BWT-RNA: A preprocessing step for RNA folding algorithms

Ben Chugg Coulter Beeson Kenny Drabble Jeffrey Jeyachandren

Abstract

1 Introduction

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][9][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 ([10],[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 ([5],[8],[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. [5] that has been optimized using the Burrow- Wheeler Transform (BWT) ([7],[6]). 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 Preliminaries

3 Results

4 Analysis

5 Extensions

6 Final Remarks

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] 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.

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

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

[8] 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.

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

[10] 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.