Cell Genomics

CellPress
OPEN ACCESS

Preview

AI techniques have facilitated the understanding of epitranscriptome distribution

Daiyun Huang,[1,2,3] Jia Meng,[4,5] and Kunqi Chen[1,6,7,*]
[1]Key Laboratory of Gastrointestinal Cancer (Fujian Medical University), Ministry of Education, School of Basic Medical Sciences, Fuzhou 350122, China
[2]Wisdom Lake Academy of Pharmacy, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[3]School of Life Sciences, Fudan University, Shanghai 200092, China
[4]Department of Biosciences and Bioinformatics, Center for Intelligent RNA Therapeutics, Suzhou Key Laboratory of Cancer Biology and Chronic Diseases, School of Science, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[5]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L7 8TX, UK
[6]Department of Medical Microbiology, Fujian Key Laboratory of Tumor Microbiology, School of Basic Medical Sciences, Fujian Medical University, Fuzhou 350122, China
[7]School of Medical Technology and Engineering, Fujian Medical University, Fuzhou 350122, China
*Correspondence: kunqi.chen@fjmu.edu.cn

https://doi.org/10.1016/j.xgen.2024.100718

N^6-methyladenosine (m6A), the most prevalent internal mRNA modification in higher eukaryotes, plays diverse roles in cellular regulation. By incorporating both sequence- and genome-derived features, Fan et al.[1] designed a novel Transformer-BiGRU framework that achieves superior performance in computational m6A identification, thus demonstrating the potential of AI in genomic studies.

N^6-methyladenosine (m6A), the most prevalent internal modification in eukaryotic mRNA, constitutes a regulatory network extensively involved in a wide array of biological processes. It alters various aspects of RNA biology, including splicing, polyadenylation, and translation, and its dysregulation has been linked to cancer progression. Functional outcomes of m6A modifications can differ across various transcripts, regions within the same transcript, and distinct cell types. Therefore, comprehensive mapping of m6A to elucidate its multitude of roles is crucial.

The most widely used m6A-mapping method relies on antibody-mediated immunoprecipitation, which has recently been shown to produce a large portion of systematic false positives.[2] Although additional solutions have been proposed, certain limitations persist, including input amounts, antibody specificities, efficiencies, predefined sequence contexts, and complicated workflows.[3] Moreover, around 5%–10% of m6A sites identified as biologically significant exhibit dynamic behavior upon varying conditions.[4] Collectively, these challenges in profiling techniques highlight the necessity for computational methods as a complement.

To date, dozens of approaches have contributed to accurate m6A prediction, ranging from early machine-learning-based predictors like SRAMP[5] to the latest models that leverage cutting-edge AI techniques. WHISTLE[6] was the first model to introduce genome-derived features, DeepPromise[7] comprehensively evaluated major AI models at the time of its proposal, and Geo2Vec[8] enabled transcript isoform-aware learning of m6A. Despite these advancements, there is still room for improvement in accuracy, highlighting the need for a more comprehensive model.

In this issue of *Cell Genomics*, Fan et al.[1] present a novel deep learning framework for high-accuracy mammalian m6A site prediction (Figure 1). This work introduces a combined framework of a Transformer architecture and a bidirectional gated recurrent unit (BiGRU), named deepSRAMP, to identify m6A sites using both sequence-derived and genome-derived features. Comparative studies across multiple benchmark datasets showed that deepSRAMP outperformed state-of-the-art m6A predictors, achieving an average 43.9% increase in area under the precision curve (AUPRC), a metric particularly useful for evaluating the model's precision in identifying minor classes. The model's promising performance was also evident when extended to 38 mammalian tissues and cells. These achievements are primarily attributed to three main contributions:

(1) Feature encodings: the authors proposed a new combination of sequence-based and genome-based encodings as the inputs of deep learning. For sequence-based encoding, they combined one-hot encoding, which converts each nucleotide into a unique vector to facilitate motif learning through subsequent convolutional layers, with a learnable embedding layer that maps nucleotides into higher-dimensional vectors to capture richer features. For genome-based encoding, they introduced a nucleotide-level feature matrix that incorporates the absolute and relative positions of each nucleotide on the transcript, as well as the type and rank of the exon it resides in, including explicit encoding of the relative distances to the start codon, stop codon, and splicing sites. Both encoding methods, when used individually, outperformed the sequence-based machine-learning model SRAMP and, when combined through a multi-input framework, showed a

© 2024 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

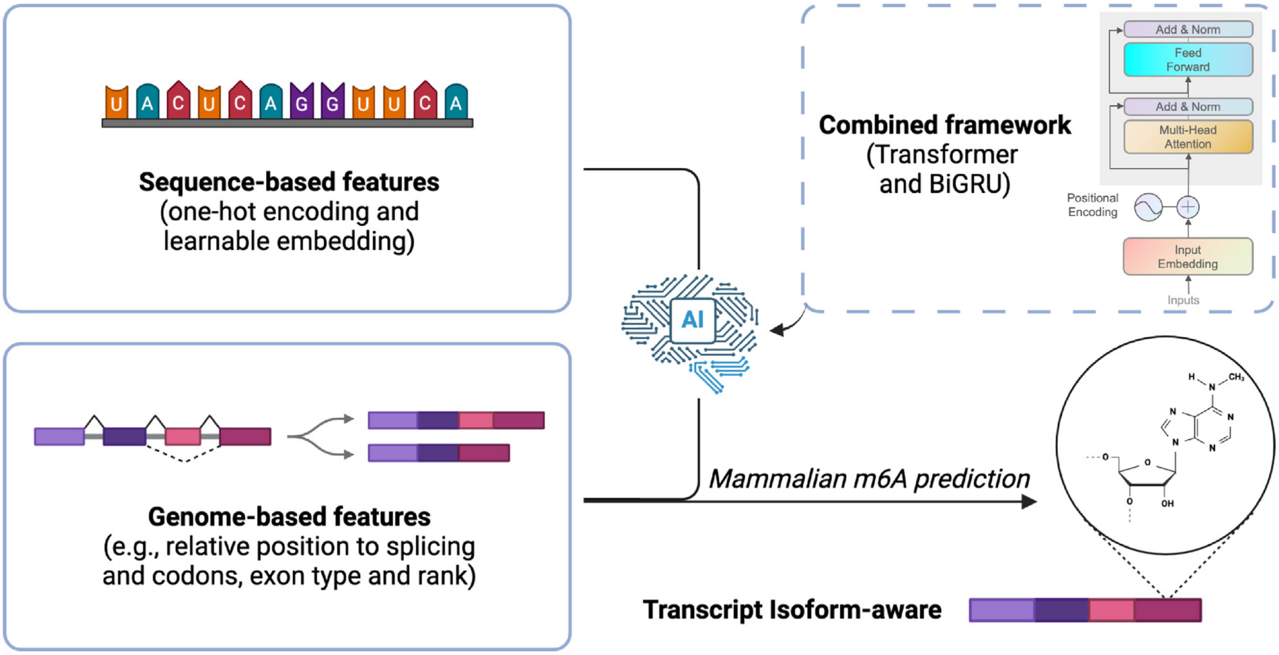

Figure 1. The deepSRAMP framework proposed by Fan et al.

significant additional improvement, indicating their synergistic effect in representing different sequence features in various regions.

(2) Novel neural network framework: to effectively capture the essential mappings from richly encoded biological information to m6A sites, a powerful neural network is required. Existing works, such as DeepPromise, have comprehensively optimized network frameworks but only for sequence inputs. WHISTLE, on the other hand, considered both sequence-based and genome-based features but used classical machine-learning methods as classifiers. In this work, Fan et al. leveraged the powerful Transformer architecture, which has proven effective in various RNA-related prediction tasks but has not been fully explored in m6A prediction. Additionally, an ablation study demonstrated that combining the Transformer encoder with a BiGRU module significantly enhances model performance. This combined Transformer-BiGRU framework, along with carefully curated features, resulted in a substantial increase in AUPRC from 0.375 in the WHISTLE model to 0.814 in deepSRAMP.

(3) Isoform-aware learning: incorporating genome-derived features not only introduces biologically meaningful information into the model but also enables the modeling of different isoforms for the same transcript, thereby advancing transcript isoform-level m6A prediction. While Geo2-Vec generates a concise feature matrix to describe the entire transcript, the genome-derived features used by deepSRAMP provide more detailed context about the specific isoform where the site of interest may be located. These higher-resolution descriptors, combined with learnable attention weights, show promise in revealing differential m6A patterns among different isoforms. Specifically, the authors used the *CALM3* gene as a case study and demonstrated that the predicted m6A distributions across isoforms can be explained by known alternative polyadenylation processes.

The success of deepSRAMP highlights the importance of feature encodings in biological prediction tasks, as well as the necessity of a congruent neural network framework. The upper limit of a machine-learning model is determined by the essential mappings hidden within the data. A powerful model can better capture these mappings, leading to a higher degree of fitting, while the quality of features determines how much of this mapping can be learned from the data. Since m6A is known to enrich at the last exon of a transcript and also exhibits functions in other regions, incorporating genome-derived features enables deepSRAMP to learn different m6A patterns in various regions of the same transcript or even different transcript isoforms. Such a domain knowledge-induced feature design method may generalize to other prediction tasks in genomics, leading to better capture of biological insights.

While deepSRAMP has achieved promising performance and a user-friendly webserver has been provided to facilitate its widespread use, key research questions remain. Current models primarily define m6A identification as a binary classification task. However, a high-accuracy quantification model that can reveal the methylation level of each site under different tissues and cells is highly desirable. This presents significant challenges, as such a model would need to consider not only sequence- and genome-derived features but also other factors, such as the expression status of m6A-related genes (e.g., METTL3/14), to accurately describe the cellular context. Addressing these challenges is essential for understanding the various functional outcomes of m6A, particularly at those dynamic sites.

In summary, the work by Fan et al. presents a novel m6A prediction framework, deepSRAMP, which substantially outperforms existing methods. This high-accuracy m6A prediction can complement experimental methods to better study the distribution of functional m6A, particularly benefiting from its potential for transcript isoform-level prediction. Additionally, the success of this combined deep learning framework highlights the promising avenues for leveraging AI techniques in genomic studies.

ACKNOWLEDGMENTS

The authors received funding from the National Natural Science Foundation of China (32100519 and 32300552), Fujian Research and Training Grants for Young and Middle-aged Leaders in Healthcare (2023WJWZQN02), the Jiangsu Science and Technology Program (BK20230279), and the Fujian Medical University High-level Talent Research Startup Funding Project (XRCZX2021019).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Fan, R., Cui, C., Kang, B., Chang, Z., Wang, G., and Cui, Q. (2024). A combined deep learning framework for mammalian m6A site prediction. Cell Genom. *4*, 100697. https://doi.org/10.1016/j.xgen.2024.100697.
2. Zhang, Z., Chen, T., Chen, H.X., Xie, Y.Y., Chen, L.Q., Zhao, Y.L., Liu, B.D., Jin, L., Zhang, W., Liu, C., et al. (2021). Systematic calibration of epitranscriptomic maps using a synthetic modification-free RNA library. Nat. Methods *18*, 1213–1222. https://doi.org/10.1038/s41592-021-01280-7.
3. Xiao, Y.L., Liu, S., Ge, R., Wu, Y., He, C., Chen, M., and Tang, W. (2023). Transcriptome-wide profiling and quantification of N^6-methyladenosine by enzyme-assisted adenosine deamination. Nat. Biotechnol. *41*, 993–1003. https://doi.org/10.1038/s41587-022-01587-6.
4. Liu, C., Sun, H., Yi, Y., Shen, W., Li, K., Xiao, Y., Li, F., Li, Y., Hou, Y., Lu, B., et al. (2023). Absolute quantification of single-base m^{6}A methylation in the mammalian transcriptome using GLORI. Nat. Biotechnol. *41*, 355–366. https://doi.org/10.1038/s41587-022-01487-9.
5. 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. https://doi.org/10.1093/nar/gkw104.
6. Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., Su, J., de Magalhães, J.P., Rigden, D.J., and Meng, J. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res. *47*, e41. https://doi.org/10.1093/nar/gkz074.
7. Chen, Z., Zhao, P., Li, F., Wang, Y., Smith, A.I., Webb, G.I., Akutsu, T., Baggag, A., Bensmail, H., and Song, J. (2020). Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief. Bioinform. *21*, 1676–1696. https://doi.org/10.1093/bib/bbz112.
8. Huang, D., Chen, K., Song, B., Wei, Z., Su, J., Coenen, F., de Magalhães, J.P., Rigden, D.J., and Meng, J. (2022). Geographic encoding of transcripts enabled high-accuracy and isoform-aware deep learning of RNA methylation. Nucleic Acids Res. *50*, 10290–10310. https://doi.org/10.1093/nar/gkac830.