Contents lists available at ScienceDirect

Genomics

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

Identifying 2′-O-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions

Wei Chen [a,*], Pengmian Feng [b], Hua Tang [c], Hui Ding [d], Hao Lin [d,*]

[a] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China
[b] School of Public Health, North China University of Science and Technology, Tangshan 063000, China
[c] Department of Pathophysiology, Sichuan Medical University, Luzhou 646000, China
[d] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics and Center for Information in Biomedicine, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

ARTICLE INFO

Article history:
Received 18 February 2016
Received in revised form 4 May 2016
Accepted 13 May 2016
Available online 15 May 2016

Keywords:
2′-O-methylationation
Nucleotide chemical property
Nucleotide composition
Support vector machine

ABSTRACT

2′-O-methylationation is an important post-transcriptional modification and plays important roles in many biological processes. Although experimental technologies have been proposed to detect 2′-O-methylationation sites, they are cost-ineffective. As complements to experimental techniques, computational methods will facilitate the identification of 2′-O-methylationation sites. In the present study, we proposed a support vector machine-based method to identify 2′-O-methylationation sites. In this method, RNA sequences were formulated by nucleotide chemical properties and nucleotide compositions. In the jackknife cross-validation test, the proposed method obtained an accuracy of 95.58% for identifying 2′-O-methylationation sites in the human genome. Moreover, the model was also validated by identifying 2′-O-methylation sites in the *Mus musculus* and *Saccharomyces cerevisiae* genomes, and the obtained accuracies are also satisfactory. These results indicate that the proposed method will become a useful tool for the research on 2′-O-methylation.

1. Introduction

The 2′-O-methylationation is a common post-transcriptional modification and has been discovered from archaea to human [1]. The 2′-O-methylationation is catalyzed by 2′-O-methylationtransferase [2]. Directed by small nucleolar RNAs [3], a methylation group is added to the 2’ hydroxyl group of the ribose moiety of a nucleotide (Fig. 1). It has been demonstrated that most 2′-O-methylated sites are clustered around functionally important regions of rRNAs and influence ribosome structure and function [4]. The 2′-O-methylation also occurs within the cap structure of mRNAs and provides a molecular signature for the distinction of self versus non-self mRNA by the RNA sensor Mda5 [5,6].

Recently, RNA was considered to be related to several diseases by the post-transcriptional regulation function [7,8]. Detection of 2′-O-methylated nucleotides and the mechanistic study of this post-transcriptional modification are important for the understanding of RNA biogenesis and function as well as its mechanisms regulating gene expression. To this end, in the past decades, many experimental techniques, such as the reverse transcription based method [9], RNase H based method [10] and RTL-P method [6], have been proposed to detect 2′-O-methylation in *Homo sapiens* (*H. sapiens*), *Mus musculus* (*M. musculus*) and *Saccharomyces cerevisiae* (*S. cerevisiae*) genomes. However, these methods are still expensive and time consuming in performing genome-wide analysis. With the rapid increasing number of sequenced genomes, it is highly desired to develop automated methods for timely identifying 2′-O-methylation sites. As excellent complements to experimental techniques, computational methods will speed up genome-wide 2′-O-methylation site detection.

Based on the experimental 2′-O-methylation data of *H. sapiens*, in the present study, a support vector machine (SVM) based model was proposed to identify 2′-O-methylation sites by encoding RNA sequences using nucleotide chemical properties and nucleotide compositions. Results from the jackknife test show that the proposed method obtained an accuracy of 95.58% for identifying 2′-O-methylation sites in *H. sapiens*. To demonstrate its effectiveness, the method trained on the *H. sapiens* was also applied to identity 2′-O-methylation sites in *M. musculus* and *S. cerevisiae* genomes, and yielded encouraging results as well.

2. Materials and methods

2.1. Dataset

RNA sequences containing experimentally validated 2′-O-methylation sites in *H. sapiens* were downloaded from RMBase [11]. All sequences are 41-nt long with the 2′-O-methylation site in the center. As elaborated in [12], a dataset including many high similar samples

* Corresponding authors.
E-mail addresses: chenweiimu@gmail.com (W. Chen), hlin@uestc.edu.cn (H. Lin).

http://dx.doi.org/10.1016/j.ygeno.2016.05.003
0888-7543/© 2016 Elsevier Inc. All rights reserved.

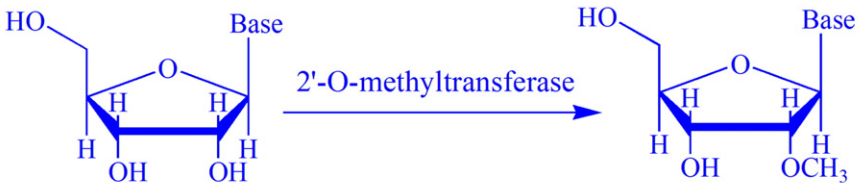

Fig. 1. Illustration of the 2′-O-methylationation. 2′-O-methylationation is catalyzed by 2′-O-methylationtransferase and a methylation group is added to the 2′ hydroxyl group of the ribose moiety of a nucleotide. The "Base" can be adenine (A), guanine (G), cytosine (C) or uracil (U).

would be lack of statistical representativeness and increase the risk of model overtraining. A predictor, if trained and tested by such a biased benchmark dataset, might yield misleading results with overestimated accuracy. To avoid redundancy and reduce the homology bias, sequences with more than 80% sequence similarity were removed by using the CD-HIT program [13]. After such a screening procedure, we obtained 147 2′-O-methylation sites in *H. sapiens*, which were deemed as positive samples.

The negative samples were obtained by choosing the 41-nt long sequences that satisfy the rule that all the nucleotides in the 41-nt long segment were not experimentally confirmed to be 2′-O-methylated. By doing so, we obtained a great number of negative samples. Therefore, the number of negative samples will be dramatically larger than those of positive samples. In machine-learning problems, imbalanced datasets can significantly affect the accuracy of learning methods. To balance out the numbers between positive and negative samples in model training, we randomly picked out 147 sequences to form the negative samples. Finally, we obtained a benchmark dataset containing 147 true 2′-O-methylation site containing sequences and 147 false 2′-O-methylation site containing sequences of *H. sapiens*, which are available in Supplementary material S1.

2.2. Support vector machine

Support vector machine (SVM) is a machine learning algorithm and has been successfully used in the realm of bioinformatics [14–21]. The basic idea of SVM is to transform the input data into a high dimensional feature space and then determine the optimal separating hyperplane. In the current study, the LibSVM package 3.18 was used to implement SVM, which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Because of its effectiveness and speed in training process, the radial basis kernel function (RBF) was used to obtain the classification hyperplane in the current study. In the SVM operation engine, the grid search method was applied to optimize the regularization parameter *C* and kernel parameter γ using a grid search approach defined by

$$\begin{cases} 2^{-5} \le C \le 2^{15} & \text{with step of } 2 \\ 2^{-15} \le \gamma \le 2^{-5} & \text{with step of } 2^{-1} \end{cases} \tag{1}$$

2.3. Chemical property

There are four kinds of nucleotides in RNA, namely, adenine (A), guanine (G), cytosine (C) and uracil (U). As shown in Fig. 2, adenine and guanine have two rings, while cytosine and uracil have only one ring. When forming secondary structures, in terms of hydrogen bond, guanine and cytosine have strong hydrogen bonds, whereas adenine and uracil have weak hydrogen bonds. In terms of chemical functionality, adenine and cytosine can be classified into the same group, called amino group, while guanine and uracil into the keto group. Accordingly, the four nucleotides can be classified into three different groups according to these three chemical properties, Fig. 3.

In order to include these chemical properties in RNA encoding, three coordinates (*x*, *y*, *z*) were used to represent the three chemical groups and assign 1 or 0 values, respectively. Each nucleotide can be encoded by the following formula [22],

$$x_i = \begin{cases} 1 & if\ s_i \in \{A, G\} \\ 0 & if\ s_i \in \{C, U\} \end{cases} \tag{2}$$
$$y_i = \begin{cases} 1 & if\ s_i \in \{A, C\} \\ 0 & if\ s_i \in \{G, U\} \end{cases}$$
$$z_i = \begin{cases} 1 & if\ s_i \in \{A, U\} \\ 0 & if\ s_i \in \{C, G\} \end{cases}$$

Thus, A can be represented by the coordinates (1, 1, 1), C by the coordinates (0, 1, 0), G by the coordinates (1, 0, 0), U by the coordinates (0, 0, 1).

2.4. Nucleotide composition

For the purpose of including nucleotide composition surrounding the 2′-O-methylation sites as well, the density d_i of any nucleotide n_j at position *i* in a RNA sequence was defined by the following formula.

$$d_i = \frac{1}{|N_i|}\sum_{j=1}^{l} f(n_j), f(n_j) = \begin{cases} 1 & if\, n_j = q \\ 0 & othercases \end{cases} \tag{3}$$

where *l* is the sequence length, $|N_i|$ is the length of the *i*-th prefix string $\{n_1, n_2, \ldots, n_i\}$ in the sequence, $q \in \{A, C, G, U\}$.

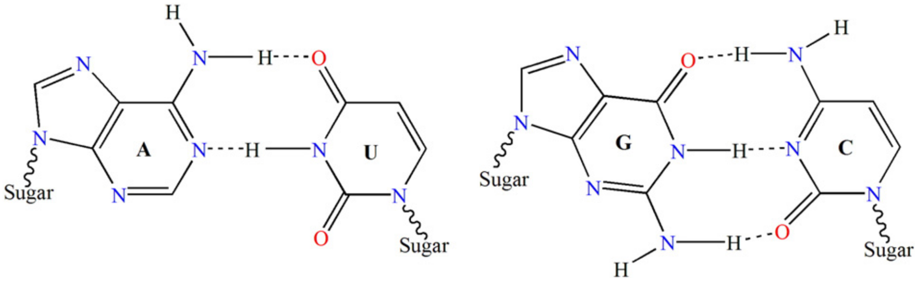

Fig. 2. Illustration to show the structure of paired nucleic acid residues. The left panel is the A-U pair which has 2 hydrogen bonds; the right panel is the G-C pair which has 3 hydrogen bonds.

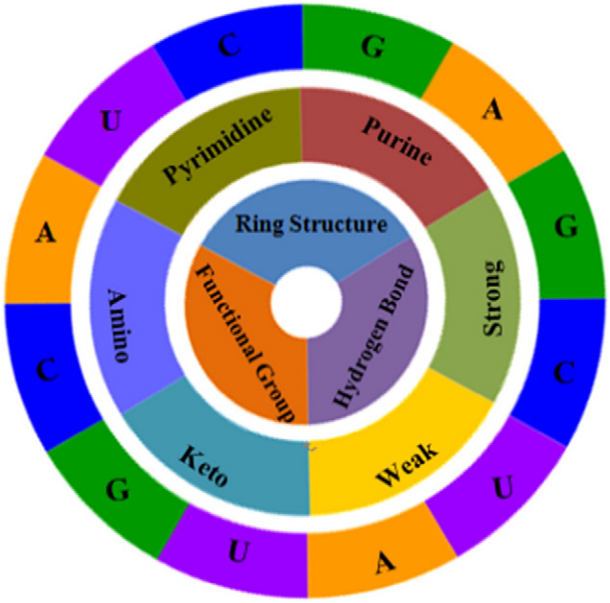

Fig. 3. Cluster of nucleotides according to their chemical properties. The first circle shows the chemical properties. The second circle shows the detailed clusters. The third circle shows the nucleotides.

Therefore, by integrating the three nucleotide chemical properties and nucleotide composition, each sample in the benchmark dataset was encoded by a 164 (4 × 41)-dimensional vector and was used as the input vector of SVM.

2.5. Performance evaluation

The performance of the proposed method was evaluated by using the following four metrics [23,24], namely sensitivity (*Sn*), specificity (*Sp*), Accuracy (*Acc*) and the Mathew's correlation coefficient (*MCC*), which are expressed as

$$\begin{cases} Sn = \dfrac{TP}{TP+FN} \times 100\% \\ Sp = \dfrac{TN}{TN+FP} \times 100\% \\ Acc = \dfrac{TP+TN}{TP+FN+TN+FP} \times 100\% \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FN) \times (TN+FP)}} \end{cases} \tag{4}$$

where *TP*, *TN*, *FP*, and *FN* represent true positive, true negative, false positive, and false negative, respectively.

3. Results and discussions

3.1. Identification of 2′-O-methylation sites

In statistical prediction, three cross-validation methods, i.e., independent dataset test, sub-sampling (or K-fold cross-validation) test, and jackknife test, are often used to evaluate the anticipated success rate of a predictor. Among the three methods, the jackknife test is deemed the least arbitrary and most objective as demonstrated by Eqs. 28–32 in [12]. Hence, it has been widely and increasingly adopted by investigators to examine and compare the quality of various predictors [25–29]. Thus, the jackknife test was used to examine the performance of the proposed model. In the jackknife test, each sample in the training dataset is in turn singled out as an independent test sample and all the properties are calculated without including the one being identified.

Table 1
Predictive results by using different features for identifying 2′-O-methylation sites in human genome.

Features	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*
Functional Group	81.63	98.64	90.14	0.81
Ring Structure	85.03	94.56	89.80	0.80
Hydrogen Bond	78.91	93.20	86.05	0.73
Nucleotide composition	54.42	76.87	65.65	0.32
Hybrid[a]	92.52	98.64	95.58	0.91

[a] Hybrid represents combining all of the four kinds of features together.

The jackknife test results obtained by the method based on the benchmark dataset are listed in Table 1. As shown in Table 1, the method obtained an accuracy of 95.58%, with the sensitivity of 92.52%, specificity of 98.64% and MCC of 0.91. In order to investigate the contribution of different features for identifying 2′-O-methylation sites, we also performed the prediction using nucleotide chemical properties and nucleotide composition on the benchmark dataset, respectively. The jackknife test results by using each kind of these features are also reported in Table 1. In comparison with nucleotide composition, nucleotide chemical properties achieved a higher accuracy, suggesting that they play more important roles for 2′-O-methylation site identification. Among the three considered nucleotide chemical properties, the functional group yields the highest accuracy (90.14%), indicating that it has the largest contribution for 2′-O-methylation site identification in the current method and the other three features (ring structure, hydrogen bond and nucleotide frequency) are complementary for the identification.

3.2. Validation on other species

To further verify the power of the proposed method trained on the *H. sapiens* genome, we also applied it to identify 2′-O-methylation sites in the *M. musculus* and *S. cerevisiae* genomes. According to the RMBase database [11], we obtained 27 and 133 experimentally confirmed 2′-O-methylation sites containing sequences in *M. musculus* and *S. cerevisiae* genomes, respectively. All sequences are also 41-nt long with the 2′-O-methylation site in the center and available in Supplementary materials S2 and S3, respectively. The proposed method correctly identified 27 and 125 2′-O-methylation sites in *M. musculus* and *S. cerevisiae* genomes, respectively. These results indicate that the proposed method is quite promising and holds the potential to become a useful tool in identifying 2′-O-methylation sites in other species.

4. Conclusions

By using nucleotide chemical properties and nucleotide composition, a support vector machine-based model was proposed to identify 2′-O-methylation sites. An overall accuracy of 95.58% was obtained for identifying 2′-O-methylation sites in the *H. sapiens* genome. To identify the key features for 2′-O-methylation sites identification, a comparison experiment was carried out among different models built by using different kinds of parameters. Jackknife test results show that the functional group has the largest contribution for 2′-O-methylation site identification.

Although the model is trained based on the data from human genome, it is encouraging that the predictive results of the method for identifying 2′-O-methylation sites in *M. musculus* and *S. cerevisiae* genomes are also quite good, indicating that our model is robust and ingenious. Therefore, we hope that our method will be helpful for identifying 2′-O-methylation sites and provide some novel insights into the research on RNA post-transcriptional modifications. We also plan to extend our method on more large scale datasets with advanced parallel computational techniques [30].

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2016.05.003.

Acknowledgements

This work was supported by National Nature Scientific Foundation of China (No. 61301260), 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), the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122), the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015J144; ZYGX2015Z006).

References

[1] J.P. Bachellerie, J. Cavaille, A. Huttenhofer, The expanding snoRNA world, Biochimie 84 (2002) 775–790.

[2] H. Wang, D. Boisvert, K.K. Kim, R. Kim, S.H. Kim, Crystal structure of a fibrillarin homologue from Methanococcus jannaschii, a hyperthermophile, at 1.6 A resolution, EMBO J. 19 (2000) 317–323.

[3] B.E. Maden, Mapping 2′-O-methyl groups in ribosomal RNA, Methods 25 (2001) 374–382.

[4] W.A. Decatur, M.J. Fournier, rRNA modifications and ribosome function, Trends Biochem. Sci. 27 (2002) 344–351.

[5] R. Zust, L. Cervantes-Barragan, M. Habjan, R. Maier, B.W. Neuman, J. Ziebuhr, K.J. Szretter, S.C. Baker, W. Barchet, M.S. Diamond, S.G. Siddell, B. Ludewig, V. Thiel, Ribose 2′-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5, Nat. Immunol. 12 (2011) 137–143.

[6] Z.W. Dong, P. Shao, L.T. Diao, H. Zhou, C.H. Yu, L.H. Qu, RTL-P: a sensitive approach for detecting sites of 2′-O-methylation in RNA molecules, Nucleic Acids Res. 40 (2012), e157.

[7] X. Zeng, X. Zhang, Q. Zou, Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks, Brief. Bioinform. 17 (2016) 193–203.

[8] Q. Zou, J. Li, Q. Hong, Z. Lin, Y. Wu, H. Shi, Y. Ju, Prediction of MicroRNA-Disease Associations Based on Social Network Analysis Methods, BioMed Research Internation, 2015, 2015 810514.

[9] P.M. Ajuh, E.B. Maden, Chemical secondary structure probing of two highly methylated regions in *Xenopus laevis* 28S ribosomal RNA, Biochim. Biophys. Acta 1219 (1994) 89–97.

[10] Y.T. Yu, M.D. Shu, J.A. Steitz, A new method for detecting sites of 2′-O-methylation in RNA molecules, RNA 3 (1997) 324–331.

[11] W.J. Sun, J.H. Li, S. Liu, J. Wu, H. Zhou, L.H. Qu, J.H. Yang, RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data, Nucleic Acids Res. 44 (2016) D259–D265.

[12] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.

[13] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[14] W. Chen, P.M. Feng, H. Ding, H. Lin, K.C. Chou, iRNA-methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33.

[15] W. Chen, P.M. Feng, E.Z. Deng, H. Lin, K.C. Chou, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83.

[16] W. Chen, P.M. Feng, H. Lin, Prediction of replication origins by calculating DNA structural properties, FEBS Lett. 586 (2012) 934–938.

[17] P.M. Feng, W. Chen, H. Lin, K.C. Chou, iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, Anal. Biochem. 442 (2013) 118–125.

[18] W. Chen, H. Tran, Z. Liang, H. Lin, L. Zhang, Identification and analysis of the N(6)-methyladenosine in the *Saccharomyces cerevisiae* transcriptome, Sci. Rep. 5 (2015) 13859.

[19] H. Lin, W. Chen, H. Ding, AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes, PLoS One 8 (2013), e75726.

[20] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, K.C. Chou, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65–W71.

[21] B. Liu, L. Fang, F. Liu, X. Wang, K.C. Chou, iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach, J. Biomol. Struct. Dyn. 34 (2016) 223–235.

[22] A.T.M. Golam Bari, M. Rokeya Reaz, B.S. Jeong, DNA encoding for splice site prediction in large DNA sequence, MATCH Commun. Math. Co. 71 (2014) 241–258.

[23] P.M. Feng, W. Chen, H. Lin, Prediction of CpG island methylation status by integrating DNA physicochemical properties, Genomics 104 (2014) 229–233.

[24] W. Chen, H. Lin, P.M. Feng, J.P. Wang, Exon skipping event prediction based on histone modifications, Interdiscip. Sci. 6 (2014) 241–249.

[25] R. Kumar, A. Srivastava, B. Kumari, M. Kumar, Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine, J. Theor. Biol. 365 (2015) 96–103.

[26] W. Chen, P.M. Feng, H. Lin, K.C. Chou, iSS-PseDNC: Identifying Splicing Sites Using Pseudo Dinucleotide Composition, BioMed Research International, 2014, 2014 623149.

[27] W. Chen, P.M. Feng, H. Lin, K.C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013), e68.

[28] W. Chen, H. Lin, Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine, Comput. Biol. Med. 42 (2012) 504–507.

[29] W. Chen, P.M. Feng, H. Lin, Prediction of ketoacyl synthase family using reduced amino acid alphabets, J. Ind. Microbiol. Biotechnol. 39 (2012) 579–584.

[30] Q. Zou, X.B. Li, W.R. Jiang, Z.Y. Lin, G.L. Li, K. Chen, Survey of MapReduce frame operation in bioinformatics, Brief. Bioinform. 15 (2014) 637–647.