AMRITA VISHWA VIDYAPEETHAM

Department of Artificial Intelligence

22BIO201: Intelligence of Biological System – I
NOV - 2024

Project Report

Predicting the Secondary Structure of RNA

Team Members:

Prisha Gupta: CB.SC.U4AIE23331
V Sai Kartikeya: CB.SC.U4AIE23345
Vedant Maheshwari: CB.SC.U4AIE23346
Y Sai Akhilesh: CB.SC.U4AIE23348
....

Date of submission: <08/11/2024>
Signature of the Project Supervisor:

Abstract

RNA secondary structure prediction is fundamental in bioinformatics, offering insights into RNA functionality, stability, and interactions within biological systems. This project presents an approach that uses machine learning to predict and visualize RNA secondary structures. Using a machine learning model optimized with focal loss, we improve prediction accuracy, especially for rare structural elements, by identifying common and unique folding patterns in RNA sequences. The model processes RNA sequences to predict their secondary structures in dot-bracket notation, which is then further analyzed using the ViennaRNA package. ViennaRNA calculates the Minimum Free Energy (MFE) of the predicted structures, providing a thermodynamically stable representation of RNA folding. The final output includes a detailed graphical visualization of the secondary structure, highlighting motifs such as hairpins, loops, and stems, which are essential for understanding RNA functionality in cellular processes.

By combining the advantages of ViennaRNA for precise energy-based modeling with machine learning for structural prediction, this method produces a comprehensive tool for RNA study. Researchers in molecular biology, bioinformatics, and drug development can benefit greatly from the combination of thermodynamic visualization and predictive modeling, which enables more in-depth investigation of RNA-based mechanisms in disease and treatment design.

1 Introduction

In the double helical structure of the DNA molecule, two complementary nucleotide strands are held together with hydrogen bonds between the Waston-Crick pairs A-T and C-G. RNA molecules usually come as single strands but left in their environment they fold themselves in their tertiary structure because of the same hydrogen bonding mechanism. Helices, also known as stems, are formed intra-molecularly.

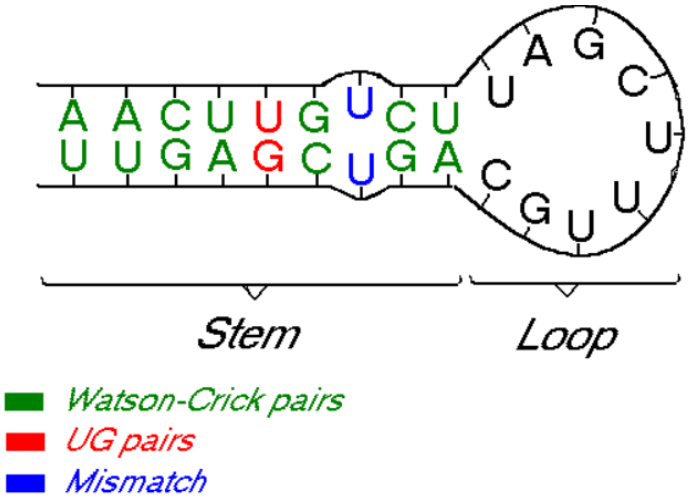


Fig:1 An RNA molecule secondary structure

Recent studies have revealed that functional non-coding RNAs (ncRNAs) play essential roles, including transcriptional regulation and guiding modification, participating in various biological processes ranging from development to cell differentiation, with defective functionality being involved in various diseases1. Because it is well-known that the functions of ncRNAs are deeply related to their structures rather than their primary sequences, discovering the structures of ncRNAs can elucidate the functions of ncRNAs. However, there are major difficulties in determining RNA tertiary structures through experimental assays such as nuclear magnetic resonance and X-ray crystal structure analysis, because of the high experimental costs and resolution limits on measurements of RNA. Although considerable advances in cryo-electron microscopy research on RNA tertiary structure determination have been achieved in recent years, these limitations have not yet been completely overcome. Therefore, instead of conducting such experimental assays, we

frequently perform computational prediction of RNA secondary structures, defined as sets of base pairs with hydrogen bonds between the nucleotides.

The most popular approach for predicting RNA secondary structures is based on thermodynamic models, such as Turner's nearest-neighbor model, in which a secondary structure is decomposed into several characteristic substructures, called nearest-neighbor loops, such as hairpin loops, internal loops, bulge loops, base-pair stackings, multi-branch loops, and external loops, as shown in Fig. 2. There are 16 possible base-pairings, however of these, only six (AU, GU, GC, UA, UG, CG) are stable enough to form actual base-pairs. The rest are called mismatches and occur at very low frequencies in helices. RNA molecules, such as ribosomal RNAs and transfer RNAs, have an important role. Their structure cannot easily be disrupted without impact on their function and lethal consequences and selection is acting to maintain the secondary structure. Yet, the primary structure of the stems (i.e., their nucleotide sequence) can still vary and we observe that RNA helical regions are quite variable in sequence.

The free energy of each nearest-neighbor loop can be calculated by summing the free energy parameters that characterize the loop. The free energy parameters have been determined in advance by experimental methods such as optical melting experiments3. The free energy of an entire RNA secondary structure is calculated by summing the free energy of the decomposed nearest-neighbor loops.

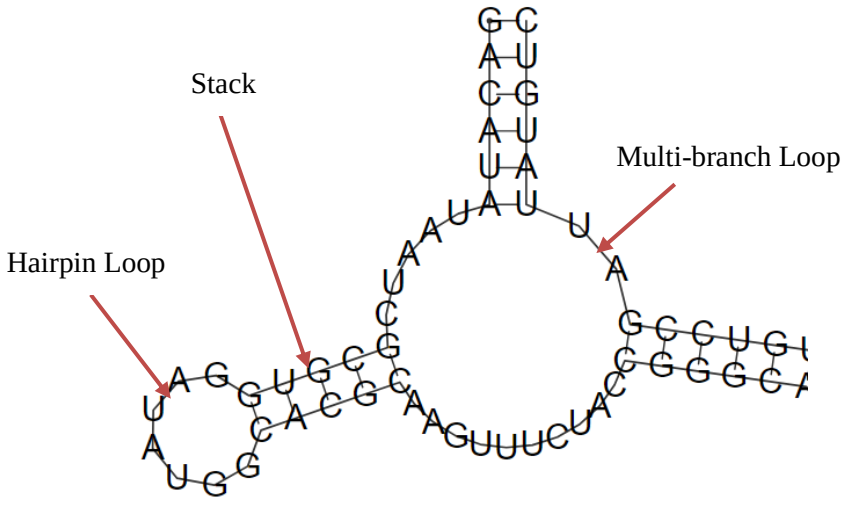


Fig:1 An RNA secondary structure can be into several types of nearest-neighbor loops

We can efficiently calculate an optimal secondary structure that has the minimum free energy using a dynamic programming (DP) technique, the well-known Zuker algorithm. Several tools, including Mfold/UNAfold, RNAfold, and RNA structure, have adopted this approach.

2 Related Work

In recent years, significant advancements have been made in RNA secondary structure prediction, a challenging task crucial for understanding RNA's biological functions and facilitating RNA-targeted drug discovery. These advances stem largely from machine learning (ML) and deep learning (DL) approaches, which have enabled more accurate and efficient models to predict RNA structure, even as the inherent complexities of RNA folding continue to pose challenges.

Sato et al. (2021) developed a deep-learning-based model that incorporates thermodynamic regularization, addressing the limitations of traditional ML approaches. Their proposed algorithm, MXfold2, combines DL capabilities with thermodynamic principles to achieve superior performance compared to prior ML-based models. This integration has set a new benchmark in RNA secondary structure prediction, demonstrating that DL models, when coupled with biophysical insights, can significantly surpass traditional ML techniques.

Further exploring the landscape of ML in RNA research, Zhao et al. (2021) provided a comprehensive review of various ML methods applied to RNA secondary structure prediction, identifying persisting limitations that prevent optimal results. They conclude that current ML models still fall short in fully resolving the complexities of RNA structures, especially those containing pseudoknots. Their review underscores the pressing need for refined and more nuanced ML models, which could pave the way for future breakthroughs.

A broader overview of recent trends in RNA informatics is provided by Sato and Hamada (2023), who highlighted the growing applications of both ML and DL in RNA structure prediction, RNA aptamer discovery, and RNA-based drug discovery.

In their review, they emphasize that accurate test data construction is essential to avoid overfitting and ensure reliable benchmarking. This work offers valuable insights into the future of RNA research, suggesting that further refinement in predictive modeling and validation practices is crucial for progress in RNA informatics and therapeutic applications.

Another contribution in DL modeling is by Chen and Chan (2023), who proposed REDfold, a deep learning model for RNA secondary structure prediction based on a residual encoder-decoder network. Utilizing a convolutional neural network (CNN), REDfold effectively learns both short- and long-range dependencies within RNA sequences, with symmetric skip connections facilitating efficient information propagation. The model demonstrates substantial improvements in both accuracy and efficiency over previous methods, underscoring the potential of CNNs to handle complex dependencies inherent to RNA.

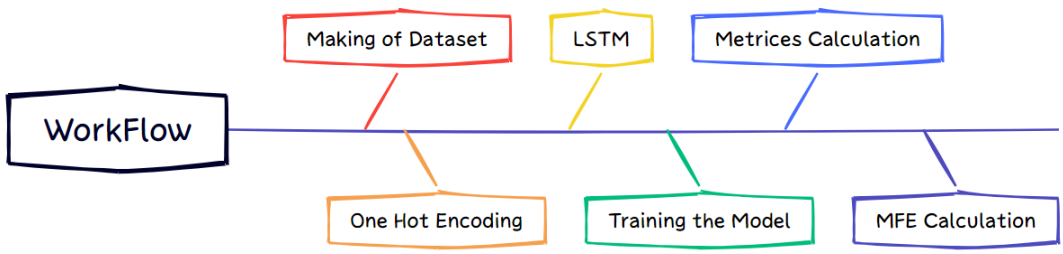
Franke et al. (2024) introduced RNAformer, a DL model designed primarily around RNA sequence inputs. By employing axial attention and latent space recycling, RNAformer can learn biophysical models of RNA folding, with prediction accuracy that scales with model size. The model outperforms existing DL approaches on standard benchmarks, highlighting the effectiveness of a simple architecture that integrates axial attention for RNA structure prediction. This approach exemplifies a scalable and powerful solution for advancing RNA modeling.

Zhou et al. (2024) explored combinatorial approaches, particularly approximation algorithms aimed at solving the Maximum Stacking Base Pairs (MSBP) and Maximum Base Pair Stackings (MBPS) problems, to address arbitrary pseudoknot structures in RNA prediction. Their study adopts a local search methodology to maximize base-pair stackings, though it simplifies the analysis to reduce time complexity. Consequently, while the proposed algorithms are computationally efficient, they might lack the precision of more complex approaches, illustrating the trade-offs between complexity and accuracy in combinatorial modeling for RNA. Finally, Zhou et al. (2024) delved into Transformer-based DL models for RNA secondary structure prediction, proposing an algorithm to extend these methods to

tertiary RNA structure modeling. Their study suggests that Transformer architectures can successfully capture complex structures within biological macromolecules, broadening the potential applications of DL in RNA prediction. This research points toward a promising future direction, with Transformer models potentially advancing RNA modeling to higher-order structural levels.

In summary, the reviewed studies underscore the transformative impact of ML and DL approaches on RNA secondary structure prediction. The field has seen the development of models that leverage novel architectures like encoder-decoder networks, Transformers, and convolutional layers, while others integrate biophysical and thermodynamic principles to enhance prediction accuracy. These works collectively suggest that by refining test data, advancing pseudoknot handling, and expanding model architectures, the field of RNA informatics will continue to evolve, unlocking new possibilities for RNA-targeted therapies and deeper biological insights.

3 Methodology



The methodology for predicting RNA secondary structure in this project involves several key steps, from data preprocessing to the application of deep learning models, and integrates both computational and biological insights.

1. **Data Collection and Preprocessing**

   Dataset Compilation: RNA sequence data is collected, including information on nucleotide sequences and their corresponding secondary

structures. This dataset typically includes dot-bracket notation to represent paired and unpaired bases within each RNA sequence.

One-Hot Encoding of Nucleotide Sequences: To prepare the RNA sequences for model training, each nucleotide (A, U, G, C) is converted into a one-hot encoded format. This encoding provides a numerical representation of RNA sequences that can be processed by deep learning models.

Sliding Window Technique: To capture local patterns within RNA sequences, a sliding window approach is applied. This allows for dividing each sequence into overlapping segments, making it easier to detect both short- and long-range dependencies critical for base pairing predictions.

Label Encoding of Secondary Structure: The secondary structure is encoded into labels based on dot-bracket notation, where each opening and closing bracket corresponds to a specific pairing in the RNA strand. Specialized loss functions, such as focal loss, are used to improve model performance on these minority class labels.

2. Feature Engineering and Integration with ViennaRNA Library

The **ViennaRNA library** plays a crucial role by introducing biophysical constraints and generating thermodynamically feasible RNA folding patterns.

Thermodynamic Folding Prediction: ViennaRNA predicts RNA folding based on minimum free energy (MFE), which aligns the folding process with the biophysical properties of RNA molecules. By considering factors like base-pairing interactions, and loop stability, it generates thermodynamically stable RNA structures.

Dot-Bracket Notation Generation: ViennaRNA outputs RNA secondary structures in dot-bracket notation, representing paired and unpaired bases in a standardized format. This notation is crucial for labeling and training the model, as it simplifies the structure into a sequence of symbols that can be readily interpreted by machine learning models.

RNA Structure Visualization: The library includes tools like RNAplot, which visually represent RNA secondary structures as diagrams, such as circular or planar graphs. This is essential for interpreting model outputs, as it allows comparison between predicted and known RNA structures, providing insight into the accuracy and plausibility of predictions.

Integration with Deep Learning Models: Using ViennaRNA to pre-generate realistic RNA secondary structures allows the model to learn folding patterns that are consistent with biological principles. This helps improve model training by grounding it in thermodynamically feasible configurations, enhancing the prediction accuracy by making use of ViennaRNA's scientifically validated algorithms.

3. Model Architecture and Training

LSTM Model Architecture: Given the sequential nature of RNA data, an LSTM (Long Short-Term Memory) model is particularly well-suited for this task. LSTMs are designed to capture long-range dependencies, which are essential for RNA structure prediction, where distant nucleotide interactions (e.g., base pairings) determine the overall secondary structure.

Training Process: The LSTM model is trained on RNA sequences and their corresponding secondary structures (in dot-bracket notation).

Loss Function: A specialized loss function, such as focal loss, can be applied to handle the imbalance in structural labels (i.e., more dots than brackets). This helps the model focus more on accurately predicting base

pairs, improving performance for paired regions.

Optimization: The model is optimized using algorithms like Adam, with adjustments to learning rates and regularization to prevent overfitting.

Thermodynamic Regularization: To improve prediction accuracy, thermodynamic constraints can be integrated by using ViennaRNA-generated structural data, guiding the model toward biologically plausible folds.

Evaluation and Validation: The LSTM model's predictions are validated against known RNA structures to assess accuracy. Performance metrics such as F1-score and structural similarity indexes evaluate the effectiveness of the model in capturing the correct base pairings and overall secondary structure.

4. **Visualization and Interpretation**

Visualization of Predicted Structures: The predicted RNA secondary structures are visualized using RNA structural tools like ViennaRNA to generate graphical representations, including dot-bracket plots and circular diagrams. This provides an intuitive view of the predicted pairings and structural folds.

4 Experiments

4.1. Dataset Overview

1. **Dataset Source and Composition**

The dataset used in this project consists of RNA sequences annotated with secondary structures in dot-bracket notation, providing a representation of paired and unpaired nucleotides. This format, widely recognized in RNA research, is ideal for translating sequence data into structural predictions. Each entry in the dataset includes:

- Header: Unique identifiers for each RNA sequence.
- Nucleotide Sequence: The string of nucleotides (A, U, G, C) forming the RNA sequence.
- Structure: The corresponding dot-bracket notation, indicating base pairing.

The dataset comprises RNA sequences of various lengths, with structures that exhibit a range of secondary motifs, such as hairpin loops, internal loops, and multi-branch loops. This diversity enables comprehensive training and testing across different RNA structural configurations.

2. **Preprocessing and Encoding**

   To prepare the dataset for machine learning, each nucleotide sequence was one-hot encoded, creating a binary vector representation that retains sequence information while making it suitable for computational models. Additionally, the structural labels in dot-bracket notation were encoded to facilitate learning, with special attention given to handling rare structural motifs to mitigate class imbalance issues.

3. **Data Splitting**

   The dataset was divided into training, validation, and test sets. The training set is used to fit the models, the validation set is used for hyperparameter tuning, and the test set evaluates the final model performance. Care was taken to ensure that each set reflects the diversity of RNA structures in the original data.

   - Training Set: 70% of the data
   - Validation Set: 15% of the data
   - Test Set: 15% of the data

4.2. Models Applied for RNA Structure Prediction

1. **Minimum Free Energy (MFE) Model Using ViennaRNA**

   The MFE model, implemented with the ViennaRNA package, calculates the secondary structure with the lowest free energy for each RNA sequence. This method is based on thermodynamic principles, predicting the most stable structure by minimizing the Gibbs free energy. Key steps involved include:

- Energy Calculation: Each base pair is evaluated based on stacking energy and loop configurations.
- Dot-Bracket Output: The model generates a secondary structure prediction in dot-bracket notation, which serves as a reference for machine learning models.
- Graphical Visualization: ViennaRNA's capabilities allow for visual output, aiding in verification and comparison with machine-predicted structures.

2. Long Short-Term Memory (LSTM) Model

LSTM networks are effective for sequential data, capturing long-range dependencies within nucleotide sequences. The LSTM model was trained to predict dot-bracket notation directly from RNA sequences. Key components of the LSTM model implementation include:

2.1. Architecture Design

- The LSTM model is structured with multiple layers to capture sequential dependencies. Each LSTM unit retains the hidden state across timesteps, capturing both local and global dependencies within RNA sequences.
- Input Layer: Takes the one-hot encoded nucleotide sequence as input.
- Hidden LSTM Layers: Stacked LSTM layers improve the model's ability to capture complex patterns in RNA sequences. Each unit's memory cell allows it to retain relevant information, essential for capturing long-range nucleotide dependencies that impact RNA folding.
- Dense Output Layer: This layer outputs predictions for each nucleotide in the sequence, labeling them as paired or unpaired according to the dot-bracket notation.

2.2. Training and Evaluation

- The LSTM model is trained using the Adam optimizer with regularization techniques like dropout to prevent overfitting.
- Loss Function: Focal loss is applied to address the class imbalance, enhancing accuracy for the minority class of paired bases.

- Evaluation Metrics: F1-score, accuracy, and structural similarity index are calculated to assess the model's prediction quality. These metrics help evaluate how well the model captures base-pair interactions and overall structure.

5 Results & Discussions

In the context of your RNA secondary structure prediction project, the model's performance is evaluated using several key metrics that help assess the accuracy, effectiveness, and reliability of the predictions made by your model (in this case, an LSTM model).

1. **Accuracy**

   Accuracy measures the percentage of correct predictions out of the total predictions made by the model. It is a simple metric that indicates how often the model's predictions are correct overall.

   The model achieved a test accuracy of **89.5%** and a test loss of **0.041**.

2. **Precision and Recall**

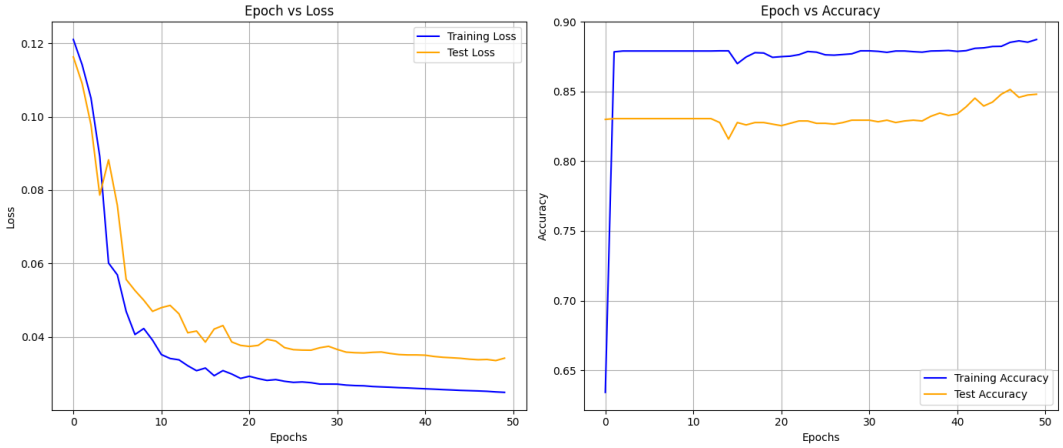
   Precision is the ratio of true positive predictions to the total number of positive predictions (i.e., how many of the predicted base pairings were correct). This is important for minimizing false positives.

   Recall is the proportion of true positive predictions out of all actual instances of a specific class.

3. **F1 Score**

   The harmonic mean of precision and recall, provides a balance between the two.

   For an LSTM (Long Short-Term Memory) model, the key indicators of performance are often shown through loss and accuracy plots over epochs



The training loss curve shows a steady decline, indicating that the model is effectively learning patterns in the training data. The loss decreases sharply at the

beginning and then stabilizes, suggesting that the LSTM model has captured the temporal dependencies in the data.

The training accuracy increases quickly and stabilizes around 87%, which is a good indicator that the LSTM model is fitting well on the training dataset.

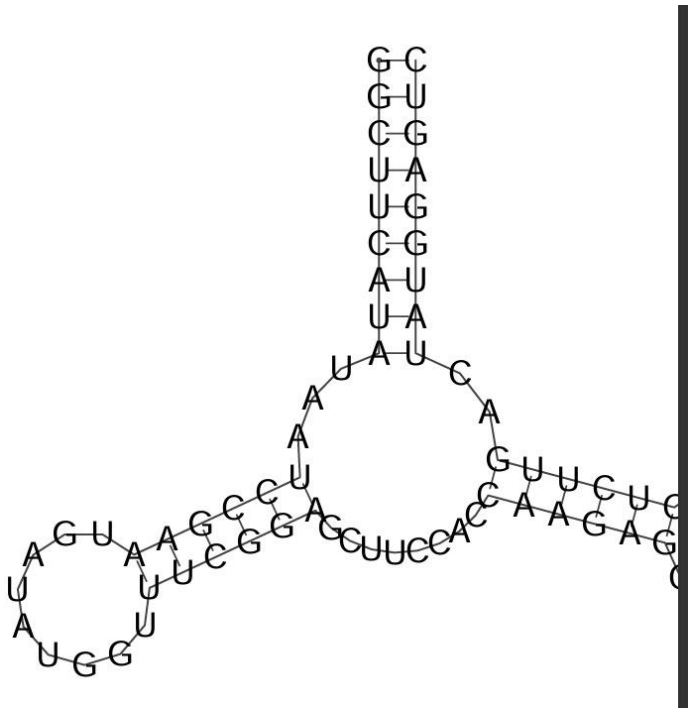
The test loss also decreases initially but starts fluctuating and even slightly increases after around epoch 20. This fluctuation suggests the model is starting to overfit; it memorizes the training data rather than generalizing to new, unseen data.

The test accuracy stabilizes at around 82%, slightly lower than the training accuracy, which is typical for a well-performing model. However, the decline towards the end hints at overfitting.

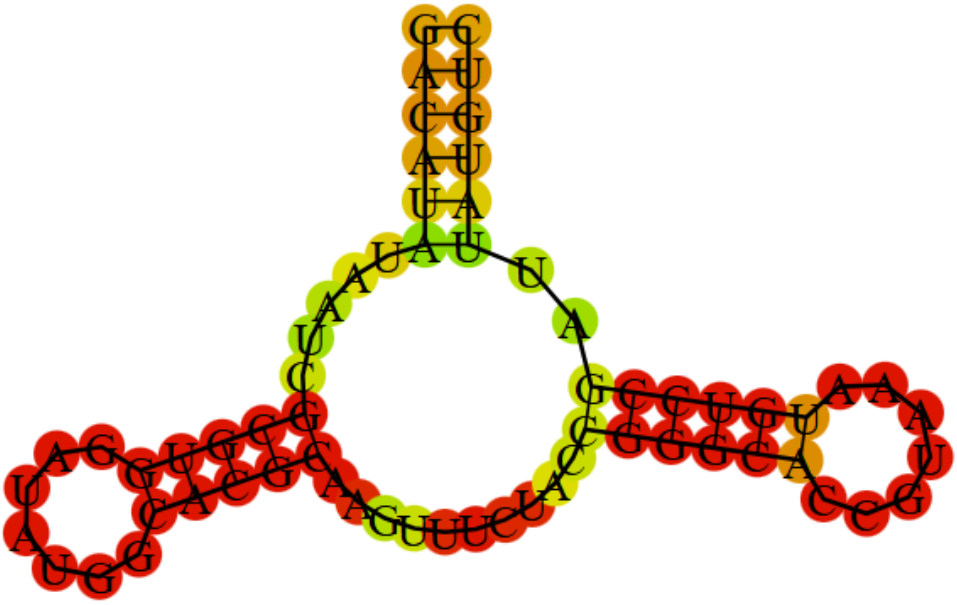
	precision	**Recall**	**F1-score**
.	**0.90**	**0.99**	**0.94**
(	**0.56**	**0.13**	**0.22**
)	**0.57**	**0.11**	**0.19**

The results show that the model performs well for non-pairing bases ('.'), with high precision, recall, and F1 scores, indicating it accurately identifies these bases in RNA sequences. However, it struggles with pairing bases ('(' and ')'), exhibiting low recall and F1 scores for both, meaning it misses a large proportion of the actual base-pairing instances. The overall accuracy of 90% is driven by the model's success with the '.' class, highlighting a potential class imbalance issue. While the weighted averages are strong, the low performance for the minority classes suggests the need for further improvements, such as adjusting for class imbalance or refining the model's ability to detect RNA base pairs.

After Visualisation, this was the output we achieved:



If we compare it to the output, we got on the RNAFold Web Server, we can say they are fairly accurate.



The optimal secondary structure in dot-bracket notation with a minimum free energy of -18.00 kcal/mol, while the one we predicted was -17.5 kcal/mol.

6 Conclusion & Future Work

1. Challenges

The primary challenges faced were related to handling complex RNA structures with intricate base pairings. The presence of uncommon base-pair structures also posed difficulties, requiring tailored loss functions to balance class representation. Additionally, visualizing predicted structures for validation against actual RNA forms remains a complex task, as minor prediction inaccuracies can lead to significant structural differences.

2. Future Work

The current results show promising accuracy, with LSTM-based architectures proving effective for RNA secondary structure prediction. However, the model can be further refined by exploring hybrid architectures, such as incorporating attention mechanisms with LSTM layers to better capture long-range dependencies. Additionally, using more advanced visualization tools and comparing with real RNA 3D structures could enhance model validation, leading to more reliable predictions for biological research and applications.

These findings underscore the potential of deep learning approaches in bioinformatics and RNA structural studies, setting a foundation for further exploration into more advanced and accurate predictive models.

References

[1] Sato, Kengo, Manato Akiyama, and Yasubumi Sakakibara. Nature communications 12.1 (2021): 941.: RNA secondary structure prediction using deep learning with thermodynamic integration.

[2] Zhao Q, Zhao Z, Fan X, Yuan Z, Mao Q, Yao Y (2021): Review of machine learning methods for RNA secondary structure prediction

[3] Kengo Sato, Michiaki Hamada (2023) : Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery

[4] Chun-Chi Chen1 and Yi-Ming Chan2 (2023): REDfold: accurate RNA secondary structure prediction using residual encoder-decoder network

[5] Jörg K.H. Franke, Frederic Runge, Ryan Köksal, Rolf Backofen, Frank Hutter (2024): RNAformer: A Simple Yet Effective Deep Learning Model for RNA Secondary Structure Prediction

[6] Aizhong Zhou, Haodi Feng, Jiong Guo, Haitao Jiang, Nan Liu, Binhai Zhu, Daming Zhu (2024): New approximation algorithms for RNA secondary structures prediction problems by local search

[7] Yanlin Zhou, Tong Zhan, Yichao Wu, Bo Song, Chenxi Shi (2024): RNA Secondary Structure Prediction Using Transformer-Based Deep Learning Models