Molecular Therapy
Nucleic Acids
Commentary

How well does a data-driven prediction method distinguish dihydrouridine from tRNA and mRNA?

Shaherin Basith[1] and Balachandran Manavalan[2]

https://doi.org/10.1016/j.omtn.2023.02.026

RNA polymers undergo extensive modifications following transcription by enzymes called “RNA modification enzymes.” More than 170 distinct post-transcriptional modifications have been identified and this number has been constantly increasing. Diverse types of RNA including ribosomal (rRNA), transfer (tRNA), messenger (mRNA), and long non-coding RNA undergo this type of post-transcriptional modifications. The enthusiasm for RNA modification research has been rekindled as more evidence demonstrates that it plays a significant role in the regulation of gene expression.[1] One of the most abundant modified bases in tRNA, dihydrouridine (D), has just entered the world of mRNA modifications, demonstrating critical physiological functions in cell growth.[2] D is a modified uridine nucleotide that is catalyzed by D synthase (DUS) enzyme and the second most prevalent modification in tRNAs. Generally, epitranscriptome profiling approaches are expensive, time-consuming, and labor-intensive. However, computational methods could offer a rapid, efficient, and affordable alternative to conventional experimental approaches. Several computational methods have been developed to predict D sites, but these tools are all trained on tRNAs, and their generalization on mRNAs is obscure. In this issue, Wang et al.,[3] develop the first computational tool, DPred, for predicting D modification sites on mRNAs in yeast using local self-attention and a convolutional framework. Their paper highlights the differences in the D site sequence motifs among mRNAs and tRNAs, suggesting putative variations observed in the formation mechanisms of D on diverse RNA types. The authors emphasize that mixed predictions based on tRNA and mRNA datasets containing D are ineffective, and that their predictions should be clearly differentiated. A major limitation of their study is that the predictions for multiple species were not conducted due to insufficient data.

The authors constructed D-containing tRNA and mRNA sequences from the literature and RMBase 2.0 (Figure 1). Due to the absence of experimentally verified unmodified sequences, they randomly selected the positive D site transcripts and considered them as negative samples. A balanced dataset was constructed, where 80% of the samples were selected for developing the prediction model, while the remaining samples were selected to test the model transferability. Using the training dataset, they assessed four feature encodings, including one hot encoding (OH), nucleotide chemical properties (NCP), nucleotide density (ND), and electron-ion interaction potential (EIIP). Instead of individual encodings, they generated four hybrid feature sets, including OH_ND, OH_EIIP, NCP_ND, and NCP_EIIP. These were trained using a deep neural network that comprises an additive local self-attention and a convolutional neural network (CNN) layer. In contrast to global attention, local attention considers only a subset of states when computing attention weights.[4] As a result of these constraints, it is easier to implement and train, particularly when dealing with smaller datasets. Among these features, NCP_ND achieved the highest performance with an area under the receiver operating curve of 0.917 and 0.903 during training and independent evaluations, respectively. In the same study, DPred outperformed four conventional machine learning classifiers, including random forest, logistic regression, extreme gradient boosting, and support vector machine. Furthermore, the authors demonstrated that the proposed DPred could accurately predict D-related tRNA and mRNA across a vast range of species, including humans and mice.

The researchers investigated whether a D-related tRNA-trained model can be applied to mRNA datasets or vice versa. Considering the results, it appears that data-driven D-prediction tools based on mRNA and tRNA cannot be used to predict each other, but should be clearly distinguished due to highly distinct sequence signatures around D. Nevertheless, this also raises concerns regarding the limitations of sequence-based computational models.[5] The data-driven model generally needs to be retrained using experimental profiles based on the new condition type, which may not always be available. To improve the generalization ability of the model and further increase our understanding of D sites, other features, such as secondary structures and cell types, could be incorporated to the predictive model in the future. Another major limitation of their study to be addressed is the lack of multi-species prediction due to the limitation of high-quality D epitranscriptome data. Overall, their study provides putative insights into the study of D in the transcriptome from two different perspectives: First, CNN and local self-attention algorithms were utilized to provide the first data-driven predictive model that can predict D modification on mRNAs. Next, it revealed that mRNAs and tRNAs have distinct nucleotide preferences around D

[1]Department of Physiology, Ajou University School of Medicine, Suwon 16499, Republic of Korea; [2]Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon 16419, Republic of Korea

Correspondence: Balachandran Manavalan, Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon 16419, Republic of Korea.
E-mail: bala2022@skku.edu

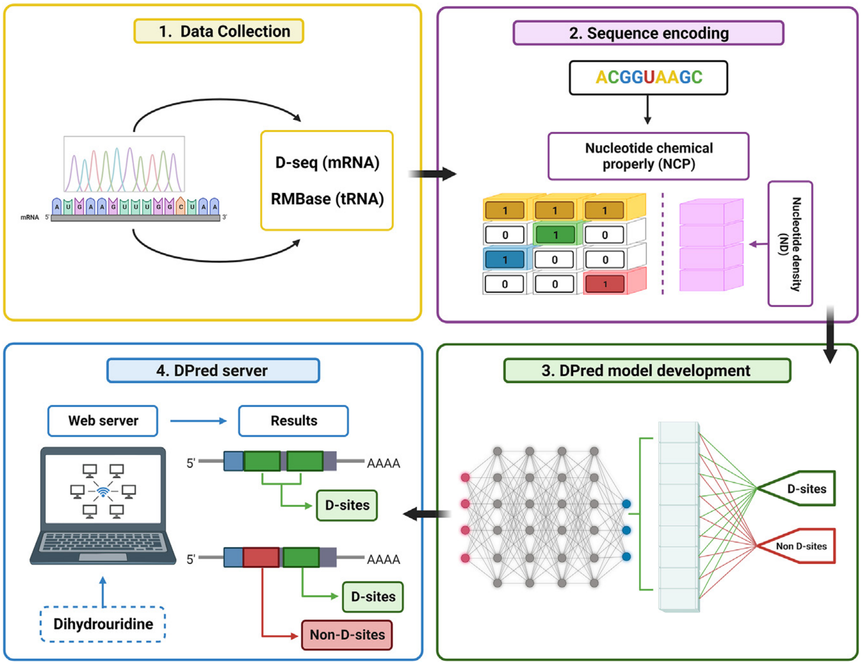

Figure 1. A general framework for DPred, a dihydrouridine prediction tool

The tool comprises four steps: data collection from the existing database, evaluation of diverse sequence encodings and selection of the appropriate one, model training using CNNs, and construction of the web server.

sites, implying that their mechanisms of formation and functions may differ.

In the coming years, the pace of diverse RNA epigenetic modification site characterization by experimental approaches is likely to increase exponentially. Consequently, sequence-based computational approaches, as presented by Wang et al., will be crucial for comprehending biological functions. Their article has an intriguing finding that the sequence signature around D sites on mRNAs differs significantly from those on tRNAs. However, future studies are essential to examine the detailed mechanisms behind the formation of modified mRNAs and their biological implications. In addition, researchers should not overlook rRNAs, as DUS enzymes do not appear to be involved in D synthesis, suggesting that another enzyme system might be responsible for its biosynthesis.[1]

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Brégeon, D., Pecqueur, L., Toubdji, S., Sudol, C., Lombard, M., Fontecave, M., de Crécy-Lagard, V., Motorin, Y., Helm, M., and Hamdane, D. (2022). Dihydrouridine in the transcriptome: new life for this ancient RNA chemical modification. ACS Chem. Biol. *17*, 1638–1657.

2. Finet, O., Yague-Sanz, C., Krüger, L.K., Tran, P., Migeot, V., Louski, M., Nevers, A., Rougemaille, M., Sun, J., Ernst, F.G.M., et al. (2022). Transcription-wide mapping of dihydrouridine reveals that mRNA dihydrouridylation is required for meiotic chromosome segregation. Mol. Cell *82*, 404–419.e9.

3. Wang, Y., Wang, X., Cui, X., Meng, J., and Rong, R. (2023). Self-attention enabled deep learning of dihydrouridine (D) modification on mRNAs unveiled a distinct sequence signature from tRNAs. Mol. Ther. Nucleic Acids *31*, 411–420.

4. Soydaner, D. (2022). Attention mechanism in neural networks: where it comes and where it goes. Neural Comput. Appl. *34*, 13371–13385.

5. Li, Z., Gao, E., Zhou, J., Han, W., Xu, X., and Gao, X. (2023). Applications of deep learning in understanding gene regulation. Cell Rep. Methods *3*, 100384.