CSC482B Project Proposal: RNA_LZW

Quinn Gieseke, Claire Champernowne

February 20, 2021

Motivation

Detecting RNA secondary structures is a process which is integral to understanding the functions of non-coding RNA sequences. Though there are existing bioinformatics tools used for detection, these programs are not without their limitations. Specifically, RNAz uses a sliding window approach in its analysis of a sequence [1], [2], meaning there may be structures with loops existing outside the scope of the window.

The Lempel-Ziv-Welch(LZW) algorithm is a universal lossless compression algorithm used for a variety of applications such as Unix file compression. The algorithm compresses data by using a dictionary to assign common sequences of characters to a fixed-length code (typically 12-bit) [3].

Given the LZW algorithm's dictionary approach to pattern-matching, we hypothesize it could be used as a novel method of detecting new RNA secondary structures which are undetectable by the sliding window method used in RNAz.

Objectives

Due to the time limitations of this project, we intend to focus solely on detection of possible structures without commenting on the structure's validity in nature. Our goal is to detect RNA secondary structures using a dictionary-based method adapted from the LZW compression algorithm.

Methods

The RNA_LZW algorithm will need to differ from normal LZW implementations in a few key ways. It is beneficial in general compression for one sequence to be matched to a large number of similar sequences, in this way maximally reducing the size of the file. However, for an RNA fold, each sequence can be paired to at most 1 other sequence. Since RNA_LZW has this restriction, it will require additional overhead in the algorithm to maintain matches and ensure that only the longest sequence matches make it to the final output.

Our approach also will be easier in some ways than general compression. Because we only have 4 base pairs to work with, instead of a full ASCII alphabet, our symbol table will have much more room for growth, allowing us to efficiently process large amounts of base pairs. Since LZW and similar algorithms like LZMA are capable of quickly processing gigabytes of data, our new algorithm should be able to handle similarly large datasets in contrast to algorithms like RNAfold, which do not scale efficiently past several thousand base pairs [4].

References

[1] A. R. Gruber S. Findeiß, S. Washietl, I. L. Hofacker, and P. F. Stadler, "RNAZ 2.0: Improved Noncoding RNA Detection," Pacific Symposium on Biocomputing, pp. 69–79, 2010.

[2] S. Washietl, I. L. Hofacker, and P. F. Stadler, "Fast and Reliable Prediction of Noncoding RNAs," Proceedings of the National Academy of Sciences, vol. 102 (7), pp. 2454–2459, 2005.

[3] T. A. Welch, "A Technique for High-Performance Data Compression," in Computer, vol. 17, no. 6, pp. 8-19, June 1984, doi: 10.1109/MC.1984.1659158.

[4] I. L. Hofacker, S. Bonhoeffer, P. F. Stadler, R. Lorenz, and W. Fontana, "RNAfold manual page for RNAfold 2.4.16," Theoretical Biochemistry Group. [Online]. Available: `https://www.tbi.univie.ac.at/RNA/RNAfold.1.html#heading7`. [Accessed: 21-Feb-2021].