Exploring End-to-End Sequence-to-Sequence Ensemble Model for Predicting RNA Secondary Structure

A Major Qualifying Project (MQP) Report
Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements
for the Degree of Bachelor of Science in

Data Science

By:

Aukkawut Ammartayakun

Project Advisors:

Prof. Dmitry Korkin

Sponsored By:

InnoTech Precision Medicine

Date: April 2024

Abstract

The problem of predicting the secondary structure of RNA has been long studied to help understand the logic and use that for many applications like designing the primer to detect specific diseases. The challenge of RNA secondary structure prediction is its search space complexity. This work will explore how to use neural networks to approximate the pairing distribution or generate the sequence from the sequence in encoder-decoder format. The former approach uses an end-to-end pre-trained Graph Convolutional Network (GCN) and convolutional neural network (CNN). This work also uses the statistical, rule-based context-free model called CONTRAfold to improve the GCN model by providing the attention-like pairing distribution as an edge feature for the GCN. The models are trained with the bpRNA dataset and evaluated on the bpRNA and bpRNA-new datasets. The GCN end-to-end model results show that graph neural networks can learn with distribution distinct to the CONTRAfold. Moreover, its performance is comparable with the State-Of-The-Art models.

Acknowledgements

I am immensely grateful to Dr. Roya Khosravi-Far from InnoTech Precision Medicine for their comments and insights into the biological view of this work.

Also, special thanks to Palawat Busaranuvong for helping design, coding, and providing insightful discussion on this model.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Introduction

The study of RNA structures in nature reveals that these nucleic acids rarely exist as mere single strands. In reality, RNA molecules spontaneously fold upon themselves, forming intricate and complex secondary structures [15]. This phenomenon is pivotal in understanding the functional dynamics of RNA within biological systems. The secondary structure of RNA, characterized by its non-linear, higher-order formations, plays a critical role in various biological processes and applications, including disease diagnosis and therapeutic interventions.

The challenge of predicting RNA's secondary structure has been a significant focus of research for decades. Early works by [18, 14] introduced foundational models and computational methods that have since evolved through the contributions of numerous studies in the field. These investigations have not only enhanced our understanding of RNA structure but have also facilitated the development of practical applications. For instance, accurately predicting RNA structures assists in designing specific primers that are crucial for detecting particular diseases, thereby advancing diagnostic techniques.

Advancements in computational biology and the introduction of sophisticated algorithms have continually improved the accuracy and efficiency of RNA structure prediction. Techniques such as stochastic context-free grammars [10], graph theoretical models, and modern machine learning approaches like MXfold2 [11] have progressively refined our capability to predict RNA structures more reliably. These developments underscore the ongoing relevance and importance of this area of research in both theoretical and applied contexts, bridging fundamental biological insights with impactful medical innovations.

1.2 Problem Statement

The problem of RNA secondary structure prediction can be conceptualized in various ways, yet the central objective remains consistent across different approaches: transforming an RNA sequence into its corresponding secondary structure. Given an RNA sequence $s \in S$, our goal is to define a mapping $\mathcal{M} : S \to \mathcal{S}$ such that the resulting secondary structure $\mathcal{M}(s) \in \mathcal{S}$ is accurately produced.

In this project, we aim to develop a sequence-to-sequence transformation model, denoted as $\hat{\mathcal{M}}$, which serves as an estimator for $\mathcal{M}$. The primary objective of this model is to ensure that the discrepancy between the predicted structure $\hat{\mathcal{M}}(s)$ and the actual structure $\mathcal{M}(s)$ is minimized. This is quantitatively

expressed as ensuring that the error measure $e(\hat{\mathcal{M}}(s), \mathcal{M}(s))$ is less than or equal to a small positive constant ϵ for all $s \in \tilde{S} \subseteq S$. The set $\tilde{S}$ represents a subset of S that the model is specifically trained and tested on.

1.3 Background

1.3.1 RNA Secondary Structure Representation

In biological systems, single-stranded RNA (ssRNA) often does not remain linear but rather folds into complex configurations, forming secondary and higher-order structures. These configurations are crucial for the RNA's biological function and are fundamentally characterized by patterns of intramolecular base pairing. The secondary structure $\mathcal{M}(s)$, in particular, can be represented as a set of ordered pairs (a, b) where a and b are indices in the ssRNA sequence. Each pair denotes a hydrogen bond between the nucleotide at position a and the nucleotide at position b, effectively stabilizing the structure.

Using this formulation, the graph theoretical framework for RNA secondary structure can be used. Specifically, the simplest case of the RNA secondary structure is linear structure (i.e., no folding). This then can be represented using an adjacency matrix. In this scenario, if self-pairing is not counted (i.e., $(i, i) \notin \mathcal{M}(s)$ for sequence s), then the adjacency matrix for sequence s would be

$$A(s) = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \ddots & \vdots \\ 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}$$

where the entries for the super- and sub-diagonal are 1 and the rest are 0. For the more complex structure, the representation (a, b) is used. Specifically, $A_{a,b}(s) = 1$ if $(a, b) \in \mathcal{M}(s)$ and 0 otherwise.

1.3.2 Machine Learning Model for RNA Secondary Structure Prediction

There are many machine learning algorithms that can be used for this problem. It can be viewed as series of regression problem, series of classification problem, or generation problem. However, most of them share the same principle on their loss functions. In this project, the energy minimization and the probabilistic based model will be discussed.

1.3.2.1 Thermodynamic Energy Minimization Based Approach

At the core of the thermodynamic energy minimization approach to RNA secondary structure prediction is the foundational principle that a theoretically stable structure is one that minimizes free energy. This concept hinges on the premise that RNA folding is thermodynamically driven towards a conformational state characterized by the lowest possible free energy within a given set of conditions [15, 17]. The computational estimation of this energy, while theoretically straightforward, presents significant complexity due to the multifaceted nature of RNA molecular interactions [18, 11, 16].

To formalize this approach within a machine learning framework, we introduce a consistent estimator, denoted as $\hat{E}(f(s))$, for the free energy associated with a predicted secondary structure s produced by a predictive algorithm $f_\theta(\cdot)$. The algorithm's objective function, parameterized by θ, can be viewed as:

$$\theta = \arg \min_{\theta \in \Theta, s \in \tilde{S}} \hat{E}(f(s, x_s|\theta)) \tag{1}$$

Here, x_s represents external features associated with the sequence s, and $\tilde{S}$ is the sample space of sequences under consideration. Notably, the additive property of the energy function allows for the utilization of the expected energy across the sampled sequence space $\tilde{S}$ as a viable optimization criterion.

The energy estimator itself is typically constructed as a linear combination of energy contributions from various structural motifs such as hairpins, loops, stacked pairs, and junctions. Each motif contributes a quantifiable energy term to the overall free energy of the structure, guided by experimentally derived or theoretically inferred thermodynamic parameters [18].

Despite its strong theoretical argument, the implementation of thermodynamic energy minimization in practical predictive models is challenging. The intrinsic complexity of accurately modeling the diverse and dynamic nature of RNA interactions poses significant computational demands. Furthermore, there are issues related to biological stability; the assumption that the lowest energy structure corresponds to the biologically active form does not always hold true, given that cellular environments and molecular kinetics can lead to the adoption of suboptimal structures *in vivo* [15].

1.3.2.2 Probabilistic Based Approach

Diverging from energy-centric paradigms, the probabilistic-based approach represents a distinctive methodology for RNA secondary structure prediction. This paradigm leverages machine learning algorithms to generate probable RNA structures, utilizing alternative loss functions that are not directly tied to struc-

tural energy considerations. A paradigmatic example of this approach is the CONTRAfold model [3], which employs a stochastic context-free grammar (SCFG) akin to those found in natural language processing.

SCFGs provide a framework for modeling the probabilistic generation of structures, where the production rules are associated with probabilities that reflect the likelihood of certain structural motifs occurring[10]. In the context of RNA, these grammatical rules are not arbitrary but are designed to encapsulate the underlying biological principles, including aspects of thermodynamic stability and biological functionality.

The objective in a probabilistic-based approach is to optimize the parameters of the SCFG in such a manner that the resultant grammar captures the essence of both thermodynamic energy minimization and biological stability. The learning process involves adjusting the rule probabilities so that the grammar can most accurately reflect the patterns observed in actual RNA structures.

In this work, we aim to refine and advance the probabilistic-based approach by drawing on the strengths of the CONTRAfold model as a prior to the prediction model. Our efforts will be directed towards enhancing the model's ability to probabilistically reason about RNA structure formation, with a particular emphasis on capturing the subtle interplay between thermodynamics and biological function.

2 Methodology

2.1 Data Descriptions

The bpRNA dataset [1] utilized in this study consists of two primary types of data: the nucleotide sequences of RNA and their experimentally validated secondary structures. The secondary structures included in the dataset have been obtained through laboratory experiments and reported in a variety of research articles. This dataset is well-established in the domain of RNA secondary structure prediction and serves as a benchmark for evaluating the performance of different predictive models.

2.2 Modeling

2.2.1 Sequence to Sequence Model

As noted in section 1.3.2.2, let's say we treat the sequence of RNA as the text. We then transform the text (RNA sequence) into another text (RNA secondary structure) like the translation task [9, 8] or generative task [5, 7]. The general architecture for text-to-text tasks is to use two sections: an Encoder to encode the text into another space and a decoder to transform the object in another space into the text (or tokenized text) again. The tokenization process in this work is to perform one-hot encoding of each nucleotide. In another word, we transform $\{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\} \overset{\text{tokenization}}{\rightarrow} \{0, 1\}^4$.

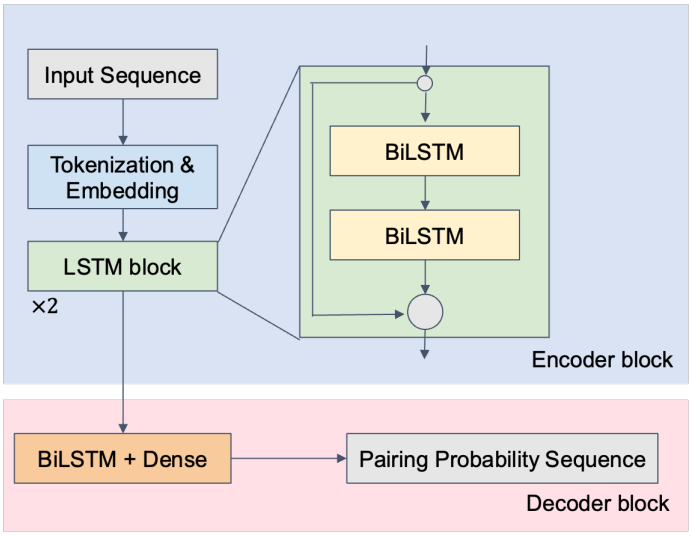

Figure 1: Sequence to Sequence Architecture

The general way of transforming tokenized text in another space is to use a Bi-directional Long

Short-Term Memory neural network (BiLSTM) which will transform the sequence in two ways: forward transformation (from position $i-1$ to i) and backward transformation (from position i to $i-1$). This way, the consistency of embedding is preserved. Then, the sequence is transformed back to secondary structure space in which the output is a stochastic quasi-adjacency matrix representing the non-self or non-linear pairing probability of the graph.

2.2.2 Graph Convolutional Neural Network

The way that traditional neural network works is that we need to create a mapping $N : \mathbb{R}^n \rightarrow O$ for output space O. To generate such a mapping where the information of the vertices and the edge connection is preserved, we introduce the embedding from the graph neural network as the input vector for the convolutional neural network (CNN) to learn to generate the random quasi-adjacency matrix that represents the probability of the existence of the edge in the graph. To tackle the problem discussed earlier, we can transform the edge embedding into the weight itself. In other words, we linearly transform the node embedding with edge embedding. That is,

$$\mathbf{z} = \tilde{A}\mathbf{x}\mathbf{w} + \mathbf{b} \tag{2}$$

for a node embedding (one-hot encoding of the sequence) $\mathbf{x}$ a stochastic quasi-adjacency matrix $\tilde{A}$ generated from pairing probability of CONTRAfold [3] model. This process can be viewed as the approximation of the convolution operator in a graph [2, 6]. Furthermore, we can expand this model with the skip connection which would be in the form of

$$\mathbf{z}_s = \tilde{A}\mathbf{x}\mathbf{w}_1 + \mathbf{x}\mathbf{w}_2 + \mathbf{b} \tag{3}$$

The activation function will then be applied and fed into another layer like a traditional neural network task. The loss for this function would be the cross-entropy of having a process of predicting the existence of pairing.

We then add another residual convolutional neural network for this work, using the input from the last layer before the output of the graph convolutional neural network to predict the stochastic quasi-adjacency matrix of the RNA sequence. The embedding is transformed by multiplying itself with its transpose to make a matrix that the convolutional neural network can work with.

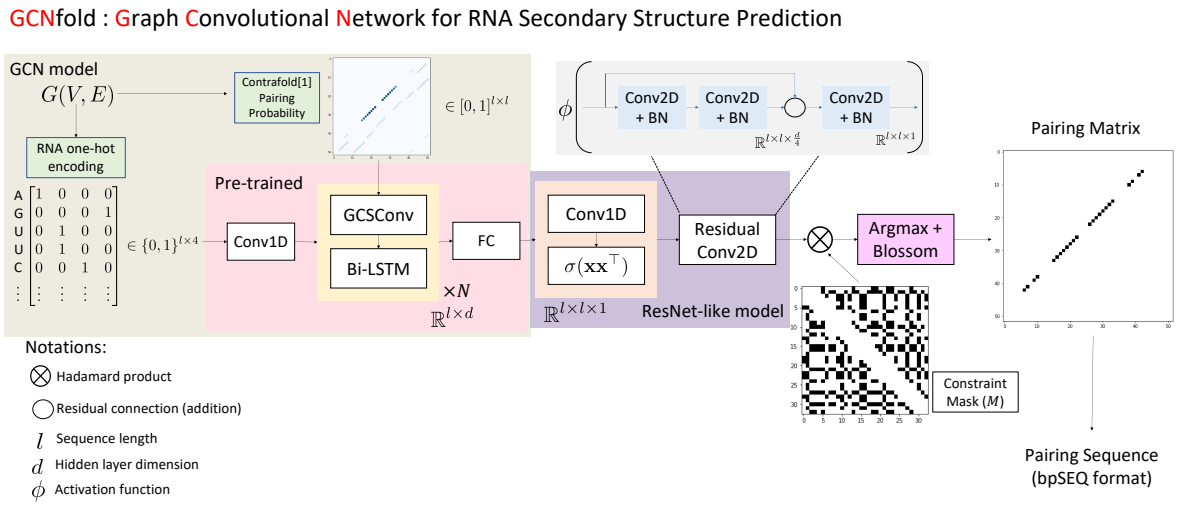

Figure 2: Graph Convolutional Neural Network Architecture

2.2.3 Post-processing

As neural network output is not guaranteed to output the quasi-adjacency matrix. The masking matrix M will be introduced. $M = [M_{ij}]$ will guarantee that

1. Only allows first three most pairing nucleotide. That is,

$$\forall i, j \in \Gamma, M_{ij} = \begin{cases} 1 & (x_i, x_j) \in \{(x_A, x_U), (x_U, x_A), (x_C, x_G), (x_G, x_C), (x_U, x_G), (x_G, x_U)\} \\ 0 & \text{else} \end{cases}$$

   for index set Γ and one-hot encoding of ith nucleotide, x_i.

2. The loops of the secondary structure must not be a sharp loop. That is,

$$\forall i, j \in \Gamma, M_{ij} = 0 \iff |i - j| < 4$$

3. Each row and column must have at most one non-zero element.

Let's say the output matrix of the neural network is Z, and the Hadamard product $Z \otimes M$ will be calculated to ensure the first two constraints that we introduced are followed. Then, the argmax and blossom algorithm [4] will be used to ensure the third constraint is followed.

3 Results and Discussion

3.1 Results

The Sequence model yields sequential output, which is not comparable with another model that yield non-sequential output and then transform into the secondary structure. The result for sequence to sequence model will be presented in Appendix A. Without sequence sequence model, we evaluate the model in two subdatasets from the bpRNA dataset similar to [12]. The sequence-wise dataset is the dataset that remove highly similar (more than 90% similarity) sequences out while the family-wise still preserve it.

	Sequence-wise			Family-wise		
	PPV	SEN	F1	PPV	SEN	F1
GCNfold (ours)	0.654	0.671	0.648	0.612	0.670	0.623
CONTRAfold [3]	0.482	0.656	0.541	0.579	0.736	0.639
RNAfold [18]	0.446	0.631	0.508	0.552	0.720	0.617
MXfold2 [11]	0.520	0.682	0.575	0.585	0.710	0.632
SPOT-RNA [12]	0.652	0.578	0.597	0.599	0.619	0.596

Table 1: Comparisons of the State-Of-The-Art model and our model

	Length ≤ 100			$100 <$ Length ≤ 200			$200 <$ Length ≤ 300			$300 <$ Length ≤ 400			Length > 400		
Model	F1	SEN	PPV	F1	SEN	PPV	F1	SEN	PPV	F1	SEN	PPV	F1	SEN	PPV
GCNfold	0.714	0.713	0.742	0.580	0.582	0.614	0.595	0.584	0.629	0.602	0.558	0.681	0.548	0.504	0.610
CONTRAfold [3]	0.581	0.654	0.548	0.482	0.568	0.436	0.458	0.539	0.410	0.445	0.483	0.429	0.448	0.479	0.426
RNAfold [18]	0.544	0.627	0.507	0.464	0.561	0.413	0.397	0.483	0.347	0.410	0.457	0.386	0.425	0.467	0.394
Mxfold2 [11]	0.628	0.707	0.600	0.495	0.577	0.454	0.439	0.510	0.397	0.493	0.528	0.483	0.475	0.505	0.455

Table 2: Comparisons of the predictive power in length of our model and State-Of-The-Art model

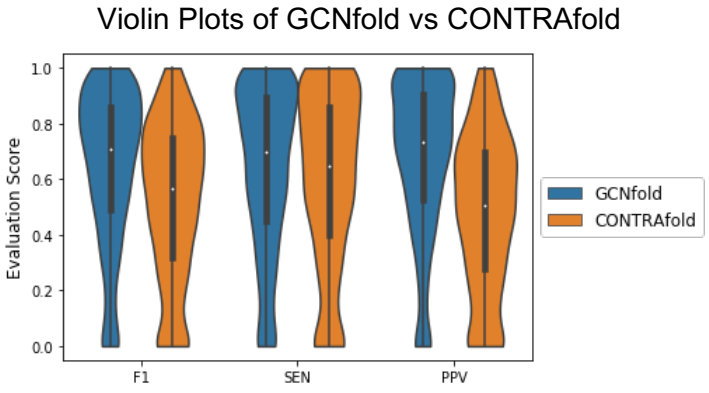

Figure 3: Violin plots comparing the distribution of the score of GCNfold vs CONTRAfold

3.2 Discussion

The result here shows that GCNfold is performing comparable with many State-Of-The-Art models. The result from GCNfold and Sequence-to-Sequence model is preserving the structure. However, the location of each feature is not accurately predicted. In Figure 4 and 5, we can see that GCNfold accurately predicts most of the shape features of the ground truth. However, the location of each pairing is not accurate. Moreover, the violin plot in Figure 3 shows that the distribution of score in our model and CONTRAfold [3] are different. Thus, our model can learn to produce distinct results compared to CONTRAfold. In other words, the dependencies of CONTRAfold prediction are less likely to influence the performance of our model, albeit it does as the baseline that the model will perform at least in the neighborhood of CONTRAfold performance.

We hypothesize that the problem with positional error happens because of the nature of the model, where it can accurately predict the structure of the shorter sequences. The dependencies of the further nucleotide should be treated as the same. However, LSTM does not compensate that. The idea of using the discounting factor similarly in reinforcement learning rewarding problem [13] can be used to possibly tackle this problem. Moreover, the redesign in architecture, especially on graph embedding, should help the performance of the model.

4 Conclusion

The Seq2Seq and the GCNfold is being explored. The performance measure suggests that the graph convolutional model performs the best in both positive, sensitivity, and F1 scores and exceeds F1 for the recent SOTA model in sequence-wise datasets. However, the results from the family are less superior to the CONTRAfold. Moreover, the inference suggests GCNfold is distinct from CONTRAfold. The future work that can be explored is to use the Seq2Seq model as the embedding for prior distribution in the stochastic quasi-adjacency matrix to see the performance change and use the discount factor to compensate for the dependencies of further nucleotides.

Appendices

A Inference Results

The inference result of predicting one structure:

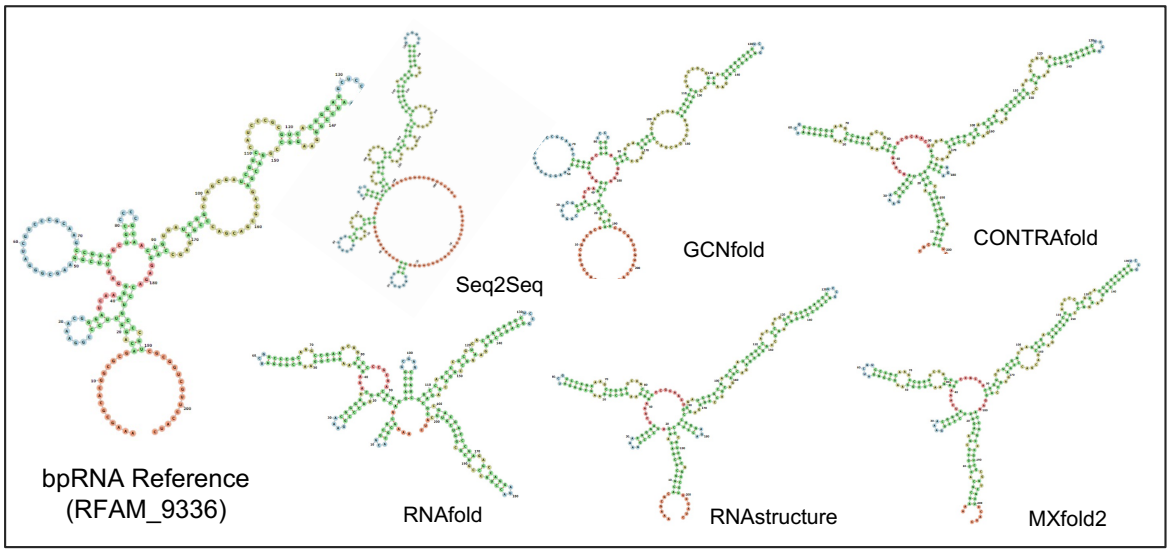

Figure 4: bpRNA:RFAM_9336 Inference

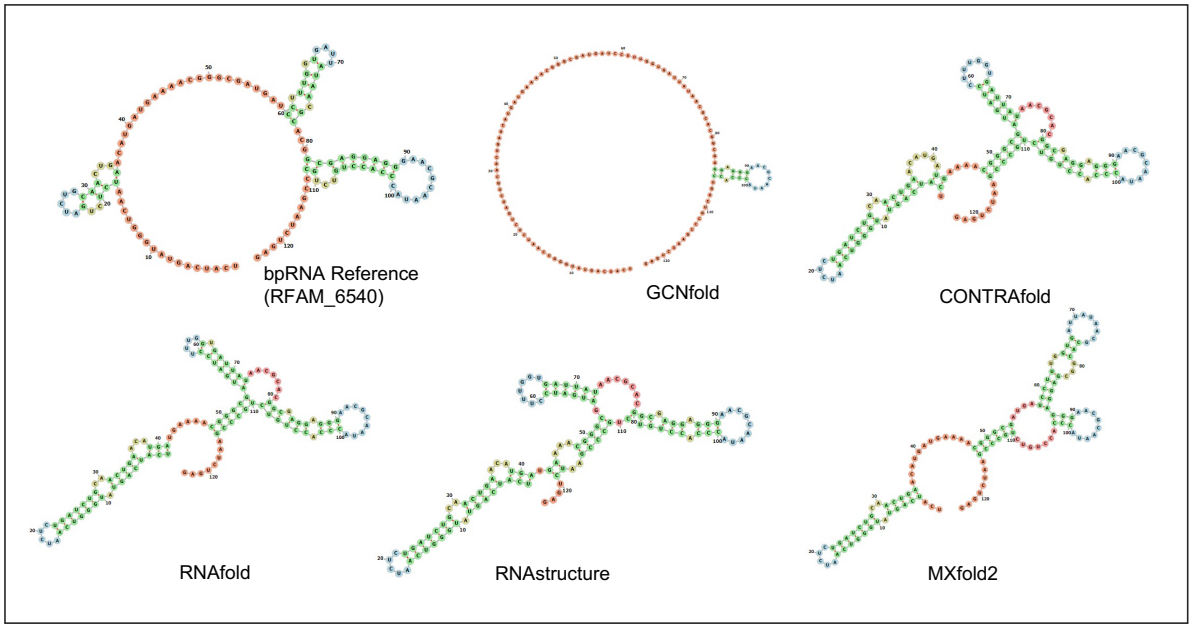

Figure 5: bpRNA:RFAM_6540 Inference

References

[1] Danaee, P., Rouches, M., Wiley, M., Deng, D., Huang, L., and Hendrix, D. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Research 46*, 11 (05 2018), 5381–5394.

[2] Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *CoRR abs/1606.09375* (2016).

[3] Do, C. B., Woods, D. A., and Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics 22*, 14 (07 2006), e90–e98.

[4] Edmonds, J. Maximum matching and a polyhedron with 0,1-vertices. *Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics* (1965), 125.

[5] Kang, D., and Hashimoto, T. B. Improved natural language generation via loss truncation. In *Association for Computational Linguistics (ACL)* (2020).

[6] Kipf, T. N., and Welling, M. Semi-supervised classification with graph convolutional networks. *CoRR abs/1609.02907* (2016).

[7] Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. Adversarial learning for neural dialogue generation. In *Empirical Methods in Natural Language Processing (EMNLP)* (2017).

[8] Luong, T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon, Portugal, Sept. 2015), Association for Computational Linguistics, pp. 1412–1421.

[9] Pryzant, R., Britz, D., and Le, Q. Effective domain mixing for neural machine translation. In *Second Conference on Machine Translation (WMT)* (2017).

[10] Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. C., and Haussler, D. Stochastic context-free grammers for tRNA modeling. *Nucleic Acids Research 22*, 23 (1994), 5112–5120.

[11] Sato, K., Akiyama, M., and Sakakibara, Y. Rna secondary structure prediction using deep learning with thermodynamic integration. *Nature Communications 12*, 1 (2021), 941.

[12] Singh, J., Hanson, J., Paliwal, K., and Zhou, Y. Rna secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications 10*, 1 (Nov 2019), 5407.

[13] Sutton, R. S., Barto, A. G., Sutton, R. S., and Barto, A. G. *Reinforcement learning: An introduction*. The MIT Press, 2020, ch. Tabular Solution Methods, p. 23–194.

[14] Takefuji, Y., Chen, L.-L., Lee, K.-C., and Huffman, J. Parallel algorithms for finding a near-maximum independent set of a circle graph. *IEEE Transactions on Neural Networks 1*, 3 (1990), 263–267.

[15] Tinoco, I., and Bustamante, C. How rna folds. *Journal of Molecular Biology 293*, 2 (1999), 271–281.

[16] Trotta, E. On the normalization of the minimum free energy of rnas by sequence length. *PLOS ONE 9*, 11 (11 2014), 1–9.

[17] Zhao, Q., Zhao, Z., Fan, X., Yuan, Z., Mao, Q., and Yao, Y. Review of machine learning methods for rna secondary structure prediction. *PLOS Computational Biology 17*, 8 (2021), e1009291.

[18] Zuker, M., and Stiegler, P. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic acids research 9*, 1 (1981), 133–148. 6163133[pmid].