Molecular Therapy
Nucleic Acids
Original Article

iRNA-m2G: Identifying N^2-methylguanosine Sites Based on Sequence-Derived Information

Wei Chen,[1,2] Xiaoming Song,[2] Hao Lv,[3] and Hao Lin[3]

[1]Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China; [2]Center for Genomics and Computational Biology, School of Life Sciences, North China University of Science and Technology, Tangshan 063000, China; [3]Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

RNA N^2-methylguanosine (m2G) is one kind of posttranscriptional modification and plays crucial roles in the control and stabilization of tRNA. However, our knowledge about the biological functions of m2G is still limited. The key step of revealing its new function is to recognize the m2G sites in the transcriptome. Since there is no effective method for detecting m2G sites, it is desirable to develop new methods to identify m2G sites. In this study, a computational predictor called iRNA-m2G was proposed to identify m2G sites in eukaryotic transcriptomes. In iRNA-m2G, the RNA sequences were encoded by using nucleotide chemical property and accumulated nucleotide frequency. iRNA-m2G was not only validated by the rigorous jackknife test on the benchmark dataset but also examined by performing cross-species validations. In addition, iRNA-m2G was also tested on an independent dataset. It was found that the accuracies obtained by iRNA-m2G were all quite promising in these tests, indicating that the proposed method could become a powerful tool for identifying m2G sites. Finally, a user-friendly web server for iRNA-m2G is freely accessible at http://lin-group.cn/server/iRNA-m2G.php.

INTRODUCTION

Since the first posttranscriptional modification was discovered in tRNA nearly 60 years ago,[1] more than 100 kinds of RNA modifications have also been found in tRNA. Besides N^6-methyladenosine,[2–4] 5-methylcytosine,[5] pseudouridine,[6] and N^1-methyladenosine,[7] N^2-methylguanosine (m2G) has also been identified in tRNA of eukaryotes and archaea.[8,9]

The formation of m2G was catalyzed by the rRNA guanine-(N^2)-methyltransferases that methylate the amino group at the C-2 position of guanine.[10] Through forming canonical or non-canonical Watson-Crick base-pairing interactions with other bases,[11] m2G plays key roles in the control and stabilization of the tertiary structure of tRNA.[12,13] In addition, it has also been reported that m2G could act as the kinetic barrier for reverse transcription.[11]

Compared with the other kinds of RNA modifications, our knowledge about the function of m2G is still in its infant stage. In order to reveal its novel biological functions, the key point is to accurately detect the positions of m2G sites in the transcriptome. Since there are no high-throughput methods for detecting m2G sites at present, it is necessary to develop an effective method for the accurate identification of m2G sites.

Keeping this in mind, in the present work, we proposed a computational method, called iRNA-m2G, to identify m2G sites in eukaryotic transcriptomes. In both the jackknife test and the independent dataset test, iRNA-m2G yielded promising predictive performances for the detection of m2G sites. For the convenience of the scientific community, a web server for iRNA-m2G is established and is freely available at http://lin-group.cn/server/iRNA-m2G.php.

RESULTS AND DISCUSSION

Nucleotide Composition Analysis

In order to find the nucleotide composition bias of m2G-site-containing sequences, the Two Sample Logo software[14] was used to calculate the differences between m2G-site-containing sequences and non-m2G-site-containing sequences. The statistically significant ($p < 0.05$, two-sample t test) nucleotides surrounding m2G sites are indicated in Figure 1. The conserved consensus motif UGGC located at positions −2 to 1 was found in *H. sapiens*, *M. musculus*, and *S. cerevisiae*. In addition, the position-specific enrichment of nucleotides was also observed both upstream and downstream of the m2G site. For example, the G was enriched at positions −9, −1, 8, and 9; the C was enriched at positions −7 and 1; the U was enriched at positions −2, 6, and 10; the A was enriched at positions 4 and 10. These position-specific enrichments of nucleotides appeared in all three species (Figures 1B–1D).

In addition to the aforementioned common pattern, the species-specific nucleotide composition bias was also observed. The enrichment

Received 3 July 2019; accepted 19 August 2019;
https://doi.org/10.1016/j.omtn.2019.08.023.

Correspondence: Wei Chen, Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China.
E-mail: chenweiimu@gmail.com
Correspondence: Hao Lin, Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China.
E-mail: hlin@uestc.edu.cn

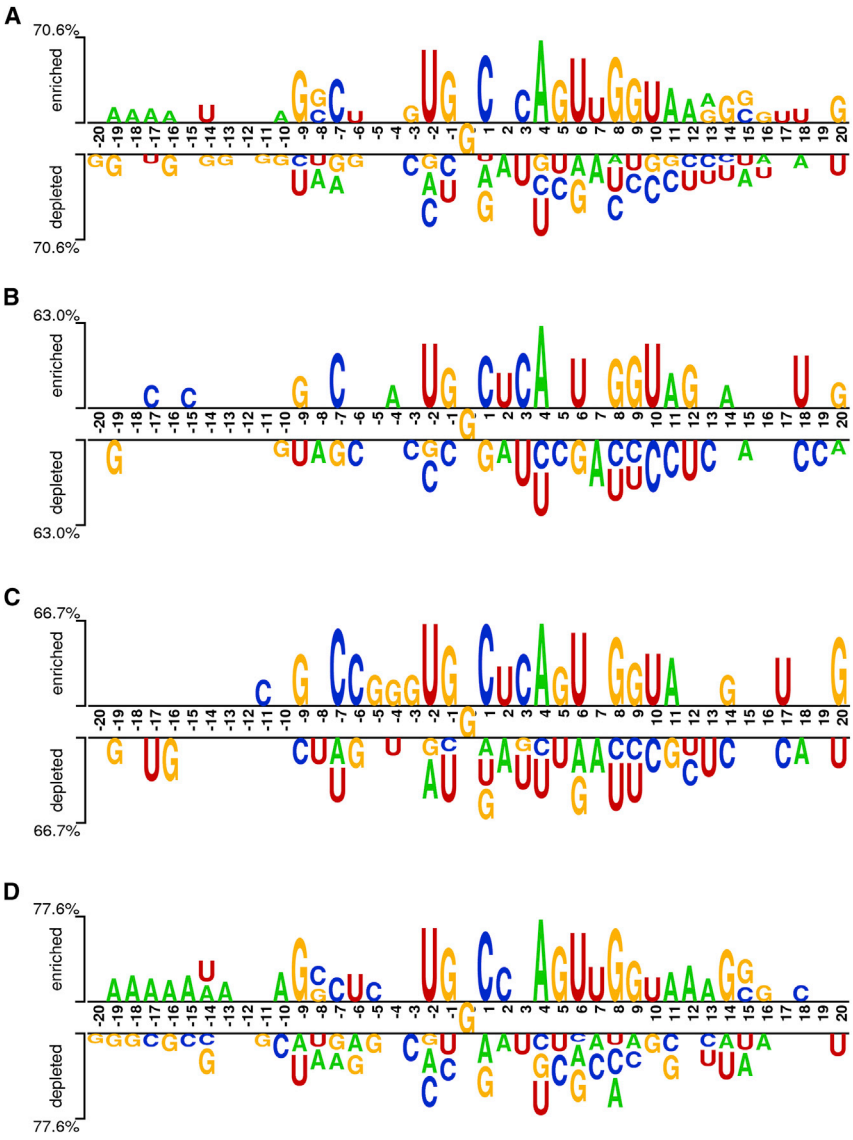

Figure 1. Nucleotide Composition Preferences of m2G-Site- and Non-m2G-Site-Containing Sequences

The m2G site (or non-m2G) site is at position 0 in the top and bottom levels of each panel. (A) Based on the sequences from dataset S1. (B–D) Based on the sequences from (B) *H. sapiens*, (C) *M. musculus*, and (D) *S. cerevisiae*, respectively.

of two consecutive nucleotides U and C at positions 2 and 3 were found in *H. sapiens* and *M. musculus* (Figures 1B and 1C). The three consecutive Gs were observed at positions −5, −4, and −3 in *M. musculus* (Figure 1C). The enrichment of A in the upstream region from −20 to −13 relative to the m2G site was observed in *S. cerevisiae*. It was also observed that, at position 18 in the downstream sequence, U was enriched in *H. sapiens* and C was enriched in *S. cerevisiae*. The observed nucleotide bias might be the signal for methyltransferases to recognize their targets and also suggest that it is reasonable to identify m2G sites based on the sequence-derived information.

Window Size Optimization

Considering the position-specific nucleotide bias, in order to obtain effective information for identifying m2G sites, it is necessary to determine the optimal window size of the flanking sequences around the m2G site. Therefore, we analyzed the effect of window size (ξ) on the accuracy of identifying m2G sites based on dataset S_1. The scope of ξ ranged from 21 to 41, with a step of 2 nt. The predictive accuracies of the 10-fold cross-validation test for the models based on different window sizes are shown in Figure 2. As indicated in Figure 2, the best predictive accuracy for identifying the m2G site was obtained when $\xi = 41$. Thus, the following analysis were all based on $\xi = 41$, i.e., all the samples in the benchmark dataset were transferred into a 164-dimensional vector.

The model thus obtained is called iRNA-m2G, where "i" stands for "identify" and "m2G" stands for "N^2-methylguanosine." To demonstrate its performance, the iRNA-m2G was evaluated by using the jackknife test, in which iRNA-m2G obtained an accuracy of 95.80% with the sensitivity of 93.00%, specificity of 98.60% and MCC of 0.92.

Robustness and Stability Analysis

In order to measure the robustness and stability of the proposed model, the following experiments were carried out. Based on the samples from *H. sapiens*, *M. musculus*, and *S. cerevisiae,* we first built species-specific models and validated their performances by using the jackknife test. The results thus obtained are reported in Table 1. As indicated in Table 1, the accuracies of the species-specific models for identifying m2G sites were 94.56%, 100%, and 96.27% in *H. sapiens*, *M. musculus*, and *S. cerevisiae*, respectively. The area under the receiver operating characteristic curve (ROC) that was used to objectively quantify a computational model was also provided.

To demonstrate to what extent a species-specific model can identify the m2G sites from other species, we evaluated the species-specific model on the data from other species. The results are indicated in Figure 3. It was found that the *H. sapiens*-based model can accurately identify the m2G sites in *M. musculus* with the accuracy of 96.67%, while its accuracy for identifying m2G sites in *S. cerevisiae* is 88.80% lower than that obtained by the model trained by using the data from *S. cerevisiae* itself. The accuracies of *M. musculus*- and

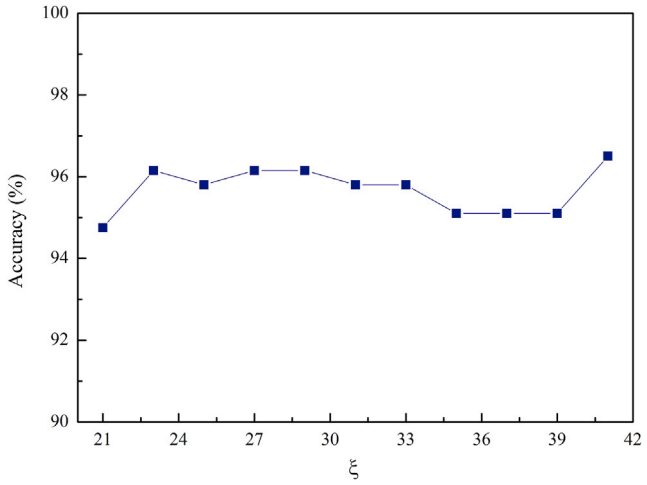

Figure 2. Predictive Performance of the Models based on Different Window Size

S. cerevisiae-specific models for identifying m2G sites in *H. sapiens* are 77.17% and 80.43%, which are still acceptable.

In addition, to illustrate that the predictive accuracy of iRNA-m2G is not sensitive to the selection of negative data, the proposed method was further evaluated on the benchmark dataset S_2. In the jackknife test, our method obtained an accuracy of 98.85%, with the sensitivity of 90.21%, specificity of 99.83%, and MCC of 0.93. It can be concluded that, even based on the dataset with the positive-to-negative ratio of approximately 1:9, the predictive accuracy of our method is still comparable with that based on dataset S_1. All these results demonstrated the robustness of the proposed method.

Web Server

Since user-friendly web servers represent the future direction for developing useful predictors,[15–19] to enable the applications of the proposed method, a publicly accessible web server for iRNA-m2G was established, which is available at http://lin-group.cn/server/iRNA-m2G.php, through which users can detect the m2G sites in transcriptomes. The user guide on how to use it is given as follows:

Step 1. Visit the homepage of iRNA-m2G at http://lin-group.cn/server/iRNA-m2G.php.

Step 2. Either type or copy or paste the query RNA sequences with a length greater than 41 nt in FASTA format. The format of the input sequences can be found by clicking the "Example" button.

Step 3. After clicking the "Submit" button, the results will be shown on the screen.

Conclusions

In this study, we developed a predictor called iRNA-m2G to identify m2G sites in *H. sapiens*, *M. musculus*, and *S. cerevisiae* in which the RNA sequences were encoded by using nucleotide chemical property and accumulated nucleotide frequency. To the best of our knowledge, iRNA-m2G is the first computational method for this aim. The jackknife test results demonstrated that iRNA-m2G is promising for identifying m2G sites.

The iRNA-m2G was also evaluated by performing cross-species validations. Although the accuracies of the cross-validation tests are a little lower, they are still acceptable. The lower accuracies of cross-species validations may be due to the limited number of samples in each species, which are too few to yield enough information to train robust models.

To demonstrate its robustness for identifying m2G sites, the method was further validated on an independent dataset with a positive-to-negative ratio of approximately 1:9. The jackknife test results obtained based on the independent dataset were also quite good, indicating that iRNA-m2G is robust and is useful for identifying m2G sites. In the future, we will collect m2G sites from different species to further improve the performance of iRNA-m2G.

MATERIALS AND METHODS

Benchmark Datasets

The positive samples (m2G-site-containing sequences) of *H. sapiens*, *M. musculus*, and *S. cerevisiae* were collected from the RMBase database[20] and were all 41 nt long, with the m2G site at the center position. In order to construct a high-quality benchmark dataset, the CD-HIT software[21] was used to remove the samples with sequence similarity greater than 90%. If the sequence similarity threshold is set to a lower value, such as 60%, the dataset will be more objective and reliable. However, in this study, such a stringent criterion was not used; otherwise, the number of samples would be too few to have statistical significance. Finally, we obtained 46, 30, and 67 m2G-site-containing sequences in *H. sapiens*, *M. musculus*, and *S. cerevisiae*, respectively.

Considering the fact that m2G modifications were mainly found in tRNA,[9] the negative samples (non-m2G-site-containing sequences) of *H. sapiens*, *M. musculus*, and *S. cerevisiae* were collected from their tRNA sequences, which are available at the GtRNAdb database.[22] By obeying the aforementioned procedures, 555, 444, and 247 negative samples were obtained for *H. sapiens*, *M. musculus*, and *S. cerevisiae*. These 1,246 sequences were also 41 nt long, with the guanine at the center, and have the similarity of less than 90%.

In order to objectively evaluate the proposed method, we built two datasets; namely, dataset S_1 and dataset S_2. S_1 is a balanced dataset including the aforementioned 143 m2G-site-containing sequences and 143 non-m2G-site-containing sequences randomly selected from the negative samples of each species (46, 30, and 67 from *H. sapiens*, *M. musculus* and *S. cerevisiae*, respectively). S_2 is an imbalanced dataset with a positive-to-negative ratio of approximately 1:9, which includes 143 m2G-site-containing sequences and 1,246 non-m2G-site-containing sequences. These datasets are available at http://lin-group.cn/server/iRNA-m2G/data.htm.

Table 1. Results of the Species-Specific Models for Identifying m2G Sites in Different Species

Species	Parameters	Sn (%)	Sp (%)	Acc (%)	MCC	auROC
H. sapiens	$C = 2, g = 0.0078125$	89.13	100.00	94.56	0.90	0.950
M. musculus	$C = 2, g = 0.0078125$	100.00	100.00	100.00	1.00	0.999
S. cerevisiae	$C = 0.5, g = 0.0078125$	92.53	100.00	96.27	0.93	0.964

auROC, area under the receiving operating characteristic; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews correlation coefficient.

Sequence Representation

In order to transfer the sequences in the benchmark dataset into vectors that can be processed by machine learning methods, they were encoded by using chemical properties and nucleotide frequency.[23–26] A brief description of this encoding scheme is introduced as follows.

Nucleotide Chemical Property

In terms of ring structures, A and G are purines containing two rings, whereas C and U are pyrimidines containing one ring. When forming secondary structures, C and G form strong hydrogen bonds, whereas A and U form weak hydrogen bonds. In terms of amino or keto bases, A and C belong to the amino group, while G and U belong to the keto group.

Accordingly, three coordinates (x, y, and z) were used to represent the chemical properties of the four nucleotides, and a value of 0 or 1 was assigned to the coordinates. If x, y and z coordinates stand for the ring structure, the hydrogen bond, and the amino or keto bases, the four nucleotides can be represented in the cartesian coordinate system. Therefore, the coordinates for A, C, G, and U are (1, 1, 1), (0, 0, 1), (1, 0, 0), and (0, 1, 0), respectively.

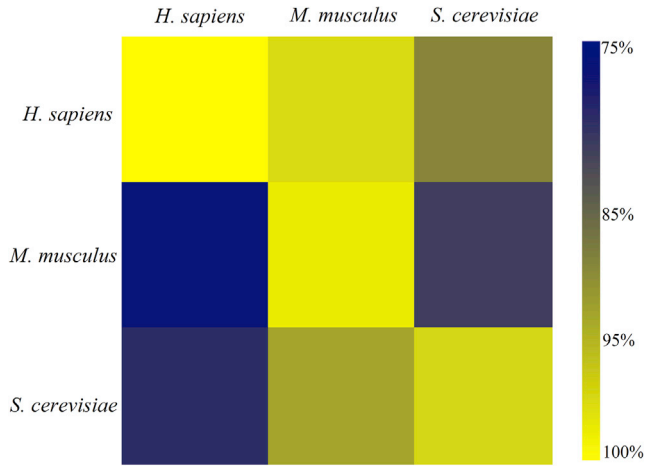

Figure 3. Heatmap Showing the Cross-Species Prediction Accuracies

Once a species-specific model was established on its own training dataset, it was tested on the data from the other seven species.

Accumulated Nucleotide Frequency

For the purpose of including nucleotide composition surrounding the m2G sites, the density d_i of nucleotide n_j at position i was defined as follows:

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

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

By integrating nucleotide chemical properties and nucleotide frequency, each nucleotide will be converted into a 4-dimensional vector, where the first three elements represent its chemical properties, and the fourth one represents the accumulated nucleotide frequency. Accordingly, an l-bp-long sequence will be encoded by a $(4 \times l)$-dimensional vector.

Support Vector Machine

The support vector machine (SVM) is a powerful and popular method for pattern recognition and has been widely used in computational genomics and computational proteomics.[27–31] In the present work, the LIBSVM package v3.18 was used to implement the SVM algorithm, which is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The basic idea of SVM is to transform the input data into a high-dimensional feature space and then determine the optimal separating hyperplane. Due to its effectiveness and speed in training process, the radial basis kernel function (RBF) was used to obtain the classification hyperplane. The grid search method was used to optimize the regularization parameter C and kernel parameter γ with the following searching spaces: $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{-5}]$, with the steps of 2 and 2^{-1}, respectively. The probability score obtained from SVM was used to make predictions. If the probability score obtained from SVM was greater than 0.5, a guanosine will be predicted as a m2G; otherwise, non-m2G.

Evaluation Metrics

In statistical prediction, three cross-validation methods—namely, independent dataset test, sub-sampling (or n-fold cross-validation) test, and jackknife test—are often used to evaluate the anticipated success rate of a predictor.[32] Among these tests, the jackknife test is deemed the least arbitrary and most objective one.[33] Accordingly, the jackknife test was used to examine the performance of the method proposed in the present study. 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.

The performance of the proposed method was evaluated by using the following four metrics—namely, sensitivity (*Sn*), specificity (*Sp*), accuracy (*Acc*) and Matthews correlation coefficient (*MCC*)—which are expressed as follows:[34–36]

$$\begin{cases} Sn = \dfrac{TP}{TP + FN} \times 100\% \\ Sp = \dfrac{TN}{TN + FP} \times 100\% \\ Acc = \dfrac{TP + TN}{TP + FN + TN + FP} \times 100\% \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \end{cases}, \quad (2)$$

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

AUTHOR CONTRIBUTIONS

W.C. and H. Lin conceived and designed the study. W.C., X.S., and H. Lv conducted the experiments. W.C. and X.S. implemented the algorithms. H. Lv established the web server. W.C., X.S., H. Lv, and H. Lin performed the analysis and wrote the paper. All authors read and approved the final manuscript.

CONFLICT OF INTERESTS

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work has been supported by the National Nature Scientific Foundation of China (61772119 and 31771471) and the Natural Science Foundation for Distinguished Young Scholars of Hebei Province (C2017209244).

REFERENCES

1. Davis, F.F., and Allen, F.W. (1957). Ribonucleic acids from yeast which contain a fifth nucleotide. J. Biol. Chem. *227*, 907–915.

2. Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.C. (2015). iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. *490*, 26–33.

3. Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA. RNA *25*, 205–218.

4. Zhou, Y., Zeng, P., Li, Y.H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res. *44*, e91.

5. Squires, J.E., Patel, H.R., Nousch, M., Sibbritt, T., Humphreys, D.T., Parker, B.J., Suter, C.M., and Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. *40*, 5023–5033.

6. Hudson, G.A., Bloomingdale, R.J., and Znosko, B.M. (2013). Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. RNA *19*, 1474–1482.

7. Dominissini, D., Nachtergaele, S., Moshitch-Moshkovitz, S., Peer, E., Kol, N., Ben-Haim, M.S., Dai, Q., Di Segni, A., Salmon-Divon, M., Clark, W.C., et al. (2016). The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. Nature *530*, 441–446.

8. Grosjean, H., Sprinzl, M., and Steinberg, S. (1995). Posttranscriptionally modified nucleosides in transfer RNA: their locations and frequencies. Biochimie *77*, 139–141.

9. Sprinzl, M., and Vassilenko, K.S. (2005). Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. *33*, D139–D140.

10. Sergiev, P.V., Bogdanov, A.A., and Dontsova, O.A. (2007). Ribosomal RNA guanine-(N2)-methyltransferases and their targets. Nucleic Acids Res. *35*, 2295–2301.

11. Bavi, R.S., Sambhare, S.B., and Sonawane, K.D. (2013). MD simulation studies to investigate iso-energetic conformational behaviour of modified nucleosides m(2)G and m(2) 2G present in tRNA. Comput. Struct. Biotechnol. J. *5*, e201302015.

12. Schneider, A., Peter, D., Schmitt, J., Leo, B., Richter, F., Rösch, P., Wöhrl, B.M., and Hartl, M.J. (2014). Structural requirements for enzymatic activities of foamy virus protease-reverse transcriptase. Proteins *82*, 375–385.

13. Limbach, P.A., Crain, P.F., and McCloskey, J.A. (1994). Summary: the modified nucleosides of RNA. Nucleic Acids Res. *22*, 2183–2196.

14. Vacic, V., Iakoucheva, L.M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics *22*, 1536–1537.

15. Yang, J., Chen, X., McDermaid, A., and Ma, Q. (2017). DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. Bioinformatics *33*, 2586–2588.

16. Ma, Q., Zhang, H., Mao, X., Zhou, C., Liu, B., Chen, X., and Xu, Y. (2014). DMINDA: an integrated web server for DNA motif identification and analyses. Nucleic Acids Res. *42*, W12–W19.

17. Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N.D., Webb, G.I., and Chou, K.C. (2019). iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief. Bioinform. *20*, 638–658.

18. Liang, Z.Y., Lai, H.Y., Yang, H., Zhang, C.J., Yang, H., Wei, H.H., Chen, X.X., Zhao, Y.W., Su, Z.D., Li, W.C., et al. (2017). Pro54DB: a database for experimentally verified sigma-54 promoters. Bioinformatics *33*, 467–469.

19. Zhao, W., Zhou, Y., Cui, Q., and Zhou, Y. (2019). PACES: prediction of N4-acetylcytidine (ac4C) modification sites in mRNA. Sci. Rep. *9*, 11112.

20. Sun, W.J., Li, J.H., Liu, S., Wu, J., Zhou, H., Qu, L.H., and Yang, J.H. (2016). RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. Nucleic Acids Res. *44* (D1), D259–D265.

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

22. Chan, P.P., and Lowe, T.M. (2016). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. *44* (D1), D184–D189.

23. Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics *33*, 3518–3523.

24. Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics *35*, 2796–2800.

25. Xu, Z.C., Feng, P.M., Yang, H., Qiu, W.R., Chen, W., and Lin, H. (2019). iRNAD: a computational tool for identifying D modification sites in RNA sequence. Bioinformatics. Published online May 11, 2019. https://doi.org/10.1093/bioinformatics/btz358.

26. Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a sequence-based predictor for identifying 2′-O-methylation sites in *Homo sapiens*. J. Comput. Biol. *25*, 1266–1277.

27. Su, Z.D., Huang, Y., Zhang, Z.Y., Zhao, Y.W., Wang, D., Chen, W., Chou, K.C., and Lin, H. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics *34*, 4196–4204.

28. Feng, P.-M., Lin, H., and Chen, W. (2013). Identification of antioxidants from sequence information using naïve Bayes. Comput. Math. Methods Med. *2013*, 567529.

29. Feng, C.Q., Zhang, Z.Y., Zhu, X.J., Lin, Y., Chen, W., Tang, H., and Lin, H. (2019). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics *35*, 1469–1477.

30. Dao, F.Y., Lv, H., Wang, F., Feng, C.Q., Ding, H., Chen, W., and Lin, H. (2019). Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. Bioinformatics *35*, 2075–2083.

31. Liao, Z.J., Li, D.P., Wang, X.R., Li, L.S., and Zou, Q. (2018). Cancer diagnosis through isomiR expression with machine learning method. Curr. Bioinform. *13*, 57–63.

32. Chen, W., Feng, P., Liu, T., and Jin, D. (2019). Recent advances in machine learning methods for predicting heat shock proteins. Curr. Drug Metab. *20*, 224–228.

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

34. Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.C. (2019). iDNA6mA-PseKNC: Identifying DNA N^6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics *111*, 96–102.

35. Lai, H.Y., Zhang, Z.Y., Su, Z.D., Su, W., Ding, H., Chen, W., and Lin, H. (2019). iProEP: a computational predictor for predicting promoter. Mol. Ther. Nucleic Acids *17*, 337–346.

36. Wei, L., Wan, S., Guo, J., and Wong, K.K. (2017). A novel hierarchical selective ensemble classifier with bioinformatics application. Artif. Intell. Med. *83*, 82–90.