

TargetM6A: Identifying N⁶-Methyladenosine Sites From RNA Sequences via Position-Specific Nucleotide Propensities and a Support Vector Machine

Guang-Qing Li, Zi Liu, Hong-Bin Shen, and Dong-Jun Yu*

Abstract—As one of the most ubiquitous post-transcriptional modifications of RNA, N⁶-methyladenosine (m⁶A) plays an essential role in many vital biological processes. The identification of m⁶A sites in RNAs is significantly important for both basic biomedical research and practical drug development. In this study, we designed a computational-based method, called TargetM6A, to rapidly and accurately target m⁶A sites solely from the primary RNA sequences. Two new features, i.e., position-specific nucleotide/dinucleotide propensities (PSNP/PSDP), are introduced and combined with the traditional nucleotide composition (NC) feature to formulate RNA sequences. The extracted features are further optimized to obtain a much more compact and discriminative feature subset by applying an incremental feature selection (IFS) procedure. Based on the optimized feature subset, we trained TargetM6A on the training dataset with a support vector machine (SVM) as the prediction engine. We compared the proposed TargetM6A method with existing methods for predicting m⁶A sites by performing stringent jackknife tests and independent validation tests on benchmark datasets. The experimental results show that the proposed TargetM6A method outperformed the existing methods for predicting m⁶A sites and remarkably improved the prediction performances, with MCC = 0.526 and AUC = 0.818. We also provided a user-friendly web server for TargetM6A, which is publicly accessible for academic use at <http://csbio.njust.edu.cn/bioinf/TargetM6A>.

Index Terms—Incremental feature selection, N⁶-methyladenosine, position-specific nucleotide propensity, RNA methylation, support vector machine.

I. INTRODUCTION

POST-translational modification (PTLM) of proteins is an important biological mechanism because it is associated

with many major diseases [1]. A similar subtle modification, the so-called post-transcriptional modification (PTCM), may also occur in RNA sequences. The PTCMs of RNA are very common and important in living organisms [2], [3]. To date, more than 100 PTCMs have been discovered in native cellular RNAs, including mRNAs and other RNAs [4], [5]. Among these modifications, N⁶-methyladenosine (m⁶A) is the most ubiquitous internal modification, which plays an important role in regulating gene expression [6]–[9]. As shown in Fig. S1 (see Supplementary Material II), methylation occurs on the sixth nitrogen atom of adenine, which is catalyzed by N⁶-methyladenosine methyltransferase (a complex composed of METTL3, METTL14, and WTAP) [10], [11], and the reverse process is catalyzed by demethylases (FTO or ALKBH5) [12], [13].

Since it was first identified in mammalian mRNA during the mid-1970s [14], m⁶A has been studied for decades to determine its exact functions and mechanisms. A series of studies have proven that m⁶A plays an essential role in diverse biological processes. Deng *et al.* discovered that m⁶A is important for respiration and stress responses in bacteria [15]; Liu *et al.* found that m⁶A can enhance the interaction between RNA and protein by altering the RNA structure [16]; Alarcon *et al.* showed that m⁶A can initialize miRNA processing and promote miRNA maturation by marking primary microRNAs (pri-miRNAs) [17]; and several studies have shown that the proteins of the YTH domain family are m⁶A-specific RNA-binding proteins that regulate both mRNA stability and localization [8], [18], [19]. All of these scientific findings indicate that knowledge of m⁶A is vitally important for both basic biomedical research and practical drug development.

Many studies have been dedicated to identifying the m⁶A sites in RNA sequences. Traditional m⁶A identification methods, such as thin-layer chromatography (TLC), high-performance liquid chromatography (HPLC), mass spectrometry, and scintillation, heavily depend on physicochemical techniques to study m⁶A distribution [20]. Accordingly, these methods are low-throughput and laborious and can only be applied to small-scale datasets [21]. Recently, based on several newly-developed high-throughput m⁶A profiling techniques (e.g., m⁶A-seq [21] and MeRIP-Seq [22]), the transcriptome-wide maps of m⁶A distributions in several species, including *Oryza sativa* [23], *Saccharomyces cerevisiae* [24],

Manuscript received March 3, 2016; revised April 22, 2016 and June 10, 2016; accepted August 8, 2016. This work was supported by the National Natural Science Foundation of China (No. 61373062 and 61222306), the Natural Science Foundation of Jiangsu (No. BK20141403), the China Postdoctoral Science Foundation (Nos. 2014T70526 and 2013M530260), the Fundamental Research Funds for the Central Universities (No. 30916011327), the Science and Technology Commission of Shanghai Municipality (No. 16JC1404300), and “The Six Top Talents” of Jiangsu Province (No. 2013-XXRJ-022). Asterisk indicates corresponding author.

G.-Q. Li and Z. Liu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing 210094, China.

D.-J. Yu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing 210094, China (e-mail: njyudj@njust.edu.cn; njyudj@njust.edu.cn).

H.-B. Shen is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNB.2016.2599115

Mus musculus [25], and *Homo sapiens* [25], have been reported. The results of these studies showed that m⁶A sites are highly conserved and are not randomly distributed as follows: m⁶A sites tend to occur near stop codons, 3' UTRs, and within long internal exons [22], [24], [25]. Although the above-mentioned methods can provide insights into the distributions and characteristics of m⁶A sites, they are both expensive and time-consuming. With the avalanche of RNA sequences released in the post-genomic era, it is highly desirable to develop computational methods to rapidly and accurately identify m⁶A sites solely from RNA sequences.

Unfortunately, to the best of our knowledge, there are currently only three computational tools, i.e., iRNAMethyl [26], m6Apre [27], and pRNAm-PC [28], available for predicting m⁶A sites in RNA sequences. The iRNAMethyl tool identifies m⁶A sites using an effective pseudo dinucleotide composition (PseDNC) feature, which integrates information about the RNA sequence order with three RNA physicochemical properties [26]. The m6Apre tool exploits nucleotide chemical category feature and accumulated nucleotide frequency feature to encode an RNA sequence [27]. The pRNAm-PC tool derives the RNA sequence feature from the physicochemical properties of the nucleotides via a series of auto-covariance and cross covariance transformations and uses support vector machine as prediction engine [28]. These pioneering studies have shown the feasibility of predicting m⁶A sites solely from RNA sequences. However, their prediction performances are still relatively low and should be further improved for potential practical applications.

This paper aims to accomplish this task using a novel feature representation method and machine-learning techniques. A new sequence-based m⁶A site predictor called TargetM6A is developed. In TargetM6A, we first encode each target RNA sequence into a fixed-length feature vector using the newly developed position-specific nucleotide/dinucleotide propensity (PSNP/PSDP) features and the nucleotide composition (NC) feature; then, we obtain an optimized feature subset by applying the incremental feature selection (IFS) algorithm [29] – [32] to the original feature space. Finally, we train a support vector machine (SVM) [33] prediction model with the optimized feature subset on the training dataset. We compared the proposed TargetM6A method with existing m⁶A site predictors via stringent jackknife tests and independent validation tests on benchmark datasets, and the results of the comparison show the superiority of the proposed method. Below, we address the detailed procedures for constructing TargetM6A.

II. MATERIALS AND METHODS

A. Benchmark Datasets

Three benchmark datasets, i.e., Met2614 [26], Train1664 [27], and Test5225 [27], were used in this study. All of the three datasets were constructed from 1183 genes in the *S. cerevisiae* genome. Because the m⁶A sites in the *S. cerevisiae* genome share a GAC consensus motif that has the potential to be methylated at the center base [24], we can formulate each RNA sample (segment),

denoted as $R_{\xi}(\text{GAC})$, in these datasets as follows:

$$R_{\xi}(\text{GAC}) = N_{-\xi} N_{-(\xi-1)} \cdots N_{-1} \text{GAC} N_1 \cdots N_{+(\xi-1)} N_{+\xi} \quad (1)$$

where $N_{-\xi}$ represents the ξ -th upstream nucleotide from the central motif GAC and $N_{+\xi}$ represents the ξ -th downstream nucleotide.

Chen *et al.* [26] constructed Met2614 based on the experimentally determined m⁶A modification data from the *S. cerevisiae* genome [24]. The positive subset of Met2614 contains 1307 methylation RNA segments, each of which has an N⁶-methylated adenosine at the center position; the negative subset is composed of 1307 non-methylation RNA segments, each of which also has an adenosine at the center position that is not N⁶-methylated. Please note that the value of ξ is 24 and thus the RNA segment length is $2 \times 24 + 3 = 51$ for Met2614. For more details about the procedure used to construct the Met2614 dataset, refer to [26].

Based on Met2614, Chen *et al.* [27] further constructed Train1664 and Test5225 as described below. Among the 1307 methylation RNA segments in Met2614, 832 segments, whose m⁶A sites were less than 10 nt from the detected m⁶A-seq peaks, were selected as the positive Train1664 samples. Then, 832 samples were randomly selected from the 33280 non-methylated RNA segments and used as the negative Train1664 samples [27]; the remaining 475 ($1307 - 832 = 475$) methylation RNA segments in Met2614 and the 4750 randomly selected samples from the 33280 non-methylation RNA segments constitute the independent validation dataset, i.e., Test5225. The value of ξ is 9 and thus the RNA segment length is $2 \times 9 + 3 = 21$ for both Train1664 and Test5225. For more details, refer to [27]. All three benchmark datasets used in this study are included in Supplementary Material I or can be freely downloaded from the web server <http://csbio.njust.edu.cn/bioinf/TargetM6A>.

Please note that in this work we first evaluated a prediction model for Met2614 compared to the jackknife test; then, we evaluated a prediction model for Train1664 compared to the jackknife test and an independent validation test (on Test5225). The final prediction model on the web server, TargetM6A, is trained on Met2614.

B. Feature Representation of the RNA Segments

One of the key problems in designing a machine-learning based m⁶A site predictor is determining how to encode an RNA sample (segment) into a fixed-length feature vector with highly discriminative information. Since the nucleotide composition (NC) feature [34] – [36] and pseudo nucleotide composition (PseNC) feature [26], [37], [38] were proposed, they have been widely utilized for nucleotide sequence representations. However, the performances of these two features for predicting m⁶A sites are still very low (see the “Results and Discussion” section). Therefore, in this study, two new features, i.e., position-specific nucleotide/dinucleotide propensity (PSNP/PSDP) features, were proposed to encode the RNA segments. Then, the two features were combined with the traditional NC feature to form the discriminative feature vector for predicting the m⁶A sites.

1) Position-Specific Nucleotide Propensity (PSNP)

Feature: Position-specific nucleotide/amino acid preferences are widely used in bioinformatics to predict functional sites in biological sequences [39] – [43]. The principle of position-specific preferences is to compute the frequency at which nucleotides/amino acids occur at certain positions and extract statistical information from sequences. Inspired by previous studies, we introduced a new position-specific nucleotide propensity (PSNP) feature in this study. Next, we describe how to encode an RNA segment into a PSNP feature vector.

Because the RNA segment expressed by (1) contains the consensus motif GAC at its center, we deleted the central GAC of the segment, and the remaining RNA segment can be reformulated as follows:

$$R_\xi = N_1 N_2 \cdots N_{2\xi} \quad (2)$$

where N_j ($j = 1, 2, \dots, 2\xi$) represents the nucleotide at the j -th position of the remaining RNA segment, and can be any one of the 4 nucleotide bases in RNA, i.e., $N_j \in \{A$ (adenine), C (cytosine), G (guanine), U (uracil)\}. We term R_ξ as a reduced RNA segment.

We first calculated the 4-dimensional position-specific occurrence frequency vector for the 4 nucleotides from the positive reduced RNA segments.

$$\mathbf{z}_j^+ = [z_{1,j}^+ \ z_{2,j}^+ \ z_{3,j}^+ \ z_{4,j}^+]^T \quad (3)$$

where $z_{i,j}^+$ is the occurrence frequency of the nucleotide base type i ($i = 1, 2, 3, 4$) at the j -th ($j = 1, 2, \dots, 2\xi$) position. Note that in this case, we use the numerical codes 1, 2, 3, and 4 to represent A (adenine), C (cytosine), G (guanine), and U (uracil), respectively.

Similarly, we can calculate the corresponding 4-dimensional position-specific occurrence frequency vector for the 4 nucleotides from the negative reduced RNA segments.

$$\mathbf{z}_j^- = [z_{1,j}^- \ z_{2,j}^- \ z_{3,j}^- \ z_{4,j}^-]^T \quad (4)$$

where $z_{i,j}^-$ is the occurrence frequency of the nucleotide base type i ($i = 1, 2, 3, 4$) at the j -th ($j = 1, 2, \dots, 2\xi$) position.

We defined the *position-specific frequency difference vector*, denoted as \mathbf{z}_j , as follows:

$$\mathbf{z}_j = \mathbf{z}_j^+ - \mathbf{z}_j^- = [z_{1,j} \ z_{2,j} \ z_{3,j} \ z_{4,j}]^T \quad (5)$$

where $z_{i,j} = z_{i,j}^+ - z_{i,j}^-$, $i = 1, 2, 3, 4$, and $j = 1, 2, \dots, 2\xi$.

Based on the position-specific frequency difference vectors, we can obtain a $4 \times 2\xi$ position-specific nucleotide propensity (PSNP) matrix as follows:

$$\mathbf{Z}_{PSNP} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_{2\xi}] = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,2\xi} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,2\xi} \\ z_{3,1} & z_{3,2} & \cdots & z_{3,2\xi} \\ z_{4,1} & z_{4,2} & \cdots & z_{4,2\xi} \end{bmatrix} \quad (6)$$

Now, an RNA segment can be encoded into a 2ξ -dimensional PSNP feature vector, denoted as \mathbf{f}_{PSNP} , by referring to the \mathbf{Z}_{PSNP} :

$$\mathbf{f}_{PSNP} = [f_1 \ f_2 \ \cdots \ f_j \ \cdots \ f_{2\xi}]^T \quad (7)$$

where each element f_j is selected from the \mathbf{Z}_{PSNP} matrix according to (8) and defined as follows:

$$f_j = \begin{cases} z_{1,j}, & \text{when } N_j = A \\ z_{2,j}, & \text{when } N_j = C \\ z_{3,j}, & \text{when } N_j = G \\ z_{4,j}, & \text{when } N_j = U \end{cases}, \quad (j = 1, 2, \dots, 2\xi). \quad (8)$$

2) Position-Specific Dinucleotide Propensity (PSDP)

Feature: We further extended the PSNP to dinucleotides (double nucleotide) to extract additional information contained in RNA segment. We termed this feature as the position-specific dinucleotide propensity (PSDP) feature. The RNA segment described in (2) could be rewritten in a dinucleotide form:

$$R_\xi = N_1 N_2 \cdots N_{2\xi} = D_1 D_2 \cdots D_{2\xi-1} \quad (9)$$

where $D_j = N_j N_{j+1}$ ($j = 1, 2, \dots, 2\xi - 1$) represents the dinucleotide at the j -th position, and can be any of the 16 types of dinucleotides, i.e., $D_j \in \{AA, AC, AG, \dots, UU\}$.

Based on the same principle used to generate the \mathbf{Z}_{PSNP} matrix, we can calculate the $16 \times (2\xi - 1)$ position-specific dinucleotide propensity (PSDP) matrix on a given dataset as follows:

$$\mathbf{Z}_{PSDP} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,2\xi-1} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,2\xi-1} \\ \vdots & \vdots & \ddots & \vdots \\ z_{16,1} & z_{16,2} & \cdots & z_{16,2\xi-1} \end{bmatrix} \quad (10)$$

where $z_{i,j} = z_{i,j}^+ - z_{i,j}^-$, and $z_{i,j}^+$ is the occurrence frequency of the dinucleotide type i ($i = 1, 2, \dots, 16$) at the j -th ($j = 1, 2, \dots, 2\xi - 1$) position calculated from the positive reduced RNA segments, whereas $z_{i,j}^-$ is the corresponding occurrence frequency derived from the negative reduced RNA segments.

Using \mathbf{Z}_{PSDP} , we can encode each RNA segment into a $(2\xi - 1)$ -dimensional PSDP feature vector, denoted as \mathbf{f}_{PSDP} , as follows:

$$\mathbf{f}_{PSDP} = [f_1 \ f_2 \ \cdots \ f_j \ \cdots \ f_{2\xi-1}]^T \quad (11)$$

where each element is obtained from the \mathbf{Z}_{PSDP} matrix in (12), which is defined as follows:

$$f_j = \begin{cases} z_{1,j}, & \text{when } D_j = AA \\ z_{2,j}, & \text{when } D_j = AC \\ \vdots & \\ z_{16,j}, & \text{when } D_j = UU \end{cases}, \quad (j = 1, 2, \dots, 2\xi - 1). \quad (12)$$

3) Nucleotide Composition (NC) Feature: Nucleotide composition (NC), i.e., the k -mer nucleotide frequency, is a classic method for representing the features of a nucleotide sequence, and has been widely used in previous studies [34]–[36]. For a given k value, a 4^k -dimensional feature vector is obtained, indicating the frequency of occurrence for each k -mer nucleotide in a nucleotide sequence. In this study, we used $k = 1, 2$, and 3, i.e., 4 types of *one*-mer nucleotide frequencies, 16 types of *two*-mer frequencies, and 64 types of

three -mer frequencies. Thus, an 84-dimensional ($4 + 16 + 64 = 84$) NC feature vector was generated for an RNA segment.

Finally, by serially combining the above-mentioned PSNP, PSDP, and NC features, an M -dimensional feature vector was obtained to represent each RNA segment, where $M = 2\zeta + (2\zeta - 1) + 84$. Because the value of ζ for Met2614 is 24, the dimensionality of the combined feature vector is 179; however, the dimensionality of the combined feature vector for Train1664 and Test5225 is 119 because the corresponding value of ζ is 9.

C. Incremental Feature Selection

The aforementioned feature representation procedure encodes each RNA segment into a fixed-length feature vector. We performed an incremental feature selection (IFS) procedure [29]–[32] based on the statistical F -test [44], [45] to identify the most prominent components of the original feature vector that are beneficial for distinguishing methylated RNA segments from non-methylated segments, as described below.

First, the statistical F -test [44], [45] was employed to measure the significance of all the extracted feature components. More specifically, by performing the F -test, we can obtain the p -value of each feature component according to the feature vectors for both the positive and negative samples in a dataset. A feature component with a smaller p -value was deemed as a more significant feature. Consequently, all the initial feature components were re-ranked by their p -values in ascending order. The set of the re-ranked feature components can be formulated as follows:

$$\{f'_1, f'_2, \dots, f'_M\} \quad (13)$$

where the p -value of f'_i is less than that of f'_j if $i < j$, and M is the dimensionality of the original feature vector, i.e., $M = 2\zeta + (2\zeta - 1) + 84$.

Second, the incremental feature selection (IFS) [29]–[32] method was employed to determine the feature components that should be selected for inclusion in the optimal feature subset based on the set of ranked feature components described in (13). After this step, we obtained the M feature subsets as follows:

$$\mathbf{f}_i = \{f'_1, \dots, f'_{i-1}, f'_i\} \quad (i = 1, 2, \dots, M) \quad (14)$$

where the i -th feature subset is composed of the top i ranked feature components defined in (13).

Finally, we evaluate the discriminative capability of each feature subset over the K -fold ($K = 10$ in this study) cross validation using a specific prediction engine, and the subset that achieves the highest MCC value will be considered the optimal feature subset. In this study, the support vector machine (SVM) was used as the prediction engine and will be briefly introduced in the next section.

D. SVM as Prediction Engine

The support vector machine (SVM), which was proposed by Cortes and Vapnik, is a machine-learning algorithm based on the statistical learning theory [33], [46]. It solves nonlinear

separable problems by kernel methods, whose basic principle is to transform the input vector into a high-dimensional Hilbert space using kernel functions and to identify the maximal separating hyperplane between classes [46], [47]. This technique has been widely used in a variety of bioinformatics areas [34], [48], [49] and has been shown to be a powerful prediction engine. In this study, we also applied SVM as the prediction engine for constructing the proposed TargetM6A method. The package LIBSVM V3.20 (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>) provided by Chang and Lin [47] was used to implement SVM. The radial basis function (RBF) kernel was chosen. The two parameters contained in the RBF kernel, i.e., the regularization parameter C and the kernel width parameter γ , were optimized based on a 10-fold cross validation using a grid search strategy. The optimized values of C and γ are both 2 for Met2614 and Train1664.

E. Evaluation Indexes

In statistical learning fields, the following three validation methods are often used to evaluate the performance of a predictor: independent dataset test, subsampling (or K -fold cross validation) test, and jackknife test [50]. In the jackknife test, all the samples in the benchmark dataset will be singled out one-by-one and tested by the predictor trained on the remaining samples [51]. Therefore, the jackknife test is considered the least arbitrary validation method that can always yield a unique outcome for a given benchmark dataset [51]. However, the complexity of the jackknife test is equal to the amount of data in the dataset, making it time-consuming to implement. Therefore, in this study, the 10-fold cross validation test was utilized to optimize the SVM parameters and select the features, whereas the jackknife test was used to evaluate the performances of the different learning algorithms and provide an unbiased assessment of the performance of the different prediction models.

In literature, Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), and the Matthews correlation coefficient (MCC) are routinely used as evaluation indexes [52], [53], which are defined as follows:

$$\begin{cases} Sen = 1 - \frac{N_-^+}{N^+} \\ Spe = 1 - \frac{N_+^-}{N^-} \\ Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \\ MCC = \frac{1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}}{\sqrt{\left(1 + \frac{N_-^+ - N_+^-}{N^+}\right) \left(1 + \frac{N_+^- - N_-^+}{N^-}\right)}} \end{cases} \quad (15)$$

where N^+ is the total number of the positive samples or true m⁶A sites, N_-^+ is the number of true m⁶A sites that were incorrectly predicted to be a non-m⁶A site, N^- is the total number of the negative samples or non-m⁶A sites, and N_+^- is the number of non-m⁶A sites that were incorrectly predicted to be an m⁶A site. Please note that this set of indexes is only valid

TABLE I
PREDICTION RESULTS OF DIFFERENT FEATURES USING DIFFERENT PREDICTION ENGINES ON THE MET2614 DATASET OVER A 10-FOLD CROSS VALIDATION¹

Feature ²	Prediction engine	<i>Sen</i> (%)	<i>Spe</i> (%)	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>
PSNP	RF	70.07 (±0.71)	75.29 (±1.13)	72.68 (±0.71)	0.454 (±0.014)	0.777 (±0.006)
	KNN	74.17 (±0.26)	73.22 (±0.46)	73.70 (±0.32)	0.473 (±0.006)	0.783 (±0.001)
	SVM	74.15 (±0.16)	74.43 (±0.12)	74.29 (±0.07)	0.485 (±0.001)	0.790 (±0.001)
PSDP	RF	70.78 (±0.25)	75.40 (±0.24)	73.09 (±0.15)	0.462 (±0.003)	0.784 (±0.001)
	KNN	72.69 (±0.36)	73.56 (±0.31)	73.12 (±0.24)	0.462 (±0.004)	0.779 (±0.001)
	SVM	72.38 (±0.25)	74.85 (±0.36)	73.62 (±0.25)	0.472 (±0.005)	0.788 (±0.001)
NC	RF	61.02 (±0.57)	62.19 (±0.93)	61.60 (±0.43)	0.232 (±0.008)	0.666 (±0.005)
	KNN	69.08 (±0.58)	54.48 (±0.55)	61.78 (±0.51)	0.238 (±0.010)	0.664 (±0.002)
	SVM	64.55 (±0.43)	63.34 (±0.60)	64.45 (±0.38)	0.288 (±0.007)	0.702 (±0.002)
PSNP+PSDP	RF	70.71 (±0.70)	75.88 (±0.44)	73.29 (±0.33)	0.466 (±0.006)	0.784 (±0.002)
	KNN	74.74 (±0.38)	72.78 (±0.19)	73.76 (±0.22)	0.475 (±0.004)	0.785 (±0.001)
	SVM	73.65 (±0.25)	75.39 (±0.24)	74.52 (±0.15)	0.490 (±0.003)	0.792 (±0.001)
PSNP+PSDP+NC	RF	71.80 (±0.80)	76.60 (±0.62)	74.20 (±0.54)	0.484 (±0.010)	0.799 (±0.003)
	KNN	75.65 (±0.26)	72.54 (±0.30)	74.10 (±0.17)	0.482 (±0.003)	0.793 (±0.001)
	SVM	74.38 (±0.28)	77.16 (±0.48)	75.77 (±0.27)	0.515 (±0.005)	0.815 (±0.001)

¹The experiment was performed 10 times for each feature and prediction engine. The average performance of each evaluation index is reported, followed by a standard deviation.

²PSNP, PSDP, and NC stand for position-specific nucleotide propensity, position-specific dinucleotide propensity, and nucleotide composition, respectively.

for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology [54] and system medicine [55], a completely different set of indexes as defined in [56] is needed.

For a soft-type classifier (e.g., the SVM used in this study), whose outputs are the continuous numerical values representing the probabilities/confidences of a feature belonging to certain classes, gradual adjustments to the discrimination threshold will produce a series of prediction confusion matrices [57]. Each of these confusion matrices generates different corresponding values of N_+^- and N_-^+ . In other words, the evaluation indexes defined in (15) are threshold-dependent. Therefore, we also employed another threshold-independent evaluation index, i.e., Area under Curve (*AUC*), which is based on the Receiver Operating Characteristic (ROC) analysis [58]. The *AUC* values range from 0 to 1. The larger the *AUC* value, the better the predictor performs. Specifically, an *AUC* = 0.5 means that the predictor generates a random prediction, whereas an *AUC* = 1 represents a perfect predictor, which always generates the correct prediction.

III. RESULTS AND DISCUSSION

A. Contributions of Different Features

A series of comparative experiments were performed using individual features (i.e., PSNP, PSDP, and NC) and their combinations (i.e., PSNP+PSDP and PSNP+PSDP+NC) to evaluate the contributions of different features to the m⁶A site predictions. We evaluated the discriminative performances of each of the five features (i.e., three individual features and the two combined features) using three popular prediction engines, i.e., Support Vector Machine (SVM) [33], Random Forests (RF) [59], and *K* Nearest Neighbor (KNN) [60], over a 10-fold cross validation. Table 1 summarizes the

prediction performances of each feature using the different prediction engines. Note that for each feature and prediction engine, the 10-fold cross validation experiment was performed 10 times. The average performance of each evaluation index is reported, followed by a standard deviation.

As shown in Table 1, several conclusions can be drawn.

First, for all five types of features, SVM consistently performed better than RF and KNN concerning the two overall evaluation indexes (i.e., *MCC* and *AUC*). This phenomenon indicates that SVM is the most appropriate engine among the three considered engines for predicting m⁶A sites using the features developed in this study.

Second, for all three prediction engines, the two newly developed features, i.e., PSNP and PSDP, performed significantly better than NC. Using the SVM prediction engine as an example, the NC feature achieved a prediction performance with an *MCC* < 0.30 and *AUC* < 0.71, whereas the proposed PSNP and PSDP features achieved much better performance with an *MCC* > 0.47 and *AUC* > 0.78.

Third, the combination of individual features can further improve prediction performance. By revisiting Table I, we found that the combined PSNP+PSDP feature consistently outperformed the individual features, i.e., PSNP and PSDP, for all three prediction engines. On the other hand, the NC feature performed significantly worse than the other two individual features. As illustrated in Table I, the NC feature only achieved an *MCC* of 0.288 and an *AUC* of 0.702 using SVM as the prediction engine. However, we found that the prediction performance can be greatly improved by adding the NC feature to the PSNP+PSDP feature. As shown in Table I, the PSNP+PSDP+NC feature (SVM as prediction engine) achieved the best performance with an *MCC* of 0.515 and an *AUC* of 0.815, which were 2.5% and 2.3%, respectively,

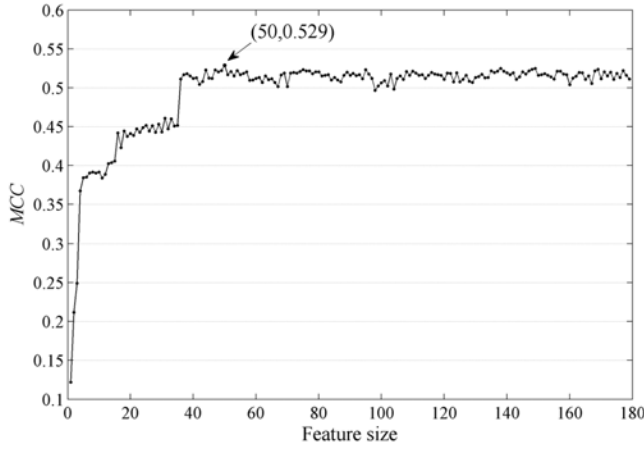


Fig. 1. The curve of the *MCC* values versus the sizes of the feature subset. The maximum *MCC* value is 0.529 and the corresponding feature size is 50.

better than the PSNP+PSDP feature.

Based on the three observations described above, we used PSNP+PSDP+NC as the input feature and SVM as the prediction engine to construct the proposed TargetM6A method.

B. Feature Selection Performance

In this section, we will try to further improve the computational efficiency and prediction performance by removing the redundant components in the PSNP+PSDP+NC feature with the incremental feature selection (IFS) procedure described in the “Incremental feature selection” section.

We first obtained the *p*-value for each component of the PSNP+PSDP+NC feature by performing *F*-test [44], [45]. Then, all the initial feature components were re-ranked by their *p*-values in ascending order. The list of the detailed re-ranked feature components for the Met2614 dataset is shown in Table S1 (see Supplementary Material II). Then, based on the ranked feature components, the IFS procedure was applied, and 179 feature subsets (for Met2614) were obtained. We evaluated the discriminative capability (measured by *MCC*) of each feature subset using the SVM prediction engine over a 10-fold cross validation. Fig. 1 plots the curve of *MCC* values versus the sizes of the feature subset.

As illustrated in Fig. 1, the *MCC* value almost consistently increases when the feature size varies from 1 to 50. The *MCC* value reaches a peak (0.529) when the feature size is 50. When the feature size is larger than 50, the *MCC* value fluctuates and no improvement was observed.

To extensively investigate the effectiveness of the feature selection procedure, we further performed the stringent jackknife test (i.e., leave-one-out cross validation) on the Met2614 dataset with SVM as prediction engine using the following two feature subsets: (1) the optimal feature subset, which consists of the top 50 feature components (denoted as Top50), and (2) the entire feature set, which consists of 179 initial feature components (denoted as All179). Table II summarizes the prediction performances of the two feature subsets on the Met2614 dataset over both the 10-fold cross validation and jackknife tests.

TABLE II

PREDICTION PERFORMANCES OF THE TOP50 AND ALL179 FEATURE SUBSETS OF THE MET2614 DATASET OVER BOTH THE 10-FOLD CROSS VALIDATION AND JACKKNIFE TESTS

Feature subset	10-fold cross-validation			jackknife test		
	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>
Top50	76.18 (± 0.28)	0.523 (± 0.005)	0.819 (± 0.001)	76.32	0.526	0.818
All179	75.77 (± 0.27)	0.515 (± 0.005)	0.815 (± 0.001)	75.67	0.513	0.816

TABLE III

COMPARISON OF THE RESULTS FROM THE PROPOSED TARGETM6A METHOD AND THE EXISTING PREDICTORS ON MET2614 USING JACKKNIFE TESTS

Prediction methods	<i>Sen</i> (%)	<i>Spe</i> (%)	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>
BLAST ¹	71.56	38.79	55.27	0.109	-
iRNAMethyl ¹	60.63	70.55	65.59	0.313	0.701
pRNAm-PC ²	69.72	69.75	69.74	0.394	0.762
m6Apred ³	73.91	74.45	74.18	0.483	0.792
TargetM6A	74.83	77.81	76.32	0.526	0.818

¹Results excerpted from [26].

²Results excerpted from [28].

³The re-implementation of m6Apred [27] on Met2614.

As shown in Table II, the Top50 feature subset consistently outperformed the All179 feature subset with regard to the three overall evaluation indexes, i.e., *Acc*, *MCC*, and *AUC*, over both the 10-fold cross validation and jackknife tests. Using the jackknife test as an example, the Top50 feature subset achieved an *Acc* of 76.32%, an *MCC* of 0.526, and an *AUC* of 0.818, which were 0.6%, 1.3%, and 0.2%, respectively, better than those of the All179 feature subset. Similar observations can also be obtained using 10-fold cross validation. These findings indicate that there is redundant information in the original 179 components, and the 50 selected feature components are truly the “significant” components for identifying the m⁶A sites. On one hand, the prediction performance can be slightly improved using the dimensionality-reduced feature subset, i.e., Top50; on the other hand, the computational efficiency is enhanced.

C. Comparisons With Existing Predictors

In this section, we will compare the proposed TargetM6A method with existing computational methods for predicting m⁶A sites, including iRNAMethyl [26], pRNAm-PC [28], and m6Apred [27], to demonstrate its efficacy.

We first compared these m⁶A site predictors using the Met2614 dataset and rigorous jackknife tests. The BLAST-based predictor [26] was used as the baseline predictor. The BLAST-based method predicts m⁶A sites by applying the sequence similarity search tool, i.e., BLAST [61], as follows: a query sample will be predicted as the true N⁶-methylation RNA segment if it is the most similar to the samples in the positive subset; otherwise, it will be predicted as a non-methylated RNA segment [26]. Because m6Apred [27] does not provide results on Met2614, we thus re-implemented it and evaluated it using Met2614. Please note that TargetM6A utilized the reduced feature subset, i.e., Top50, as the feature

TABLE IV
PREDICTION RESULTS OF DIFFERENT FEATURES USING DIFFERENT PREDICTION ENGINES ON
THE MET2614 DATASET OVER A 10-FOLD CROSS VALIDATION*

Methods	Train1664					Test5225				
	<i>Sen</i> (%)	<i>Spe</i> (%)	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>	<i>Sen</i> (%)	<i>Spe</i> (%)	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>
m6Apred *	79.21	77.04	78.12	0.563	0.840	53.89	79.07	76.78	0.222	-
TargetM6A	79.45	77.88	78.67	0.573	0.847	53.68	78.02	75.81	0.210	0.708

*Results excerpted from [27].

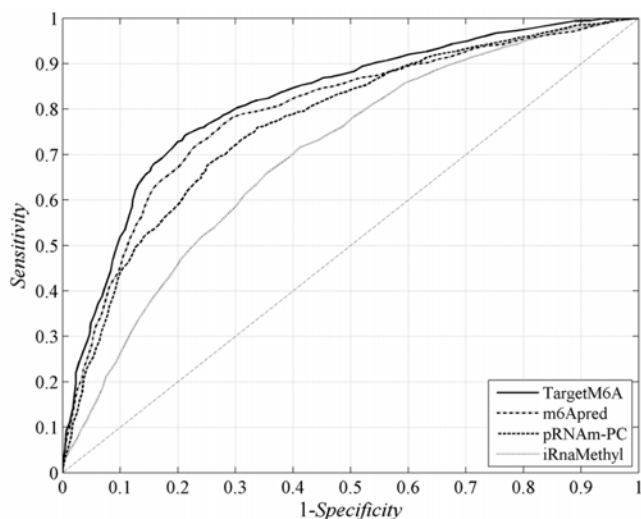


Fig. 2. The ROC curves of iRNAMethyl, m6Apred, pRNAm-PC, and the proposed TargetM6A method. The *AUC* values of these methods are 0.701, 0.762, 0.792, and 0.818, respectively.

representation and SVM as the prediction engine. In addition, the default threshold, i.e., 0.5, was used for TargetM6A and the re-implemented m6Apred to produce the *Sen*, *Spe*, *Acc*, and *MCC* values. Table III summarizes the results of the comparison of the different predictors on the Met2614 dataset using the jackknife test, and Fig. 2 compares their ROC curves.

As shown in Table III, we can find that the proposed TargetM6A method significantly outperformed the baseline predictor; the *MCC* value was remarkably improved from 0.109 to 0.526. TargetM6A also outperformed iRNAMethyl [26], pRNAm-PC [28], and m6Apred [27] with respect to all of the five evaluation indexes. Using *MCC* as an example, 21.3%, 13.2%, and 4.3% improvements were observed compared with iRNAMethyl [26], pRNAm-PC [28], and m6Apred [27].

In addition, the ROC curves plotted in Fig. 2, together with the *AUC* values listed in Table III, indicate that TargetM6A achieved the best performance regarding *AUC*. TargetM6A outperformed iRNAMethyl, pRNAm-PC, and m6Apred for the *AUC*, with improvements of 11.7%, 5.6%, and 2.6%, respectively. Considering that the jackknife test is the most rigorous cross validation method, these experimental results indeed show the superiority of the proposed TargetM6A method over the existing m⁶A site predictors.

We further evaluated the efficacy of the proposed method using two other benchmark datasets, i.e., Train1664 and Test5225. Please note that the original feature vector extracted

from the Train1664 dataset is 119-D (18-D PSNP, 17-D PSDP, and 84-D NC features), as described in the “Feature representation of RNA segments” section. Then, the IFS procedure was performed and an optimized feature subset of 65 feature components (see Table S2 in Supplementary Material II for details) was obtained to construct the SVM-based prediction model.

We first evaluated TargetM6A on Train1664 using the jackknife test. Then, we trained TargetM6A on Train1664 and tested the trained model with the independent validation dataset, i.e., Test5225. Because only m6Apred [27] currently provides results on Train1664 and Test5225, we only compared the proposed TargetM6A method with m6Apred. The performances of m6Apred were excerpted from a previous study [27]. Table IV summarizes the comparison of the results from m6Apred and the proposed TargetM6A method using the Train1664 and Test5225 datasets.

As shown in Table IV, the proposed TargetM6A method achieved very comparable performances with m6Apred for both Train1664 and Test5225. It has not escaped our notice that TargetM6A achieved much better performance than m6Apred for Met2614 (refer to Table III). However, TargetM6A failed to improve the prediction performance and only achieved comparable performance with m6Apred for Train1664 and Test5225 (refer to Table IV). We speculate that the differences in the RNA segment length and amount of data between Met2614 and Train1664 may account for this phenomenon: the RNA segment length ($\zeta = 24$) and data volume (2614 samples) of Met2614 are much larger than those of the Train1664 dataset ($\zeta = 9$, 1664 samples). Because the proposed method extracts the features according to the distribution of nucleotides in the central GAG motif, a dataset with larger RNA segment length and data volume will contain much more information about the distribution of nucleotides and thus will be beneficial for extracting more discriminative features. We believe that the prediction performance of TargetM6A can be further improved by increasing the number of available methylated and non-methylated RNA samples.

D. Web Server Implementation

A web server has been placed online at <http://csbio.njust.edu.cn/bioinf/TargetM6A> to enhance the applicability of the proposed TargetM6A method. The final online TargetM6A method was trained on Met2614, where each training sample is a 51-nt RNA segment. The TargetM6A method predicts m⁶A sites as described below.

For a query RNA sequence submitted by a user, TargetM6A first identifies all the GAC motifs in the query sequence; then,

for each GAC motif, a corresponding 51-nt RNA segment is constructed by placing a sliding window centered on the GAC motif. Based on the constructed 51-nt RNA segment, the features are extracted and then fed to the SVM classification engine to perform the prediction. Please note that to construct the corresponding 51-nt segment for a GAC motif with an insufficient number of nucleotides around the sequence (e.g., a motif near sequence terminals or a query RNA sequence shorter than 51-nt), the previously proposed “mirror image” technique [26] is used to fill in the missing nucleotides.

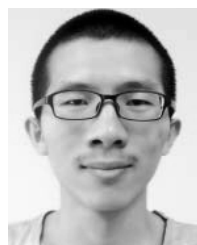
IV. CONCLUSIONS

In this study, we established a computational predictor for targeting the N⁶-methyladenosine sites in RNA sequences. Inspired by the successful applications of the position-specific amino acid propensity features in proteomics, we introduced the new PSNP and PSDP features to encode RNA sequences. Based on the newly developed feature, we implemented an m⁶A site predictor called TargetM6A. We compared the proposed TargetM6A method with the most recently released m⁶A site predictors by performing stringent jackknife tests and independent validation tests on the benchmark datasets. The experimental results show that our TargetM6A method achieved high prediction performance and outperformed the existing predictors. A web-server has been placed online at <http://csbio.njust.edu.cn/bioinf/TargetM6A> to enhance the applicability of the proposed method. The findings of this study enrich our understanding of sequence-based m⁶A sites and can potentially be applied to other nucleotide sequence-related prediction problems.

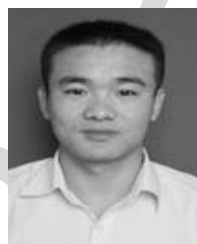
REFERENCES

- [1] K.-C. Chou, “Impacts of bioinformatics to medicinal chemistry,” *Med. Chem.*, vol. 11, no. 3, pp. 218–234, 2015.
- [2] C. Yi and T. Pan, “Cellular dynamics of RNA modification,” *Accounts Chem. Res.*, vol. 44, no. 12, pp. 1380–1388, Dec. 2011.
- [3] C. He, “Grand challenge commentary: RNA epigenetics?” *Nature Chem. Biol.*, vol. 6, pp. 863–865, Dec. 2010.
- [4] W. A. Cantara *et al.*, “The RNA modification database, RNAMDB: 2011 update,” *Nucl. Acids Res.*, vol. 39, no. 1, pp. D195–D201, Jan. 2011.
- [5] A. Czerwonec *et al.*, “MODOMICS: A database of RNA modification pathways. 2008 update,” *Nucl. Acids Res.*, vol. 37, pp. D118–D121, Jan. 2009.
- [6] Y. Fu, D. Dominissini, G. Rechavi, and C. He, “Gene expression regulation mediated through reversible m⁶A RNA methylation,” *Nature Rev. Genet.*, vol. 15, pp. 293–306, May 2014.
- [7] G. Jia, Y. Fu, and C. He, “Reversible RNA adenosine methylation in biological regulation,” *Trends Genet.*, vol. 29, no. 2, pp. 108–115, Feb. 2013.
- [8] X. Wang *et al.*, “N⁶-methyladenosine-dependent regulation of messenger RNA stability,” *Nature*, vol. 505, pp. 117–120, Jan. 2014.
- [9] K. D. Meyer and S. R. Jaffrey, “The dynamic epitranscriptome: N⁶-methyladenosine and gene expression control,” *Nature Rev. Mol. Cell Biol.*, vol. 15, pp. 313–326, May 2014.
- [10] J. Liu *et al.*, “A METTL3–METTL14 complex mediates mammalian nuclear RNA N⁶-adenosine methylation,” *Nature Chem. Biol.*, vol. 10, pp. 93–95, Feb. 2014.
- [11] X.-L. Ping *et al.*, “Mammalian WTAP is a regulatory subunit of the RNA N⁶-methyladenosine methyltransferase,” *Cell Res.*, vol. 24, pp. 177–189, Feb. 2014.
- [12] G. Jia *et al.*, “N⁶-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO,” *Nature Chem. Biol.*, vol. 7, pp. 885–887, Dec. 2011.
- [13] G. Zheng *et al.*, “ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility,” *Mol. Cell*, vol. 49, no. 1, pp. 18–29, Jan. 2013.
- [14] R. Desrosiers, K. Friderici, and F. Rottman, “Identification of methylated nucleosides in messenger RNA from novikoff hepatoma cells,” *Proc. Natl. Acad. Sci. USA*, vol. 71, no. 10, pp. 3971–3975, Oct. 1974.
- [15] X. Deng *et al.*, “Widespread occurrence of N⁶-methyladenosine in bacterial mRNA,” *Nucl. Acids Res.*, vol. 43, no. 13, pp. 6557–6567, Jul. 2015.
- [16] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan, “N⁶-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions,” *Nature*, vol. 518, pp. 560–564, Feb. 2015.
- [17] C. R. Alarcón, H. Lee, H. Goodarzi, N. Halberg, and S. F. Tavazoie, “N⁶-methyladenosine marks primary microRNAs for processing,” *Nature*, vol. 519, pp. 482–485, Mar. 2015.
- [18] C. Xu *et al.*, “Structural basis for selective binding of m⁶A RNA by the YTHDC1 YTH domain,” *Nature Chem. Biol.*, vol. 10, pp. 927–929, Nov. 2014.
- [19] D. Theler, C. Dominguez, M. Blatter, J. Boudet, and F. H.-T. Allain, “Solution structure of the YTH domain in complex with N⁶-methyladenosine RNA: A reader of methylated RNA,” *Nucl. Acids Res.*, vol. 42, no. 22, pp. 1399–13911, Dec. 2014.
- [20] S. Kellner, J. Burhenne, and M. Helm, “Detection of RNA modifications,” *RNA Biol.*, vol. 7, pp. 237–247, Mar./Apr. 2010.
- [21] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, and G. Rechavi, “Transcriptome-wide mapping of N⁶-methyladenosine by m⁶A-seq based on immunocapturing and massively parallel sequencing,” *Nature Protocols*, vol. 8, pp. 176–189, Jan. 2013.
- [22] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, “Comprehensive analysis of mRNA methylation reveals enrichment in 3’ UTRs and near stop codons,” *Cell*, vol. 149, no. 7, pp. 1635–1646, Jun. 2012.
- [23] Y. Li, X. Wang, C. Li, S. Hu, J. Yu, and S. Song, “Transcriptome-wide N⁶-methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification,” *RNA Biol.*, vol. 11, no. 9, pp. 1180–1188, 2014.
- [24] S. Schwartz *et al.*, “High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis,” *Cell*, vol. 155, no. 6, pp. 1409–1421, 2013.
- [25] D. Dominissini *et al.*, “Topology of the human and mouse m⁶A RNA methylomes revealed by m⁶A-seq,” *Nature*, vol. 485, pp. 201–206, May 2012.
- [26] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, “iRNA-Methyl: Identifying N⁶-methyladenosine sites using pseudo nucleotide composition,” *Anal. Biochem.*, vol. 490, pp. 26–33, Dec. 2015.
- [27] W. Chen, H. Tran, Z. Liang, H. Lin, and L. Zhang, “Identification and analysis of the N⁶-methyladenosine in the *Saccharomyces cerevisiae* transcriptome,” *Sci. Rep.*, vol. 5, p. 13859, Sep. 2015.
- [28] Z. Liu, X. Xiao, D.-J. Yu, J. Jia, W.-R. Qiu, and K.-C. Chou, “pRNAm-PC: Predicting N⁶-methyladenosine sites in RNA sequences via physical–chemical properties,” *Anal. Biochem.*, vol. 497, pp. 60–67, Mar. 2016.
- [29] Y. Liu, W. Gu, W. Zhang, and J. Wang, “Predict and analyze protein glycation sites with the mRMR and IFS methods,” *Biomed. Res. Int.*, vol. 2015, Dec. 2015, Art. no. 561547.
- [30] J. Wang, D. Zhang, and J. Li, “PREAL: Prediction of allergenic protein by maximum relevance minimum redundancy (mRMR) feature selection,” *BMC Syst. Biol.*, vol. 7, no. 5, p. S9, 2013.
- [31] C. Zou, J. Gong, and H. Li, “An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis,” *BMC Bioinform.*, vol. 14, p. 90, Mar. 2013.
- [32] H. Liu and R. Setiono, “Incremental feature selection,” *Appl. Intell.*, vol. 9, no. 3, pp. 217–230, 1998.
- [33] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [34] J. Brayet, F. Zehraoui, L. Jeanson-Leh, D. Israeli, and F. Tahi, “Towards a piRNA prediction using multiple kernel fusion and support vector machine,” *Bioinformatics*, vol. 30, pp. i364–i370, Sep. 2014.
- [35] E. K. M. Hashim and R. Abdullah, “Rare K-mer DNA: Identification of sequence motifs and prediction of CPG island and promoter,” *J. Theor. Biol.*, vol. 387, pp. 88–100, Dec. 2015.
- [36] H. Vinje, K. H. Liland, T. Almøy, and L. Snipen, “Comparing K-mer based methods for improved classification of 16S sequences,” *BMC Bioinform.*, vol. 16, p. 205, Jul. 2015.
- [37] S. H. Guo *et al.*, “iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition,” *Bioinformatics*, vol. 30, no. 11, pp. 1522–1529, 2014.

- [38] H. Lin, E. Z. Deng, H. Ding, W. Chen, and K. C. Chou, "iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucl. Acids Res.*, vol. 42, pp. 12961–12972, Oct. 2014.
- [39] K. C. Chou, "A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins," *J. Biol. Chem.*, vol. 268, pp. 16938–16948, Aug. 1993.
- [40] M. C. Frith, U. Hansen, and Z. Weng, "Detection of CIS-element clusters in higher eukaryotic DNA," *Bioinformatics*, vol. 17, pp. 878–889, Oct. 2001.
- [41] Y. R. Tang, Y. Z. Chen, C. A. Canchaya, and Z. Zhang, "GANNPhos: A new phosphorylation site predictor based on a genetic algorithm integrated neural network," *Protein Eng. Design Sel.*, vol. 20, pp. 405–412, Aug. 2007.
- [42] A. M. Thangakani, S. Kumar, R. Nagarajan, D. Velmurugan, and M. M. Gromiha, "GAP: Towards almost 100 percent prediction for SS-strand-mediated aggregating peptides with distinct morphologies," *Bioinformatics*, vol. 30, pp. 1983–1990, Jul. 2014.
- [43] Y. Xu, Y. X. Ding, J. Ding, L. Y. Wu, and N. Y. Deng, "Phogly-PseAAC: Prediction of lysine phosphoglycerylation in proteins incorporating with position-specific propensity," *J. Theor. Biol.*, vol. 379, pp. 10–15, Aug. 2015.
- [44] A. Golugula, G. Lee, and A. Madabhushi, "Evaluating feature selection strategies for high dimensional, small sample size datasets," in *Proc. Conf. Proc. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 949–952.
- [45] Y. Mao, X. Zhou, D. Pi, Y. Sun, and S. T. Wong, "Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection," *J. Biomed. Biotechnol.*, Jun. 2005, pp. 160–171.
- [46] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
- [47] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2006, Art. no. 27.
- [48] D. J. Yu, J. Hu, J. Yang, H. B. Shen, J. Tang, and J. Y. Yang, "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 10, no. 4, pp. 994–1008, Jul./Aug. 2013.
- [49] D. J. Yu, J. Hu, H. Yan, X. B. Yang, J. Y. Yang, and H. B. Shen, "Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble," *BMC Bioinform.*, vol. 15, p. 297, Sep. 2014.
- [50] K. C. Chou and C. T. Zhang, "Prediction of protein structural classes," *Critical Rev. Biochem. Mol. Biol.*, vol. 30, no. 4, pp. 275–349, 1995.
- [51] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, 2011.
- [52] K. C. Chou, "Using subsite coupling to predict signal peptides," *Protein Eng.*, vol. 14, pp. 75–79, Feb. 2001.
- [53] W. Chen, H. Ding, P. Feng, H. Lin, and K. C. Chou, "iACP: A sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, pp. 16895–16909, Mar. 2016.
- [54] K. C. Chou, Z. C. Wu, and X. Xiao, "iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Mol. Biosyst.*, vol. 8, pp. 629–641, Feb. 2012.
- [55] X. Xiao, P. Wang, W. Z. Lin, J. H. Jia, and K. C. Chou, "iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Anal. Biochem.*, vol. 436, pp. 168–177, May 2013.
- [56] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Mol. Biosyst.*, vol. 9, pp. 1092–1100, Jun. 2013.
- [57] D. J. Yu *et al.*, "TargetATPsite: A template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble," *J. Comput. Chem.*, vol. 34, pp. 974–985, Apr. 2013.
- [58] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
- [59] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [60] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
- [61] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, Oct. 1990.
- [62] J. O. Williams, "Narrow-band analyzer," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, USA, 1993.



Guang-Qing Li received his B.S. degree in computer science and technology from Nanjing University of Science and Technology, China, in 2014. Currently, he is working towards the M.S. degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include bioinformatics, data mining, and pattern recognition.



Zi Liu received his M.S. degree in computer science and technology from Jingdezhen Ceramic Institute, China, in 2015. Currently, he is working towards the Ph.D. degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include bioinformatics, biomedical image processing, and pattern recognition.



Hong-Bin Shen received his Ph.D. degree from Shanghai Jiaotong University, China, in 2007. He was a Postdoctoral Research Fellow of Harvard Medical School from 2007 to 2008. Currently, he is a Professor of Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University. His research interests include data mining, pattern recognition, and bioinformatics. Dr. Shen has published more than 60 papers and constructed 20 bioinformatics servers in these areas and he serves the editorial members of several international journals.



Dong-Jun Yu (M'12) received the B.S. degree in computer science and the M.S. degree in artificial intelligence from Jiangsu University of Science and Technology, China, in 1997 and 2000, respectively, and the Ph.D. degree in pattern analysis and machine intelligence from Nanjing University of Science and Technology, China, in 2003. In 2008, he acted as an academic visitor at University of York, U.K. He is currently a Professor in the School of Computer Science and Engineering of Nanjing University of Science and Technology. His current interests include bioinformatics, pattern recognition, and data mining. He is a Senior Member of CCF.

TargetM6A: Identifying N⁶-Methyladenosine Sites From RNA Sequences via Position-Specific Nucleotide Propensities and a Support Vector Machine

Guang-Qing Li, Zi Liu, Hong-Bin Shen, and Dong-Jun Yu*

Abstract—As one of the most ubiquitous post-transcriptional modifications of RNA, N⁶-methyladenosine (m⁶A) plays an essential role in many vital biological processes. The identification of m⁶A sites in RNAs is significantly important for both basic biomedical research and practical drug development. In this study, we designed a computational-based method, called TargetM6A, to rapidly and accurately target m⁶A sites solely from the primary RNA sequences. Two new features, i.e., position-specific nucleotide/dinucleotide propensities (PSNP/PSDP), are introduced and combined with the traditional nucleotide composition (NC) feature to formulate RNA sequences. The extracted features are further optimized to obtain a much more compact and discriminative feature subset by applying an incremental feature selection (IFS) procedure. Based on the optimized feature subset, we trained TargetM6A on the training dataset with a support vector machine (SVM) as the prediction engine. We compared the proposed TargetM6A method with existing methods for predicting m⁶A sites by performing stringent jackknife tests and independent validation tests on benchmark datasets. The experimental results show that the proposed TargetM6A method outperformed the existing methods for predicting m⁶A sites and remarkably improved the prediction performances, with MCC = 0.526 and AUC = 0.818. We also provided a user-friendly web server for TargetM6A, which is publicly accessible for academic use at <http://csbio.njust.edu.cn/bioinf/TargetM6A>.

Index Terms—Incremental feature selection, N⁶-methyladenosine, position-specific nucleotide propensity, RNA methylation, support vector machine.

I. INTRODUCTION

POST-translational modification (PTLM) of proteins is an important biological mechanism because it is associated

with many major diseases [1]. A similar subtle modification, the so-called post-transcriptional modification (PTCM), may also occur in RNA sequences. The PTCMs of RNA are very common and important in living organisms [2], [3]. To date, more than 100 PTCMs have been discovered in native cellular RNAs, including mRNAs and other RNAs [4], [5]. Among these modifications, N⁶-methyladenosine (m⁶A) is the most ubiquitous internal modification, which plays an important role in regulating gene expression [6]–[9]. As shown in Fig. S1 (see Supplementary Material II), methylation occurs on the sixth nitrogen atom of adenine, which is catalyzed by N⁶-methyladenosine methyltransferase (a complex composed of METTL3, METTL14, and WTAP) [10], [11], and the reverse process is catalyzed by demethylases (FTO or ALKBH5) [12], [13].

Since it was first identified in mammalian mRNA during the mid-1970s [14], m⁶A has been studied for decades to determine its exact functions and mechanisms. A series of studies have proven that m⁶A plays an essential role in diverse biological processes. Deng *et al.* discovered that m⁶A is important for respiration and stress responses in bacteria [15]; Liu *et al.* found that m⁶A can enhance the interaction between RNA and protein by altering the RNA structure [16]; Alarcon *et al.* showed that m⁶A can initialize miRNA processing and promote miRNA maturation by marking primary microRNAs (pri-miRNAs) [17]; and several studies have shown that the proteins of the YTH domain family are m⁶A-specific RNA-binding proteins that regulate both mRNA stability and localization [8], [18], [19]. All of these scientific findings indicate that knowledge of m⁶A is vitally important for both basic biomedical research and practical drug development.

Many studies have been dedicated to identifying the m⁶A sites in RNA sequences. Traditional m⁶A identification methods, such as thin-layer chromatography (TLC), high-performance liquid chromatography (HPLC), mass spectrometry, and scintillation, heavily depend on physicochemical techniques to study m⁶A distribution [20]. Accordingly, these methods are low-throughput and laborious and can only be applied to small-scale datasets [21]. Recently, based on several newly-developed high-throughput m⁶A profiling techniques (e.g., m⁶A-seq [21] and MeRIP-Seq [22]), the transcriptome-wide maps of m⁶A distributions in several species, including *Oryza sativa* [23], *Saccharomyces cerevisiae* [24],

Manuscript received March 3, 2016; revised April 22, 2016 and June 10, 2016; accepted August 8, 2016. This work was supported by the National Natural Science Foundation of China (No. 61373062 and 61222306), the Natural Science Foundation of Jiangsu (No. BK20141403), the China Postdoctoral Science Foundation (Nos. 2014T70526 and 2013M530260), the Fundamental Research Funds for the Central Universities (No. 30916011327), the Science and Technology Commission of Shanghai Municipality (No. 16JC1404300), and “The Six Top Talents” of Jiangsu Province (No. 2013-XXRJ-022). Asterisk indicates corresponding author.

G.-Q. Li and Z. Liu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing 210094, China.

D.-J. Yu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing 210094, China (e-mail: njyudj@njust.edu.cn; njyudj@njust.edu.cn).

H.-B. Shen is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNB.2016.2599115

Mus musculus [25], and *Homo sapiens* [25], have been reported. The results of these studies showed that m⁶A sites are highly conserved and are not randomly distributed as follows: m⁶A sites tend to occur near stop codons, 3' UTRs, and within long internal exons [22], [24], [25]. Although the above-mentioned methods can provide insights into the distributions and characteristics of m⁶A sites, they are both expensive and time-consuming. With the avalanche of RNA sequences released in the post-genomic era, it is highly desirable to develop computational methods to rapidly and accurately identify m⁶A sites solely from RNA sequences.

Unfortunately, to the best of our knowledge, there are currently only three computational tools, i.e., iRNAMethyl [26], m6ApreD [27], and pRNAm-PC [28], available for predicting m⁶A sites in RNA sequences. The iRNAMethyl tool identifies m⁶A sites using an effective pseudo dinucleotide composition (PseDNC) feature, which integrates information about the RNA sequence order with three RNA physicochemical properties [26]. The m6ApreD tool exploits nucleotide chemical category feature and accumulated nucleotide frequency feature to encode an RNA sequence [27]. The pRNAm-PC tool derives the RNA sequence feature from the physicochemical properties of the nucleotides via a series of auto-covariance and cross covariance transformations and uses support vector machine as prediction engine [28]. These pioneering studies have shown the feasibility of predicting m⁶A sites solely from RNA sequences. However, their prediction performances are still relatively low and should be further improved for potential practical applications.

This paper aims to accomplish this task using a novel feature representation method and machine-learning techniques. A new sequence-based m⁶A site predictor called TargetM6A is developed. In TargetM6A, we first encode each target RNA sequence into a fixed-length feature vector using the newly developed position-specific nucleotide/dinucleotide propensity (PSNP/PSDP) features and the nucleotide composition (NC) feature; then, we obtain an optimized feature subset by applying the incremental feature selection (IFS) algorithm [29] – [32] to the original feature space. Finally, we train a support vector machine (SVM) [33] prediction model with the optimized feature subset on the training dataset. We compared the proposed TargetM6A method with existing m⁶A site predictors via stringent jackknife tests and independent validation tests on benchmark datasets, and the results of the comparison show the superiority of the proposed method. Below, we address the detailed procedures for constructing TargetM6A.

II. MATERIALS AND METHODS

A. Benchmark Datasets

Three benchmark datasets, i.e., Met2614 [26], Train1664 [27], and Test5225 [27], were used in this study. All of the three datasets were constructed from 1183 genes in the *S. cerevisiae* genome. Because the m⁶A sites in the *S. cerevisiae* genome share a GAC consensus motif that has the potential to be methylated at the center base [24], we can formulate each RNA sample (segment),

denoted as $R_{\xi}(\text{GAC})$, in these datasets as follows:

$$R_{\xi}(\text{GAC}) = N_{-\xi} N_{-(\xi-1)} \cdots N_{-1} \text{GAC} N_1 \cdots N_{+(\xi-1)} N_{+\xi} \quad (1)$$

where $N_{-\xi}$ represents the ξ -th upstream nucleotide from the central motif GAC and $N_{+\xi}$ represents the ξ -th downstream nucleotide.

Chen *et al.* [26] constructed Met2614 based on the experimentally determined m⁶A modification data from the *S. cerevisiae* genome [24]. The positive subset of Met2614 contains 1307 methylation RNA segments, each of which has an N⁶-methylated adenosine at the center position; the negative subset is composed of 1307 non-methylation RNA segments, each of which also has an adenosine at the center position that is not N⁶-methylated. Please note that the value of ξ is 24 and thus the RNA segment length is $2 \times 24 + 3 = 51$ for Met2614. For more details about the procedure used to construct the Met2614 dataset, refer to [26].

Based on Met2614, Chen *et al.* [27] further constructed Train1664 and Test5225 as described below. Among the 1307 methylation RNA segments in Met2614, 832 segments, whose m⁶A sites were less than 10 nt from the detected m⁶A-seq peaks, were selected as the positive Train1664 samples. Then, 832 samples were randomly selected from the 33280 non-methylated RNA segments and used as the negative Train1664 samples [27]; the remaining 475 ($1307 - 832 = 475$) methylation RNA segments in Met2614 and the 4750 randomly selected samples from the 33280 non-methylation RNA segments constitute the independent validation dataset, i.e., Test5225. The value of ξ is 9 and thus the RNA segment length is $2 \times 9 + 3 = 21$ for both Train1664 and Test5225. For more details, refer to [27]. All three benchmark datasets used in this study are included in Supplementary Material I or can be freely downloaded from the web server <http://csbio.njust.edu.cn/bioinf/TargetM6A>.

Please note that in this work we first evaluated a prediction model for Met2614 compared to the jackknife test; then, we evaluated a prediction model for Train1664 compared to the jackknife test and an independent validation test (on Test5225). The final prediction model on the web server, TargetM6A, is trained on Met2614.

B. Feature Representation of the RNA Segments

One of the key problems in designing a machine-learning based m⁶A site predictor is determining how to encode an RNA sample (segment) into a fixed-length feature vector with highly discriminative information. Since the nucleotide composition (NC) feature [34] – [36] and pseudo nucleotide composition (PseNC) feature [26], [37], [38] were proposed, they have been widely utilized for nucleotide sequence representations. However, the performances of these two features for predicting m⁶A sites are still very low (see the “Results and Discussion” section). Therefore, in this study, two new features, i.e., position-specific nucleotide/dinucleotide propensity (PSNP/PSDP) features, were proposed to encode the RNA segments. Then, the two features were combined with the traditional NC feature to form the discriminative feature vector for predicting the m⁶A sites.

1) Position-Specific Nucleotide Propensity (PSNP)

Feature: Position-specific nucleotide/amino acid preferences are widely used in bioinformatics to predict functional sites in biological sequences [39] – [43]. The principle of position-specific preferences is to compute the frequency at which nucleotides/amino acids occur at certain positions and extract statistical information from sequences. Inspired by previous studies, we introduced a new position-specific nucleotide propensity (PSNP) feature in this study. Next, we describe how to encode an RNA segment into a PSNP feature vector.

Because the RNA segment expressed by (1) contains the consensus motif GAC at its center, we deleted the central GAC of the segment, and the remaining RNA segment can be reformulated as follows:

$$R_\xi = N_1 N_2 \cdots N_{2\xi} \quad (2)$$

where N_j ($j = 1, 2, \dots, 2\xi$) represents the nucleotide at the j -th position of the remaining RNA segment, and can be any one of the 4 nucleotide bases in RNA, i.e., $N_j \in \{A$ (adenine), C (cytosine), G (guanine), U (uracil)\}. We term R_ξ as a reduced RNA segment.

We first calculated the 4-dimensional position-specific occurrence frequency vector for the 4 nucleotides from the positive reduced RNA segments.

$$\mathbf{z}_j^+ = [z_{1,j}^+ \quad z_{2,j}^+ \quad z_{3,j}^+ \quad z_{4,j}^+]^T \quad (3)$$

where $z_{i,j}^+$ is the occurrence frequency of the nucleotide base type i ($i = 1, 2, 3, 4$) at the j -th ($j = 1, 2, \dots, 2\xi$) position. Note that in this case, we use the numerical codes 1, 2, 3, and 4 to represent A (adenine), C (cytosine), G (guanine), and U (uracil), respectively.

Similarly, we can calculate the corresponding 4-dimensional position-specific occurrence frequency vector for the 4 nucleotides from the negative reduced RNA segments.

$$\mathbf{z}_j^- = [z_{1,j}^- \quad z_{2,j}^- \quad z_{3,j}^- \quad z_{4,j}^-]^T \quad (4)$$

where $z_{i,j}^-$ is the occurrence frequency of the nucleotide base type i ($i = 1, 2, 3, 4$) at the j -th ($j = 1, 2, \dots, 2\xi$) position.

We defined the *position-specific frequency difference vector*, denoted as \mathbf{z}_j , as follows:

$$\mathbf{z}_j = \mathbf{z}_j^+ - \mathbf{z}_j^- = [z_{1,j} \quad z_{2,j} \quad z_{3,j} \quad z_{4,j}]^T \quad (5)$$

where $z_{i,j} = z_{i,j}^+ - z_{i,j}^-$, $i = 1, 2, 3, 4$, and $j = 1, 2, \dots, 2\xi$.

Based on the position-specific frequency difference vectors, we can obtain a $4 \times 2\xi$ position-specific nucleotide propensity (PSNP) matrix as follows:

$$\mathbf{Z}_{PSNP} = [\mathbf{z}_1 \mathbf{z}_2 \cdots \mathbf{z}_{2\xi}] = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,2\xi} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,2\xi} \\ z_{3,1} & z_{3,2} & \cdots & z_{3,2\xi} \\ z_{4,1} & z_{4,2} & \cdots & z_{4,2\xi} \end{bmatrix} \quad (6)$$

Now, an RNA segment can be encoded into a 2ξ -dimensional PSNP feature vector, denoted as \mathbf{f}_{PSNP} , by referring to the \mathbf{Z}_{PSNP} :

$$\mathbf{f}_{PSNP} = [f_1 \quad f_2 \quad \cdots \quad f_j \quad \cdots \quad f_{2\xi}]^T \quad (7)$$

where each element f_j is selected from the \mathbf{Z}_{PSNP} matrix according to (8) and defined as follows:

$$f_j = \begin{cases} z_{1,j}, & \text{when } N_j = A \\ z_{2,j}, & \text{when } N_j = C \\ z_{3,j}, & \text{when } N_j = G \\ z_{4,j}, & \text{when } N_j = U \end{cases}, \quad (j = 1, 2, \dots, 2\xi). \quad (8)$$

2) Position-Specific Dinucleotide Propensity (PSDP)

Feature: We further extended the PSNP to dinucleotides (double nucleotide) to extract additional information contained in RNA segment. We termed this feature as the position-specific dinucleotide propensity (PSDP) feature. The RNA segment described in (2) could be rewritten in a dinucleotide form:

$$R_\xi = N_1 N_2 \cdots N_{2\xi} = D_1 D_2 \cdots D_{2\xi-1} \quad (9)$$

where $D_j = N_j N_{j+1}$ ($j = 1, 2, \dots, 2\xi - 1$) represents the dinucleotide at the j -th position, and can be any of the 16 types of dinucleotides, i.e., $D_j \in \{AA, AC, AG, \dots, UU\}$.

Based on the same principle used to generate the \mathbf{Z}_{PSNP} matrix, we can calculate the $16 \times (2\xi - 1)$ position-specific dinucleotide propensity (PSDP) matrix on a given dataset as follows:

$$\mathbf{Z}_{PSDP} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,2\xi-1} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,2\xi-1} \\ \vdots & \vdots & \ddots & \vdots \\ z_{16,1} & z_{16,2} & \cdots & z_{16,2\xi-1} \end{bmatrix} \quad (10)$$

where $z_{i,j} = z_{i,j}^+ - z_{i,j}^-$, and $z_{i,j}^+$ is the occurrence frequency of the dinucleotide type i ($i = 1, 2, \dots, 16$) at the j -th ($j = 1, 2, \dots, 2\xi - 1$) position calculated from the positive reduced RNA segments, whereas $z_{i,j}^-$ is the corresponding occurrence frequency derived from the negative reduced RNA segments.

Using \mathbf{Z}_{PSDP} , we can encode each RNA segment into a $(2\xi - 1)$ -dimensional PSDP feature vector, denoted as \mathbf{f}_{PSDP} , as follows:

$$\mathbf{f}_{PSDP} = [f_1 \quad f_2 \quad \cdots \quad f_j \quad \cdots \quad f_{2\xi-1}]^T \quad (11)$$

where each element is obtained from the \mathbf{Z}_{PSDP} matrix in (12), which is defined as follows:

$$f_j = \begin{cases} z_{1,j}, & \text{when } D_j = AA \\ z_{2,j}, & \text{when } D_j = AC \\ \vdots & \\ z_{16,j}, & \text{when } D_j = UU \end{cases} \quad (j = 1, 2, \dots, 2\xi - 1). \quad (12)$$

3) Nucleotide Composition (NC) Feature: Nucleotide composition (NC), i.e., the k -mer nucleotide frequency, is a classic method for representing the features of a nucleotide sequence, and has been widely used in previous studies [34]–[36]. For a given k value, a 4^k -dimensional feature vector is obtained, indicating the frequency of occurrence for each k -mer nucleotide in a nucleotide sequence. In this study, we used $k = 1, 2$, and 3, i.e., 4 types of *one*-mer nucleotide frequencies, 16 types of *two*-mer frequencies, and 64 types of

three -mer frequencies. Thus, an 84-dimensional ($4 + 16 + 64 = 84$) NC feature vector was generated for an RNA segment.

Finally, by serially combining the above-mentioned PSNP, PSDP, and NC features, an M -dimensional feature vector was obtained to represent each RNA segment, where $M = 2\zeta + (2\zeta - 1) + 84$. Because the value of ζ for Met2614 is 24, the dimensionality of the combined feature vector is 179; however, the dimensionality of the combined feature vector for Train1664 and Test5225 is 119 because the corresponding value of ζ is 9.

C. Incremental Feature Selection

The aforementioned feature representation procedure encodes each RNA segment into a fixed-length feature vector. We performed an incremental feature selection (IFS) procedure [29]–[32] based on the statistical F -test [44], [45] to identify the most prominent components of the original feature vector that are beneficial for distinguishing methylated RNA segments from non-methylated segments, as described below.

First, the statistical F -test [44], [45] was employed to measure the significance of all the extracted feature components. More specifically, by performing the F -test, we can obtain the p -value of each feature component according to the feature vectors for both the positive and negative samples in a dataset. A feature component with a smaller p -value was deemed as a more significant feature. Consequently, all the initial feature components were re-ranked by their p -values in ascending order. The set of the re-ranked feature components can be formulated as follows:

$$\{f'_1, f'_2, \dots, f'_M\} \quad (13)$$

where the p -value of f'_i is less than that of f'_j if $i < j$, and M is the dimensionality of the original feature vector, i.e., $M = 2\zeta + (2\zeta - 1) + 84$.

Second, the incremental feature selection (IFS) [29]–[32] method was employed to determine the feature components that should be selected for inclusion in the optimal feature subset based on the set of ranked feature components described in (13). After this step, we obtained the M feature subsets as follows:

$$\mathbf{f}_i = \{f'_1, \dots, f'_{i-1}, f'_i\} \quad (i = 1, 2, \dots, M) \quad (14)$$

where the i -th feature subset is composed of the top i ranked feature components defined in (13).

Finally, we evaluate the discriminative capability of each feature subset over the K -fold ($K = 10$ in this study) cross validation using a specific prediction engine, and the subset that achieves the highest MCC value will be considered the optimal feature subset. In this study, the support vector machine (SVM) was used as the prediction engine and will be briefly introduced in the next section.

D. SVM as Prediction Engine

The support vector machine (SVM), which was proposed by Cortes and Vapnik, is a machine-learning algorithm based on the statistical learning theory [33], [46]. It solves nonlinear

separable problems by kernel methods, whose basic principle is to transform the input vector into a high-dimensional Hilbert space using kernel functions and to identify the maximal separating hyperplane between classes [46], [47]. This technique has been widely used in a variety of bioinformatics areas [34], [48], [49] and has been shown to be a powerful prediction engine. In this study, we also applied SVM as the prediction engine for constructing the proposed TargetM6A method. The package LIBSVM V3.20 (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>) provided by Chang and Lin [47] was used to implement SVM. The radial basis function (RBF) kernel was chosen. The two parameters contained in the RBF kernel, i.e., the regularization parameter C and the kernel width parameter γ , were optimized based on a 10-fold cross validation using a grid search strategy. The optimized values of C and γ are both 2 for Met2614 and Train1664.

E. Evaluation Indexes

In statistical learning fields, the following three validation methods are often used to evaluate the performance of a predictor: independent dataset test, subsampling (or K -fold cross validation) test, and jackknife test [50]. In the jackknife test, all the samples in the benchmark dataset will be singled out one-by-one and tested by the predictor trained on the remaining samples [51]. Therefore, the jackknife test is considered the least arbitrary validation method that can always yield a unique outcome for a given benchmark dataset [51]. However, the complexity of the jackknife test is equal to the amount of data in the dataset, making it time-consuming to implement. Therefore, in this study, the 10-fold cross validation test was utilized to optimize the SVM parameters and select the features, whereas the jackknife test was used to evaluate the performances of the different learning algorithms and provide an unbiased assessment of the performance of the different prediction models.

In literature, Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), and the Matthews correlation coefficient (MCC) are routinely used as evaluation indexes [52], [53], which are defined as follows:

$$\begin{cases} Sen = 1 - \frac{N_{-}^{+}}{N_{+}^{+}} \\ Spe = 1 - \frac{N_{+}^{-}}{N_{-}^{-} + N_{+}^{-}} \\ Acc = 1 - \frac{N_{-}^{+} + N_{+}^{-}}{N_{+}^{+} + N_{-}^{-}} \\ MCC = \frac{1 - \frac{N_{-}^{+} + N_{+}^{-}}{N_{+}^{+} + N_{-}^{-}}}{\sqrt{\left(1 + \frac{N_{-}^{-} - N_{+}^{-}}{N_{+}^{+}}\right) \left(1 + \frac{N_{-}^{+} - N_{+}^{+}}{N_{-}^{-}}\right)}} \end{cases} \quad (15)$$

where N_{+}^{+} is the total number of the positive samples or true m⁶A sites, N_{-}^{+} is the number of true m⁶A sites that were incorrectly predicted to be a non-m⁶A site, N_{-}^{-} is the total number of the negative samples or non-m⁶A sites, and N_{+}^{-} is the number of non-m⁶A sites that were incorrectly predicted to be an m⁶A site. Please note that this set of indexes is only valid

TABLE I
PREDICTION RESULTS OF DIFFERENT FEATURES USING DIFFERENT PREDICTION ENGINES ON THE MET2614 DATASET OVER A 10-FOLD CROSS VALIDATION¹

Feature ²	Prediction engine	<i>Sen</i> (%)	<i>Spe</i> (%)	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>
PSNP	RF	70.07 (±0.71)	75.29 (±1.13)	72.68 (±0.71)	0.454 (±0.014)	0.777 (±0.006)
	KNN	74.17 (±0.26)	73.22 (±0.46)	73.70 (±0.32)	0.473 (±0.006)	0.783 (±0.001)
	SVM	74.15 (±0.16)	74.43 (±0.12)	74.29 (±0.07)	0.485 (±0.001)	0.790 (±0.001)
PSDP	RF	70.78 (±0.25)	75.40 (±0.24)	73.09 (±0.15)	0.462 (±0.003)	0.784 (±0.001)
	KNN	72.69 (±0.36)	73.56 (±0.31)	73.12 (±0.24)	0.462 (±0.004)	0.779 (±0.001)
	SVM	72.38 (±0.25)	74.85 (±0.36)	73.62 (±0.25)	0.472 (±0.005)	0.788 (±0.001)
NC	RF	61.02 (±0.57)	62.19 (±0.93)	61.60 (±0.43)	0.232 (±0.008)	0.666 (±0.005)
	KNN	69.08 (±0.58)	54.48 (±0.55)	61.78 (±0.51)	0.238 (±0.010)	0.664 (±0.002)
	SVM	64.55 (±0.43)	63.34 (±0.60)	64.45 (±0.38)	0.288 (±0.007)	0.702 (±0.002)
PSNP+PSDP	RF	70.71 (±0.70)	75.88 (±0.44)	73.29 (±0.33)	0.466 (±0.006)	0.784 (±0.002)
	KNN	74.74 (±0.38)	72.78 (±0.19)	73.76 (±0.22)	0.475 (±0.004)	0.785 (±0.001)
	SVM	73.65 (±0.25)	75.39 (±0.24)	74.52 (±0.15)	0.490 (±0.003)	0.792 (±0.001)
PSNP+PSDP+NC	RF	71.80 (±0.80)	76.60 (±0.62)	74.20 (±0.54)	0.484 (±0.010)	0.799 (±0.003)
	KNN	75.65 (±0.26)	72.54 (±0.30)	74.10 (±0.17)	0.482 (±0.003)	0.793 (±0.001)
	SVM	74.38 (±0.28)	77.16 (±0.48)	75.77 (±0.27)	0.515 (±0.005)	0.815 (±0.001)

¹The experiment was performed 10 times for each feature and prediction engine. The average performance of each evaluation index is reported, followed by a standard deviation.

²PSNP, PSDP, and NC stand for position-specific nucleotide propensity, position-specific dinucleotide propensity, and nucleotide composition, respectively.

for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology [54] and system medicine [55], a completely different set of indexes as defined in [56] is needed.

For a soft-type classifier (e.g., the SVM used in this study), whose outputs are the continuous numerical values representing the probabilities/confidences of a feature belonging to certain classes, gradual adjustments to the discrimination threshold will produce a series of prediction confusion matrices [57]. Each of these confusion matrices generates different corresponding values of N_+^- and N_-^+ . In other words, the evaluation indexes defined in (15) are threshold-dependent. Therefore, we also employed another threshold-independent evaluation index, i.e., Area under Curve (*AUC*), which is based on the Receiver Operating Characteristic (ROC) analysis [58]. The *AUC* values range from 0 to 1. The larger the *AUC* value, the better the predictor performs. Specifically, an *AUC* = 0.5 means that the predictor generates a random prediction, whereas an *AUC* = 1 represents a perfect predictor, which always generates the correct prediction.

III. RESULTS AND DISCUSSION

A. Contributions of Different Features

A series of comparative experiments were performed using individual features (i.e., PSNP, PSDP, and NC) and their combinations (i.e., PSNP+PSDP and PSNP+PSDP+NC) to evaluate the contributions of different features to the m⁶A site predictions. We evaluated the discriminative performances of each of the five features (i.e., three individual features and the two combined features) using three popular prediction engines, i.e., Support Vector Machine (SVM) [33], Random Forests (RF) [59], and *K* Nearest Neighbor (KNN) [60], over a 10-fold cross validation. Table 1 summarizes the

prediction performances of each feature using the different prediction engines. Note that for each feature and prediction engine, the 10-fold cross validation experiment was performed 10 times. The average performance of each evaluation index is reported, followed by a standard deviation.

As shown in Table 1, several conclusions can be drawn.

First, for all five types of features, SVM consistently performed better than RF and KNN concerning the two overall evaluation indexes (i.e., *MCC* and *AUC*). This phenomenon indicates that SVM is the most appropriate engine among the three considered engines for predicting m⁶A sites using the features developed in this study.

Second, for all three prediction engines, the two newly developed features, i.e., PSNP and PSDP, performed significantly better than NC. Using the SVM prediction engine as an example, the NC feature achieved a prediction performance with an *MCC* < 0.30 and *AUC* < 0.71, whereas the proposed PSNP and PSDP features achieved much better performance with an *MCC* > 0.47 and *AUC* > 0.78.

Third, the combination of individual features can further improve prediction performance. By revisiting Table I, we found that the combined PSNP+PSDP feature consistently outperformed the individual features, i.e., PSNP and PSDP, for all three prediction engines. On the other hand, the NC feature performed significantly worse than the other two individual features. As illustrated in Table I, the NC feature only achieved an *MCC* of 0.288 and an *AUC* of 0.702 using SVM as the prediction engine. However, we found that the prediction performance can be greatly improved by adding the NC feature to the PSNP+PSDP feature. As shown in Table I, the PSNP+PSDP+NC feature (SVM as prediction engine) achieved the best performance with an *MCC* of 0.515 and an *AUC* of 0.815, which were 2.5% and 2.3%, respectively,

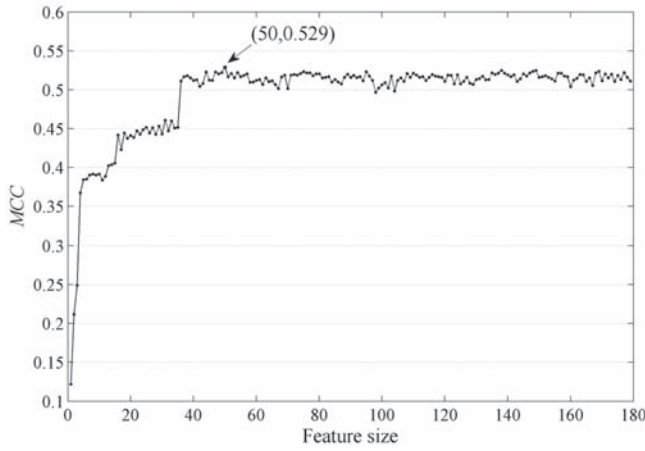


Fig. 1. The curve of the *MCC* values versus the sizes of the feature subset. The maximum *MCC* value is 0.529 and the corresponding feature size is 50.

better than the PSNP+PSDP feature.

Based on the three observations described above, we used PSNP+PSDP+NC as the input feature and SVM as the prediction engine to construct the proposed TargetM6A method.

B. Feature Selection Performance

In this section, we will try to further improve the computational efficiency and prediction performance by removing the redundant components in the PSNP+PSDP+NC feature with the incremental feature selection (IFS) procedure described in the “Incremental feature selection” section.

We first obtained the *p*-value for each component of the PSNP+PSDP+NC feature by performing *F*-test [44], [45]. Then, all the initial feature components were re-ranked by their *p*-values in ascending order. The list of the detailed re-ranked feature components for the Met2614 dataset is shown in Table S1 (see Supplementary Material II). Then, based on the ranked feature components, the IFS procedure was applied, and 179 feature subsets (for Met2614) were obtained. We evaluated the discriminative capability (measured by *MCC*) of each feature subset using the SVM prediction engine over a 10-fold cross validation. Fig. 1 plots the curve of *MCC* values versus the sizes of the feature subset.

As illustrated in Fig. 1, the *MCC* value almost consistently increases when the feature size varies from 1 to 50. The *MCC* value reaches a peak (0.529) when the feature size is 50. When the feature size is larger than 50, the *MCC* value fluctuates and no improvement was observed.

To extensively investigate the effectiveness of the feature selection procedure, we further performed the stringent jackknife test (i.e., leave-one-out cross validation) on the Met2614 dataset with SVM as prediction engine using the following two feature subsets: (1) the optimal feature subset, which consists of the top 50 feature components (denoted as Top50), and (2) the entire feature set, which consists of 179 initial feature components (denoted as All179). Table II summarizes the prediction performances of the two feature subsets on the Met2614 dataset over both the 10-fold cross validation and jackknife tests.

TABLE II

PREDICTION PERFORMANCES OF THE TOP50 AND ALL179 FEATURE SUBSETS OF THE MET2614 DATASET OVER BOTH THE 10-FOLD CROSS VALIDATION AND JACKKNIFE TESTS

Feature subset	10-fold cross-validation			jackknife test		
	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>
Top50	76.18 (± 0.28)	0.523 (± 0.005)	0.819 (± 0.001)	76.32	0.526	0.818
All179	75.77 (± 0.27)	0.515 (± 0.005)	0.815 (± 0.001)	75.67	0.513	0.816

TABLE III

COMPARISON OF THE RESULTS FROM THE PROPOSED TARGETM6A METHOD AND THE EXISTING PREDICTORS ON MET2614 USING JACKKNIFE TESTS

Prediction methods	<i>Sen</i> (%)	<i>Spe</i> (%)	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>
BLAST ¹	71.56	38.79	55.27	0.109	-
iRNAMethyl ¹	60.63	70.55	65.59	0.313	0.701
pRNAm-PC ²	69.72	69.75	69.74	0.394	0.762
m6Apred ³	73.91	74.45	74.18	0.483	0.792
TargetM6A	74.83	77.81	76.32	0.526	0.818

¹Results excerpted from [26].

²Results excerpted from [28].

³The re-implementation of m6Apred [27] on Met2614.

As shown in Table II, the Top50 feature subset consistently outperformed the All179 feature subset with regard to the three overall evaluation indexes, i.e., *Acc*, *MCC*, and *AUC*, over both the 10-fold cross validation and jackknife tests. Using the jackknife test as an example, the Top50 feature subset achieved an *Acc* of 76.32%, an *MCC* of 0.526, and an *AUC* of 0.818, which were 0.6%, 1.3%, and 0.2%, respectively, better than those of the All179 feature subset. Similar observations can also be obtained using 10-fold cross validation. These findings indicate that there is redundant information in the original 179 components, and the 50 selected feature components are truly the “significant” components for identifying the m⁶A sites. On one hand, the prediction performance can be slightly improved using the dimensionality-reduced feature subset, i.e., Top50; on the other hand, the computational efficiency is enhanced.

C. Comparisons With Existing Predictors

In this section, we will compare the proposed TargetM6A method with existing computational methods for predicting m⁶A sites, including iRNAMethyl [26], pRNAm-PC [28], and m6Apred [27], to demonstrate its efficacy.

We first compared these m⁶A site predictors using the Met2614 dataset and rigorous jackknife tests. The BLAST-based predictor [26] was used as the baseline predictor. The BLAST-based method predicts m⁶A sites by applying the sequence similarity search tool, i.e., BLAST [61], as follows: a query sample will be predicted as the true N⁶-methylation RNA segment if it is the most similar to the samples in the positive subset; otherwise, it will be predicted as a non-methylated RNA segment [26]. Because m6Apred [27] does not provide results on Met2614, we thus re-implemented it and evaluated it using Met2614. Please note that TargetM6A utilized the reduced feature subset, i.e., Top50, as the feature

TABLE IV
PREDICTION RESULTS OF DIFFERENT FEATURES USING DIFFERENT PREDICTION ENGINES ON
THE MET2614 DATASET OVER A 10-FOLD CROSS VALIDATION*

Methods	Train1664					Test5225				
	<i>Sen</i> (%)	<i>Spe</i> (%)	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>	<i>Sen</i> (%)	<i>Spe</i> (%)	<i>Acc</i> (%)	<i>MCC</i>	<i>AUC</i>
m6Apred *	79.21	77.04	78.12	0.563	0.840	53.89	79.07	76.78	0.222	-
TargetM6A	79.45	77.88	78.67	0.573	0.847	53.68	78.02	75.81	0.210	0.708

*Results excerpted from [27].

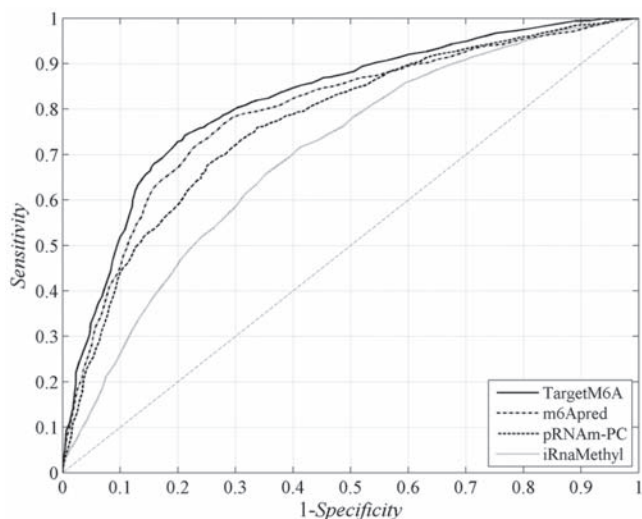


Fig. 2. The ROC curves of iRNAMethyl, m6Apred, pRNAm-PC, and the proposed TargetM6A method. The *AUC* values of these methods are 0.701, 0.762, 0.792, and 0.818, respectively.

representation and SVM as the prediction engine. In addition, the default threshold, i.e., 0.5, was used for TargetM6A and the re-implemented m6Apred to produce the *Sen*, *Spe*, *Acc*, and *MCC* values. Table III summarizes the results of the comparison of the different predictors on the Met2614 dataset using the jackknife test, and Fig. 2 compares their ROC curves.

As shown in Table III, we can find that the proposed TargetM6A method significantly outperformed the baseline predictor; the *MCC* value was remarkably improved from 0.109 to 0.526. TargetM6A also outperformed iRNAMethyl [26], pRNAm-PC [28], and m6Apred [27] with respect to all of the five evaluation indexes. Using *MCC* as an example, 21.3%, 13.2%, and 4.3% improvements were observed compared with iRNAMethyl [26], pRNAm-PC [28], and m6Apred [27].

In addition, the ROC curves plotted in Fig. 2, together with the *AUC* values listed in Table III, indicate that TargetM6A achieved the best performance regarding *AUC*. TargetM6A outperformed iRNAMethyl, pRNAm-PC, and m6Apred for the *AUC*, with improvements of 11.7%, 5.6%, and 2.6%, respectively. Considering that the jackknife test is the most rigorous cross validation method, these experimental results indeed show the superiority of the proposed TargetM6A method over the existing m⁶A site predictors.

We further evaluated the efficacy of the proposed method using two other benchmark datasets, i.e., Train1664 and Test5225. Please note that the original feature vector extracted

from the Train1664 dataset is 119-D (18-D PSNP, 17-D PSDP, and 84-D NC features), as described in the “Feature representation of RNA segments” section. Then, the IFS procedure was performed and an optimized feature subset of 65 feature components (see Table S2 in Supplementary Material II for details) was obtained to construct the SVM-based prediction model.

We first evaluated TargetM6A on Train1664 using the jackknife test. Then, we trained TargetM6A on Train1664 and tested the trained model with the independent validation dataset, i.e., Test5225. Because only m6Apred [27] currently provides results on Train1664 and Test5225, we only compared the proposed TargetM6A method with m6Apred. The performances of m6Apred were excerpted from a previous study [27]. Table IV summarizes the comparison of the results from m6Apred and the proposed TargetM6A method using the Train1664 and Test5225 datasets.

As shown in Table IV, the proposed TargetM6A method achieved very comparable performances with m6Apred for both Train1664 and Test5225. It has not escaped our notice that TargetM6A achieved much better performance than m6Apred for Met2614 (refer to Table III). However, TargetM6A failed to improve the prediction performance and only achieved comparable performance with m6Apred for Train1664 and Test5225 (refer to Table IV). We speculate that the differences in the RNA segment length and amount of data between Met2614 and Train1664 may account for this phenomenon: the RNA segment length ($\xi = 24$) and data volume (2614 samples) of Met2614 are much larger than those of the Train1664 dataset ($\xi = 9$, 1664 samples). Because the proposed method extracts the features according to the distribution of nucleotides in the central GAG motif, a dataset with larger RNA segment length and data volume will contain much more information about the distribution of nucleotides and thus will be beneficial for extracting more discriminative features. We believe that the prediction performance of TargetM6A can be further improved by increasing the number of available methylated and non-methylated RNA samples.

D. Web Server Implementation

A web server has been placed online at <http://csbio.njust.edu.cn/bioinf/TargetM6A> to enhance the applicability of the proposed TargetM6A method. The final online TargetM6A method was trained on Met2614, where each training sample is a 51-nt RNA segment. The TargetM6A method predicts m⁶A sites as described below.

For a query RNA sequence submitted by a user, TargetM6A first identifies all the GAC motifs in the query sequence; then,

for each GAC motif, a corresponding 51-nt RNA segment is constructed by placing a sliding window centered on the GAC motif. Based on the constructed 51-nt RNA segment, the features are extracted and then fed to the SVM classification engine to perform the prediction. Please note that to construct the corresponding 51-nt segment for a GAC motif with an insufficient number of nucleotides around the sequence (e.g., a motif near sequence terminals or a query RNA sequence shorter than 51-nt), the previously proposed “mirror image” technique [26] is used to fill in the missing nucleotides.

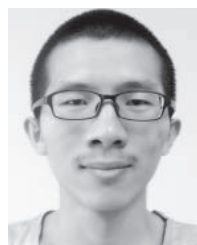
IV. CONCLUSIONS

In this study, we established a computational predictor for targeting the N⁶-methyladenosine sites in RNA sequences. Inspired by the successful applications of the position-specific amino acid propensity features in proteomics, we introduced the new PSNP and PSDP features to encode RNA sequences. Based on the newly developed feature, we implemented an m⁶A site predictor called TargetM6A. We compared the proposed TargetM6A method with the most recently released m⁶A site predictors by performing stringent jackknife tests and independent validation tests on the benchmark datasets. The experimental results show that our TargetM6A method achieved high prediction performance and outperformed the existing predictors. A web-server has been placed online at <http://csbio.njust.edu.cn/bioinf/TargetM6A> to enhance the applicability of the proposed method. The findings of this study enrich our understanding of sequence-based m⁶A sites and can potentially be applied to other nucleotide sequence-related prediction problems.

REFERENCES

- [1] K.-C. Chou, “Impacts of bioinformatics to medicinal chemistry,” *Med. Chem.*, vol. 11, no. 3, pp. 218–234, 2015.
- [2] C. Yi and T. Pan, “Cellular dynamics of RNA modification,” *Accounts Chem. Res.*, vol. 44, no. 12, pp. 1380–1388, Dec. 2011.
- [3] C. He, “Grand challenge commentary: RNA epigenetics?” *Nature Chem. Biol.*, vol. 6, pp. 863–865, Dec. 2010.
- [4] W. A. Cantara *et al.*, “The RNA modification database, RNAMDB: 2011 update,” *Nucl. Acids Res.*, vol. 39, no. 1, pp. D195–D201, Jan. 2011.
- [5] A. Czerwionec *et al.*, “MODOMICS: A database of RNA modification pathways. 2008 update,” *Nucl. Acids Res.*, vol. 37, pp. D118–D121, Jan. 2009.
- [6] Y. Fu, D. Dominissini, G. Rechavi, and C. He, “Gene expression regulation mediated through reversible m⁶A RNA methylation,” *Nature Rev. Genet.*, vol. 15, pp. 293–306, May 2014.
- [7] G. Jia, Y. Fu, and C. He, “Reversible RNA adenosine methylation in biological regulation,” *Trends Genet.*, vol. 29, no. 2, pp. 108–115, Feb. 2013.
- [8] X. Wang *et al.*, “N⁶-methyladenosine-dependent regulation of messenger RNA stability,” *Nature*, vol. 505, pp. 117–120, Jan. 2014.
- [9] K. D. Meyer and S. R. Jaffrey, “The dynamic epitranscriptome: N⁶-methyladenosine and gene expression control,” *Nature Rev. Mol. Cell Biol.*, vol. 15, pp. 313–326, May 2014.
- [10] J. Liu *et al.*, “A METTL3–METTL14 complex mediates mammalian nuclear RNA N⁶-adenosine methylation,” *Nature Chem. Biol.*, vol. 10, pp. 93–95, Feb. 2014.
- [11] X.-L. Ping *et al.*, “Mammalian WTAP is a regulatory subunit of the RNA N⁶-methyladenosine methyltransferase,” *Cell Res.*, vol. 24, pp. 177–189, Feb. 2014.
- [12] G. Jia *et al.*, “N⁶-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO,” *Nature Chem. Biol.*, vol. 7, pp. 885–887, Dec. 2011.
- [13] G. Zheng *et al.*, “ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility,” *Mol. Cell*, vol. 49, no. 1, pp. 18–29, Jan. 2013.
- [14] R. Desrosiers, K. Friderici, and F. Rottman, “Identification of methylated nucleosides in messenger RNA from novikoff hepatoma cells,” *Proc. Natl. Acad. Sci. USA*, vol. 71, no. 10, pp. 3971–3975, Oct. 1974.
- [15] X. Deng *et al.*, “Widespread occurrence of N⁶-methyladenosine in bacterial mRNA,” *Nucl. Acids Res.*, vol. 43, no. 13, pp. 6557–6567, Jul. 2015.
- [16] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan, “N⁶-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions,” *Nature*, vol. 518, pp. 560–564, Feb. 2015.
- [17] C. R. Alarcón, H. Lee, H. Goodarzi, N. Halberg, and S. F. Tavazoie, “N⁶-methyladenosine marks primary microRNAs for processing,” *Nature*, vol. 519, pp. 482–485, Mar. 2015.
- [18] C. Xu *et al.*, “Structural basis for selective binding of m⁶A RNA by the YTHDC1 YTH domain,” *Nature Chem. Biol.*, vol. 10, pp. 927–929, Nov. 2014.
- [19] D. Theler, C. Dominguez, M. Blatter, J. Boudet, and F. H.-T. Allain, “Solution structure of the YTH domain in complex with N⁶-methyladenosine RNA: A reader of methylated RNA,” *Nucl. Acids Res.*, vol. 42, no. 22, pp. 1399–13911, Dec. 2014.
- [20] S. Kellner, J. Burhenne, and M. Helm, “Detection of RNA modifications,” *RNA Biol.*, vol. 7, pp. 237–247, Mar./Apr. 2010.
- [21] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, and G. Rechavi, “Transcriptome-wide mapping of N⁶-methyladenosine by m⁶A-seq based on immunocapturing and massively parallel sequencing,” *Nature Protocols*, vol. 8, pp. 176–189, Jan. 2013.
- [22] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, “Comprehensive analysis of mRNA methylation reveals enrichment in 3’ UTRs and near stop codons,” *Cell*, vol. 149, no. 7, pp. 1635–1646, Jun. 2012.
- [23] Y. Li, X. Wang, C. Li, S. Hu, J. Yu, and S. Song, “Transcriptome-wide N⁶-methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification,” *RNA Biol.*, vol. 11, no. 9, pp. 1180–1188, 2014.
- [24] S. Schwartz *et al.*, “High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis,” *Cell*, vol. 155, no. 6, pp. 1409–1421, 2013.
- [25] D. Dominissini *et al.*, “Topology of the human and mouse m⁶A RNA methylomes revealed by m⁶A-seq,” *Nature*, vol. 485, pp. 201–206, May 2012.
- [26] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, “iRNA-Methyl: Identifying N⁶-methyladenosine sites using pseudo nucleotide composition,” *Anal. Biochem.*, vol. 490, pp. 26–33, Dec. 2015.
- [27] W. Chen, H. Tran, Z. Liang, H. Lin, and L. Zhang, “Identification and analysis of the N⁶-methyladenosine in the *Saccharomyces cerevisiae* transcriptome,” *Sci. Rep.*, vol. 5, p. 13859, Sep. 2015.
- [28] Z. Liu, X. Xiao, D.-J. Yu, J. Jia, W.-R. Qiu, and K.-C. Chou, “pRNAm-PC: Predicting N⁶-methyladenosine sites in RNA sequences via physical–chemical properties,” *Anal. Biochem.*, vol. 497, pp. 60–67, Mar. 2016.
- [29] Y. Liu, W. Gu, W. Zhang, and J. Wang, “Predict and analyze protein glycation sites with the mRMR and IFS methods,” *Biomed. Res. Int.*, vol. 2015, Dec. 2015, Art. no. 561547.
- [30] J. Wang, D. Zhang, and J. Li, “PREAL: Prediction of allergenic protein by maximum relevance minimum redundancy (mRMR) feature selection,” *BMC Syst. Biol.*, vol. 7, no. 5, p. S9, 2013.
- [31] C. Zou, J. Gong, and H. Li, “An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis,” *BMC Bioinform.*, vol. 14, p. 90, Mar. 2013.
- [32] H. Liu and R. Setiono, “Incremental feature selection,” *Appl. Intell.*, vol. 9, no. 3, pp. 217–230, 1998.
- [33] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [34] J. Brayet, F. Zehraoui, L. Jeanson-Leh, D. Israeli, and F. Tahi, “Towards a pRNA prediction using multiple kernel fusion and support vector machine,” *Bioinformatics*, vol. 30, pp. i364–i370, Sep. 2014.
- [35] E. K. M. Hashim and R. Abdullah, “Rare K-mer DNA: Identification of sequence motifs and prediction of CPG island and promoter,” *J. Theor. Biol.*, vol. 387, pp. 88–100, Dec. 2015.
- [36] H. Vinje, K. H. Liland, T. Almøy, and L. Snipen, “Comparing K-mer based methods for improved classification of 16S sequences,” *BMC Bioinform.*, vol. 16, p. 205, Jul. 2015.
- [37] S. H. Guo *et al.*, “iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition,” *Bioinformatics*, vol. 30, no. 11, pp. 1522–1529, 2014.

- [38] H. Lin, E. Z. Deng, H. Ding, W. Chen, and K. C. Chou, "iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucl. Acids Res.*, vol. 42, pp. 12961–12972, Oct. 2014.
- [39] K. C. Chou, "A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins," *J. Biol. Chem.*, vol. 268, pp. 16938–16948, Aug. 1993.
- [40] M. C. Frith, U. Hansen, and Z. Weng, "Detection of CIS-element clusters in higher eukaryotic DNA," *Bioinformatics*, vol. 17, pp. 878–889, Oct. 2001.
- [41] Y. R. Tang, Y. Z. Chen, C. A. Canchaya, and Z. Zhang, "GANNPhos: A new phosphorylation site predictor based on a genetic algorithm integrated neural network," *Protein Eng. Design Sel.*, vol. 20, pp. 405–412, Aug. 2007.
- [42] A. M. Thangakani, S. Kumar, R. Nagarajan, D. Velmurugan, and M. M. Gromiha, "GAP: Towards almost 100 percent prediction for SS-strand-mediated aggregating peptides with distinct morphologies," *Bioinformatics*, vol. 30, pp. 1983–1990, Jul. 2014.
- [43] Y. Xu, Y. X. Ding, J. Ding, L. Y. Wu, and N. Y. Deng, "Phogly-PseAAC: Prediction of lysine phosphoglycerylation in proteins incorporating with position-specific propensity," *J. Theor. Biol.*, vol. 379, pp. 10–15, Aug. 2015.
- [44] A. Golugula, G. Lee, and A. Madabhushi, "Evaluating feature selection strategies for high dimensional, small sample size datasets," in *Proc. Conf. Proc. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 949–952.
- [45] Y. Mao, X. Zhou, D. Pi, Y. Sun, and S. T. Wong, "Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection," *J. Biomed. Biotechnol.*, Jun. 2005, pp. 160–171.
- [46] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
- [47] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2006, Art. no. 27.
- [48] D. J. Yu, J. Hu, J. Yang, H. B. Shen, J. Tang, and J. Y. Yang, "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 10, no. 4, pp. 994–1008, Jul./Aug. 2013.
- [49] D. J. Yu, J. Hu, H. Yan, X. B. Yang, J. Y. Yang, and H. B. Shen, "Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble," *BMC Bioinform.*, vol. 15, p. 297, Sep. 2014.
- [50] K. C. Chou and C. T. Zhang, "Prediction of protein structural classes," *Critical Rev. Biochem. Mol. Biol.*, vol. 30, no. 4, pp. 275–349, 1995.
- [51] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, 2011.
- [52] K. C. Chou, "Using subsite coupling to predict signal peptides," *Protein Eng.*, vol. 14, pp. 75–79, Feb. 2001.
- [53] W. Chen, H. Ding, P. Feng, H. Lin, and K. C. Chou, "iACP: A sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, pp. 16895–16909, Mar. 2016.
- [54] K. C. Chou, Z. C. Wu, and X. Xiao, "iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Mol. Biosyst.*, vol. 8, pp. 629–641, Feb. 2012.
- [55] X. Xiao, P. Wang, W. Z. Lin, J. H. Jia, and K. C. Chou, "iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Anal. Biochem.*, vol. 436, pp. 168–177, May 2013.
- [56] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Mol. Biosyst.*, vol. 9, pp. 1092–1100, Jun. 2013.
- [57] D. J. Yu *et al.*, "TargetATPsite: A template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble," *J. Comput. Chem.*, vol. 34, pp. 974–985, Apr. 2013.
- [58] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
- [59] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [60] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
- [61] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, Oct. 1990.
- [62] J. O. Williams, "Narrow-band analyzer," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, USA, 1993.



Guang-Qing Li received his B.S. degree in computer science and technology from Nanjing University of Science and Technology, China, in 2014. Currently, he is working towards the M.S. degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include bioinformatics, data mining, and pattern recognition.



Zi Liu received his M.S. degree in computer science and technology from Jingdezhen Ceramic Institute, China, in 2015. Currently, he is working towards the Ph.D. degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include bioinformatics, biomedical image processing, and pattern recognition.



Hong-Bin Shen received his Ph.D. degree from Shanghai Jiaotong University, China, in 2007. He was a Postdoctoral Research Fellow of Harvard Medical School from 2007 to 2008. Currently, he is a Professor of Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University. His research interests include data mining, pattern recognition, and bioinformatics. Dr. Shen has published more than 60 papers and constructed 20 bioinformatics servers in these areas and he serves the editorial members of several international journals.



Dong-Jun Yu (M'12) received the B.S. degree in computer science and the M.S. degree in artificial intelligence from Jiangsu University of Science and Technology, China, in 1997 and 2000, respectively, and the Ph.D. degree in pattern analysis and machine intelligence from Nanjing University of Science and Technology, China, in 2003. In 2008, he acted as an academic visitor at University of York, U.K. He is currently a Professor in the School of Computer Science and Engineering of Nanjing University of Science and Technology. His current interests include bioinformatics, pattern recognition, and data mining. He is a Senior Member of CCF.