Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Predicting miRNA-mRNA using Sequences and RNA Secondary Structure

Haiyan Hu[1*], Sanjeda Sara Jennifer[1]

[1]Department of Computer Science and Engineering

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Understanding the complex interactions between microRNAs (miRNAs), their isoforms (isomiRs), and messenger RNAs (mRNAs) is critical for decoding post-transcriptional gene regulation. However, accurate prediction of miRNA/isomiR–mRNA binding sites at base-pair resolution remains a major challenge due to the diversity of isomiRs and the dynamic nature of RNA structure. This work proposes a deep learning framework that first works with the miRNA/isomiR–mRNA interaction features directly from sequence data.
Afterwards, the next phase is to get and incorporate RNA secondary structure into the model to enhance predictive accuracy and capture the biophysical context of binding. By integrating self-attention and convolutional mechanisms with structural features, the model learns context-aware representations that reflect both sequence and structural determinants of interaction. This approach advances the precision of interaction prediction while offering biological insights into the structural underpinnings of post-transcriptional regulation.

Results: The results achieved shows promise that this is a good research direction with good potential to revolutionize the healthcare domain and understanding of the gene expressions, and is worth investigating
Availability: The data and the relevant code are available at https://github.com/sanjeda-sara/miRNA_mRNA_interaction_prediction.
Contact: sanjedasara.jennifer@ucf.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

MicroRNAs (miRNAs) are short (~22 nucleotides), single-stranded non-coding RNAs that regulate gene expression post-transcriptionally by binding to complementary sequences in messenger RNAs (mRNAs), leading to mRNA degradation or translational repression. They play pivotal roles in diverse biological processes and disease mechanisms, including cancer, neurological disorders, and immune responses [1]. The identification of precise miRNA–mRNA interaction sites is thus essential to understanding gene regulation and developing RNA-based therapeutics.

Adding to the complexity of this regulatory layer is the presence of miRNA isoforms, or isomiRs—variants of canonical miRNAs generated by imprecise cleavage, nucleotide additions or deletions, and post-transcriptional modifications. These isomiRs often differ in sequence from their parental miRNAs, particularly at the 5′ and 3′ ends, which can shift seed regions and alter target specificity. Accumulating evidence suggests that isomiRs are not sequencing artifacts but are biologically functional and expressed in a condition-specific and tissue-specific manner. This diversity in miRNA and isomiR expression under different physiological conditions underscores the need to model their interactions with mRNAs comprehensively.

RNA structure plays a critical role in determining the accessibility of binding sites and the stability of RNA molecules. Secondary structures, such as hairpins and loops, can either shield or expose target regions, thereby influencing miRNA/isomiR binding and regulatory efficacy [2]. Although thermodynamic RNA folding algorithms can predict these structures, they often operate in isolation from learning-based methods and lack the ability to model intricate interactions across diverse biological contexts.

To address these challenges, this paper proposes a deep learning framework that integrates sequence-based learning of miRNA and isomiR interactions with mRNAs and incorporates RNA secondary structure to enhance prediction accuracy and biological relevance. This work leverages convolutional and self-attention mechanisms to capture both local sequence motifs and long-range structural dependencies, enabling the identification of functional interaction sites at nucleotide resolution. By incorporating predicted RNA secondary structure as an additional input layer, our framework models not only sequence complementarity but also the biophysical landscape that governs miRNA targeting.

2 Methods

The methodology for this experimentation was divided into two main tasks:

(1) Work with the Sequences (miRNA and mRNA) to get the initial prediction, target interaction, result and visualize how it is performing [3].

(2) Incorporate the RNA secondary structure with the same ultimate goal of predicting the interaction and going to the interpretation [4].

2.1 Dataset Description

The dataset used for this study is a novel dataset that has been created and curated by the “hulab” [https://hulab.ucf.edu] at the University of Central Florida. A portion of the dataset has been used for this research and a small snapshot is presented in Figure 01:

miRNA_isomiR	mRNA	label
GAGTTCTACAGTCCGAC	CAAAAGAATTTTAAAGTACCTCGGTTTATAGGGATTTACAATTCAGTAGAAA	1
GAGTTCTACAGTCCGAC	ACGGATGTAATAAATTTTTAAGACAATTGCTTTTATTATCATGATATTT	1
ACCCCACTCCTGGTAC	CTGCATAATTTGTGGTAATGGGGGACTGTGTTCCTGCTTTTACCTGGTTACAAAAAACA	1
ACCCCACTCCTGGTACCA	AAACCCAGGAGGCGGAGGTTGCAGTGAGCTGAGATCATGCCATTGCACTCCAGTCTGGGC	0
AGCGGCCCCCGG	CGGCGCGCCGGGCCCGGGTCCCCCACCCGCAGCGCCCCGCCGGGC	0
CCTCTGTCTCCC	GGTGCAATCTTGGCTCACTGCAACCTCCACCTCCAGTTCAAGCAGGTCTCAT	0

Figure 01: Dataset Snippet

In this dataset, the labels are present where label=1 means that the miRNA and mRNA interacted, whereas label=0 means that they did not. While preprocessing, sequences were encoded using a fixed-length representation of up to 50 nucleotides. Each nucleotide (A, C, G, T) was mapped to an integer token (A=1, C=2, G=3, T=4), with padding (0) applied as needed to cover the difference in the length, as shown in Figure 02.

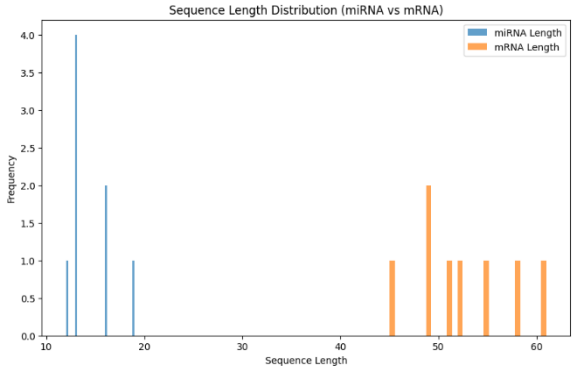

Figure 02: Sequence length Distribution

2.2 Implementation Details

In order to run the experiments, Google Colab was used, which has 15GB RAM and 122.6 Disk Space.

The code was implemented in two phases: the first phases were to clean the dataset and complete the data analysis part, and then with the processed data, the second phase was implemented where the interaction prediction and processing for the secondary structure was done.

Importantly, multiple tools, such as ViennaRNA [5], RNAfold, RNAformer, Scanfold, IFold, PRIMITI [6], BiBiServ-2, RNAeval were tried out. All these tools, webpages and work were done to deal with the RNA secondary structure [7][8] and its different parameters, such as minimum free energy (MFE), Binding Energy etc. The tools have many more different features to work with. The visualizations that have been done to get the hairpin loop structure of the dot-bracket notation were tried out using all these tools.

To predict interactions between miRNAs/isomiRs and mRNAs, a Transformer-based deep learning framework was implemented using PyTorch, with support from PyTorch Geometric and ViennaRNA for future structure-aware extensions. After the equences were encoded, they were split into training and test sets using an 80/20 stratified split to preserve class balance. Data were wrapped into PyTorch Dataset and DataLoader objects to enable batch training.

The model is a dual-sequence Transformer encoder that processes miRNA/isomiR and mRNA sequences independently, before combining the learned representations for classification. Each sequence is first embedded into a 32-dimensional trainable vector space. Both miRNA and mRNA sequences are then passed through separate Transformer encoders, consisting of two layers, four attention heads, and 64 hidden units, with positional attention applied to capture sequence dependencies. The resulting token embeddings are averaged to generate fixed-length representations for each sequence. These representations are concatenated and passed through a feedforward neural network with ReLU activation, followed by a sigmoid function to output the probability of interaction. The model is trained using Binary Cross-Entropy (BCE) loss and optimized with the AdamW optimizer (learning rate = 0.001), running for 10 epochs with a batch size of 64. Training progress is monitored through batch-wise and epoch-wise loss values.

3 Results

This section contains the results achieved and a detailed analysis has been provided using the different sub-sections.

3.1 Structural Characterization of mRNA and Binding Site Accessibility

To investigate potential binding interactions between microRNAs (miRNAs) and target messenger RNAs (mRNAs), we first analyzed the secondary structure of candidate mRNA sequences using RNAfold from the ViennaRNA Package. The predicted secondary structure was represented using dot-bracket notation, where paired and unpaired bases are denoted by matching parentheses and dots, respectively, as shown in Table 01.

Table 1. mRNA sequence and its dot-bracket notation

mRNA Sequence	AGAAGGAGTGAGAAGATTTGGCCGAGGGATAGAGTG CAAGTAGAGAGAGGAGCTTGTCGCCCGGGT
Secondary Structure	.....(((((((((.....((((((.....)))))).....)))))))...)))

This structure was visualized using 2D colored RNA plots, which clearly revealed canonical structural motifs such as hairpin loops, internal loops, and stem regions. Figure 03 and 04 have been drawn using the same sequence, but they folded differently when the option was set to MFE and BE revealing the different sites having binding affinity.

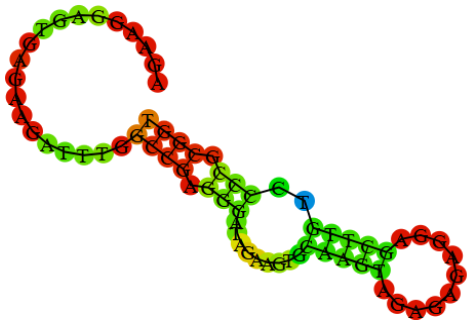

Figure 03: Secondary structure fold with minimum free energy

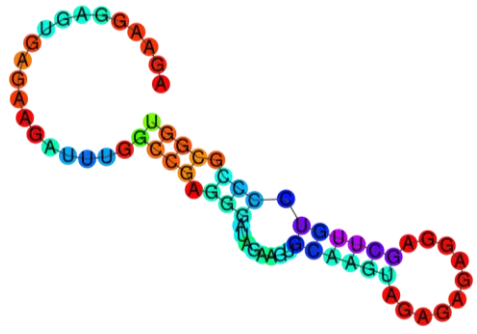

Figure 04: Secondary Structure fold with binding energy

The accessibility of target regions within these structures was further evaluated using a mountain plot as shown in the figure 05, which depicts the cumulative base-pairing height along the sequence, and an entropy profile, indicating positional uncertainty in the predicted structure.

The entropy plot revealed that specific regions within the mRNA have high structural entropy, suggestive of single-strandedness and flexibility—conditions that are favorable for miRNA binding. These unstructured regions are particularly important for miRNA targeting, as base pairing requires physical accessibility of the target nucleotides.

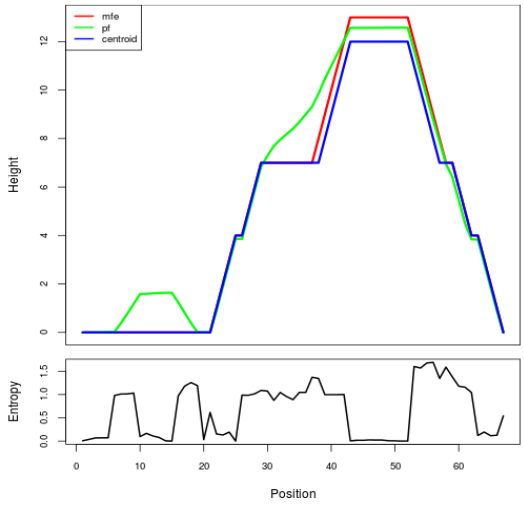

Figure 05: Mountain Plot

3.2 Integration of miRNA Structural Simplicity and Graph-Based Representations

In contrast to the relatively complex mRNA structures, the corresponding miRNAs typically exhibited minimal or no secondary structure, as confirmed by their dot-bracket representations (e.g., "....."). This structural simplicity aligns with biological expectations, given that miRNAs are short (22 nt), single-stranded RNA molecules that act as guides for RNA-induced silencing complexes (RISC) and remain largely unstructured to facilitate binding.

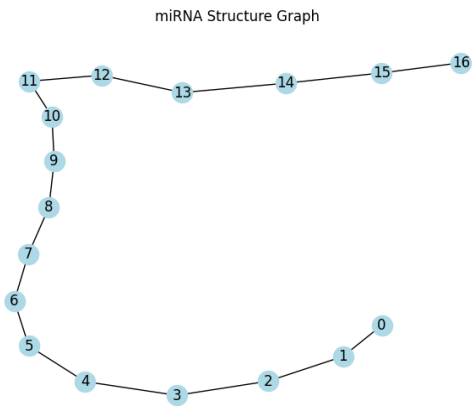

Figure 06: Graphical Representation

To further model the secondary structure in a form amenable to machine learning, graph representations of the RNAs were constructed, as shown in figure 06. Nodes represent nucleotides, and edges are formed based on both backbone connectivity and base-pairing interactions inferred from the dot-bracket notation. This graph formulation enables the future use of Graph Neural Networks (GNNs) to encode structural features alongside sequence context, providing a powerful, structure-aware input for predictive modeling tasks.

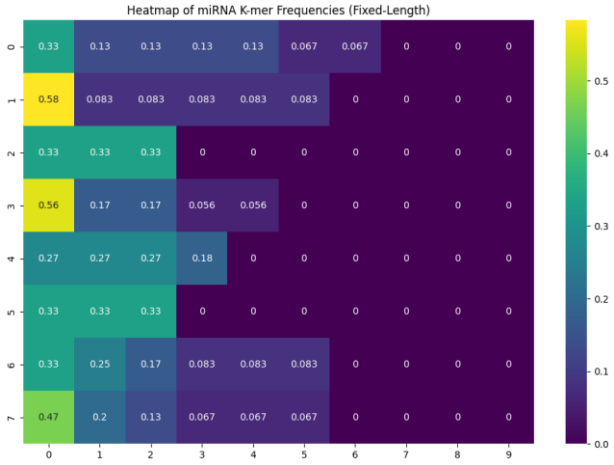

Figure 07: Heatmap of miRNA K-mer Frequencies

After training, the model demonstrated moderate performance in distinguishing between interacting and non-interacting miRNA/isomiR–mRNA pairs. The model was able to capture patterns in the sequence data, leading to average predictions. The training loss curve showed a steady decrease across epochs, indicating stable convergence without overfitting. t-SNE embedding visualization of the latent representations revealed no clear clustering of miRNA/isomiR–mRNA pairs based on their interaction labels, without any well-separated clusters in the two-dimensional space, as shown in figure 08.

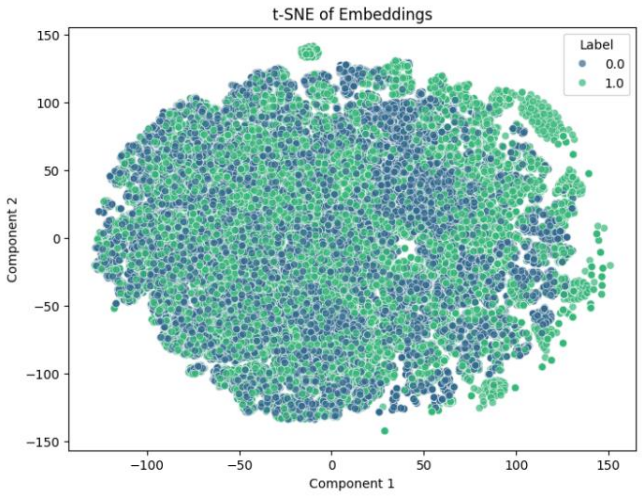

Figure 08: t-SNE representation of the Embeddings

The model was evaluated using key metrics. On the test set, it achieved moderate precision, recall, F1-score, and accuracy, in detecting miRNA/isomiR–mRNA interactions. On the ROC Curve, the model showed above average discriminative power with an AUC 0.75, as shown in figure 09. The accuracy achieved for this was 66.8%.

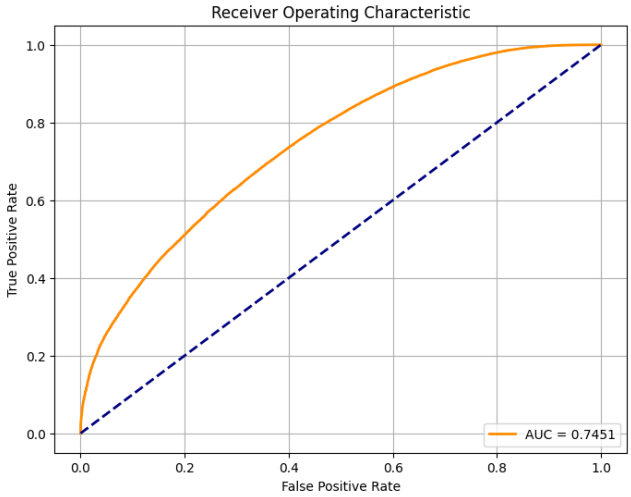

Figure 09: ROC Curve

The prediction probability distribution, as shown in figure 10 shows that while the model can confidently classify some miRNA–mRNA interaction pairs (as seen by the left-side peak near 0.0 for interactions), a large portion of the predictions cluster around the 0.5 range, indicating considerable uncertainty. This central peak suggests that the model struggles to confidently differentiate between interacting and non-interacting pairs in many cases.

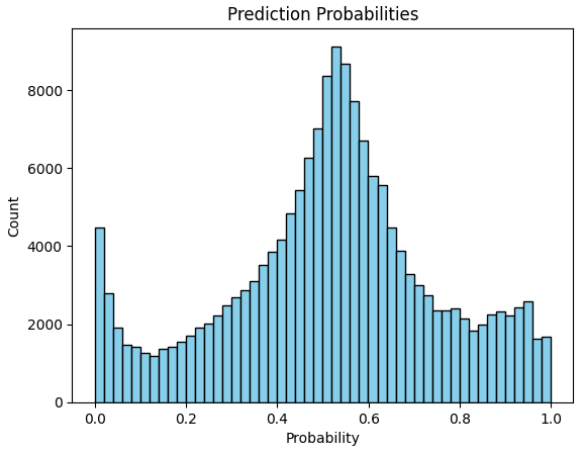

Figure 10: Probability distribution of the predictions

Table 2. Evaluation Matrices

	Precision	Recall	F1-score
Label 1 (interaction)	0.647	0.739	0.689
Label 0 (no interaction)	0.696	0.596	0.642

4 Conclusion and Future Work

This study presents an integrative framework combining RNA sequence and secondary structure to characterize miRNA–mRNA interactions. Visualization techniques such as the dot-bracket notation, entropy profiles, mountain plots, and base pairing matrices highlight structurally accessible and energetically favorable target regions. The Transformer-based model utilizing self-attention and independent sequence embeddings distinguishes interacting from non-interacting pairs, supported by moderate precision, recall, F1-score, and AUC values. Interpretability analyses reveal biologically relevant sequence features influencing predictions. RNAhybrid-derived mfe values and structural alignments further validate interaction plausibility. Graph-based representation of secondary structures facilitates future integration into graph neural networks. This approach establishes a strong foundation for accurate prediction of functional interactions, accounting for isomiRs, non-canonical binding, and structure-driven targeting.

As for future work, there are a lot of areas to explore, experiment and implement. Enhancing the model's interpretability and biological relevance by incorporating RNA secondary structure into the learning framework is one of them. RNA secondary structures are known to influence the stability and binding affinity of RNA molecules, which may provide additional context for understanding interaction dynamics. Additionally, exploring graph-based models could further improve the representation of RNA sequences, allowing for the incorporation of spatial relationships and more complex dependencies between nucleotides. A hybrid Transformer-GNN architecture will be developed to capture both sequence and structural dependencies, validated across species and tissue types using datasets such as miRTarBase, TarBase, and CLASH. Finally, a web-based tool or Python API will be deployed to enable interactive prediction and visualization of miRNA–mRNA interactions, advancing the interpretability and accessibility of post-transcriptional regulatory modeling.

Acknowledgements

This project has been done under the guidance of Dr. Haiyan (Nancy) Hu, for the CAP5510 (Bioinformatics) course at the University of Central Florida. I would like to thank Dr Haiyan Hu for providing the data and for showing the research direction, without which, this would have been difficult.

Funding

This work has been supported by Dr Haiyan Hu and the University of Central Florida

Conflict of Interest: none declared.

References

Talukder, A., Zhang, W., Li, X. *et al.* A deep learning method for miRNA/isomiR target detection. *Sci Rep* **12**, 10618 (2022). https://doi.org/10.1038/s41598-022-14890-8

He, S., Gao, B., Sabnis, R., & Sun, Q. (2023). RNAdegformer: accurate prediction of mRNA degradation at nucleotide resolution with deep learning. *Briefings in bioinformatics*, *24*(1), bbac581. https://doi.org/10.1093/bib/bbac581

Caroline Diener, Andreas Keller, Eckart Meese, The miRNA–target interactions: An underestimated intricacy, *Nucleic Acids Research*, Volume 52, Issue 4, 28 February 2024, Pages 1544–1557, https://doi.org/10.1093/nar/gkad1142

Sato, K., Akiyama, M. & Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat Commun* **12**, 941 (2021). https://doi.org/10.1038/s41467-021-21194-4

Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C. et al. ViennaRNA Package 2.0. Algorithms Mol Biol 6, 26 (2011). https://doi.org/10.1186/1748-7188-6-26

Korawich, U., de Sá, A. G. C., Alavi, A., Pires, D. E. V., & Ascher, D. B. (2024). PRIMITI: A computational approach for accurate prediction of miRNA-target mRNA interaction. *Computational and Structural Biotechnology Journal*, *23*, 3030–3039. https://doi.org/10.1016/j.csbj.2024.06.030

Tan, C., Gao, Z., Cao, H., Chen, X., Wang, G., Wu, L., Xia, J., Zheng, J., & Li, S. Z. (2024). *Deciphering RNA secondary structure prediction: A probabilistic K-rook matching perspective*. arXiv. https://arxiv.org/abs/2212.14041

Singh, J., Hanson, J., Paliwal, K. *et al.* RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat Commun* **10**, 5407 (2019). https://doi.org/10.1038/s41467-019-13395-9