BWT-RNA: A preprocessing step for RNA folding algorithms

Ben Chugg Coulter Beeson Kenny Drabble Jeffrey Jeyachandren

1 Introduction and Background

Ribonucleic acids (RNA) play a crucial role in all living organisms, serving both as information storage as well as providing catalytic activity. Given their diverse functions, RNA come in many different varieties, such as mRNA encoding genetic information for translation into proteins, tRNA for the mapping of codons to amino acids, and ribozymes with catalytic activity such as ribosomes and spliceosomes ([2][10][1]), as well as numerous other less understood forms. As opposed to DNA, which is double stranded, RNA is often single stranded and forms - usually complex - three dimensional structures by pairing with itself. As with proteins, the three dimensional structure of RNA is critical to its function, and structural prediction is a natural first step when aiming to ascribe function to a given RNA, as well as in the construction of synthetic sequences with novel properties ([11],[4]).

To predict the three dimensional structure of a given RNA sequence it is often necessary to first determine the secondary structure. RNA, as with proteins, will adopt structure(s) that minimize their total energy. The major stabilizing interaction for RNA comes from their intramolecular base pairing. That is, sequences of similar length base pair internally with other near palindromic sequences [3]. Accordingly, most algorithmic approaches seek to maximize the number of these base pairings. Alternatively some approaches aim to measure other energetic interactions between bases, such as base stacking, and search for a structure of minimal global energy ([6],[9],[2]). Regardless of the approach used, most modern RNA folding algorithms use a similar recurrence that is amenable to dynamic programming.

Although dynamic programming is optimal under a specific scoring model, for many biological applications it is still too slow to be practical. While heuristic approaches are often faster they find only approximate solutions. Base pairing and energy minimization models both use the same recurrence, so any improvements at this level will likely find widespread use. Given that the same recurrence is used in both base pairing and energy minimization models any improvements at this level will likely find widespread use.

Our proposal is to use BWT-SW: a new variant of the Smith-Waterman (SW) local alignment algorithm developed by Lam et. al. [6] that has been optimized using the Burrow- Wheeler Transform (BWT) ([8],[7]). By first considering local alignments of the RNA sequence with itself as a preprocessing step we hope to identify highly probable regions of intramolecular base pairing to limit the search space for subsequent secondary structure prediction techniques.

2 Proposal

To test the efficacy of this approach we intend to implement variant of RNA secondary structure prediction, BWT-RNA, that aims to maximize the number of complementary base pairing interactions. It will carry out local alignment on a single RNA strand by pairing the strand with itself, with the scoring matrix chosen to induce palindromic-like matches. A certain number of the highest scoring matching regions will be assumed to be in a hairpin structure, and the rest of the strand

will then be analyzed using a typical RNA folding algorithm. Local alignment will be implemented with the BWT-SW - a subroutine to speed up Smith-Waterman. The goal BWT-RNA is to prune the search space by pairing obvious regions in order to speed up the execution time of typical RNA folding algorithms.

BWT-RNA will be tested against a control which will implement the typical dynamic programming recurrence used for RNA pairing ([3]). These approaches will be compared both in their ability to find the optimal structure, and on how quickly they can produce structural predictions for sequences of varying length. While the main interest lies in examining the effectiveness of BWT-RNA, we will also implement a version of Smith-Waterman which does not use RNA in order to determine if the BWT significantly speeds up the local alignment preprocessing.

It is possible that the globally optimal structure does not include any of the highly probable regions, and thus our approach may best be considered a heuristic approach. Additionally, the preprocessing may take more time than is saved in the subsequent structure search. To benchmark the efficacy of this approach we will compare the time it takes to analyze datasets of various size both with and without BWT-SW pre-processing.

3 Analyses and Expected Results

Given the complexity of RNA folding, heuristic and approximation algorithms do not easily lend themselves to provable results. Therefore, most of our comparisons will be experimental in nature. However, analytic methods will also be applied in an attempt to estimate runtime analysis and generate approximation ratios.

3.1 Experimental Analysis

In BWT-RNA, there will a question of how many (disjoint) local alignments should be matched as hairpin structures before passing off the data to another folding algorithm. By comparing against known RNA foldings, we will attempt to determine this number experimentally. Presumably, this number will be a function of the strand length and potentially the relative cardinalities of the base pairs. Furthermore, typical RNA folding algorithms are unable to predict pseudoknots ([5]) so we will examine whether our suggested approach is able to do so.

3.2 Analytic Analysis

Of course, the main question is whether BWT-RNA is provably optimal (or provably worse). We will also consider the possibility that it is within an approximation factor of the control. Furthermore, we will attempt to determine the asymptotic difference between local alignment using and not using BWT.

If time allows, we would also like to consider the following question: Given a data set of RNA which is too big to hold on one machine, does BWT-RNA lend itself to a map-reduce model of parallel computing? If so, does it require a randomized approach or can it be done deterministically?

4 Concluding Remarks

It is our prediction that this approach will not be optimal, and would best be considered a heuristic approach. However, this is likely not an issue as most scoring models fail to capture all interactions and are not guaranteed to produce the correct biologically active structure regardless.

References

[1] Jamie H Cate, Anne R Gooding, Elaine Podell, Kaihong Zhou, and et al. Crystal structure of a group i ribozyme domain: Principles of rna packing. *Science*, 273:5282:1678–1685, 1996.

[2] Jennifer A Doudna and Jon R Lorsch. Ribozyme catalysis: not different, just worse. *Nature Structural and Molecular Biology*, 12:395–402, 2005.

[3] Sean R Eddy. How do rna folding algorithms work? *Nature Biotechnology*, 22:1457–1458, 2004.

[4] Zemora Georgeta and Christina Waldsich. Rna folding in living cells. *RNA Biology*, 7.6:634641, 2010.

[5] Gultyaev, Olsthoorn Alexander P, Ren CL. Pleij, Cornelis WA, and Westhof. Rna structure: Pseudoknots. *eLS John Wiley and Sons Ltd, Chichester.*

[6] T.W. Lam, W.K. Sung, S.L. Tam, C.K. Wong, and S.M. Yiu. Compressed indexing and local alignment of dna. *Bioinformatics*, 24:6:791–797, 2008.

[7] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 36:5:589–595, 2010.

[8] Giovanni Manzini. An analysis of the burrows-wheeler transform. *Journal of the Association for Computing Machinery*, 48:3:407–430, 2001.

[9] David H Mathews and Douglas H Turner. Prediction of rna secondary structure by free energy minimization. *Current Opinion in Structural Biology*, 16:3:270–278, 2006.

[10] Markus C Wahl, Cindy L Will, and Reinhard Lurhmann. The spliceosome: Design principles of a dynamic rnp machine. *Cell*, 136:4:701–718, 2009.

[11] Christian Hner zu Siederdissena, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker. A folding algorithm for extended rna secondary structures. *Bioinformatics*, 27, 2011.