Using stochastic context-free grammars to predict RNA secondary structure: Project preliminary report

Sudhanshu Bharadwaj
Undergraduate, IISc
sudhanshub@iisc.ac.in

I Introduction

I-A. *Problem Relevance*

RNA, being a single-stranded molecule, can undergo complex and intricate base-pairing interactions. These interactions lead to various secondary structures which are important for its functionality. RNA nucleotides interact to form secondary structure motifs such as stems, loops, bulges, and pseudo-knots [2]. These motifs help RNA adapt its 3D conformation which is important for its functionality. Therefore better modelling such sequences will help us predict RNA function, identify novel RNA sequences, and recognise evolutionary conserved motifs better. Recently, researchers had to design efficient guide RNAs for CRISPR systems [5] which required an intricate understanding of the structural stability of RNA sequences. Therefore there is no doubt why this problem is in important one.

I-B. *Algorithm*

These pairings often have a nested structure and complicated algorithms are needed to predict or model such structures. The most popular methodology to do this is by thermodynamic energy minimization [10]. But often scalable algorithms for the same are not accurate enough [8]. In addition, there is a lack of interpret-ability in such methods.

Stochastic context-free grammars are one way to model such systems, leveraging regularities and base-pairing rules associated with the secondary structures. Mathematically the extend the Hidden Markov Models, which are common in modelling systems like this [1]. The high level idea is that we generate a structure using a set of grammatical rules, and assign a likelihood to the structure based on the probabilities of the grammar rules used to construct it. The probabilities of such grammar rules can either be learnt or can incorporate biological/chemical reality. With the correct set of rules, we can use comparative sequence analysis methods to predict new structures.

Pfold [4] is a popular algorithm that uses SCFGs. I will mainly be sticking to many of the choices made by this model.

II Deliverables and targets

II-A. *Topics to be studied and presented in class*

- A brief introduction transformations grammars and their types
- RNA secondary structure and some approaches to predict its secondary structures.
- Stochastic context-free grammars, and how they're relevant to RNA secondary structure modelling. Deeper analysis of algorithms that have been implemented.
- Pitfalls, and comparisons with current state of the art approaches.

II-B. *Algorithms to be implemented*

- I will start of by implementing some general algorithms that extend to this problem too, like the CYK algorithm and the Inside-Out Algorithm.
- The main attempt will be to recreate Pfold, the most popular software using this algorithm
- Other algorithms(used either in pre-processing or for comparison), will be direct (appropriately cited) library/ popular implementations.

**I will try to implement most code in C++, but depending on time constraints some code might be in Python*

III Reference material to be used

- Several review papers such as [9] and [3] and Wikipedia to get a broad understanding of the topics
- Lecture slides by Prof. Anthony Gitter from the university of Wisconsin
- Bioinformatics: the machine learning approach [2]
- Two foundational papers on the topics: [4] and [6]

IV Data-set

I will use popular data-sets and implementations to benchmark and test my code. This is a popular topic, and there are several publicly available data-sets, one of which is cited here [7].

References

[1] P Baldi, Y Chauvin, T Hunkapiller, and M A McClure. Hidden markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences*, 91(3):1059–1063, February 1994.
[2] Pierre Baldi and Søren Brunak. *Bioinformatics: The Machine Learning Approach*. 01 2001.
[3] Robin D Dowell and Sean R Eddy. *BMC Bioinformatics*, 5(1):71, 2004.
[4] Bjarne Knudsen and Jotun Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13):3423–3428, 07 2003.
[5] D. Dewran Kocak, Eric A. Josephs, Vidit Bhandarkar, Shaunak S. Adkar, Jennifer B. Kwon, and Charles A. Gersbach. Increasing the specificity of crispr systems with engineered rna secondary structures. *Nature News*, Apr 2019.

[6] Yasubumi Sakakibara, Michael Brown, Richard Hughey, I. Saira Mian, Kimmen Sjölander, Rebecca C. Underwood, and David Haussler. Stochastic context-free grammers for tRNA modeling. *Nucleic Acids Research*, 22(23):5112–5120, 11 1994.

[7] Kengo Sato. The datasets used in "RNA secondary structure prediction using deep learning with thermodynamic integration", January 2021.

[8] Lisa Yu. *Study of RNA Secondary Structure Prediction Algorithms*. PhD thesis.

[9] Qi Zhao, Zheng Zhao, Xiaoya Fan, Zhengwei Yuan, Qian Mao, and Yudong Yao. Review of machine learning methods for rna secondary structure prediction. *PLOS Computational Biology*, 17(8):1–22, 08 2021.

[10] Michael Zuker. [20] computer prediction of rna structure. In *Methods in enzymology*, volume 180, pages 262–288. Elsevier, 1989.