Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

RNA-MethylPred: A high-accuracy predictor to identify N6-methyladenosine in RNA

Cang-Zhi Jia*, Jia-Jia Zhang, Wei-Zhen Gu

Department of Mathematics, Dalian Maritime University, Dalian 116026, China

ARTICLE INFO

Article history:
Received 11 March 2016
Received in revised form
7 June 2016
Accepted 8 June 2016
Available online 20 June 2016

Keywords:
RNA methylation
Bi-profile Bayes
Dinucleotide frequency
k nearest neighbor
Support vector machine

ABSTRACT

N6-methyladenosine (m^6A) is present ubiquitously in the RNA of living organisms from *Escherichia coli* to humans. Nonetheless, the exact molecular mechanism of this modification remains unclear. The experimental identification of m^6A modification is time-consuming and expensive; therefore, bioinformatics tools with high accuracy represent desirable alternatives for the large-scale, rapid identification of N6-methyladenosine sites. In this study, RNA-MethylPred, a new bioinformatics model, was developed by incorporating bi-profile Bayes, dinucleotide composition, and *k* nearest neighbor (KNN) scores for three feature extractions. RNA-MethylPred yielded a Matthew's correlation coefficient (MCC) of 0.53 in a jackknife test, which was 0.24 higher than that of iRNA-Methyl and 0.13 higher than that of pRNAm-PC. The obvious improvements demonstrated that RNA-MethylPred might be a powerful and complementary tool for further experimental investigation of N6-methyladenosine modification.

N6-methyladenosine (m^6A) is the most ubiquitous and abundant modification present in mRNA and long noncoding RNA across eukaryotes, and it functions in various pathways of RNA metabolism, including mRNA splicing, nuclear export, translation, and degradation. It also has a vital role in the regulation of gene expression [1]. The m^6A modification is involved in diverse life activities such as embryo development, cell apoptosis, spermatogenesis, and the circadian clock. As a reversible and dynamic modification, m^6A is co-regulated by methyltransferases and demethylases [1–3].

The development of high-throughput sequencing technology, particularly methylated RNA immunoprecipitation sequencing (MeRIP-seq), has enabled high-efficiency detection of whole transcriptomes with different RNA methylation patterns. Currently, the genome-wide distribution of m^6A is available for several species, such as *Saccharomyces cerevisiae* [4], *Mus musculus* [5], and *Homo sapiens* [5], providing an important basis for methylation research. However, a major limitation of this approach is the relatively low resolution, which is restricted by an average 100-nucleotide (nt) length of the RNA fragments used for immunoprecipitation [1].

The bioinformatics approach, as a powerful auxiliary tool for experimental verification, has been widely applied to other types of post-translational modifications of proteins and RNA [6–10]. To the best of our knowledge, there are only two site prediction tools for the m^6A modification. One of them is iRNA-Methyl, which incorporates three physicochemical properties for 16 different dinucleotides into the general form of the dinucleotide composition as feature matrix and uses support vector machine (SVM) as the classifier [11]. It is worth noting that the prediction model iRNA-Methyl achieved a Matthew's correlation coefficient (MCC) of only 0.29 in a jackknife test. The other method is pRNAm-PC [12], in which 10 physicochemical properties for 16 different dinucleotides are mapped into a pseudo dinucleotide composition via a series of auto-covariance and cross-covariance transformations. This approach attained a MCC of 0.40 when evaluated on the same dataset as used for iRNA-Methyl [11]. The overall performances of the aforementioned two predictors are not fully satisfactory; therefore, there is scope to improve the predictive accuracy.

Advance machine learning techniques were employed in the protein or RNA identification, including secondary structure features [13], ensemble learning [14,15], feature ranking and reducing [16], and sample selection [17,18]. In this work, three kinds of feature extraction strategy were used to improve the detection accuracy of m^6A sites. First, bi-profile Bayes, as an effective feature

Abbreviations used: m^6A, N6-methyladenosine; nt, nucleotide; SVM, support vector machine; MCC, Matthew's correlation coefficient; KNN, *k* nearest neighbor; BPB, bi-profile Bayes; DNC, dinucleotide composition; PseAAC, pseudo amino acid composition; Sn, sensitivity; Sp, specificity; Acc, accuracy; ROC, receiver operating characteristic.

* Corresponding author.
E-mail address: cangzhijia@dlmu.edu.cn (C.-Z. Jia).

http://dx.doi.org/10.1016/j.ab.2016.06.012
0003-2697/© 2016 Elsevier Inc. All rights reserved.

extraction method, was employed [19–21] to reflect the posterior probability of positive and negative samples. Second, two forms of dinucleotide composition were applied to reflect sequence order information. Third, the k nearest neighbor (KNN) scores were applied to measure whether the local sequence surrounding m^6A is more similar to the methylation segments or to non-methylation segments in the RNA. These three features were combined with an SVM classifier to build our prediction model, which was named RNA-MethylPred. Comparisons with the above-mentioned two methods revealed that RNA-MethylPred is generally useful for identification of m^6A modification.

Materials and methods

Datasets

To compare our approach with other available approaches objectively and comprehensively, we used the same datasets used in [11,12]. The length of sequence segments from 49 to 59 were optimized, and the detailed results are shown in Table S1 of the online supplementary material. Finally, sequence segments of 51 nt around the methylation sites and non-methylation sites in RNA were extracted as positive and negative training datasets, respectively. The benchmark dataset contained 1307 positive samples and 1307 negative samples with less than 85% pairwise sequence identity (see Supplementary Materials S1 and S2).

Sequence feature representation

Bi-profile Bayes (BPB)

BPB, recently proposed in [19], outperformed other methods because of its consideration of information from both positive and negative training samples. It has been applied successfully in the fields of predicting protein methylation sites [19], caspase cleavage sites [20], mitochondrial proteins of malaria parasite [21], and type III secreted effectors [22], among others.

Given a peptide sequence S, we encoded this sequence into a probability vector $V = (p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_{2n})$, where $p_i(i = 1, 2, \ldots, n)$ denotes the posterior probability of each nucleic acid at the i-th position in the positive samples and $p_i(i = n + 1, n + 2, \ldots, 2n)$ denotes the posterior probability of each nucleic acid at the i-th position in the negative samples (n is the length of peptide sequences and $n = 51$ in the current study). The posterior probability of positive and negative samples was calculated as the occurrence of each nucleotide at each position in the positive training dataset and negative training dataset, respectively [19]. The feature vectors simultaneously contain positive and negative information, the dimension of which is $2n$.

Dinucleotide composition (DNC)

The concept of pseudo amino acid composition or Chou's PseAAC was proposed in 2001 and has penetrated rapidly into nearly all fields of computational proteomics [6,23–25]. For a brief introduction to Chou's PseAAC and its recent development and applications, see the comprehensive review in [26]. Recently, the concept of the pseudo component approach was further employed in the fields of computational genetics and genomics [27].

To avoid losing much important information hidden in RNA sequences or RNA fragments, in this study the adjoining dinucleotides composition was adopted to stand for an RNA sequence. We defined the two adjoining dinucleotide compositions that can be formulated as follows:

$$P_{ab} = \frac{N_{ab}}{N_{a\bullet}} \tag{1}$$

$$P'_{ab} = \frac{N_{ab}}{n-1}, \tag{2}$$

where ab stands for the adjoining dinucleotides, N_{ab} stands for the number of the adjoining dinucleotides in an RNA segment sample, $a\bullet$ stands for the adjoining dinucleotides, $\bullet$ stands for any nucleotide, and $n = 51$ is the length of RNA sample. The dimension of the feature vector is $4 \times 4 + 4 \times 4 = 32$.

KNN score

In spite of its simplicity, the KNN algorithm has been widely applied to classification problems. Recently, KNN has also been shown to be powerful in feature extraction [28–30].

For a query sequence, we first found its k nearest neighbors in both the positive and negative sets according to the RNA local sequence similarity. Given two local sequence fragments A and B, $S(A,B)$ represented the similarity scores of the two fragments. The similarity score is calculated as follows:

$$S(A,B) = \sum_{1 \le i \le 51} Score(A[i], B[i]), \tag{3}$$

where $A[i]$ stands for a nucleotide at position i in the RNA sequence fragment.

For two nucleotides a and b, the similarity score between them is defined as

$$Score = \begin{cases} +2, & if\ a = b \\ -1, & others \end{cases}. \tag{4}$$

We then calculated the percentage of the positive neighbors in its k nearest neighbors as the KNN score. In this study, the considered Ks were 10, 20, 30, ..., 200.

From previous work [28–30], we knew that informative KNN scores play an important role in the prediction of phosphorylation sites, ubiquitylation sites, and succinylation sites.

SVM implementation and performance evaluation

The SVM classification method has proven to be powerful in many fields of bioinformatics [11,12,19–22]. In this work, the SVM was trained with the LIBSVM package (version 3.0) [31] to build the model and perform the predictions. The radial basis kernel function $k(x_i, x_j) = \exp\{-\gamma \left\| x_i - x_j \right\|^2\}$ was selected, and the parameters $c = 128$, $\gamma = 0.03125$, optimized by the SVMcgForClass program, were downloaded from http://www.matlabsky.com.

The jackknife test is deemed as the least arbitrary test that can always yield a unique outcome for a given benchmark dataset [32]. Thus, we used the jackknife test to select important features and optimize all parameters.

To evaluate the proposed predictor, four measurements are calculated: sensitivity (Sn), specificity (Sp), accuracy (Acc), and MCC, which were defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{5}$$

$$Sp = \frac{TN}{TN + FP} \tag{6}$$