Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions†

Pengmian Feng, a Hui Ding, b Wei Chen* c and Hao Lin* b

RNA 5-methylcytosine (m\(^5\)C) plays an important role in numerous biological processes. Accurate identification of the m\(^5\)C site is helpful for a better understanding of its biological functions. However, the drawbacks of the experimental methods available preclude progress towards the identification of the m\(^5\)C site. As an excellent complement to experimental techniques, computational methods will facilitate the identification of m\(^5\)C. In the present study, a support vector machine based-method is proposed to identify m\(^5\)C sites in \textit{Homo sapiens}. In this method, RNA sequences are encoded using the pseudo dinucleotide composition in which three RNA physiochemical properties are incorporated. It was observed by the jackknife cross-validation that the overall success rate achieved by the proposed model is 90.42%. This result indicates that the proposed model holds the potential to become a useful tool for the identification of m\(^5\)C sites.

1. Introduction

The chemical modification of RNA is emerging as an important factor in the regulation of gene expression. Among the ~150 types of RNA modifications, RNA 5-methylcytosine (m\(^5\)C) has been discovered in archaea to eukaryotes, which is catalyzed by RNA methyltransferase (Fig. 1). Thus far, a series of biological functions of the m\(^5\)C modification in RNA have been explored. In tRNA, m\(^5\)C sites play key roles in stabilizing the tRNA secondary structure, thus affecting aminoacylation and codon recognition, and conferring metabolic stability. In rRNA, m\(^5\)C is shown to play important roles in translational fidelity and tRNA recognition.

Although the functions of m\(^5\)C in tRNA and rRNA have been well studied, progress in mRNA is precluded due to the lack of effective methods for the detection of m\(^5\)C in mRNA. Therefore, knowledge about the positions of the m\(^5\)C site is important to understand the mechanisms and functions of this post-transcriptional modification. To this end, many experimental techniques, such as bisulfite sequencing, m5C-RIP, Aza-IP and miCLIP, have been proposed to identify m\(^5\)C. However, as pointed out in a recent review, these experimental approaches still have drawbacks to detect m\(^5\)C. Therefore, it is necessary to develop new techniques to study the distribution and function of m\(^5\)C in RNA. As an excellent complement to experimental techniques, computational methods will speed up genome-wide m\(^5\)C detection.

The experimental methods have yielded quite encouraging results and provide m\(^5\)C site data in the \textit{H. sapiens} transcriptome, which give us an unprecedented opportunity to develop a computational model for identifying the m\(^5\)C site. To the best of our knowledge, there is no computational tool available for the discovery of m\(^5\)C in RNA. Therefore, in the present study, we propose a support vector machine based method to identify the m\(^5\)C sites in the \textit{H. sapiens} transcriptome. By using the pseudo nucleotide composition, long-range sequence-order effects and RNA physicochemical properties were integrated to formulate the samples. In the jackknife cross-validation, the proposed model obtained an overall accuracy of 90.42% for identifying m\(^5\)C sites.
2. Materials and methods

2.1. Benchmark datasets

By combining RNA bisulfite conversion with next-generation sequencing, m^5C sites have been detected in *H. sapiens* and have been manually checked and deposited in RMBase. Therefore, from RMBase, we obtained 680 m^5C site containing sequences in *H. sapiens*. All of these sequences in RMBase are 41 nt long with the m^5C site in the center. Preliminary tests indicate that the best prediction results are achieved when the sequence is 41 nt long. To avoid redundancy and reduce homology bias, sequences with more than 70% sequence similarity were removed using the CD-HIT program.

Negative samples were obtained by choosing 41 nt long sequences that satisfy the following rules: (1) cytosine is located at the center of the sequence; (2) the centered cytosine is not methylated, as proven in experiments. By doing so, we obtained a great number of negative samples. Therefore, the number of negative samples will be dramatically larger than those of the positive samples. In machine-learning problems, imbalanced datasets can significantly affect the accuracy of learning methods. To balance the numbers between positive and negative samples in model training, we randomly picked 120 cytosine containing sequences to form the negative samples. All negative samples were also 41 nt long. After following these procedures, we obtained a benchmark dataset including 120 m^5C site containing sequences and 120 non-m^5C site containing sequences, which are available in ESI† S1.

To examine whether the predictive accuracy is sensitive to the selection of negative data, we repeated the random sampling procedure ten times and obtained nine other random samples of negative datasets for model validation. These sequences are also 41 nt long and with a sequence identity of less than 70%. The other nine negative datasets are available in ESI† S2.

2.2. RNA sequence formulation

Suppose an RNA sequence with L nucleic acid residues, R_1 R_2 R_3 ... R_L, where R_i is the residue at position i and it can be adenine (A), cytosine (C), guanine (G), or urine (U), the straightforward method to formulate the RNA sequence is using its nucleic acid composition as given below,

\[ R = [f(A), f(C), f(G), f(U)]^T \]  

(1)

where, \( f(A), f(C), f(G), \) and \( f(U) \) are the normalized occurrence frequencies of adenine, cytosine, guanine, and urine in the RNA sequence. However, the sequence order information is missing. If the dinucleotide composition, i.e. \( f(AA), f(AC), f(AG), ... , f(UU) \), is used to represent the RNA sequence, although the most contiguous local sequence order information is included, none of the global sequence order information is reflected by the formulation.

To incorporate both the local and global sequence pattern information of the RNA sequences, the pseudo dinucleotide composition (PseDNC) was proposed to represent RNA sequences by incorporating the global sequence order information into the feature vector. Recently, some web-servers were developed to generate pseudo nucleotide compositions. The pseudo nucleotide composition has been widely used in the realm of computational genomics. Therefore, PseDNC was used to encode RNA sequences in the current study. According to eqn (5) and (6) in ref. 22, the pseudo dinucleotide composition can be defined as,

\[ D = [d_1, d_2, \ldots, d_{16}, d_{16+1}, \ldots, d_{16+\lambda}]^T \]  

(2)

where,

\[ d_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (1 \leq u \leq 16) \\ \frac{w \theta_{m-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (16 < u \leq 16 + \lambda) \end{cases} \]  

(3)

In eqn (3), \( f_u \) \((u = 1, 2, \ldots, 16)\) is the normalized occurrence frequency of the non-overlapping dinucleotides in the RNA sequence. \( \lambda \) is the number of total counted ranks (or tiers) of the correlations along an RNA sequence, and \( w \) is the weight factor. It is through the \( \lambda \) correlation factors that not only considerable global sequence-order effects can be incorporated but also RNA sequences with extreme differences in length can be converted into a set of feature vectors with the same dimension.

The concrete values for \( \lambda \) and \( w \) will be further discussed in Section 3.1 and

\[ \theta_j = \frac{1}{L-j+1} \sum_{i=1}^{L-j+1} \Theta(D_i, D_{i+j}) \quad (j = 1, 2, \ldots, \lambda; \quad \lambda < L) \]  

(4)

where, \( \theta_1 \) is called the first-tier correlation factor which reflects the sequence order correlation between all the most contiguous dinucleotides along an RNA sequence (Fig. 2a); \( \theta_2 \), the second-tier correlation factor between all the most contiguous...
dinucleotides (Fig. 2b); \( \theta_3 \), the third-tier correlation factor between all the third most contiguous dinucleotides (Fig. 2c); and so forth. The correlation function \( \Theta(D_i, D_{ij}) \) is given by

$$\Theta(D_i, D_{ij}) = \frac{1}{v} \sum_{u=1}^{v} \left[ P_u(D_i) - P_u(D_{ij}) \right]^2$$

(5)

where, \( v \) is the number of RNA physicochemical properties.

2.3. RNA structural properties

Since RNA 5-methylcytosine is closely correlated with the RNA secondary structure,\(^7\) the following three physicochemical properties, enthalpy,\(^23\) entropy\(^23\) and free energy\(^24\) that can quantify the RNA secondary structures,\(^25\)\( \text{ }-\)27 are used to calculate the global or long-range sequence-order effects via eqn (4) and (5). Therefore, \( v \) equal to 3 reflects the number of RNA physicochemical properties considered. The concrete values of these three physicochemical properties are given in Table 1. Note that before substituting them into eqn (5), all the original values \( P_u(D_i) \) \((u = 1, 2, 3)\) were subjected to a standard conversion, as described by the following equation,

$$P_u(D_i) = \frac{\left[ P_u(D_i) - \langle P_u(D_i) \rangle \right]}{SD(P_u(D_i))}$$

(6)

where, the symbol \( \langle \rangle \) means taking the average of the quantity therein over the 16 different dinucleotides, and SD means the corresponding standard deviation.

2.4 Support vector machine

Since it has been widely used in the realm of bioinformatics,\(^14\)\( \text{ }-\)20,\( \text{ }28\)\( \text{ }-\)33 the support vector machine was employed to perform the classifications in the current study. Due to its effectiveness and speed in the training process, the radial basis kernel function (RBF) was used to obtain the classification hyperplane of the SVM operation engine. In the current study, the LibSVM package 3.18 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used to implement the SVM. The grid search method was applied to optimize the regularization parameter \( C \) and kernel parameter \( \gamma \) of SVM using a grid search approach as defined by

$$2^{-5} \leq C \leq 2^{15} \text{ with step of 2}$$

$$2^{-15} \leq \gamma \leq 2^{-5} \text{ with step of } 2^{-1}$$

(7)

2.5 Evaluation metrics

The performance of the proposed method was evaluated by sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthew’s correlation coefficient (MCC), which are expressed as

$$\text{Sn} = \frac{TP}{TP + FN} \times 100\%$$

(8)

$$\text{Sp} = \frac{TN}{TN + FP} \times 100\%$$

(9)

$$\text{Acc} = \frac{TP + TN}{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)} \times 100\%$$

(10)

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{(FP + TN) \times (FN + TP) \times (FP + FN) \times (FN + TN)}$$

(11)

where, TP represents the number of correctly recognized m\(^5\)C-containing sequences, TN represents the number of correctly recognized non-m\(^5\)C-containing sequences, FP represents the number of non-m\(^5\)C containing sequences recognized as m\(^5\)C-containing sequences and FN represents the number of m\(^5\)C-containing sequences recognized as non-m\(^5\)C-containing sequences.

2.6 Cross-validation

Three cross-validation methods, i.e., independent dataset test, K-fold cross-validation test and jackknife cross-validation, are often used to evaluate the performance of a predictor. Among these three methods, however, jackknife cross-validation is deemed the least arbitrary and the most objective, as demonstrated by eqn 28\( \text{ }-\)32 in a current review,\(^34\) and hence has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors.\(^35\)\( \text{ }-\)38 Accordingly, jackknife cross-validation was also used to examine the performance of the method proposed in the current study. In jackknife cross-validation, each sample in the training dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the one being identified.

3. Results and discussion

3.1 Parameter optimization

As we can see from eqn (2) and (3), the present model depends on the two parameters \( w \) and \( \lambda \). The former is the weight factor usually within the range from 0 to 1, while the latter is the number of correlation tiers considered to reflect the global sequence pattern effect. Generally speaking, the greater the \( \lambda \), the more global sequence-order information the model contains.
However, if $l$ is too large, it would reduce the cluster-tolerant capacity so as to lower the cross-validation accuracy due to over-fitting or a “high dimension disaster” problem. Therefore, our search for the optimal values of the two parameters is in the range of $w \in [0, 1]$ and $l \in [0, 10]$ with the steps of 0.1 and 1, respectively.

At this step, in order to reduce computational time, 5-fold cross validation was used to investigate the performance of candidate predictive models on the benchmark dataset. The results obtained are illustrated in Fig. 3, from which we can see that, when $l = 2$, the predictor yields the best predictive accuracy (90.49%) and the accuracy is irrespective of the value of $w$. Therefore, the two numerical values, $l = 2$ and $w = 0.5$, were used for the two uncertain parameters in the further analysis and performance evaluation.

### 3.2 Prediction quality

The prediction quality of the present model in identifying m$^5$C in the benchmark dataset via rigorous jackknife cross-validation is listed in Table 2. As shown in Table 2, an accuracy of 90.42% was obtained with the Sn of 85.00%, Sp of 95.83% and MCC of 0.81.

In addition, to demonstrate that the predictive accuracy is not sensitive to the selection of negative data, the method was also tested on nine other negative datasets. The predictive results of the jackknife cross-validation are shown in Fig. 4. As indicated in Fig. 4, we found that the predictive accuracy is not affected by the selection of negative data. These results demonstrate the reliability of the model developed in this study.

### 3.3 Comparison with other classifiers

In order to further prove its superiority, the predictive results of the proposed method were also compared with that of other commonly used classifiers, i.e., Naïve Bayes, Random forest, BayesNet and J48 Tree as implemented in WEKA. The jackknife cross-validation results of different classifiers for identifying m$^5$C using the same benchmark dataset and same parameters ($l = 2$ and $w = 0.5$) used in the current study are reported in Table 2.

It is shown that the four metrics (eqn (8)–(11)) of the proposed SVM model are all higher than that of Random forest, BayesNet and J48 Tree. Although the sensitivity of the proposed method is a bit lower than that of Naïve Bayes, its specificity, accuracy and MCC are all significantly higher than those of Naïve Bayes, indicating that the proposed SVM model can be effectively used to identify m$^5$C sites.

### 4. Conclusions

Although our knowledge of m$^5$C is still in its infancy, it is clear that m$^5$C participates in many aspects of gene expression. Therefore, the identification of m$^5$C will facilitate our understanding of its regulatory roles.

In this work, a support vector machine based model was developed to identify m$^5$C sites in the *H. sapiens* transcriptome using pseudo dinucleotide compositions. Jackknife cross-validation results show that our model is helpful for m$^5$C identification.

In addition, we also compared the predictive accuracy of SVM with that of the other four commonly used classification methods. We found that the predictive result of SVM is better than that of Naïve Bayes, Random forest, BayesNet and J48 Tree for m$^5$C site identification. Therefore, we hope that our method will become a useful tool for identifying m$^5$C sites.

### Conflicts of interest

The authors declare that they have no competing interests.

### Authors’ contributions

WC and HL conceived and designed the experiments. FP and HD performed the analysis. FP, HD, WC and HL wrote the paper. All authors read and approved the final manuscript.
Acknowledgements

This work was supported by Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028), the Outstanding Youth Foundation of North China University of Science and Technology (No. JP201502), China Postdoctoral Science Foundation (No. 2015M582533), the Applied Basic Research Program of Sichuan Province (No. 2015JY0100) and the Fundamental Research Funds for the Central Universities (No. ZYGX2015J144, ZYGX2015Z006).

References