Review

Recent Advances in Identification of RNA Modifications

Wei Chen [1,*] and Hao Lin [2,*]

[1] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China

[2] 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

[*] Correspondence: chenweiimu@gmail.com (W.C.); hlin@uestc.edu.cn (H.L.); Tel.: +86-315-3725715 (W.C.)

Abstract: RNA modifications are involved in a broad spectrum of biological and physiological processes. To reveal the functions of RNA modifications, it is important to accurately predict their positions. Although high-throughput experimental techniques have been proposed, they are cost-ineffective. As good complements of experiments, many computational methods have been proposed to predict RNA modification sites in recent years. In this review, we will summarize the existing computational approaches directed at predicting RNA modification sites. We will also discuss the challenges and future perspectives in developing reliable methods for predicting RNA modification sites.

1. Introduction

Since the first kind of RNA modification was discovered 60 years ago [1], more than 100 kinds of RNA modifications have been reported in different RNA species [2]. RNA modifications have been found to participate in various biological activities [3–11]. Therefore, the knowledge about their accurate positions in transcriptome is important for understanding the mechanisms and functions of these post-transcriptional modifications.

Due to the lack of effective methods, studies on RNA modifications have been hindered for a long period of time. Recent advances in next-generation sequencing technology have opened doors for the detection of RNA modifications. By using high-throughput sequencing methods [12–18], RNA modifications have been detected in various species of RNAs. The details of these experimental techniques have been summarized in a recent review [19]. These experimental methods indeed played key roles in promoting research progress on the biological functions of RNA modifications. However, because of the labor-intensive nature of experiments, the gap between the number of transcriptomes from different cell lines or organisms and the number of known modification sites is widening rapidly. Hence, the development of computational methods to accurately predict post-transcriptional modification sites from sequence information is urgent for the biological community.

A high quality dataset is the primary requirement for developing machine learning models. Although various high-throughput experimental techniques have been developed to predict diverse RNA modifications, the generated data are scattered separately and thus inconvenient for the scientific community until the appearance of the RMBase [20]. By collecting and integrating experimental data from high-throughput modification sequencing methods, RMBase provides information for N^6-methyladenosine (m^6A), pseudouridine (ψ), 5-methylcytosine (m^5C), and other types of RNA modifications [20], which is invaluable for the development of computational models.

Based on high-throughput experimental data and RMBase, a host of computational approaches have been developed for the identification of RNA modifications in the past four years [21–36].

This review will summarize the representative computational approaches developed for the identification of RNA modifications that have been mapped transcriptome-wide, i.e., m^6A, ψ, and N^1-methyladenosine (m^1A). Current challenges facing the computational prediction of RNA modifications and future perspectives are also discussed.

2. Computational Models for N^6-methyladenosine (m^6A)

m^6A occurs at the 6th N position of the adenosine residue in the consensus sequence motif RRACH (R is purine and H is either A, C, or U) [37,38]. As a dynamic chemical modification, m^6A is catalyzed by multicomponent methyltransferase complex, i.e. methyltransferase like 3 (METTL3) methyltransferase like 14 (METTL14) and Wilm's tumor 1 associating protein (WTAP) and is reversed by demethylases fat mass and obesity-associated protein (FTO) and alkylation repair homologue protein 5 (ALKBH5) [39,40].

Although a broad set of its biological functions have been revealed, how m^6A controls these processes is still a major challenge. Therefore, knowledge about the positions of m^6A site will be helpful for understanding its mechanisms and functions.

Based on the data from m^6A-seq experiment, Schwartz et al. proposed the first computational model to predict m^6A site in the yeast transcriptome [38]. Although this work plays a role in promoting the research progress on computationally predicting the distribution of m^6A, no web server was provided for their method.

In 2014, Chen et al. established two efficient computational models to predict the m^6A sites in yeast, namely, iRNA-Methyl [23] and m^6Apred [29], respectively. The m^6Apred not only considers the accumulated frequency of nucleotides surrounding m^6A sites, but also integrates nucleotide chemical properties. For the convenience of biologists, a web server for m^6Apred has been provided at http://lin.uestc.edu.cn/server/m6Apred.php, which is the first free online tool for predicting m^6A sites. Later on, by incorporating RNA sequences using the pseudo nucleotide composition [41,42], the iRNA-Methyl was developed and is freely accessible at http://lin.uestc.edu.cn/server/iRNA-Methyl. Both methods have become a useful starting point for developing computational tools for predicting m^6A sites.

Inspired by Chen et al.'s works [23,29], Zhou and his colleagues developed a random forest-based method called SRAMP (http://www.cuilab.cn/sramp) to predict m^6A sites [31], in which both sequence information and the RNA secondary structures were used to encode RNA sequences. The overall performance of SRAMP is comparable with that of iRNA-Methyl and m^6Apred. In addition, SRAMP is not only applicable for predicting m^6A sites in yeast, but is also able to predict m^6A sites in human and mouse transcriptomes, which is superior to both m^6Apred and iRNA-Methyl.

Benefiting from the RMBase [20], Chen et al. proposed the MethyRNA (http://lin.uestc.edu.cn/server/methyrna) to predict m^6A sites in both human and mouse transcriptomes [28], in which RNA sequences are encoded using the nucleotide-accumulated frequency and chemical properties. MethyRNA obtained accuracies of 90.38% and 88.39% for human and mouse [28], respectively.

Considering that there was no computational tool available for predicting m^6A sites in plants until 2016, by using the same encoding scheme as the previous work [28], an online tool called M6ATH was developed to predict m^6A sites in the *Arabidopsis thaliana* transcriptome [26]. The online web server for M6ATH is available at http://lin.uestc.edu.cn/server/M6ATH.

3. Computational models for Pseudouridine (ψ)

ψ is the isomer of uridine [43], which has been found in rRNAs, tRNAs, snoRNAs, and mRNA [11]. Unlike m^6A modification, ψ is not reversible [43]. In 2015, Li and his colleagues performed a pioneering work and developed the first ψ site predictor called PPUS [32], which is available at http://lyh.pkmu.cn/ppus/. Each RNA sequence sample in PPUS was formulated using the orthogonal binary coding

scheme [32], i.e., the classic four nucleotides A, C, G, and U, and the dummy nucleotide X were encoded as {1,0,0,0,0}, {0,1,0,0,0}, {0,0,1,0,0}, {0,0,0,1,0}, and {0,0,0,0,1}, respectively. It has been shown that PPUS exhibit considerable accuracy for predicting ψ sites in fivefold cross-validation tests in both human and yeast transcriptomes. However, the accuracy for predicting ψ sites still needs to be improved.

For predicting the Ψ modification sites with higher success rates and being able to cover more species, Chen et al. developed a more powerful predictor called iRNA-PseU [27], which is available at http://lin.uestc.edu.cn/server/iRNA-PseU. iRNA-PseU is trained based on the data collected from the RMBase [20], and in which RNA sequences are also encoded using the nucleotide accumulated frequency and chemical properties. iRNA-PseU is not only able to predict ψ sites in humans and yeast, but is also applicable to the mouse transcriptome. In addition, the performance of iRNA-PseU is better than that of PPUS when tested on the same independent dataset, indicating that iRNA-PseU will play a complementary role to the existing PPUS for predicting ψ sites.

4. Computational Model for N^1-Methyladenosine (m^1A)

m^1A was first discovered in RNA 50 years ago [44]. However, research on m^1A has lagged behind as a result of lack of effective methods for detecting m^1A sites. Recently, two high-throughput experimental techniques, i.e., MeRIP-seq [12] and m^1A-ID-seq [13], have been developed. These methods promote the research progress on predicting m^1A sites. However, their resolutions are not fully satisfactory, as they cannot pinpoint which adenosine residue is modified. Therefore, it is necessary to develop new methods for studying the distribution of m^1A sites.

Experimental data has provided unprecedented opportunities. Based on the data from the MeRIP-seq and m^1A-ID-seq experiments, Chen et al. proposed the RAMPred server for predicting m^1A sites [22], which is the first computational tool for predicting m^1A sites so far and is freely accessible at http://lin.uestc.edu.cn/server/RAMPred. The RAMPred is able to predict m^1A sites in human, mouse, and yeast transcriptomes [22]. It has been shown that RAMPred achieves promising performances in the rigorous jackknife tests and cross cell line tests, indicating that RAMPred holds a very high potential to become a useful tool for predicting m^1A sites.

5. Current Challenges and Future Directions

As a new level of gene regulation, RNA modifications participate in diverse biological processes, ranging from regulating RNA splicing, inducing RNA decay to altering the genetic code. Although high-throughput sequencing methods have been proposed to detect RNA modification sites, most of them are still cost-ineffective and could not precisely pinpoint out which ribonucleic acid is chemically modified.

In the past four years, computational approaches have provided useful strategies for efficiently detecting RNA modification sites. Although impressive progress has been achieved by computational methods, there still exist some challenges that need to be considered in future work.

Since most of the existing approaches are based on sequence context, they cannot accurately predict changes in position and stoichiometry of RNA modifications. As an epigenetic modification, RNA methylation is also a complicated progress. Besides sequence context and nucleotide chemical properties, other factors may also be helpful for RNA modification site identification. Recently, Patil et al. reported that RNA-binding motif protein 15 (RBM15) and its paralogue RBM15B binding sites are in near the methylated m6A sites in DRACH consensus, while the non-methylated DRACH neighborhood is poor in the RBM15/15B binding sites [5]. These results indicate the consensus motifs surrounding m^6A sites may be important for identifying m^6A sites. This point has been proved in our recent study [45]. By integrating the consensus motif in the computational model, the predictive performance for identifying m^6A sites was improved [45]. We believe that the consensus motif will also be helpful for identifying other kinds of RNA modifications. Therefore, for improving the performance for RNA modification sites identification, it is wise to combine all the above-mentioned factors together

when developing new models in the future work. Nevertheless, as the accuracy and predictive power of computational approaches in identification of RNA modification sites improve, they are expected to provide valuable avenues for understanding the biological functions of RNA modifications.

Reminiscent of the regulation of gene expression by histone modifications, it is also possible that a combination of different types of RNA modifications might mediate biological functions together [46]. Thus, it is necessary to develop a platform that can be used to deal with the system that simultaneously contains several different types of RNA modifications.

Author Contributions: W.C. and H.L. performed the analysis and wrote the paper. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cohn, W.E. Some results of the applications of ion-exchange chromatography to nucleic acid chemistry. *J. Cell. Physiol. Suppl.* **1951**, *38*, 21–40. [CrossRef] [PubMed]

2. Machnicka, M.A.; Milanowska, K.; Osman Oglou, O.; Purta, E.; Kurkowska, M.; Olchowik, A.; Januszewski, W.; Kalinowski, S.; Dunin-Horkawicz, S.; Rother, K.M.; et al. MODOMICS: A database of RNA modification pathways—2013 update. *Nucleic Acids Res.* **2013**, *41*, D262–D267. [CrossRef] [PubMed]

3. Meyer, K.D.; Jaffrey, S.R. The dynamic epitranscriptome: N^6-methyladenosine and gene expression control. *Nat. Rev. Mol. Cell Biol.* **2014**, *15*, 313–326. [CrossRef] [PubMed]

4. Nilsen, T.W. Molecular biology. Internal mRNA methylation finally finds functions. *Science* **2014**, *343*, 1207–1208. [CrossRef] [PubMed]

5. Patil, D.P.; Chen, C.K.; Pickering, B.F.; Chow, A.; Jackson, C.; Guttman, M.; Jaffrey, S.R. m^6A RNA methylation promotes XIST-mediated transcriptional repression. *Nature* **2016**, *537*, 369–373. [CrossRef] [PubMed]

6. Zhao, B.S.; Roundtree, I.A.; He, C. Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell Biol.* **2016**, *18*, 31–42. [CrossRef] [PubMed]

7. Chen, T.; Hao, Y.J.; Zhang, Y.; Li, M.M.; Wang, M.; Han, W.; Wu, Y.; Lv, Y.; Hao, J.; Wang, L.; et al. m^6A RNA methylation is regulated by micrornas and promotes reprogramming to pluripotency. *Cell Stem Cell* **2015**, *16*, 289–301. [CrossRef] [PubMed]

8. Chan, C.T.; Dyavaiah, M.; DeMott, M.S.; Taghizadeh, K.; Dedon, P.C.; Begley, T.J. A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. *PLoS Genet.* **2010**, *6*, e1001247. [CrossRef] [PubMed]

9. Helm, M.; Alfonzo, J.D. Posttranscriptional RNA modifications: Playing metabolic games in a cell's chemical legoland. *Chem. Biol.* **2014**, *21*, 174–185. [CrossRef] [PubMed]

10. Peifer, C.; Sharma, S.; Watzinger, P.; Lamberth, S.; Kotter, P.; Entian, K.D. Yeast Rrp8p, a novel methyltransferase responsible for m^1A 645 base modification of 25s rRNA. *Nucleic Acids Res.* **2013**, *41*, 1151–1163. [CrossRef] [PubMed]

11. Ge, J.; Yu, Y.T. RNA pseudouridylation: New insights into an old modification. *Trends Biochem. Sci.* **2013**, *38*, 210–218. [CrossRef] [PubMed]

12. 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. The dynamic N^1-methyladenosine methylome in eukaryotic messenger RNA. *Nature* **2016**, *530*, 441–446. [CrossRef] [PubMed]

13. Li, X.; Xiong, X.; Wang, K.; Wang, L.; Shu, X.; Ma, S.; Yi, C. Transcriptome-wide mapping reveals reversible and dynamic N^1-methyladenosine methylome. *Nat. Chem. Biol.* **2016**, *12*, 311–316. [CrossRef] [PubMed]

14. Carlile, T.M.; Rojas-Duran, M.F.; Zinshteyn, B.; Shin, H.; Bartoli, K.M.; Gilbert, W.V. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* **2014**, *515*, 143–146. [CrossRef] [PubMed]

15. Edelheit, S.; Schwartz, S.; Mumbach, M.R.; Wurtzel, O.; Sorek, R. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m^{5}C within archaeal mRNAs. *PLoS Genet.* **2013**, *9*, e1003602. [CrossRef] [PubMed]

16. Khoddami, V.; Cairns, B.R. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat. Biotechnol.* **2013**, *31*, 458–464. [CrossRef] [PubMed]

17. Chen, K.; Lu, Z.; Wang, X.; Fu, Y.; Luo, G.Z.; Liu, N.; Han, D.; Dominissini, D.; Dai, Q.; Pan, T.; et al. High-resolution N^6-methyladenosine (m^{6}A) map using photo-crosslinking-assisted m^{6}A sequencing. *Angew. Chem.* **2015**, *54*, 1587–1590. [CrossRef] [PubMed]

18. Linder, B.; Grozhik, A.V.; Olarerin-George, A.O.; Meydan, C.; Mason, C.E.; Jaffrey, S.R. Single-nucleotide-resolution mapping of m^{6}A and m^{6}Am throughout the transcriptome. *Nat. Methods* **2015**, *12*, 767–772. [CrossRef] [PubMed]

19. Shafik, A.; Schumann, U.; Evers, M.; Sibbritt, T.; Preiss, T. The emerging epitranscriptomics of long noncoding RNAs. *Biochim. Biophys. Acta* **2016**, *1859*, 59–70. [CrossRef] [PubMed]

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

21. Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K.C. iRNA-Pseu: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* **2016**, *5*, e332.

22. Chen, W.; Feng, P.; Tang, H.; Ding, H.; Lin, H. Rampred: Identifying the N^1-methyladenosine sites in eukaryotic transcriptomes. *Sci. Rep.* **2016**, *6*, 31080. [CrossRef] [PubMed]

23. Chen, W.; Feng, P.; Ding, H.; Lin, H.; Chou, K.C. iRNA-methyl: Identifying N^6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* **2015**, *490*, 26–33. [CrossRef] [PubMed]

24. Chen, W.; Feng, P.; Tang, H.; Ding, H.; Lin, H. Identifying 2′-*O*-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions. *Genomics* **2016**, *107*, 255–258. [CrossRef] [PubMed]

25. Chen, W.; Feng, P.; Ding, H.; Lin, H. PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions. *Sci. Rep.* **2016**, *6*, 35123. [CrossRef] [PubMed]

26. Chen, W.; Feng, P.; Ding, H.; Lin, H. Identifying N^6-methyladenosine sites in the *Arabidopsis thaliana* transcriptome. *Mol. Genet. Genom. MGG* **2016**, *291*, 2225–2229. [CrossRef] [PubMed]

27. Chen, W.; Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chou, K.C. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget* **2017**. [CrossRef] [PubMed]

28. Chen, W.; Tang, H.; Lin, H. MethyRNA: A web server for identification of N^6-methyladenosine sites. *J. Biomol. Struct. Dyn.* **2016**, 1–5. [CrossRef] [PubMed]

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

30. Feng, P.; Ding, H.; Chen, W.; Lin, H. Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. *Mol. Biosyst.* **2016**, *12*, 3307–3311. [CrossRef] [PubMed]

31. Zhou, Y.; Zeng, P.; Li, Y.H.; Zhang, Z.; Cui, Q. SRAMP: Prediction of mammalian Nn^6-methyladenosine (m^{6}A) sites based on sequence-derived features. *Nucleic Acids Res.* **2016**, *44*, e91. [CrossRef] [PubMed]

32. Li, Y.H.; Zhang, G.; Cui, Q. PPUS: A web server to predict PUS-specific pseudouridine sites. *Bioinformatics* **2015**, *31*, 3362–3364. [CrossRef] [PubMed]

33. Xiang, S.; Yan, Z.; Liu, K.; Zhang, Y.; Sun, Z. AthMethPre: A web server for the prediction and query of mRNA m^{6}A sites in *Arabidopsis thaliana*. *Mol. Biosyst.* **2016**, *12*, 3333–3337. [CrossRef] [PubMed]

34. Zhang, M.; Sun, J.W.; Liu, Z.; Ren, M.W.; Shen, H.B.; Yu, D.J. Improving N^6-methyladenosine site prediction with heuristic selection of nucleotide physical-chemical properties. *Anal. Biochem.* **2016**, *508*, 104–113. [CrossRef] [PubMed]

35. Panwar, B.; Raghava, G.P. Prediction of uridine modifications in tRNA sequences. *BMC Bioinform.* **2014**, *15*, 326. [CrossRef] [PubMed]

36. Li, G.Q.; Liu, Z.; Shen, H.B.; Yu, D.J. TargetM^{6}A: Identifying M^6-methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. *IEEE Trans. Nanobiosci.* **2016**. [CrossRef] [PubMed]

37. Dominissini, D.; Moshitch-Moshkovitz, S.; Schwartz, S.; Salmon-Divon, M.; Ungar, L.; Osenberg, S.; Cesarkas, K.; Jacob-Hirsch, J.; Amariglio, N.; Kupiec, M.; et al. Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. *Nature* **2012**, *485*, 201–206. [CrossRef] [PubMed]

38. Schwartz, S.; Agarwala, S.D.; Mumbach, M.R.; Jovanovic, M.; Mertins, P.; Shishkin, A.; Tabach, Y.; Mikkelsen, T.S.; Satija, R.; Ruvkun, G.; et al. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* **2013**, *155*, 1409–1421. [CrossRef] [PubMed]

39. Roundtree, I.A.; He, C. RNA epigenetics—Chemical messages for posttranscriptional gene regulation. *Curr. Opin. Chem. Biol.* **2016**, *30*, 46–51. [CrossRef] [PubMed]

40. Hoernes, T.P.; Erlacher, M.D. Translating the epitranscriptome. *Wiley Interdiscip. Rev. RNA* **2017**, *8*, e1375. [CrossRef] [PubMed]

41. Chen, W.; Lei, T.Y.; Jin, D.C.; Lin, H.; Chou, K.C. PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. *Analy. Biochem.* **2014**, *456*, 53–60. [CrossRef] [PubMed]

42. Chen, W.; Lin, H.; Chou, K.C. Pseudo nucleotide composition or pseKNC: An effective formulation for analyzing genomic sequences. *Mol. BioSyst.* **2015**, *11*, 2620–2634. [CrossRef] [PubMed]

43. Charette, M.; Gray, M.W. Pseudouridine in RNA: What, where, how, and why. *IUBMB Life* **2000**, *49*, 341–351. [PubMed]

44. Dunn, D.B. The occurrence of 1-methyladenine in ribonucleic acid. *Biochim. Biophys. Acta* **1961**, *46*, 198–200. [CrossRef]

45. Chen, W.; Xing, P.W.; Zou, Q. Detecting N^6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. *Sci. Rep.* **2017**. [CrossRef]

46. Hoernes, T.P.; Huttenhofer, A.; Erlacher, M.D. mRNA modifications: Dynamic regulators of gene expression? *RNA Biol.* **2016**, *13*, 760–765. [CrossRef] [PubMed]