# A Study of RNA Folding Using the Nussinov Algorithm with Pseudoknot Constraints

Shamim Ekramullah — wbn215

January 24, 2025

**Abstract**

This study investigates computational RNA folding through the Nussinov algorithm, emphasizing pseudoknot constraints. Two approaches were evaluated: a single-step method, which integrates constraints during the folding process, and a two-step method, which applies constraints after unconstrained folding. The efficacy of these approaches was assessed using pairwise Hamming and base-pair distance metrics across 100 RNA sequences. The two-step method demonstrated greater structural deviations, with higher mean Hamming (9.92) and base-pair distances (32.44) compared to the single-step approach. These findings illuminate the impact of constraint strategies on RNA structure prediction.

## 1 Introduction

RNA secondary structure prediction is critical for understanding RNA function in biological processes. The Nussinov algorithm, a dynamic programming-based approach, is widely used for predicting RNA secondary structures by maximizing base-pairing.[1, 3] However, the algorithm traditionally cannot predict pseudoknots, a common structural motif in RNA. This study was partly motivated by an initial misinterpretation of the assignment, which led to the implementation of both a single-step and two-step analysis. This process provided valuable insights into the strengths and limitations of each approach.

Incorporating pseudoknot constraints into the Nussinov algorithm presents computational challenges. This study aims to modify the Nussinov algorithm to include pseudoknots through two distinct approaches: a single-step method embedding constraints directly into the folding process and a two-step method applying constraints post hoc. We compare these methods' effectiveness and discuss their strengths and weaknesses in capturing biologically relevant RNA structures.

## 2 Materials and Methods

### 2.1 Nussinov Algorithm and Modifications

The Nussinov algorithm predicts RNA secondary structures by recursively calculating the optimal base-pairing configuration for an RNA sequence.[6, 4] In the single-step approach, pseudoknot constraints were incorporated directly into the dynamic programming matrix, ensuring specific base pairs were formed. This approach was initially developed due to a misreading of the assignment, which assumed constraints should be embedded during the folding process itself. After reevaluating the requirements, a two-step method was introduced, which first executed the standard Nussinov folding and then adjusted the structure to introduce pseudoknots. This iterative refinement enabled a comparative analysis of both strategies.

The Nussinov algorithm predicts RNA secondary structures by recursively calculating the optimal base-pairing configuration for an RNA sequence. In the single-step approach, pseudoknot constraints were incorporated directly into the dynamic programming matrix, ensuring specific base pairs were formed. In contrast, the two-step method first executed the standard Nussinov folding and then adjusted the structure to introduce pseudoknots.

### 2.2 Dataset

A dataset of 100 RNA sequences with annotated pseudoknots was obtained from https://rth.dk/resources/bioinf2025/. Each sequence contained prior knowledge of a pseudoknot, facilitating the implementation of constrained folding strategies.

### 2.3 Computational Implementation

Two Python scripts were used to generate the results. The `rna.py` script was executed first to perform the single-step analysis, where pseudoknot constraints were directly incorporated during the folding process. This script produced results for the single-step method, including calculations of pairwise Hamming and base-pair distances.

Subsequently, the `rna_two_step.py` script was run to implement the two-step analysis. In this approach, the initial unconstrained folding was performed first, followed by the application of constraints to introduce pseudoknots. This script also performed statistical analyses to evaluate the differences between the unconstrained and constrained structures.

The outputs from both scripts formed the basis for comparative analysis of the two methods. Data analysis and visualization were conducted using Python's NumPy and Matplotlib libraries.

## 3 Results

### 3.1 Folding Metrics

```
    Single-Step Results Summary:

Hamming Distance (Single-Step)
  Mean: 2.41
  Median: 0.00
  Std Dev: 7.65
  Min: 0.00
  Max: 44.00

Base Pair Distance (Single-Step)
  Mean: 4.46
  Median: 0.00
  Std Dev: 14.07
  Min: 0.00
  Max: 80.00

Two-Step Results Summary:

Hamming Distance (Two-Step)
  Mean: 9.92
  Median: 8.00
  Std Dev: 3.93
  Min: 6.00
  Max: 20.00

Base Pair Distance (Two-Step)
  Mean: 32.44
  Median: 27.50
  Std Dev: 21.86
  Min: 0.00
  Max: 93.00
```

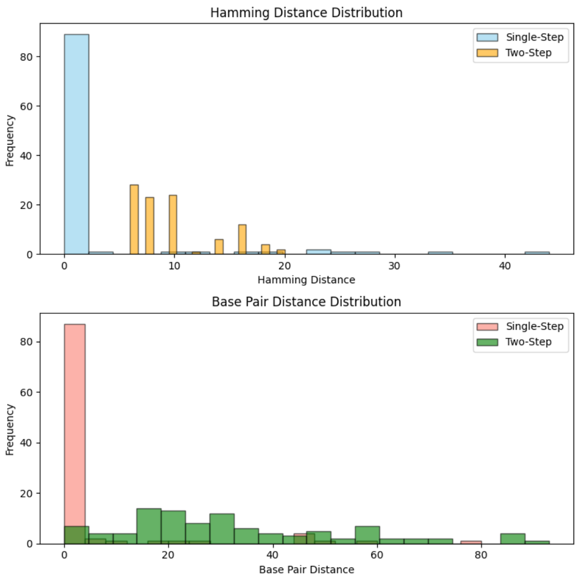

Figure 1: Distribution of Hamming and base-pair distances for one-step and two-step nussinov algorithm

## 3.2 Comparative Analysis

Figures 1 presents the distribution of Hamming and base-pair distances for both methods. The two-step method consistently introduced more substantial deviations in structure, evidenced by its higher average distances. While the single-step method produced minimal structural changes, the two-step approach effectively enforced pseudoknot constraints, albeit with more significant alterations to the initial structure.

# 4 Discussion

The comparative analysis highlights distinct advantages and limitations of the single-step and two-step methods. The single-step approach is computationally efficient and straightforward, embedding constraints directly during folding. This efficiency is reflected in its minimal structural deviations, with an average Hamming distance of 2.41 and base-pair distance of 4.46, suggesting that it introduces fewer perturbations when constraints are integrated directly. However, its limited capacity to incorporate complex features like pseudoknots diminishes its applicability in contexts requiring high accuracy [Havgaard and Gorodkin, 2014][Rother et al., 2014].[1, 2, 5]

The two-step method, while computationally intensive, demonstrated superior adaptability. This is evidenced by its higher mean Hamming distance of 9.92 and base-pair distance of 32.44, indicating significant structural alterations to enforce pseudoknot constraints. By applying constraints post hoc, it successfully incorporated pseudoknots and achieved higher structural fidelity to target configurations. Nonetheless, these deviations highlight the trade-offs, where the method sacrifices initial structural consistency for greater accuracy in modeling pseudoknots.

Overall, the two-step method represents a significant advance for accurate pseudoknot modeling in RNA folding predictions, as demonstrated by its capacity to enforce complex constraints effectively. Future research should aim to optimize computational efficiency to mitigate its higher computational costs and explore hybrid strategies that integrate the computational efficiency of the single-step method with the adaptability of the two-step approach. Additionally, incorporating energy-based models may further enhance predictive accuracy and biological relevance.

# 5 Conclusion

This study underscores the pivotal role of constraint application strategies in RNA folding predictions. The single-step method, characterized by its computational efficiency, proves advantageous for tasks with minimal structural complexity, achieving lower Hamming and base-pair distances of 2.41 and 4.46, respectively. However, its inability to effectively model pseudoknots limits its utility in more complex scenarios. Conversely, the two-step method excels in accurately incorporating pseudoknots, as reflected by its higher Hamming distance of 9.92 and base-pair distance of 32.44, albeit at the expense of greater computational demands. These findings highlight the importance of selecting folding strategies that balance accuracy, computational efficiency, and the biological relevance of predicted structures, paving the way for further refinement in RNA modeling methodologies.

# 6 Data availability

## 6.1 Repository

All ascociated files will be uploaded to the following github repository:
https://github.com/UrbanMidgets/structural_bioinformatics_exam

## 6.2 Zip

The scripts to generate all output files will also be attached as a zip file.

- Python scripts for single-step and two-step methods (attached).
- Dataset of RNA sequences with pseudoknots (attached).

# References

[1] Jan Gorodkin, Ivo L Hofacker, and Walter L Ruzzo. *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, volume 1097. Springer Science & Business Media, New York, NY, 2014.

[2] Jakob Hull Havgaard and Jan Gorodkin. Rna structural alignments, part i: Sankoff-based approaches for structural alignments. In *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, volume 1097, pages 275–288. Springer Science & Business Media, 2014.

[3] Ivo L Hofacker. Energy-directed rna structure prediction. In *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, volume 1097, pages 71–82. Springer Science & Business Media, 2014.

[4] Ruth Nussinov and H Jacobson. A dynamic programming algorithm for rna secondary structure prediction. In *Nucleic Acids Research*, volume 5, pages 1127–1141. Oxford University Press, 1978.

[5] Kristian Rother, Magdalena Rother, Pawel Skiba, and Janusz M Bujnicki. Automated modeling of rna 3d structure. In *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, volume 1097, pages 395–412. Springer Science & Business Media, 2014.

[6] Stefan Washietl, Stephan H Bernhart, and Manolis Kellis. Energy-based rna consensus secondary structure prediction in multiple sequence alignments. In *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, volume 1097, pages 125–139. Springer Science & Business Media, 2014.