summary of the state-of-the-art computational predictors for RNA m6A sites of *S. cerevisiae*

Predictor	Training dataset	Independent test set	Feature representation	Classifier	Cross validation	Web server	Ref
iRNA-Methyl	Met2614	None	Pseudo dinucleotide composition (PseDNC)	SVM	10-fold and Jackknife	http://lin.uestc.edu.cn/server/iRNA-Methyl	[33]
m6Apred	Met1664[a]	Met950[a]	Chemical property with density	SVM	Jackknife	http://lin.uestc.edu.cn/server/m6Apred	[34]
pRNAm-PC	Met2614	None	Auto-covariance and cross-covariance transformations of physical chemical property	SVM	Jackknife	http://www.jci-bioinfo.cn/pRNAm-PC	[36]
RNA-MethylPred	Met2614	None	Bi-profile Bayes; dinucleotides composition; KNN score	SVM	Jackknife	None	[38]
	Met1664[a]	Met950[a]					
TargetM6A	Met2614	None	Nucleotide composition; position-specific nucleotide/dinucleotide propensity	SVM	10-fold and Jackknife	http://csbio.njust.edu.cn/bioinf/TargetM6A	[42]
	Met1664[a]	Met950[a]					
M6A-HPCS	Met2614	None	PseDNC; auto-covariance and cross-covariance transformations of physical–chemical property	SVM	10-fold and Jackknife	http://csbio.njust.edu.cn/bioinf/M6A-HPCS	[39]
RAM-ESVM	Met2614	None	PseDNC; motif features; gapped K-mers	SVM	10-fold and Jackknife	http://server.malab.cn/RAM-ESVM/	[44]
RAM-NPPS	Met2614	None	Position-specific condition propensity	SVM	Jackknife	http://server.malab.cn/RAM-NPPS/	[45]
M6APred-EL	Met2614	None	Position-specific nucleotide propensity; physical–chemical properties; ring-function-hydrogen-chemical properties	SVM	10-fold	http://server.malab.cn/M6APred-EL/	[47]
iMethyl-STTNC	Met2614	None	Split-tetra-nucleotide-composition	SVM	Jackknife	None	[48]
iRNA(m6A)-PseDNC	Met2614	None	PseDNC	SVM	10-fold	http://lin-group.cn/server/iRNA(m6A)-PseDNC.php	[50]
BERMP	Met2200[a]	Met414[a]	Enhanced nucleic acid composition	RF BGRU LR	5-fold and 10-fold	http://www.bioinfogo.org/bermp	[51]
M6AMRFS	Met2614	None	Dinucleotide binary encoding; local position-specific dinucleotide frequency	XGBoost	10-fold and Jackknife	http://server.malab.cn/M6AMRFS/	[52]

[a]The datasets are derived from Met2614, so all the datasets are not overlapped with M6Atest6540.
LR, logistic regression; SVM, support vector machine; RF, random forest; BGRU, bidirectional Gated Recurrent Unit; XGBoost, eXtreme Gradient Boosting.

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019

portals. As shown in Table 1, totally 11 predictors have web services as reported in their original papers; however, two of them, TargetM6A [42] and M6A-HPCS [39], are not open-access now. Therefore, we tested and compared the rest nine m6A prediction webservers in this study. An independent test set, M6Atest6540 including 3270 positive examples, is collected from RMBase [58]. Our comparative results indicate that RAM-ESVM achieved the best performance on M6Atest6540; however, the generalization and robustness of all nine methods are not good compared with the performances on Met2614. The RNA segments in Met2614 were further compared with the segments around m6A sites recorded in RMBase, and the results indicate the bias of Met2614 compared with the segments recorded in RMBase.

To further analyze if the bias of Met2614 affects the generalization of the models, we built new datasets by randomly selecting 1307 nonredundant positive examples from RMBase, which was called newMet2614. Then, three different kinds of features were extracted from the RNA segments, namely nucleotide composition-based features, position-specific propensity-based features and physical–chemical properties-based features. With these features, the performances of the models built on newMet2614 were compared with the models built on Met2614. The results indicate that generalization of the models based on newMet2614 is better than the models based on Met2614. Our results also indicate the position-specific propensity-based features outperform other features; however, they are also easily over-fitted on a biased dataset.

Materials and methods

Datasets

The dataset Met2614 [33] contains total 1307 positive RNA segments with experimentally verified m6A sites at the center, which were collected from Schwartz *et al.*'s work [57]. In order to fairly compare the predictive performance of different web tools, we constructed an independent dataset, M6Atest6540, which contains 3270 experimentally verified m6A sites and other remaining non-m6A sites. Note that benchmark dataset Met2614 and M6Atest6540 are mutually exclusive.

The following seven steps were conducted to create the benchmark dataset M6Atest6540 (Figure 1): (i) collecting all RNA segments of 51-tuple nucleotides with exactly the same RGAC consensus motif at the center position by sliding a flexible window along each RNA sequences transcribed from the *S. cerevisiae* genome. This set of segments obtained here is called T1. (ii) Collecting all RNA segments whose center positions were recorded in RMBase [58], this set of segments is called P1 because all examples are positive. Then P1 was removed from T1 to obtain the set of segments N1, which contains all the negative examples. (iii) A subset of P1 with 'SupportNum' > 15 and 'Region' == 'cds' was selected to increase the reliability of positive examples. 'SuppotNum' is a term in RMBase to describe the number of supporting experiments or studies for that position, and 'Region' is another term to describe the region the nucleotide located. This subset is called P2. (iv) CD-Hit [59] was used to remove the redundancy [60] of all the positive examples of P2 and Met2614 with the sequence identity cutoff set as 75%. A total of 3270 positive examples were obtained from this step. (v) The negative examples of Met2614 were removed from N1 to get a new negative dataset N2. (vi) A total of 3270 negative examples were randomly selected from N2. (vii) The 3270 positive samples and 3270 negative samples were assembled into the benchmark

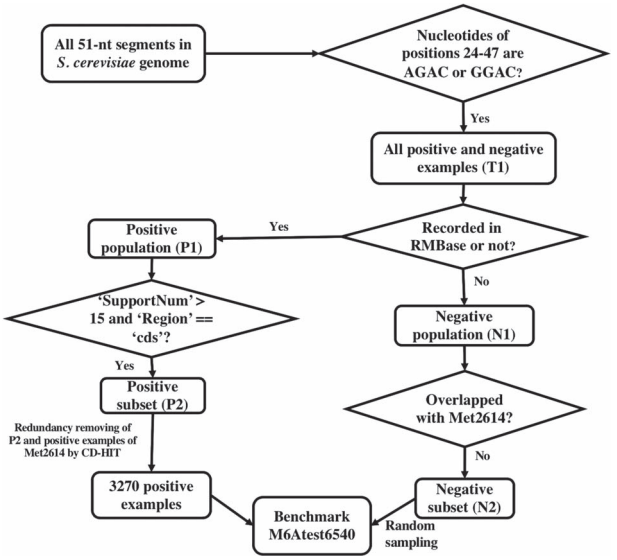

Figure 1. The flowchart for generating the benchmark dataset M6Atest6540.

dataset M6Atest6540. All the RNA segments of M6Atest6540, P1 and N1 can be found in Table S1, S2, S3, respectively.

Web-accessible prediction tools

As shown in Table 1, there are totally 13 methods for predicting m6A sites of *S. cerevisiae*, and 11 of them provide web servers. Currently, only 9 of these 11 web servers are open-access for users to predict m6A sites, namely iRNA-Methyl [33], m6Apred [34], pRNAm-PC [36], RAM-ESVM [44], RAM-NPPS [45], M6APred-EL [47], iRNA(m6A)-PseDNC [50], BERMP [51] and M6AMRFS [52]. Table 1 lists the basic information of these methods. As shown in Table 1, the validation and test datasets are all based on Met2614. Notably, the outputs of the servers pRNAm-PC and RAM-ESVM are color labeled, which makes it a little bit hard to collect the results.

Features extraction for machine learning-based predictors

In the process of training a model, it is crucial to extract informative, discriminative and independent features for converting the RNA sequence into the numeric vector. Among the past studies, many features have been extracted to interpret and keep the sequence information [44, 61–66]. In this study, three groups of features were gathered to analyze the possible reasons that caused the weak generalization of the web tools and to study the effects of benchmark dataset bias on the performances of different features for m6A sites prediction. The three groups of features are nucleotide composition-based features, position-specific propensity-based features and physical–chemical properties-based features (Table 2). The detailed descriptions of all the features can be found in Supplementary Data.

Support vector machine

The basic idea of the support vector machine (SVM) is to determine the optimal separating hyperplane that can correctly

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019

Table 2. Three groups of features that had been extracted for predicting m6A sites

Group	Abbreviation	Full name	Dimension	Reference
Nucleotide composition based	NC	Nucleotide composition	4	[42]
	DC	Dinucleotide composition	16	[38, 42, 43, 46, 49]
	TC	Trinucleotide composition	64	[42, 43, 46, 49]
	KSNPF	K-spaced nucleotide pair frequencies	16	[46]
	PseKNC	Pseudo K-tuple nucleotide composition	$4+\alpha$	[39, 50, 60]
Position-specific propensity based	PSNP	Position-specific nucleotide propensity	ξ	[42, 46, 47]
	PSDP	Position-specific dinucleotide propensity	$\xi-\lambda-1$	[42, 46]
	PSCP	Position-specific condition propensity	$\xi-\lambda-1$	[45]
Physical–chemical properties based	CPD	Chemical property with density	4ξ	[34, 37, 40, 47, 49]
	PCP	Physical–chemical properties	$100\lambda_m$	[36, 39, 47]

ξ represents the length of the RNA sequence; λ is the symbol of the intervals between two individual nucleotides in the nucleotide pair; λ_m is maximum values of the λ; α is the number of the total pseudo components used to show the long-range sequence effect.

divide the training data set and have the largest interval for linearly separable data [67, 68]. In terms of the nonlinearly separable data, SVM maps the original data to high-dimensional space by using the specific kernel, which transforms the problem into linear one. Kernel functions include linear kernels, polynomial kernels, Gaussian kernels, sigmoid kernels and so forth. Among them, Gaussian kernels, also called Radial Basis Function (RBF), are most commonly used [69–72], which can map data to infinite dimensions. There are two indispensable parameters for SVM models with RBF, which are C and gamma. The former indicates the tolerance of the model to the error, and the latter implicitly determines the distribution of the data after mapping to the new feature space. As shown in Table 1, most of the models were trained by using SVM.

Performance evaluation parameters

According to the previous related studies [73, 74], Sensitivity (SN), Specificity (SP), Accuracy (ACC) and Matthews correlation coefficient (MCC) are the mostly used performance evaluation parameters for the computational predictors, which are defined as follows:

$$\begin{cases} SN = \dfrac{TP}{TP+FN} \\ SP = \dfrac{TN}{TN+FP} \\ ACC = \dfrac{TP+TN}{TP+FP+TN+FN} \\ MCC = \dfrac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \end{cases} \tag{1}$$

where, TP, TN, FP and FN represent the number of true positive, true negative, false positive and false negative, respectively. SN and SP show the ability of the model to correctly predict positive samples and negative samples, respectively. Furthermore, it is worth noting that ACC and MCC are the most important ones of the aforementioned four metrics, because the former reflects the overall accuracy of the predictor, while the latter signified the overall stability.

Position bias index

Here, we introduced a position bias index (PBI) to evaluate the bias at a specified position between two groups of sequences.

Table 3. Performance of the nine accessible RNA M6A predictors on the dataset M6Atest6540

Prediction tools	SN (%)	SP (%)	ACC (%)	MCC
BERMP	26.54	**86.76**	56.65	0.167
M6AMRFS	37.13	72.11	54.62	0.099
M6APred-EL	40.83	75.38	58.10	0.173
RAM-NPPS	34.59	71.07	52.83	0.061
RAM-ESVM	59.27	64.53	**61.90**	**0.238**
pRNAm-PC	55.72	65.84	60.78	0.217
iRNA(m6A)-PseDNC	**83.85**	19.91	51.88	0.049
m6Apred	30.31	78.90	54.60	0.105
iRNA-Methyl	59.82	61.68	60.75	0.215

The biggest values for different evaluation parameters are shown in bold.

Suppose we have two groups of sequences with m and n sequences of the same length, respectively; the PBI was defined as follows:

$$PBI(k) = \sum_{i=1}^{t} \left(P^{i}_{1,k} - P^{i}_{2,k}\right)^2 \tag{2}$$

where k is the specified sequence position, t is the number of nucleotide or amino acid types and $P^{i}_{g,k}$ $(g \in \{1,2\})$ is the probability of the ith type residue or nucleotide at the specified position k of the sequence group g, which was defined as follows:

$$P^{i}_{g,k} = N^{i}_{g,k}/N_g \tag{3}$$

where N_g $(N_1 = m, N_2 = n)$ is the number of sequences in group g and $N^{i}_{g,k}$ is the number of ith type residue at the specified position k of the sequence group g.

Generally speaking, the bigger the PBI is, the more biased the corresponding position is. However, the bias is also determined by the distribution of PBIs at the specified position.

Results and Discussion

Comparative results on the independent dataset M6Atest6540

Our original objective of this study is to conduct a fair comparison of existing prediction methods with the help of an independent dataset. We first built the independent dataset M6Atest6540 as described in the Material and Methods section. Then, we compared the nine existing prediction methods with

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019

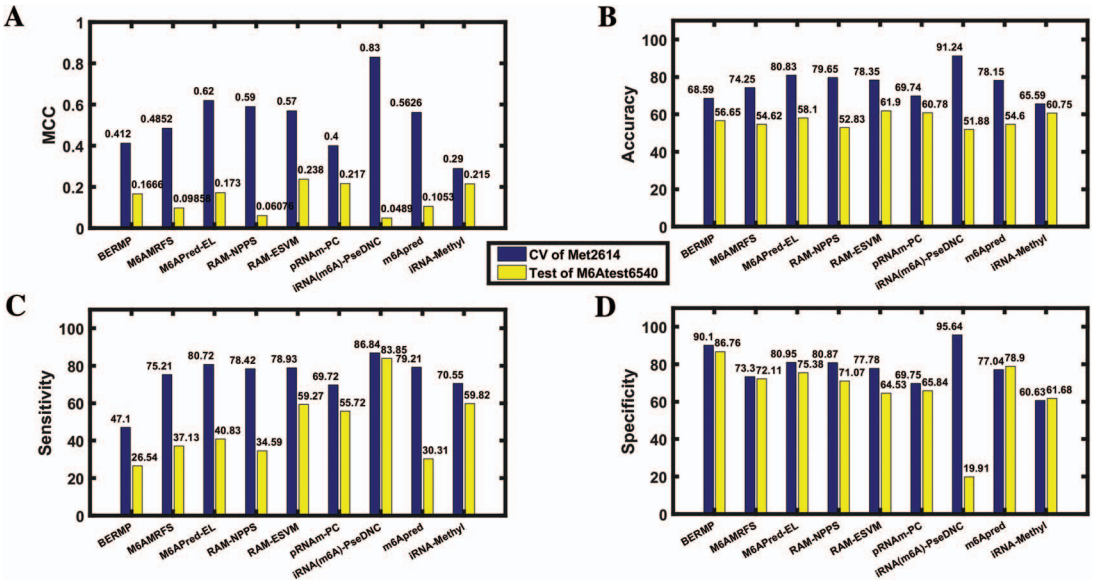

Figure 2. The performances of different models on Met2614 and M6Atest6540.

available web servers, namely, M6APred-EL [47], RAM-NPPS [45], RAM-ESVM [44], pRNAm-PC [36], iRNA-Methyl [33], BERMP [51], M6AMRFS [52], iRNA(m6A)-PseDNC [50] and m6Apred [34]. The predictive results of these methods on the independent dataset are shown in Table 3 and Figure 2, and the predictive labels of these methods for the RNA segments in M6Atest6540 are shown in Table S1.

As shown in Table 3, RAM-ESVM achieved the best performance among the nine tested predictors, with the highest ACC of 61.90% and MCC of 0.238. pRNAm-PC and iRNA-Methyl are the 2nd and 3rd best performers, which achieved ACC of 60.78%, 60.75% and MCC of 0.217, 0.215, respectively. Besides, BERMP achieved the highest specificity of 86.76%, and iRNA(m6A)-PseDNC achieved the highest sensitivity of 83.85%.

Moreover, we also tried to plot the receiver operating characteristics (ROCs) curves for all the predictive results of the nine servers; however, only six of them outputted scores or probabilities. So, we only plotted the ROC curves (Figure 3) for the six methods. The area under the curves are 0.652, 0.649, 0.638, 0.577, 0.574 and 0.533 for M6APred-EL, iRNA-Methyl, BERMP, M6AMRFS, m6Apred and iRNA(m6A)-PseDNC, respectively, which is basically in line with the ACCs in Table 3, except M6APred-EL.

In addition, we compared the performances of these models on their original training dataset and the independent dataset M6Atest6540 to evaluate the generalization of these models. Figure 2 shows the comparison results. MCCs of these models on the training dataset Met2614 and the independent test set M6Atest6540 were shown in Figure 2A, and it is surprising that the generalization of most models is not good, especially those models with high MCCs. If the difference of MCCs between the training dataset and the independent dataset is considered as the parameter to evaluate the generalization, the three best performing models are iRNA-Methyl, pRNAm-PC and BERMP, with differences of MCCs of 0.075, 0.183 and 0.245, respectively. Because MCC is an integral parameter for model evaluation, the sensitivities and specificities of models on the training dataset and the independent dataset were also compared to further explore the cause of the poor generalization. Figure 2C and D show that, for all models except iRNA(m6A)-PseDNC, the low sensitivity on the independent dataset is the major cause of

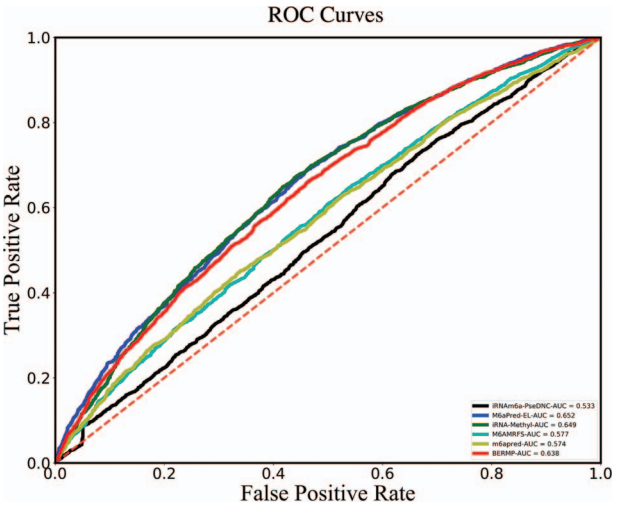

Figure 3. The ROC curves for the predictive results of the six servers on M6Atest6540.

the poor generalization. For iRNA(m6A)-PseDNC, the low specificity on the independent dataset is the main reason. Compared with training datasets used in other methods, the training dataset of iRNA(m6A)-PseDNC contains different negative examples, which may explain the different causes of poor generalization of iRNA(m6A)-PseDNC.

Analysis of the bias of the training dataset Met2614

Given that Met2614 is an early small dataset which was collected from Schwartz *et al.*'s work [57], we analyzed if the bias between the positive and negative samples in Met2614 is significantly different from the bias between the positive and negative samples recorded in RMBase. By using the PBI defined in the Material and Methods section, the position bias of the training dataset Met2614 was analyzed. As shown in Figure 4, three positions of the dataset Met2614, 22, 24 and 30, are substantially biased. Intuitively, the two positions, 24 and 30, are more seriously

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019

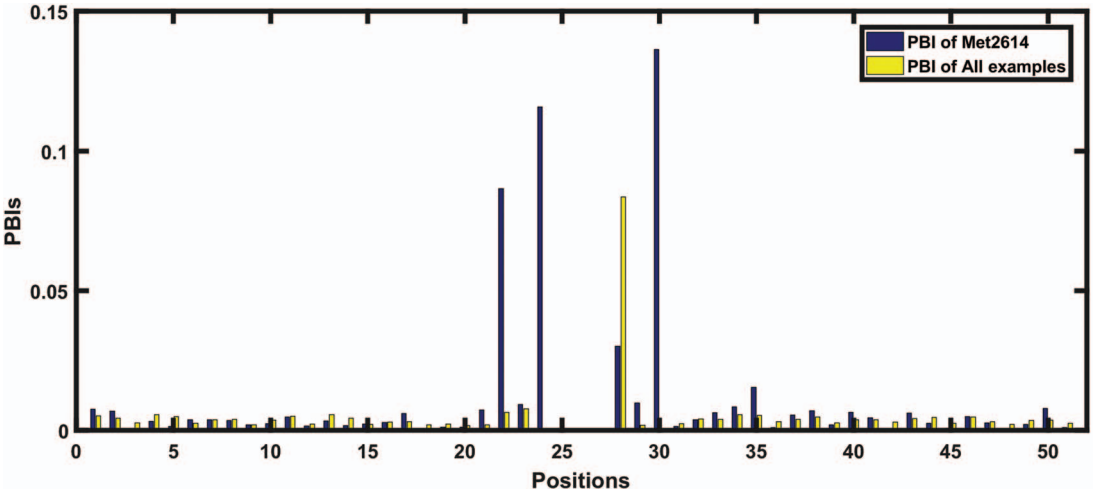

Figure 4. The PBIs for 51 positions of the RNA segments from Met2614 and the *S. cerevisiae* genome.

Table 4. The statistics and parameters of the PBI distribution at positions 24 and 30 of the length 51 sequence segment

Position	Mean of PBI	Std of PBI	Scale parameter[a]	Shape parameter[a]
24	7.5199E-04	0.0010	0.0006	0.6827
30	0.0012	0.0010	0.0013	1.2678

[a]These are the two parameters for Weibull distribution.

biased. To show if the bias is statistically significant, we focused on these two positions.

PBIs were first calculated based on all the positive and negative examples in the whole *S. cerevisiae* genome. As mentioned in the Material and Methods section, all the positive and negative examples were generated during generation of the dataset M6Atest6540. As shown in Figure 4, the PBIs at positions 24 and 30 of Met2614 are substantially higher than the corresponding PBIs for the whole genome.

Furthermore, 5000 different datasets were created by random selection of 1307 positive examples and 1307 negative examples from the genome, respectively. The PBIs at each position of the 5000 datasets were calculated, and the distributions of PBIs at positions 24 and 30 are shown in Figure S1. The distributions were fitted by the Weibull distribution. The statistics and parameters of the distributions for positions 24 and 30 are summarized in Table 4. The cumulative distribution function of these two distributions are showed in Figure S2. Based on the fitted distribution, the *P*-values of PBIs at positions 24 and 30 of Met2614 were calculated to be 1.652E-16 and 1.432E-158, respectively, which proved the bias of the dataset Met2614.

The data were further analyzed to explore the possible causes of the bias. Firstly, the nucleotide composition of the positive examples of Met2614 were compared with that of the positive examples recorded in RMBase [58] at positions 24 and 30. Figure 5A shows that the nucleotide composition of positive examples of Met2614 is substantially different from that of the positive examples recorded in RMBase at these two positions, while Figure 5B shows that the nucleotide composition of the negative examples of Met2614 is similar to that of the negative examples recorded in the genome at these two positions. Thus, the bias is more likely to be caused by the positive examples of Met2614.

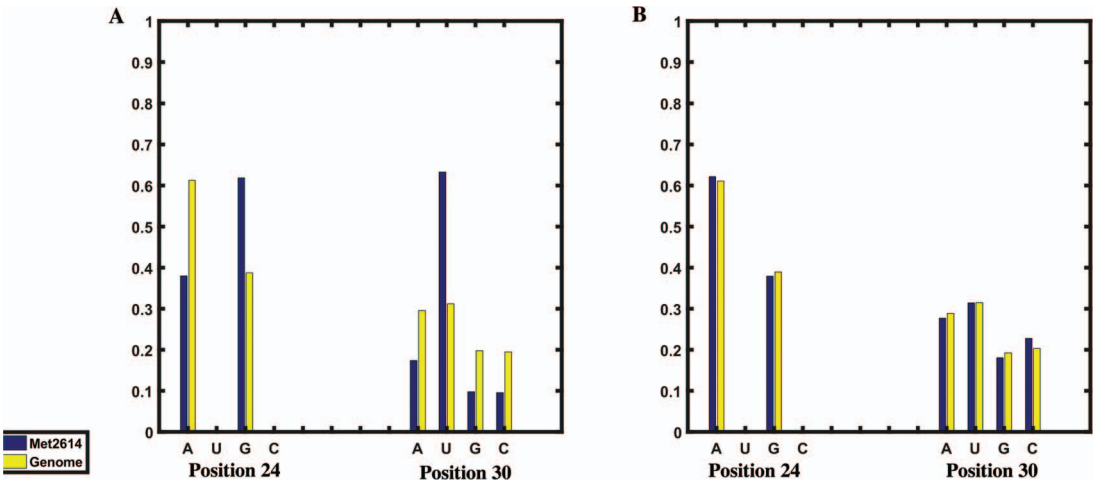

Figure 5. The nucleotide composition of positions 24 and 30 of RNA segments in Met2614 (blue) and the *S. cerevisiae* genome (yellow). (A) Positive examples; (B) negative examples.

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019

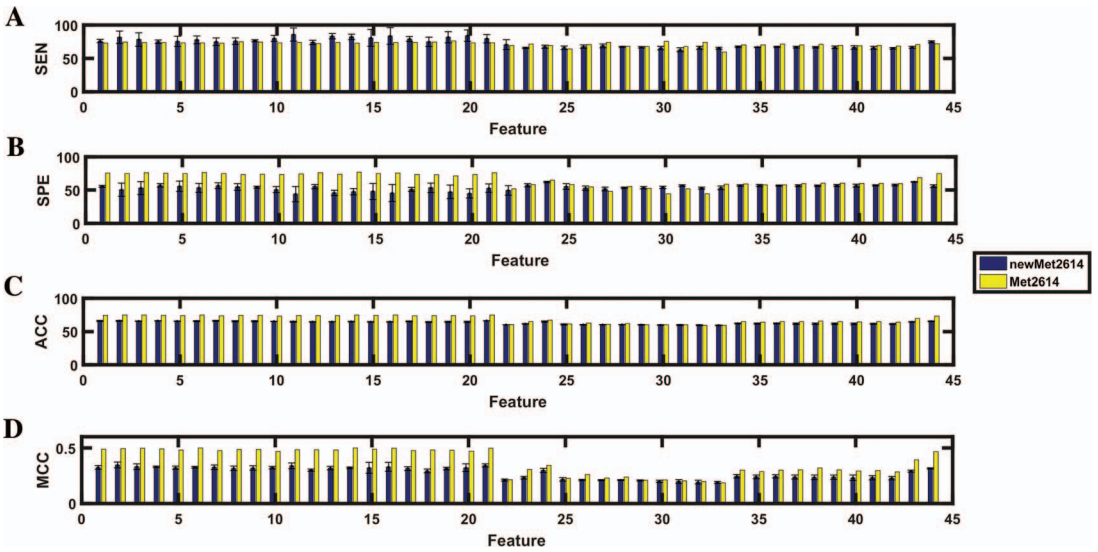

Figure 6. The cross validation performances of the models built on newMet2614 and Met2614 with 44 features (blue, the models built on newMet2614; yellow, the models built on Met2614). (A) The cross validation sensitivity (SEN); (B) the cross validation specificity (SPE); (C) the cross validation accuracies (ACC); (D) the cross validation MCCs.

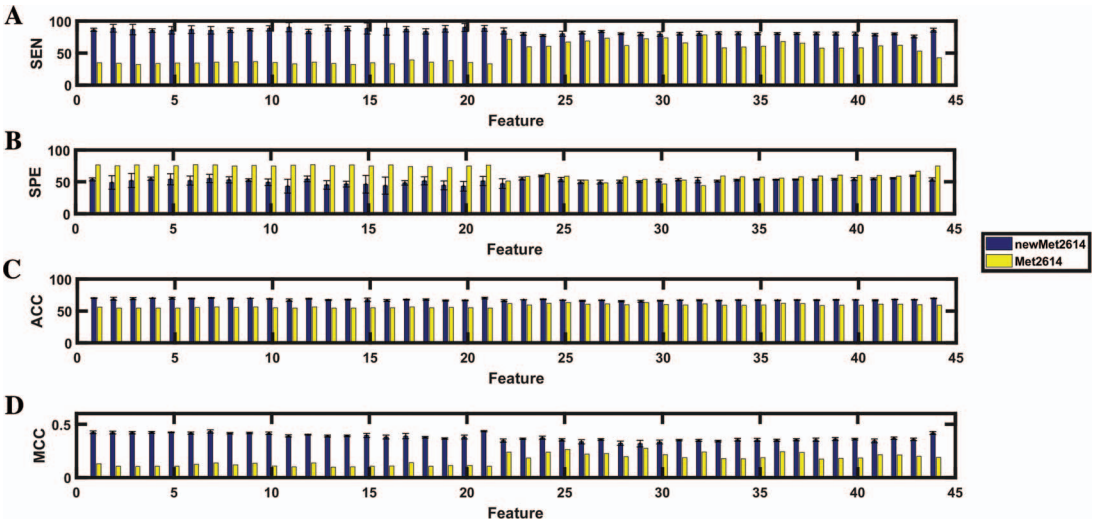

Figure 7. The predictive performances for M6Atest6540 by the models built on newMet2614 and Met2614 with 44 features (blue, the models built on newMet2614; yellow, the models built on Met2614). (A) The cross validation sensitivity (SEN); (B) the cross validation specificity (SPE); (C) the cross validation accuracies (ACC); (D) the cross validation MCCs.

Performances of the models built on Met2614 and newMet2614 with different features

To further explore if the bias of Met2614 affects the generalization of the models, we collected a new dataset, newMet2614, by randomly selecting positive examples from a non-redundant dataset based on RMBase. The newMet2614 also contains 1307 positive and negative examples, respectively. Three groups of features as shown in Table 2 were extracted from the RNA segments of both Met2614 and newMet2614. With each feature, the models were built on Met2614 and newMet2614 by using SVM, respectively. The two parameters, KernelScale and BoxConstraint, of the MATLAB function FITCSVM were selected by a grid search according to the results of the leave-one-out cross validation. The range of KernelScale is from 2^{-10} to 2^{6}, and the range of BoxConstraint is from 2^{-5} to 2^{15}. Note that all the works for newMet2614, including the generation of newMet2614, were repeated three times. Figure 6 shows the cross validation results of the models on Met2614 and newMet2614, respectively. Figure 7 shows the predictive results of the models on dataset M6Atest6540. For the models based on newMet2614, it indicates that the performances on M6Atest6540 are better than the cross validation results on newMet2614; in other words, the ACCs and MCCs on M6Atest6540 are higher than the cross validation results (Figures 6C and D, 7C and D and Table S4). These models show good performances on the positive examples of M6Atest6540; in other words, the sensitivities on M6Atest6540 are higher than the cross validation results (Figures 6A and 7A and Table S4). This is normal because the positive examples in M6Atest6540 are typical for their experimental 'SupportNum' > 15 (see Datasets). For the models based on Met2614, especially the models built with position-specific propensity-based features (Features 1–21 and 44), the performances on M6Atest6540 are substantially worse than the cross validation

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019

results on Met2614 (Figures 6C and D, 7C and 7D and Table S5), especially for sensitivities (Figures 6A and 7A and Table S5). Intuitively, the models built with position-specific propensity-based features are easily affected by the position bias of Met2614, which is in line with our results.

In addition, we compared the performances of the models based on newMet2614 with the models based on Met2614 on the independent test set M6Atest6540. Figure 7C and D shows that the models built on newMet2614 outperform the models based on Met2614; in other words, the ACCs and MCCs of the models based on newMet2614 are higher than the corresponding values of the models based on Met2614; these values are also higher than the corresponding values of the nine previous developed models. Although the models based on newMet2614 show better generalization than the models based on Met2614, it is better to use a large dataset to build the models to predict RNA m6A of *S. cerevisiae*.

Moreover, the ratio of m6A sites and non-m6A sites in both Met2614 and M6Atest6540 is set to 1.0, which creates the balanced datasets; however, there are more non-m6A sites in the genome than the m6A sites realistically. In other words, the datasets for this classification problem should be imbalanced. Random under-sampling is one of the ways to solve the imbalanced learning problem; however, the important information of negative examples may be lost, and the model may be easily to predict more examples as positive, such that the sensitivity will be overestimated. So, other methods such as informed under-sampling, cost-sensitive learning and kernel-based method could be used to deal this problem.

Conclusion

In this study, a benchmark dataset M6Atest6540 was compiled to evaluate and compare the performances of nine available methods for predicting m6A sites. Most of the methods performed substantially worse on the independent dataset M6Atest6540 than on the training dataset Met2614. Note that the Met2614 is a benchmark dataset used for most method building, and this dataset is a small dataset based on the early work of Schwartz *et al.* [57]. Further analysis shows that Met2614 is a biased dataset compared with the examples recorded in RMBase. In addition, the impact of the dataset bias on the generalization and robustness of models built on different features was further analyzed. The position-specific propensity-based features were found to be easily over-fitted on a biased dataset. Due to the more data we can obtain for RMBase, it is better to use a large benchmark to build RNA m6A sites prediction models of *S. cerevisiae*. And other kinds of modification [75, 76] would be studied in the future work.

Key Points

- We comprehensively review a variety of existing computational methods for the prediction of RNA m6A sites of *S. cerevisiae* and conduct a comparative study of available web servers.
- Benchmarking results demonstrate the substantially worse generalization of most of the nine available m6A prediction servers compared with the performances reported in the original papers.
- A PBI was introduced to analyze the possible biased positions between two sets of sequences. And the statistical analysis demonstrates that the widely used dataset Met2614 is significantly biased compared with the examples recorded in RMBase.
- The performances on M6Atest6540 indicate that the models based on newMet2614 that was randomly selected from RMBase outperform the models based on Met2614.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bfg.

Funding

This work was supported by National Natural Science Foundation of China (grant numbers 21403002, 31601074 and 61872094).

References

1. Desrosiers R, Friderici K, Rottman F. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc Natl Acad Sci U S A* 1974;**71**(10):3971–5.
2. Perry RP, Kelley DE. Existence of methylated messenger RNA in mouse L cells. *Cell* 1974;**1**(1):37–42.
3. Beemon K, Keith J. Localization of N6-methyladenosine in the Rous sarcoma virus genome. *J Mol Biol* 1977;**113**(1):165–79.
4. Aloni Y, Dhar R, Khoury G. Methylation of nuclear simian virus 40 RNAs. *J Virol* 1979;**32**(1):52–60.
5. Clancy MJ, Shambaugh ME, Timpte CS, *et al.* Induction of sporulation in Saccharomyces cerevisiae leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic Acids Res* 2002;**30**(20):4509–18.
6. Bodi Z, Button JD, Grierson D, *et al.* Yeast targets for mRNA methylation. *Nucleic Acids Res* 2010;**38**(16):5327–35.
7. Kennedy TD, Lane BG. Wheat embryo ribonucleates. XIII. Methyl-substituted nucleoside constituents and 5′-terminal dinucleotide sequences in bulk poly (AR)-rich RNA from imbibing wheat embryos. *Can J Biochem* 1979;**57**(6):927–31.
8. Zhong S, Li H, Bodi Z, *et al.* MTA is an Arabidopsis messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *Plant Cell* 2008;**20**(5):1278–88.
9. Wei CM, Gershowitz A, Moss B. 5'-Terminal and internal methylated nucleotide sequences in HeLa cell mRNA. *Biochemistry* 1976;**15**(2):397–401.
10. Adams JM, Cory S. Modified nucleosides and bizarre 5′-termini in mouse myeloma mRNA. *Nature* 1975;**255**(5503):28–33.
11. Perry RP, Kelley DE, Friderici K, *et al.* The methylated constituents of L cell messenger RNA: evidence for an unusual cluster at the 5′ terminus. *Cell* 1975;**4**(4):387–94.
12. Levis R, Penman S. 5′-Terminal structures of poly(A)+ cytoplasmic messenger RNA and of poly(A)+ and poly(A)- heterogeneous nuclear RNA of cells of the dipteran Drosophila melanogaster. *J Mol Biol* 1978;**120**(4):487–515.
13. Yue Y, Liu J, He C. RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes Dev* 2015;**29**(13):1343–55.

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019

14. Liu N, Pan T. N6-methyladenosine-encoded epitranscriptomics. *Nat Struct Mol Biol* 2016;**23**(2):98–102.
15. Lin Z, Hsu PJ, Xing X, *et al*. Mettl3−/Mettl14-mediated mRNA N(6)-methyladenosine modulates murine spermatogenesis. *Cell Res* 2017;**27**(10):1216–30.
16. Edupuganti RR, Geiger S, Lindeboom RG, *et al*. N(6)-methyladenosine (m(6)A) recruits and repels proteins to regulate mRNA homeostasis. 2017;**24**(10):870–8.
17. Slobodin B, Han R, Calderone V, *et al*. Transcription impacts the efficiency of mRNA translation via co-transcriptional N6-adenosine methylation. *Cell* 2017;**169**(2):326–337.e12.
18. Maity A, Das B. N6-methyladenosine modification in mRNA: machinery, function and implications for health and diseases. 2016;**283**(9):1607–30.
19. Zhao X, Yang Y, Sun BF, *et al*. FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Res* 2014;**24**(12):1403–19.
20. Liu N, Zhou KI, Parisien M, *et al*. N6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein. *Nucleic Acids Res* 2017;**45**(10):6051–63.
21. Heiliger KJ, Hess J, Vitagliano D, *et al*. Novel candidate genes of thyroid tumourigenesis identified in Trk-T1 transgenic mice. *Endocr Relat Cancer* 2012;**19**(3):409–21.
22. Machiela MJ, Lindstrom S, Allen NE, *et al*. Association of type 2 diabetes susceptibility variants with advanced prostate cancer risk in the Breast and Prostate Cancer Cohort Consortium. *Am J Epidemiol* 2012;**176**(12):1121–9.
23. Akilzhanova A, Nurkina Z, Momynaliev K, *et al*. Genetic profile and determinants of homocysteine levels in Kazakhstan patients with breast cancer. *Anticancer Res* 2013;**33**(9):4049–59.
24. Reddy SM, Sadim M, Li J, *et al*. Clinical and genetic predictors of weight gain in patients diagnosed with breast cancer. *Br J Cancer* 2013;**109**(4):872–81.
25. Long J, Zhang B, Signorello LB, *et al*. Evaluating genome-wide association study-identified breast cancer risk variants in African-American women. *PLoS One* 2013;**8**(4):e58350.
26. Lin Y, Ueda J, Yagyu K, *et al*. Association between variations in the fat mass and obesity-associated gene and pancreatic cancer risk: a case-control study in Japan. *BMC Cancer* 2013;**13**:337.
27. Pierce BL, Austin MA, Ahsan H. Association study of type 2 diabetes genetic susceptibility variants and risk of pancreatic cancer: an analysis of PanScan-I data. *Cancer Causes Control* 2011;**22**(6):877–83.
28. Casalegno-Garduno R, Schmitt A, Wang X, *et al*. Wilms' tumor 1 as a novel target for immunotherapy of leukemia. *Transplant Proc* 2010;**42**(8):3309–11.
29. Keith G. Mobilities of modified ribonucleotides on two-dimensional cellulose thin-layer chromatography. *Biochimie* 1995;**77**(1–2):142–4.
30. Zheng G, Dahl JA, Niu Y, *et al*. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol Cell* 2013;**49**(1):18–29.
31. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, *et al*. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 2012;**485**(7397):201–6.
32. Meyer KD, Saletore Y, Zumbo P, *et al*. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 2012;**149**(7):1635–46.
33. Chen W, Feng P, Ding H, *et al*. iRNA-methyl: identifying N 6-methyladenosine sites using pseudo nucleotide composition. *Anal Biochem* 2015;**490**:26–33.
34. Chen W, Tran H, Liang Z, *et al*. Identification and analysis of the N(6)-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci Rep* 2015;**5**:13859.
35. Zhou Y, Zeng P, Li YH, *et al*. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res* 2016;**44**(10):e91.
36. Liu Z, Xiao X, Yu DJ, *et al*. pRNAm-PC: predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. *Anal Biochem* 2016;**497**:60–7.
37. Chen W, Tang H, Lin H. MethyRNA: a web server for identification of N(6)-methyladenosine sites. *J Biomol Struct Dyn* 2017;**35**(3):683–7.
38. Jia CZ, Zhang JJ, Gu WZ. RNA-MethylPred: a high-accuracy predictor to identify N6-methyladenosine in RNA. *Anal Biochem* 2016;**510**:72–5.
39. Zhang M, Sun JW, Liu Z, *et al*. Improving N(6)-methyladenosine site prediction with heuristic selection of nucleotide physical-chemical properties. *Anal Biochem* 2016;**508**:104–13.
40. Chen W, Feng P, Ding H, *et al*. Identifying N (6)-methyladenosine sites in the *Arabidopsis thaliana* transcriptome. *Mol Genet Genomics* 2016;**291**(6):2225–9.
41. Xiang S, Liu K, Yan Z, *et al*. RNAMethPre: a web server for the prediction and query of mRNA m6A sites. *PLoS One* 2016;**11**(10):e0162707.
42. Li GQ, Liu Z, Shen HB, *et al*. TargetM6A: identifying N(6)methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. *IEEE Trans Nanobioscience* 2016;**15**(7):674–82.
43. Xiang S, Yan Z, Liu K, *et al*. AthMethPre: a web server for the prediction and query of mRNA m(6) A sites in *Arabidopsis thaliana*. *Mol Biosyst* 2016;**12**(11):3333–7.
44. Chen W, Xing P, Zou Q. Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. *Sci Rep* 2017;**7**:40242.
45. Xing P, Su R, Guo F, *et al*. Identifying N(6)-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Sci Rep* 2017;**7**:46757.
46. Wang X, Yan R. RFAthM6A: a new tool for predicting m(6) A sites in *Arabidopsis thaliana*. *Plant Mol Biol* 2018;**96**(3):327–37.
47. Wei L, Chen H, Su R. M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol Ther Nucleic Acids* 2018;**12**:635–44.
48. Akbar S, Hayat M. iMethyl-STTNC: identification of N(6)-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. *J Theor Biol* 2018;**455**:205–11.
49. Zhao Z, Peng H, Lan C, *et al*. Imbalance learning for the prediction of N(6)-methylation sites in mRNAs. *BMC Genomics* 2018;**19**(1):574.
50. Chen W, Ding H, Zhou X, *et al*. iRNA(m6A)-PseDNC: identifying N(6)-methyladenosine sites using pseudo dinucleotide composition. *Anal Biochem* 2018;**561–562**:59–65.
51. Huang Y, He N, Chen Y, *et al*. BERMP: a cross-species classifier for predicting m(6) A sites by integrating a deep learning algorithm and a random forest approach. *Int J Biol Sci* 2018;**14**(12):1669–77.
52. Qiang X, Chen H, Ye X, *et al*. M6AMRFS: robust prediction of N6-methyladenosine sites with sequence-based features in multiple species. *Front Genet* 2018;**9**:495.
53. Zou Q, Xing P, Wei L, *et al*. Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. *RNA* 2019;**25**(2):205–18.

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019

54. Chen K, Wei Z, Zhang Q, *et al*. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res* 2019;**47**(7):e41.
55. Zhang Y, Hamada M. DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. *BMC Bioinformatics* 2018;**19**(Suppl 19):524.
56. Wei L, Su R, Wang B, *et al*. Integration of deep feature representations and handcrafted features to improve the prediction of N 6-methyladenosine sites. *Neurocomputing* 2019;**324**:3–9.
57. Schwartz S, Agarwala SD, Mumbach MR, *et al*. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* 2013;**155**(6):1409–21.
58. Xuan JJ, Sun WJ, Lin PH, *et al*. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res* 2018;**46**(D1):D327–34.
59. Fu L, Niu B, Zhu Z, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**(23): 3150–2.
60. Zou, Q., Lin G, Jiang X, *et al*., Sequence clustering in bioinformatics: an empirical study. *Brief Bioinform*, 2019; bby090, doi:10.1093/bib/bby090.
61. He, J., Fang T, Zhang Z, *et al*., PseUI:pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics*, 2018, **19**:306. https://doi.org/10.1186/s12859-018-2321-0.
62. Xiong Y, Wang Q, Yang J, *et al*. PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front Microbiol* 2018;**9**:2571.
63. Li D, Luo L, Zhang W, *et al*. A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinformatics* 2016;**17**(1):329.
64. Zhang W, Yue X, Tang G, *et al*. SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PLoS Comput Biol* 2018;**14**(12):e1006616.
65. Tang G, Shi J, Wu W, *et al*. Sequence-based bacterial small RNAs prediction using ensemble learning strategies. *BMC Bioinformatics* 2018;**19**(Suppl 20):503.
66. Chen W, Lei TY, Jin DC, *et al*. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal Biochem* 2014;**456**:53–60.
67. Vapnik VN. *The Nature of Statistical Learning Theory*. Springer, 1995, 333.
68. Vapnik VN. An overview of statistical learning theory. *IEEE Trans Neural Netw* 1999;**10**(5):988–99.
69. Chen C, Tian YX, Zou XY, *et al*. Using pseudo-amino acid composition and support vector machine to predict protein structural class. *J Theor Biol* 2006;**243**(3):444–8.
70. Manavalan B, Shin TH, Lee G. PVP-SVM: sequence-based prediction of phage Virion proteins using a support vector machine. *Front Microbiol* 2018;**9**:476.
71. Xia J, Caragea D, Brown SJ. Prediction of alternatively spliced exons using support vector machines. *Int J Data Min Bioinform* 2010;**4**(4):411.
72. Vieira L, Grativol C, Thiebaut F, *et al*. PlantRNA_Sniffer: a SVM-based workflow to predict Long Intergenic non-coding RNAs in plants. *Non-Coding RNA* 2017;**3**(1):11.
73. Zhang Y, Xie R, Wang J, *et al*. Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform* 2018.
74. Chen Z, Liu X, Li F, *et al*. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief Bioinform* 2018.
75. He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* 2019;**35**(4):593–601.
76. Wei L, Su R, Luan S, *et al*. Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 2019.

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/elz018/5586923 by Shanghai Jiao Tong University user on 28 October 2019