Methods xxx (xxxx) xxx

Contents lists available at ScienceDirect

Methods

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

iRNA-m5U: A sequence based predictor for identifying 5-methyluridine modification sites in *Saccharomyces cerevisiae*

Pengmian Feng [a], Wei Chen [a,b,c,*]

[a] School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China
[b] Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China
[c] School of Life Sciences, North China University of Science and Technology, Tangshan 063000, China

ARTICLE INFO

Keywords:
RNA modification
5-methyluridine
Feature representation
Support vector machine

ABSTRACT

The 5-methyluridine (m^5U) modification plays important roles in a series of biological processes. Accurate identification of m^5U sites will be helpful to decode its biological functions. Although experimental techniques have been proposed to detect m^5U, they are still expensive and time consuming. In the present work, a support vector machine based method, called iRNA-m5U, was developed to identify the m^5U sites in the *Saccharomyces cerevisiae* transcriptome. The performance of iRNA-m5U was validated based on different datasets. The accuracies obtained by iRNA-m5U is promising, indicating that it holds the potential to become an useful tool for the identification of m^5U sites.

1. Introduction

Over the past six decades, approximately 170 kinds of RNA modifications have been reported in the three kingdoms of life [1]. These covalent post-transcriptional modifications not only enriched the genetic information, but also participate in a series of biological processes. For example, by regulating RNA splicing [2], RNA stability [3,4] and protein translation efficiency [5], they were reported to play important roles in cell differentiation and reprogramming [6], immune tolerance [7], and even diseases [8].

Owing to the development of high throughput sequencing technology, the transcriptome-wide profiles were available for the common RNA modifications, such as N^6-methyladenosine (m^6A) [9], N^1-methyladenosine (m^1A) [10], 5-methylcytosine (m^5C) [11], etc. Compared with those modifications, researches on 5-methyluridine (m^5U) are extremely deficient. Therefore, it's necessary to develop new methods for identifying m^5U sites.

In 2019, the fluorouracil induced catalytic crosslinking sequencing technique was proposed to identify m^5U site in *Homo sapiens* [12]. However, this experimental method is still cost ineffective for transcriptome-wide detections. These dilemma was also faced by other kinds of modifications. To solve this problem, a series of in silico methods have been proposed to detect the RNA modifications in different species [13–20].

To the best of our knowledge, m5UPred is the only computational method for identifying m^5U site [21]. However, m5UPred is trained based on the data from human, and its performance is still not satisfactory for identifying the m^5U sites in *Saccharomyces cerevisiae*.

In the present work, we present a new predictor, called iRNA-m5U, to identify the m^5U site in *S. cerevisiae*. In this predictor, the nucleotide chemical property and nucleotide density were used to convert the RNA sequences into discrete feature vectors. iRNA-m5U obtained an accuracy of 98.82% for identifying m^5U site in the benchmark dataset, which is better than that of m5UPred.

2. Materials and methods

2.1. Benchmark dataset

The 263 m^5U site containing sequences in *S. cerevisiae* were obtained from the RMBase database [22]. These sequences were all 41 nt with the m^5U site in their center positions. Our series of works [23,24] have proved that the 41-nt long sequence with the modification site in the center is the optimal window size for RNA modification site identification. To construct a high quality benchmark dataset, the CD-HIT tool [25] was used to remove samples with the sequence similarity greater than 90%. Accordingly, 49 m^5U site containing sequences were retained and deemed as the positive dataset.

* Corresponding author at: School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China.
E-mail address: greatchen@ncst.edu.cn (W. Chen).

https://doi.org/10.1016/j.ymeth.2021.04.013
Received 22 March 2021; Received in revised form 11 April 2021; Accepted 15 April 2021
Available online 18 April 2021

Since the m^{5}U in RMBase database were all from tRNA, the negative samples were collected from the tRNAs of *S. cerevisiae* by obeying the following criteria. The sequences should be 41 nt with the unmodified uridine sites in the center. By doing so, 205 negative samples with the sequence similarity no greater than 90% were obtained. Therefore, the benchmark dataset (namely tRNA_Dataset) containing 49 samples and 205 negative samples was obtained.

In addition, to further evaluate the performance of the proposed method, we built another 10 negative datasets by harvesting the unmodified uridine sites from the transcripts of *S. cerevisiae*. By doing so, a huge number of 41-nt long sequences could be obtained. To balance the samples in the positive and negative datasets, we randomly selected out 490 negative samples and averagely divided them into 10 groups. Accordingly, another 10 negative datasets (namely dataset 1, dataset 2, …, dataset 10) with a 1:1 positive-to-negative ratio were constructed. All these data were provided in Supplementary Material.

2.2. Sequence encoding scheme

Since the effectiveness of nucleotide chemical property and nucleotide density have been proved for identifying nucleotide modification sites [26], they were also used to encode the samples in the present work.

2.2.1. Nucleotide chemical property

The four components of RNA, namely adenine (A), guanine (G), cytosine (C) and uridine (U), have different chemical structures. They could be categorized into three different groups in terms of the number of rings, strong or weak hydrogen bonds, and existence of amino or keto group. In order to include the different chemical properties, the nucleotides in RNA were projected into a three dimensional Cartesian coordinate system, where the $\times$, y and z coordinates stand for the ring structure, the hydrogen bond, and the amino/keto group, and were defined by the following formula [27],

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

Therefore, A, C, G and U in the sequence can be represented by (1, 1, 1), (0, 0, 1), (1, 0, 0) and (0, 1, 0), respectively.

2.2.2. Nucleotide density

In order to include the sequence order information surrounding the modification sites, the density d_i for nucleotide n_j at position i was defined as following [27],

$$d_i = \frac{1}{|N_i|}\sum_{j=1}^{i} f(n_j), \qquad f(n_j) = \begin{cases} 1 & if \quad n_j = q \\ 0 & other \quad cases \end{cases} \tag{2}$$

$|N_i|$ is the length of the prefix string containing i nucleotides. Take "A" in "AGCAUGCGA" as an example, the density of A at the 1st, 4th, and 9th positions were 1, 0.5 and 0.33, respectively.

By combining the nucleotide chemical property and nucleotide density, each nucleotide in the sequence will be encoded by a discrete vector containing 4 elements, 3 of them represent the nucleotide chemical property, and the other one represents the nucleotide density.

2.3. Classification algorithm

In the present work, the popular and powerful machine learning algorithm, namely support vector machine (SVM), was used to perform the classification [28–31]. The LibSVM package 3.18 downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/ was used to implement the SVM with the radial basis kernel function. The regularization parameter C and kernel parameter γ of SVM was determined by using the grid search method.

The predictions were made based on the probability score obtained from SVM. If the score is greater than 0.5, a uridine will be predicted as a m^{5}U, otherwise, non-m^{5}U.

2.4. Evaluation metrics

The jackknife test was used to examine the performance of the proposed method. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample. The performance of the proposed model was measured by using the Sn (Sensitivity), Sp (Specificity), ACC (accuracy), MCC (Mathew's correlation coefficient), which have been widely used to evaluate computational models in bioinformatics and were defined as following [32–34].

$$\begin{cases} Sn = \dfrac{TP}{TP + FN} \times 100\% \\ Sp = \dfrac{TN}{TN + FP} \times 100\% \\ Acc = \dfrac{TP + TN}{TP + FN + TN + FP} \times 100\% \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \end{cases} \tag{3}$$

where TP, TN, FP and FN represent the number of true positive, true negative, false positive, and false negative, respectively.

3. Result

3.1. Nucleotide composition analysis

In order to discover whether these exists nucleotide composition bias surrounding the m^{5}U sites, the sequence logo of the positive samples was plotted by using the WebLogo tool [35]. The motif "GUUCGA" located at the positions ranging from −1 to 4 relative to the m^{5}U site was detected (Fig. 1), which is exactly the consensus sequence to be recognized by m^{5}U methyltransferases [36,37].

3.2. m^{5}U sites identification

The above analysis indicates that the sequence based information holds the potential for the identification of m^{5}U sites. Thus, we encoded the RNA samples by using the scheme described in section 2.2. Accordingly, each of the 41-nt long sequence in the dataset was converted into a 164-dimensional vector and was used as the input of SVM. The regularization parameter and kernel parameters of SVM were 2 and 0.0078125, respectively. The model thus obtained is called iRNA-m5U. In the jackknife test, iRNA-m5U obtained an accuracy of 98.82% with the sensitivity of 93.88%, specificity of 100% and MCC of 0.96 for identifying the m^{5}U site in the tRNA_Dataset (Table 1). In addition, the receiver operating characteristic (ROC) curve was also plotted as shown in Fig. 2. It was found that iRNA-m5U obtained an AUC of (area under curve) 0.969, indicating its excellent performance for identifying m^{5}U sites.

To demonstrate whether the performance of the proposed method is depended on the negative samples, we further evaluated it by using the 10 negative datasets derived from the transcript. The performances of the proposed method for identifying the m^{5}U sites in the 10 datasets are still excellent, and are comparable with that based on the tRNA_Dataset (Table 1). This result demonstrates the robustness of the proposed method, and also indicates that its performance is independent of the negative sample selection bias.

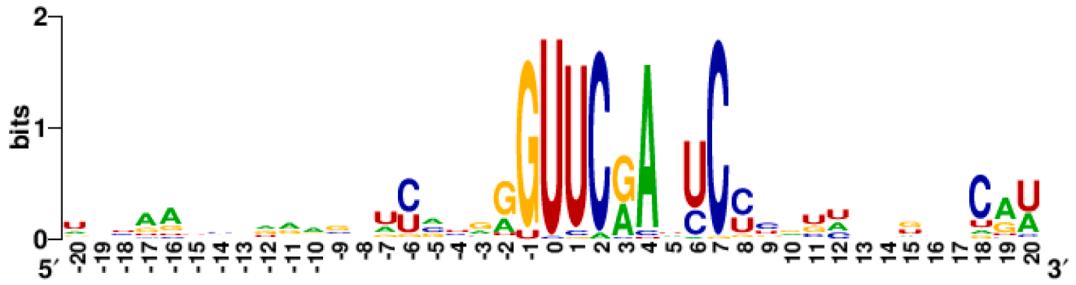

Fig. 1. Nucleotide composition analysis for sequences surrounding the m^5U sites.

Table 1
Predictive results for identifying m^5U site based on different dataset.

Source	Dataset	Sn (%)	Sp (%)	Acc (%)	MCC
tRNA	tRNA_Dataset	93.88	100	98.82	0.96
Transcriptome	Dataset 1	93.88	100	96.94	0.94
	Dataset 2	93.88	100	96.94	0.94
	Dataset 3	93.88	100	96.94	0.94
	Dataset 4	93.88	100	96.94	0.94
	Dataset 5	93.88	100	96.94	0.94
	Dataset 6	91.84	100	95.92	0.92
	Dataset 7	95.92	97.96	96.94	0.94
	Dataset 8	93.88	97.96	95.92	0.92
	Dataset 9	93.88	100	96.94	0.94
	Dataset 10	93.88	100	96.94	0.94

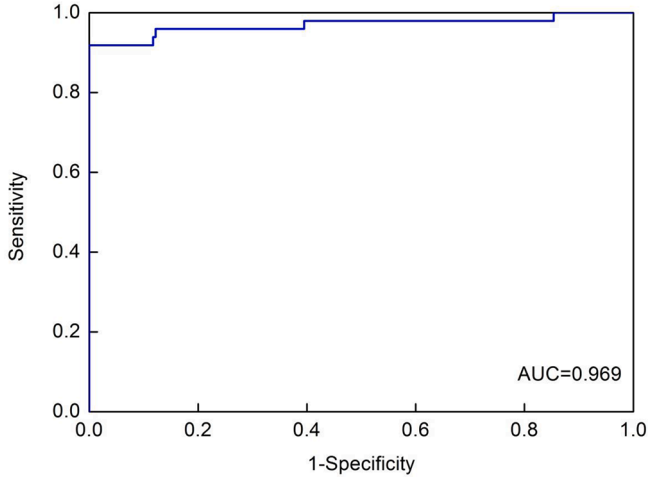

Fig. 2. The ROC curve of iRNA-m5U for identifying m^5U.

3.3. Comparison with existing methods

To the best of our knowledge, m5UPred [21] is the only computational method for identifying m^5U sites in human transcriptome. Therefore, we compared the performance of iRNA-m5U with that of m5UPred for identifying the m^5U sites in the tRNA_Dataset. It was found that m5UPred only obtained an accuracy of 64.96% for identifying the m^5U sites in the *S. cerevisiae* transcriptome (Table 2).

At the same time, we also validated the iRNA-m5U based on the independent test dataset built by Jiang et al [21], which includes 245 human m^5U sites containing sequences. However, iRNA-m5U only correctly identified 55 m^5U sites with the accuracy of 22.45%, which is lower than that of m5UPred.

The above results indicate that both m5UPred and iRNA-m5U are species specific. Therefore, it's necessary to develop species-specific methods for identifying m^5U sites.

Table 2
Comparative results for identifying m^5U site based on the tRNA_Dataset.

Dataset	Sn (%)	Sp (%)	Acc (%)	MCC
m5UPred	61.22	65.85	64.96	0.22
Our work	93.88	100	98.82	0.96

4. Discussion

By performing the nucleotide composition analysis, the motif "GUUCGA" were detected surrounding the m^5U sites. Accordingly, based on the sequence-derived information, namely nucleotide chemical property and nucleotide density, the iRNA-m5U was proposed to identify the m^5U sites in the *S. cerevisiae* transcriptome. The performances of iRNA-m5U were validated by using different datasets. The jackknife test results indicate that iRNA-m5U is promising and smarter than m5UPred for identifying m^5U sites in *S. cerevisiae*.

In addition, iRNA-m5U was also applied to identify the m^5U sites in human transcriptome. However, its performance was lower than that of m5UPred. This might be due to the small size of the benchmark dataset used to train iRNA-m5U. The features used in the present work were not informative enough to capture the key information to represent m^5U site containing sequences in all species.

In order to enhance the generalization ability of iRNA-m5U, in future work, we will try to enlarge the dataset by harvesting much more data, and optimize the model by integrating features from different sources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Nature Science Foundation of China (No. 31771471).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymeth.2021.04.013.

References

[1] M. Frye, B.T. Harada, M. Behm, C. He, RNA modifications modulate gene expression during development, Science 361 (6409) (2018) 1346–1349.
[2] N. Guzzi, M. Cieśla, P.C.T. Ngoc, S. Lang, S. Arora, M. Dimitriou, K. Pimková, M.N. E. Sommarin, R. Munita, M. Lubas, Y. Lim, K. Okuyama, S. Soneji, G. Karlsson, J. Hansson, G. Jönsson, A.H. Lund, M. Sigvardsson, E. Hellström-Lindberg, A. C. Hsieh, C. Bellodi, Pseudouridylation of tRNA-Derived Fragments Steers Translational Control in Stem Cells, Cell 173 (5) (2018) 1204–1216.e26.

[3] X. Wang, Z. Lu, A. Gomez, G.C. Hon, Y. Yue, D. Han, Y.e. Fu, M. Parisien, Q. Dai, G. Jia, B. Ren, T. Pan, C. He, N6-methyladenosine-dependent regulation of messenger RNA stability, Nature 505 (7481) (2014) 117–120.
[4] Z.C. Xu, P.M. Feng, H. Yang, W.R. Qiu, W. Chen, H. Lin, iRNAD: a computational tool for identifying D modification sites in RNA sequence, Bioinformatics 35 (23) (2019) 4922–4929.
[5] S.Y. Hwang, H. Jung, S. Mun, S. Lee, K. Park, S.C. Baek, H.C. Moon, H. Kim, B. Kim, Y. Choi, Y.H. Go, W. Tang, J. Choi, J.K. Choi, H.J. Cha, H.Y. Park, P. Liang, V. N. Kim, K. Han, K. Ahn, L1 retrotransposons exploit RNA m(6)A modification as an evolutionary driving force, Nat. Commun. 12 (1) (2021) 880.
[6] S. Delaunay, M. Frye, RNA modifications regulating cell fate in cancer, Nat. Cell Biol. 21 (5) (2019) 552–559.
[7] X. Lou, J.J. Wang, Y.Q. Wei, J.J. Sun, Emerging role of RNA modification N6-methyladenosine in immune evasion, Cell Death Dis. 12 (4) (2021) 300.
[8] N. Jonkhout, J. Tran, M.A. Smith, N. Schonrock, J.S. Mattick, E.M. Novoa, The RNA modification landscape in human disease, RNA 23 (12) (2017) 1754–1769.
[9] H. Liu, O. Begik, M.C. Lucas, J.M. Ramirez, C.E. Mason, D. Wiener, S. Schwartz, J. S. Mattick, M.A. Smith, E.M. Novoa, Accurate detection of m(6)A RNA modifications in native RNA sequences, Nat. Commun. 10 (1) (2019) 4079.
[10] D. Dominissini, S. Nachtergaele, S. Moshitch-Moshkovitz, E. Peer, N. Kol, M.S. Ben-Haim, Q. Dai, A. Di Segni, M. Salmon-Divon, W.C. Clark, G. Zheng, T. Pan, O. z. Solomon, E. Eyal, V. Hershkovitz, D. Han, L.C. Doré, N. Amariglio, G. Rechavi, C. He, The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA, Nature 530 (7591) (2016) 441–446.
[11] S. Edelheit, S. Schwartz, M.R. Mumbach, O. Wurtzel, R. Sorek, V. de Crécy-Lagard, Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs, PLoS genetics 9 (6) (2013) e1003602.
[12] J.M. Carter, W. Emmett, I.R. Mozos, A. Kotter, M. Helm, J. Ule, S. Hussain, FICC-Seq: a method for enzyme-specified profiling of methyl-5-uridine in cellular RNA, Nucleic acids research 47(19) (2019) e113.
[13] K. Chen, Z. Wei, Q. Zhang, X. Wu, R. Rong, Z. Lu, J. Su, J.P. de Magalhaes, D.J. Rigden, J. Meng, WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach, Nucleic acids research 47(7) (2019) e41.
[14] K. Liu, W. Chen, iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications, Bioinformatics 36 (11) (2020) 3336–3342.
[15] Y. Zhou, P. Zeng, Y.H. Li, Z. Zhang, Q. Cui, SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features, Nucleic acids research 44(10) (2016) e91.
[16] Q. Zou, P. Xing, L. Wei, B. Liu, Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA, RNA 25 (2) (2019) 205–218.
[17] B. Song, Y. Tang, Z. Wei, G. Liu, J. Su, J. Meng, K. Chen, PIANO: A Web Server for Pseudouridine-Site (Psi) Identification and Functional Annotation, Front. Genet. 11 (2020) 88.
[18] C. Dai, P. Feng, L. Cui, R. Su, W. Chen, L. Wei, Iterative feature representation algorithm to improve the predictive performance of N7-methylguanosine sites, Briefings Bioinf. (2020), https://doi.org/10.1093/bib/bbaa278.
[19] K. Liu, W. Chen, H. Lin, XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites, Mol. Genet. Genomics 295 (1) (2020) 13–21.
[20] Z. Lv, J. Zhang, H. Ding, Q. Zou, RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites, Front. Bioengineering Biotechnol. 8 (2020) 134.
[21] J. Jiang, B. Song, Y. Tang, K. Chen, Z. Wei, J. Meng, m5UPred: A Web Server for the Prediction of RNA 5-Methyluridine Sites from Sequences, Molecular Therapy-Nucleic acids 22 (2020) 742–747.
[22] J.J. Xuan, W.J. Sun, P.H. Lin, K.R. Zhou, S. Liu, L.L. Zheng, L.H. Qu, J.H. Yang, RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data, Nucleic Acids Res. 46 (D1) (2018) D327–D334.
[23] W. Chen, P. Feng, X. Song, H. Lv, H. Lin, iRNA-m7G: Identifying N(7)-methylguanosine Sites by Fusing Multiple Features, Molecular therapy, Nucleic acids 18 (2019) 269–274.
[24] W. Chen, P. 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.
[25] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (23) (2012) 3150–3152.
[26] W. Chen, H. Yang, P. Feng, H. Ding, H. Lin, iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, Bioinformatics 33 (22) (2017) 3518–3523.
[27] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, K.C. Chou, iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC, Molecular therapy. Nucleic acids 7 (2017) 155–163.
[28] J.-H. Kuo, C.-C. Chang, C.-W. Chen, H.-H. Liang, C.-Y. Chang, Y.-W. Chu, Sequence-based Structural B-cell Epitope Prediction by Using Two Layer SVM Model and Association Rule Features, Curr. Bioinform. 15 (3) (2020) 246–252.
[29] M. Tahir, A. Idris, MD-LBP: An Efficient Computational Model for Protein Subcellular Localization from HeLa Cell Lines Using SVM, Curr. Bioinform. 15 (3) (2020) 204–211.
[30] Y. Zou, H. Wu, X. Guo, L. Peng, F. Guo, MK-FSVM-SVDD: A Multiple Kernel-based Fuzzy SVM Model for Predicting DNA-binding Proteins via Support Vector Data Description, Curr. Bioinform. 16 (2) (2020) 274–283.
[31] Z.Y. Zhang, Y.H. Yang, H. Ding, D. Wang, W. Chen, H. Lin, Design powerful predictor for mRNA subcellular location prediction in Homo sapiens, Briefings Bioinf. 22 (1) (2021) 526–535.
[32] W. Chen, P. Feng, F. Nie, iATP: A Sequence Based Method for Identifying Anti-tubercular Peptides, Med. Chem. 16 (5) (2020) 620–625.
[33] Z. Lv, P. Wang, Q. Zou, Q. Jiang, Identification of Sub-Golgi Protein Localization by Use of Deep Representation Learning Features, Bioinformatics (2021), https://doi.org/10.1093/bioinformatics/btaa1074.
[34] D. Wang, Z. Zhang, Y. Jiang, Z. Mao, D. Wang, H. Lin, D. Xu, DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism, Nucleic Acids Res. (2021), https://doi.org/10.1093/nar/gkab016.
[35] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (6) (2004) 1188–1190.
[36] A. Alian, T.T. Lee, S.L. Griner, R.M. Stroud, J. Finer-Moore, Structure of a TrmA-RNA complex: A consensus RNA fold contributes to substrate selectivity and catalysis in m5U methyltransferases, PNAS 105 (19) (2008) 6876–6881.
[37] K.M. McKenney, M.A.T. Rubio, J.D. Alfonzo, The Evolution of Substrate Specificity by tRNA Modification Enzymes, The Enzymes 41 (2017) 51–88.