Molecular Therapy
Nucleic Acids
Original Article

iRNA-m7G: Identifying N^7-methylguanosine Sites by Fusing Multiple Features

Wei Chen,[1,2] Pengmian Feng,[1] 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

As an essential post-transcriptional modification, N^7-methylguanosine (m7G) regulates nearly every step of the life cycle of mRNA. Accurate identification of the m7G site in the transcriptome will provide insights into its biological functions and mechanisms. Although the m7G-methylated RNA immunoprecipitation sequencing (MeRIP-seq) method has been proposed in this regard, it is still cost-ineffective for detecting the m7G site. Therefore, it is urgent to develop new methods to identify the m7G site. In this work, we developed the first computational predictor called iRNA-m7G to identify m7G sites in the human transcriptome. The feature fusion strategy was used to integrate both sequence- and structure-based features. In the jackknife test, iRNA-m7G obtained an accuracy of 89.88%. The superiority of iRNA-m7G for identifying m7G sites was also demonstrated by comparing with other methods. We hope that iRNA-m7G can become a useful tool to identify m7G sites. A user-friendly web server for iRNA-m7G is freely accessible at http://lin-group.cn/server/iRNA-m7G/.

INTRODUCTION

Besides N^1-methyladenosine (m^1A), N^7-methylguanosine (m7G) is another kind of positively charged RNA modification.[1] m7G is added to the 5′ end co-transcriptionally during transcription, and it is essential for efficient gene expression and cell viability.[2] It has been found that m7G is required for nearly all phages of the mRNA cycles, such as RNA splicing,[3] polyadenylation,[4] nuclear export of mRNA,[5] translation,[6] and so on. Although studies on m7G have been carried out for a long time, the knowledge about its function is still limited. The key step of revealing the functions of m7G is to determine its accurate position in the transcriptome.

By using the mass spectrometry quantification and m7G-methylated RNA immunoprecipitation sequencing (MeRIP-seq) method,[7] Zhang et al. not only detected the m7G sites in *Homo sapiens* and *Mus. Musculus* but also provided the base resolution m7G sites in human HeLa and HepG2 cells. However, the MeRIP-seq method still has its own limitations,[7] and it is cost-ineffective for performing transcriptome-wide detections. Therefore, it is necessary to develop computational methods for identifying m7G sites.

To the best of our knowledge, there are no computational methods available for this aim. Inspired by the wide application of machine-learning methods for identifying RNA modification sites,[8,9] in this study, we developed a support vector machine (SVM)-based method, called iRNA-m7G, to identify m7G sites. To extract informative features to encode the RNA sequence, the feature fusion strategy was used to integrate three kinds of features, including nucleotide property and frequency, pseudo nucleotide composition, and secondary structure component. Experiments exhibited that the feature fusion strategy is superior to the single kind of features for identifying m7G sites. Moreover, a user-friendly web server for iRNA-m7G has been provided at http://lin-group.cn/server/iRNA-m7G/. We expect that the proposed predictor will speed up the detection of the m7G site.

RESULTS AND DISCUSSION

Performance of Each Kind of Feature

We built three models based on the three kinds of features (nucleotide property and frequency [NPF], pseudo nucleotide composition [PseDNC], and secondary structure component [SSC]), and we compared their performances for identifying m7G sites. As indicated in Equations 4 and 5, the PseDNC model is dependent on two parameters, w and λ. Hence, we first optimized the parameters of PseDNC. In general, the greater the λ value is, the more global sequence-order information the model contains. However, a larger λ would reduce the cluster-tolerant capacity so as to lower the cross-validation accuracy due to an overfitting problem. Therefore, the search ranges for

Received 8 July 2019; accepted 19 August 2019;
https://doi.org/10.1016/j.omtn.2019.08.022.

Correspondence: Wei Chen, Center for Genomics and Computational Biology, School of Life Sciences, North China University of Science and Technology, Tangshan 063000, 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

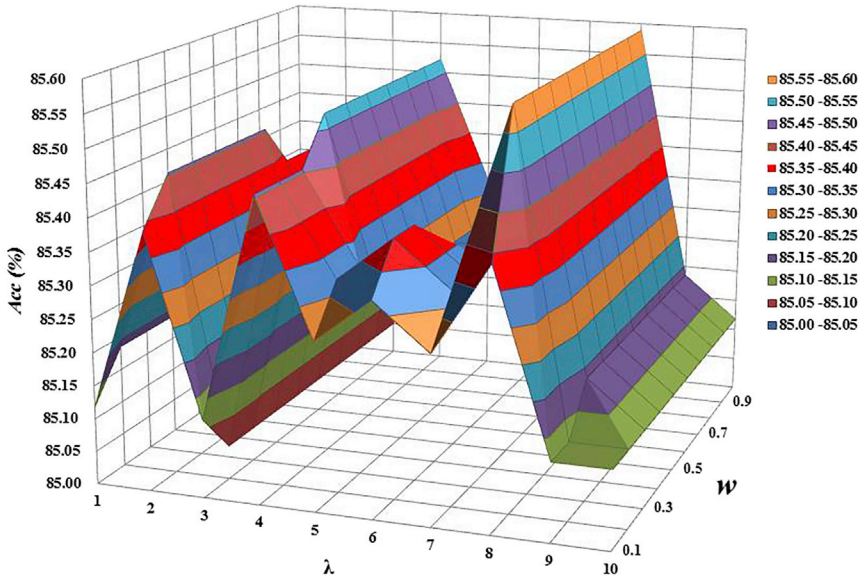

Figure 1. Determining the Optimal Values for the Two Parameters *w* and λ of PseDNC

w and λ were set in [0, 1] and [1, 10] with a step of 0.1 and 1, respectively. As shown in Figure 1, the PseDNC-based model yielded the best results when $w = 0.8$ and $\lambda = 8$.

The k-fold cross-validation test method was often used to examine the quality of various predictors.[10] For saving computational time, in the current study, the 10-fold cross-validation test was used to evaluate the performance of these models. Their predictive results were reported in Table 1. Among the three models, the NPF-based model obtained the highest accuracy of 89.14%, which is approximately 5% and 14% higher than that of the PseDNC- and SSC-based models, respectively, for identifying m7G sites in the dataset.

To objectively compare their performances, the area under the receiver operating characteristic curve (auROC) of these methods was also calculated. The NPF-based model obtained an auROC of 0.899, higher than the 0.841 and 0.776 obtained by the PseDNC- and SSC-based models, respectively.

Performance of Fusing Multiple Features

To investigate whether the feature fusion strategy could improve the performance, we built another model by fusing the NPF, PseDNC, and SSC features. The framework of how to build the model is shown in Figure 2. The model thus obtained was then evaluated by using the 10-fold cross-validation test. The detailed results are provided in the last row of Table 1. As indicated in Table 1, the sensitivity (Sn), specificity (Sp), accuracy (Acc), and Mathew's correlation coefficient (MCC) were all improved compared with those obtained by the NPF-, PseDNC-, and SSC-based models.

To intuitively compare the performance of the models based on different features, their ROC curves from the 10-fold cross-validation test were plotted in Figure 3. The fusion strategy-based model obtained an auROC of 0.946, which is higher than those of the NPF-, PseDNC-, and SSC-based models.

Moreover, to further demonstrate its stability for identifying m7G sites, the fusion strategy-based model was also evaluated by the jackknife test, in which 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. In the jackknife test, the fusion strategy-based model obtained an accuracy of 89.88% with the sensitivity of 89.07%, specificity of 90.69%, and MCC of 0.80, which is comparable to those from the 10-fold cross-validation test. These results indicate that the feature fusion strategy is effective and the model is robust for identifying m7G sites.

Comparison of SVM and Other Classifiers

Since there is no computational method that has been proposed for identifying m7G sites, to demonstrate its effectiveness, we compared the performance of the current SVM-based model with those of the Naive Bayes-, Random Forest-, LogitBoost-, and BayesNet-based models. The Naive Bayes, Random Forest, LogitBoost, and BayesNet were implemented by using WEKA.[11] For a fair comparison, all the models were built by using the the feature fusion strategy and tested on the same dataset. The 10-fold cross-validation test results of these models are reported in Table 2. As shown in Table 2, the SVM-based model obtained the best results in terms of the four metrics defined in Equation 9. The predictive accuracy of the SVM-based model is 9.7%, 3.3%, 6.1%, and 7.7% higher than those of the Naive Bayes-, Random Forest-, LogitBoost-, and BayesNet-based models, respectively. This result demonstrates that the SVM is more effective than other classification algorithms for identifying m7G sites.

Conclusions

In this study, we proposed iRNA-m7G, the first computational method to identify m7G sites. In this predictor, the feature fusion strategy was used to represent RNA sequences. Comparative results demonstrated that the feature fusion strategy is much more effective for identifying m7G sites than a single kind of feature.

Moreover, we also compared iRNA-m7G with the other four machine-learning algorithm-based methods, and we found that the SVM-based model achieves the best performance for identifying m7G sites.

Table 1. Predictive Results for Identifying m7G Sites by Using Different Features

Features	Sn (%)	Sp (%)	Acc (%)	MCC	auROC
NPF	88.12	90.15	89.14	0.78	0.899
PseDNC	81.92	87.99	84.95	0.70	0.841
SSC	73.11	78.71	75.91	0.52	0.776
Fusion	88.66	90.96	89.81	0.80	0.946

Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Mathew's correlation coefficient; auROC, area under the receiver operating characteristic curve; NPF, nucleotide property and frequency; PseDNC, pseudo nucleotide composition; SSC, secondary structure component.

For the convenience of the scientific community, a publicly accessible web server called iRNA-m7G that allows the prediction of m7G sites in RNA was established at http://lin-group.cn/server/iRNA-m7G/. We anticipate that iRNA-m7G will become a useful tool for identifying m7G sites. In future works, we will collect more m7G data and use powerful methods such as deep learning[12–15] to improve the performance of computationally identifying m7G sites.

MATERIALS AND METHODS

Benchmark Datasets

By using the MeRIP-seq method, Zhang et al.[7] detected 801 base-resolution m7G sites that appeared in human HeLa and HepG2 cells. By mapping these sites to the human genome (hg19), 801 m7G sites containing sequences were obtained. Preliminary tests indicated that the best predictive result was achieved when the sequence length is 41 bp with the m7G site in the center. To build a high-quality dataset, the CD-HIT software with the threshold of 80% was used to remove redundant sequences.[16,17] Accordingly, we obtained 741 m7G site-containing sequences.

The non-m7G site-containing sequences were obtained by choosing 41-bp-long sequences with the intermediate guanosine not detected as m7G by the MeRIP-seq method. By doing so, a huge number of negative samples is obtained. Since imbalanced datasets affect the performance evaluation of computational methods, to balance out the numbers between positive and negative samples in model training, we randomly picked out 741 non-m7G site sequences with the sequence similarity less than 80% to form the negative samples.

Sequence Representation

NPF

The NPF is an effective sequence-encoding scheme for computationally identifying nucleotide modification sites.[18–21] According to NPF, the i-th nucleotide n_i in RNA sequence can be represented by a four-dimensional vector (x_i, y_i, z_i, d_i), in which the elements are defined as follows:

$$x_i = \begin{cases} 1 & \text{if } n_i \in \{A, G\} \\ 0 & \text{otherwise} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{if } n_i \in \{A, U\} \\ 0 & \text{otherwise} \end{cases}$$

$$z_i = \begin{cases} 1 & \text{if } n_i \in \{A, C\} \\ 0 & \text{otherwise} \end{cases}, \qquad \text{(Equation 1)}$$

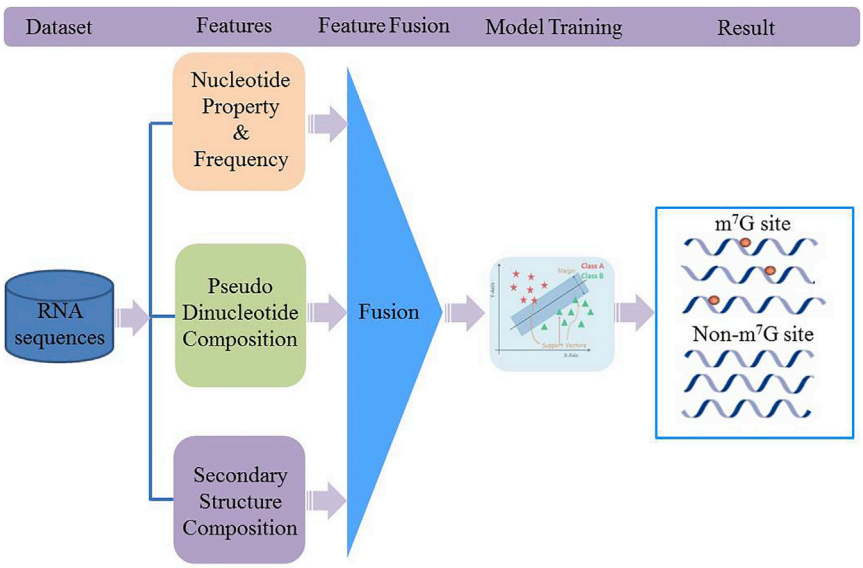

Figure 2. Framework of Developing iRNA-m7G

For an RNA sequence, it is converted into a feature vector by fusing nucleotide property and frequency, pseudo nucleotide composition, and secondary structure component. The support vector machine was used to build the classification model.

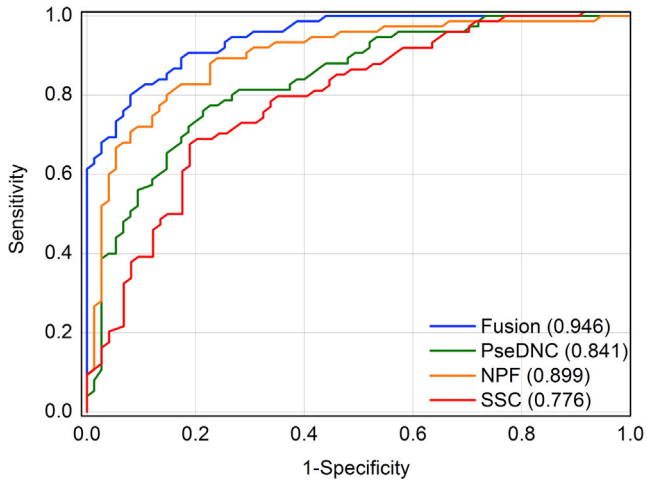

Figure 3. The Receiver Operating Characteristic Curves of the Models Based on Different Features Identifying m7G sites

SSC is the abbreviation for secondary structure component, NPF is for nucleotide property and frequency, PseDNC is for pseudo nucleotide composition, and fusion is the combination of the abovementioned three kinds of features. The auROC values were provided in brackets.

where the x, y, and z coordinates stand for the ring structure, hydrogen bond, and chemical functionality, respectively; d_i is the accumulated frequency and is defined as

$$d_i = \frac{1}{|N_i|}\sum_{j=1}^{l} f(n_j),\ f(n_j) = \begin{cases} 1 & if\ \ n_j = n_i \\ 0 & \end{cases}, \qquad \text{(Equation 2)}$$

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.

According to NPF, an RNA sequence with a length of l bp will be encoded by the following vector:

$$\mathbf{R} = \left[x_1\ y_1\ z_1\ d_1 \cdots x_i\ y_i\ z_i\ d_i \cdots\ x_l\ y_l\ z_l\ d_l\right]^{\mathrm{T}}. \qquad \text{(Equation 3)}$$

PseDNC

Besides the local sequence order information, the global sequence order effect is also important for computationally identifying RNA modification sites. Accordingly, in the current study, the PseDNC was also used to encode the RNA sequences,[22] which can be calculated by using PseKNC[23] and PseKNC-General.[24] Based on PseDNC, the RNA sequence is converted into a discrete vector defined as follows:

$$\mathrm{R} = [d_1\ d_2 \cdots d_{16}\ d_{16+1} \cdots d_{16+\lambda}]^{\mathrm{T}}, \qquad \text{(Equation 4)}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 16) \\ \dfrac{w\theta_{u-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (16 < u \le 16+\lambda) \end{cases} \qquad \text{(Equation 5)}$$

f_u $(u = 1, 2, \cdots, 16)$ is the occurrence frequency of the u-th non-overlapping dinucleotide in the RNA sequence, and

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1} \mathrm{C}_{i,\,i+j}\ (j = 1, 2, \cdots, \lambda; \lambda < L), \qquad \text{(Equation 6)}$$

where θ_j is the j-tier correlation factor that reflects the sequence order correlation between all the j-th most contiguous dinucleotide, and $\mathrm{C}_{i,i+j}$ is defined as

$$\mathrm{C}_{i,i+j} = \frac{1}{\mu}\sum_{g=1}^{\mu}\left[P_g(\mathrm{D}_i) - P_g\left(\mathrm{D}_{i+j}\right)\right]^2, \qquad \text{(Equation 7)}$$

where μ is the number of RNA physicochemical properties considered, $P_g(\mathrm{D}_i)$ is the normalized numerical value of the g-th (g = 1, 2, 3, ..., μ) RNA local structural property for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$ at position i, and $P_g(\mathrm{D}_{i+j})$ is the corresponding value for the dinucleotide $\mathrm{R}_{i+j}\mathrm{R}_{i+j+1}$ at position $i + j$.

In the current work, the enthalpy, entropy, and free energy were used to define PseDNC, which have been used to identify other kinds of RNA modifications. The values for the three physicochemical properties of the 16 different RNA dinucleotides were obtained from previous works.[25,26] Thus, μ in Equation 7 is equal to 3.

Table 2. Performance Comparison of Different Classifiers for Identifying m7G Sites by the 10-Fold Cross-Validation Test

Classifiers	Sn (%)	Sp (%)	Acc (%)	MCC
Naive Bayes	72.47	87.85	80.16	0.61
Random Forest	83.27	89.88	86.57	0.73
LogitBoost	81.38	86.23	83.81	0.68
BayesNet	77.19	87.04	82.12	0.65
SVM	88.66	90.96	89.81	0.80

Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Mathew's correlation coefficient; SVM, support vector machine.

SSC

The formation of RNA modification is affected by RNA structures. Hence, the RNAfold tool in the ViennaRNA package[27] was used to predict the secondary structure of the RNA sequences in the dataset. For each position in the RNA, the paired nucleotide was represented by a parenthesis ("(" or ")"), while the unpaired one was represented by a dot ("."). In the current study, we do not distinguish "(" and ")" and use "(" for both statuses. For a given tri-nucleotide, there are eight (2^3) possible structure statuses (i.e., "(((," "((.," "(..," "(.(," ".((," ".(.," "..(," and "..."). Together with the first nucleotide of the tri-nucleotide,

there will be 32 (4×8) possible sequence-structure modes denoted as "A-(((," "A-((.," "A-(.." ..., and "U-...".[28] Therefore, by using the sequence-structure mode, an RNA sequence can be represented as follows:

$$R = \left[f^{A}_{(((}, f^{A}_{((.}, f^{A}_{(..}, \ldots, f^{A}_{...}, f^{C}_{(((}, \ldots, f^{U}_{...}\right]^{T}. \quad \text{(Equation 8)}$$

SVM

In the current study, the LibSVM package 3.18, which is available at https://www.csie.ntu.edu.tw/~cjlin/libsvm/, was used to perform the classification task. The basic idea of SVM is to transform the input data into a high-dimensional feature space and then determine the optimal separating hyperplane. Because of its better performance, the radial basis kernel function (RBF) was used to obtain the separating hyperplane. The regularization parameter C and kernel parameter γ of the SVM operation engine were optimized in the ranges of $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{-5}]$ with the steps of 2 and 2^{-1}, respectively. The final prediction was made according to the probability obtained by SVM.[29–33] If its probability is >0.5, a guanine will be predicted as an m7G site.

Evaluation Metrics

In this study, the four metrics,[34–40] namely, Sn, Sp, Acc, and MCC, were used to measure the performance of the proposed methods, which are defined as follows:

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N^{+}_{-}}{N^{+}} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N^{-}_{+}}{N^{-}} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = 1 - \dfrac{N^{+}_{-} + N^{-}_{+}}{N^{+} + N^{-}} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N^{+}_{-}}{N^{+}} + \dfrac{N^{-}_{+}}{N^{-}}\right)}{\sqrt{\left(1 + \dfrac{N^{-}_{+} - N^{+}_{-}}{N^{+}}\right)\left(1 + \dfrac{N^{+}_{-} - N^{-}_{+}}{N^{-}}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases}, \quad \text{(Equation 9)}$$

where N^{+} represents the m7G site-containing sequence, while N^{+}_{-} is the number of m7G site-containing sequences incorrectly predicted to be of false m7G site-containing sequences; N^{-} is the total number of false m7G site-containing sequences, while N^{-}_{+} is the number of the false m7G site-containing sequences incorrectly predicted to be of m7G site-containing sequences.

Moreover, by plotting the sensitivity against (1-specificity) with the varying of the threshold, the ROC curve[41,42] was generated to evaluate the performance of the proposed method. The auROC is an indicator of the performance of the method. An auROC value of 0.5 is equivalent to random prediction while an auROC of 1 represents a perfect one.

AUTHOR CONTRIBUTIONS

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

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

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

REFERENCES

1. Furuichi, Y. (2015). Discovery of m(7)G-cap in eukaryotic mRNAs. Proc. Jpn. Acad., Ser. B, Phys. Biol. Sci. *91*, 394–409.
2. Cowling, V.H. (2009). Regulation of mRNA cap methylation. Biochem. J. *425*, 295–302.
3. Lindstrom, D.L., Squazzo, S.L., Muster, N., Burckin, T.A., Wachter, K.C., Emigh, C.A., McCleery, J.A., Yates, J.R., 3rd, and Hartzog, G.A. (2003). Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. Mol. Cell. Biol. *23*, 1368–1378.
4. Drummond, D.R., Armstrong, J., and Colman, A. (1985). The effect of capping and polyadenylation on the stability, movement and translation of synthetic messenger RNAs in Xenopus oocytes. Nucleic Acids Res. *13*, 7375–7394.
5. Lewis, J.D., and Izaurralde, E. (1997). The role of the cap structure in RNA processing and nuclear export. Eur. J. Biochem. *247*, 461–469.
6. Murthy, K.G., Park, P., and Manley, J.L. (1991). A nuclear micrococcal-sensitive, ATP-dependent exoribonuclease degrades uncapped but not capped RNA substrates. Nucleic Acids Res. *19*, 2685–2692.
7. Zhang, L.S., Liu, C., Ma, H., Dai, Q., Sun, H.L., Luo, G., Zhang, Z., Zhang, L., Hu, L., Dong, X., and He, C. (2019). Transcriptome-wide Mapping of Internal N7-Methylguanosine Methylome in Mammalian mRNA. Mol. Cell *74*, 1304–1316.e8.
8. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.C. (2018). iRNA-3typeA: Identifying Three Types of Modification at RNA's Adenosine Sites. Mol. Ther. Nucleic Acids *11*, 468–474.
9. 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.
10. Zhao, W., Zhou, Y., Cui, Q., and Zhou, Y. (2019). PACES: prediction of N4-acetylcytidine (ac4C) modification sites in mRNA. Sci. Rep. *9*, 11112.
11. Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I.H. (2004). Data mining in bioinformatics using Weka. Bioinformatics *20*, 2479–2481.
12. Hou, J., Wu, T., Cao, R., and Cheng, J. (2019). Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins. Published online April 15, 2019. https://doi.org/10.1002/prot.25697.
13. Patel, S., Tripathi, R., Kumari, V., and Varadwaj, P. (2017). DeepInteract: Deep Neural Network Based Protein-Protein Interaction Prediction Tool. Curr. Bioinform. *12*, 551–557.
14. Cao, R., Bhattacharya, D., Hou, J., and Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics *17*, 495.
15. Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., Zhang, J., Chan, L., and Cao, R. (2019). Survey of Machine Learning Techniques in Drug Discovery. Curr. Drug Metab. *20*, 185–193.

16. Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). Sequence clustering in bioinformatics: an empirical study. Brief. Bioinform. Published online September 18, 2018. https://doi.org/10.1093/bib/bby1090.

17. 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.

18. 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.

19. 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.

20. 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, btz358.

21. He, W., Jia, C., and Zou, Q. (2019). 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics *35*, 593–601.

22. Yang, H., Qiu, W.R., Liu, G., Guo, F.B., Chen, W., Chou, K.C., and Lin, H. (2018). iRSpot-Pse6NC: Identifying recombination spots in *Saccharomyces cerevisiae* by incorporating hexamer composition into general PseKNC. Int. J. Biol. Sci. *14*, 883–891.

23. Chen, W., Lei, T.-Y., Jin, D.-C., Lin, H., and Chou, K.-C. (2014). PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. *456*, 53–60.

24. Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.-C. (2015). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics *31*, 119–120.

25. Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T., and Turner, D.H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. Proc. Natl. Acad. Sci. USA *83*, 9373–9377.

26. Xia, T., SantaLucia, J., Jr., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., and Turner, D.H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry *37*, 14719–14735.

27. Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

28. Xue, C., Li, F., He, T., Liu, G.P., Li, Y., and Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics *6*, 310.

29. Tang, H., Chen, W., and Lin, H. (2016). Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. Mol. Biosyst. *12*, 1269–1275.

30. Zhu, P.P., Li, W.C., Zhong, Z.J., Deng, E.Z., Ding, H., Chen, W., and Lin, H. (2015). Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. Mol. Biosyst. *11*, 558–563.

31. Ding, H., and Li, D. (2015). Identification of mitochondrial proteins of malaria parasite using analysis of variance. Amino Acids *47*, 329–333.

32. Manavalan, B., Shin, T.H., and Lee, G. (2018). PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. Front. Microbiol. *9*, 476.

33. Cao, R., Wang, Z., Wang, Y., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC Bioinformatics *15*, 120.

34. Zhu, X.J., Feng, C.Q., Lai, H.Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowl. Based Syst. *163*, 787–793.

35. Tan, J.X., Li, S.H., Zhang, Z.M., Chen, C.X., Chen, W., Tang, H., and Lin, H. (2019). Identification of hormone binding proteins based on machine learning methods. Math. Biosci. Eng. *16*, 2466–2480.

36. Lv, H., Zhang, Z.M., Li, S.H., Tan, J.X., Chen, W., and Lin, H. (2019). Evaluation of different computational methods on 5-methylcytosine sites identification. Brief. Bioinform. Published online June 3, 2019. https://doi.org/10.1093/bib/bbz048.

37. Manavalan, B., Subramaniyam, S., Shin, T.H., Kim, M.O., and Lee, G. (2018). Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. J. Proteome Res. *17*, 2715–2726.

38. Tang, H., Cao, R.Z., Wang, W., Liu, T.S., Wang, L.M., and He, C.M. (2017). A two-step discriminated method to identify thermophilic proteins. Int. J. Biomath. *10*, 1750050.

39. Liu, B., Han, L., Liu, X., Wu, J., and Ma, Q. (2019). Computational prediction of sigma-54 promoters in bacterial genomes by integrating motif finding and machine learning strategies. IEEE/ACM Trans. Comput. Biol. Bioinform. *16*, 1211–1218.

40. Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif. Intell. Med. *83*, 67–74.

41. 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.

42. 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.