The RNA Folding Problem

Max Ward
School of Computer Science & Software Engineering
The University of Western Australia

May 30, 2014

Abstract

Ribonucleic Acid (RNA) is an important biological molecule with myriad functions. For example, it drives developmental processes, regulates the expression of genes, enables to synthesis of proteins, and catalyses important biological reactions. Because of this, algorithms for predicting RNA structures have been proposed as early as the 1970s. The problem they attempt to solve is called the RNA folding problem. In this report, I formally describe a simplified version of this problem, which I call the RNA bond problem. I show that this is equivalent to finding the maximum weight independent set of a circle graph, which is a common problem in VLSI design, bioinformatics, and register allocation for optimizing compilers. I then describe the Nussinov algorithm, which efficiently solves the RNA bond problem. Finally, I compare this algorithm to existing algorithms for solving the maximum weight independent set of a circle graph. It is noteworthy that nobody has recognized that these problems are equivalent, this highlights the issues associated with cliquing within scientific communities.

Keywords: Ribonucleic acid, structure, prediction, empirical, comparison.

CR Classification: J.3 Biology and genetics.

1 Introduction

Ribonucleic acid (RNA) is at the core of many biological processes. Traditionally it has been described as the messenger molecule of DNA, faithfully carrying code from DNA to the site of protein synthesis. However, in a recent landmark paper, Amaral et al. [1] described our genome, and those of other eukaryotes, as being driven by a RNA machine. They noted that most of the eukaryote genome is transcribed into RNA, despite little of it coding for protein. It seems that much of our genome, originally called 'junk DNA', codes for functional RNA molecules. These RNAs can interact with DNA, affecting gene expression. This allows DNA to essentially regulate itself. For example, Makeyev & Maniati [8] reported that microRNAs affect the expression of genes by interfering with the translation of protein. They also argued that microRNAs, and other regulatory RNAs, explain the vast differences between organisms with similar genomes. To put this idea into perspective, we share roughly 90% of our genes with the domestic cat [14]. Mattick [11] has suggested that the process of development—from embryo to adult—is encoded in the interactions of such RNAs.

A widely held axiom is that chemical structure is tantamount to biological function. With increasingly important biological functions being associated with RNA, it is essential to accurately predict its structure. The purpose of this paper is to provide a survey of some widely used RNA structure prediction algorithms. In the interest of succinctness, I review only algorithms based on a thermodynamic model. Other algorithms often use machine learned parameters; these shall not be explored here as they represent fringe areas of research. In essence, all the algorithms I test take a single RNA primary sequence as input, and produce a predicted structure as output. I hypothesize that newer algorithms should have improved accuracy compared to older algorithms. In addition, I aim to empirically verify the time complexities of tested algorithms. The Zuker algorithm was the first thermodynamic algorithm to achieve usable prediction accuracy, and is the oldest algorithm I tested.

1.1 The RNA Bond Problem

1.1.1 Description

Before the RNA bond problem can be explained, some terminology and background information must be covered. A RNA molecule comprises a sequence of nucleotides connected in sequence. This is often described as a chain. The four RNA nucleotides are Adenosine (A), Uracil (U), Cytosine (C), and Guanine (G). Chemical bonds form between Adenosine and Uracil, such a bond is called an A-U bond. Other valid bonds are G-U, and G-C. Every bond causes the nucleotide chain to fold on itself. Successive folds lead to the complex, paper-clip like structures common to RNA molecules. For the sake of illustration, let use imagine a RNA nucleotide chain as being connected end to end, thus forming a circle. An example of this is shown visually in Figure **??**. A valid bond is any chord crossing from one nucleotide to another nucleotide, such that they form a chemically valid bond (A-U, G-C, or G-U). In addition, a valid bond must not cross any other chord in the circle. Finally,

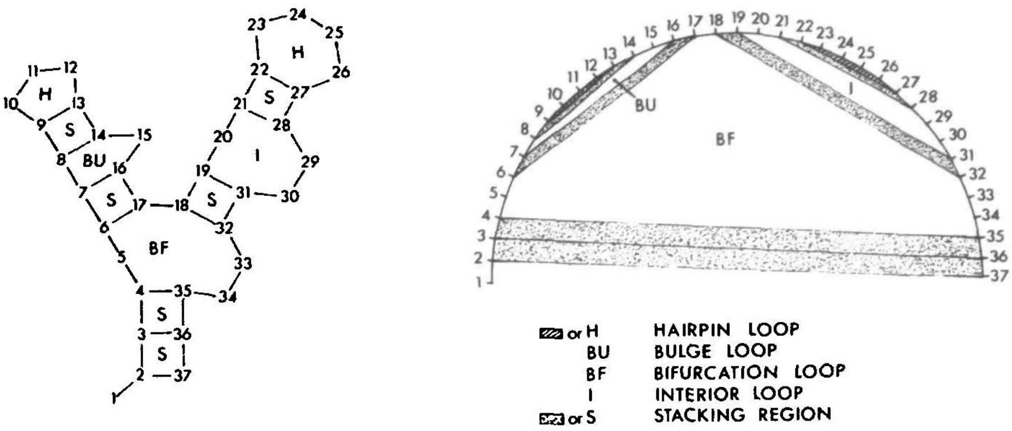

Figure 1: Substructures used in the Zuker algorithm. On the left is a diagram of a RNA structure. On the right is the same structure laid out on a semi-circle. Bonds are represented as lines crossing the semi-circle. Taken from original publication [19].

every bond has a weight, which indicates the strength of the bond. Examples of valid and invalid bonds can be seen in Figure **??**. Given a RNA sequence, the RNA bond problem involves finding a set of valid bonds having maximum weight. In other words, we must find a collection of mutually valid chords such that their sum weight is maximized. I shall now outline an algorithm which solves this problem.

1.1.2 The Nussinov Algorithm

In 1973 Nussinov & someass described an algorithm that can efficiently find solutions to the RNA bond problem. However, this is not by design. The algorithm was originally introduced to solve the RNA folding problem. The intuition is that every bond increases the stability of the RNA's structure, so a structure with maximum bonds should be very stable. The Nussinov algorithm is no longer used in practice, as better algorithms now exist for solving the RNA folding problem. I describe it here only because it solves the RNA bond problem.

The nucleotides which make up a RNA sequence can be indexed from zero to $n-1$ where n is the length of the RNA sequence. In addition, let the weight of a bond between nucleotides i and j be defined by $W(i,j)$. Also, the function $M(i,j)$ returns the solution to the RNA bond problem considering only nucleotides between i and j inclusive. This function is defined by the recurrence relation presented below.

$$
\begin{aligned}
M(i,j) &= \max\{A, B, C, D\} \\
A &= M(i, j-1) \\
B &= M(i+1, j) \\
C &= M(i+1, j-1) + W(i,j) \\
D &= \max\{M(i,k) + M(k+1,j)\} \ for\, all\, k\, where\, i < k < j
\end{aligned} \tag{1}
$$

The recurrence relation in Equation 1 can be used to solve the RNA bond problem by calling $M(0, \texttt{RNA\_Length} - 1)$. The first two cases ($A$ and B) find the score associated with not allowing the bases corresponding to indexes i and j to bond. Case C conversely determines the score given that i and j are bonded. The final case D computes the score associated with a bifurcation. A bifurcation is the decomposition of RNA into two separate structures. This recurrence relation is solved using dynamic programming. As such, it implies a $O(N^3)$ worst case time complexity and a $O(N^2)$ space complexity, as a $O(N^2)$ state space (all combinations of i and j) is explored with a linear time recurrence relation.

2 Maximum Weight Independent Sets For Circle Graphs

2.1 Description

All software was run on the Debian 7.5 ”wheezy” operating system using the default configuration. Debian was run on top of an Intel i7-4770k processor with 32 gigabytes of RAM. The GNU C Compiler version 4.8.2 was used to compile all the required code. Though the processor used was multicore, OpenMP was disabled at compile time to prevent the use of multiple cores during testing. The source code for all the algorithms tested was compiled using makefiles provided as part of their source. The latest version of the ViennaRNA suite [7] (version 2.1.7, released April 13th 2014 [6]) was used as the reference implementation of a MEA algorithm and the Zuker algorithm. This was because it contains widely used versions of both. The module in ViennaRNA that implements the Zuker algorithm is called RNAfold, and I shall hereafter refer to it as such. Likewise, the module for MEA is simply called MEA. The entire ViennaRNA suite was compiled from source, then was linked as a static library at compile time for testing. The latest version of CoFold was downloaded from the CoFold webserver [15]. Because CoFold is based on an older version of RNAfold, it was compiled separately, and linked as a separate static library. In addition, the GNU Regression, Econometrics, and Time-series Library [3] was used for all statistical tests, and to produce accompanying graphics.

2.2 Graphs and RNA

The RNA structures used to test algorithms presented in this paper were taken from the RNA STRAND database [2]. The RNA STRAND database is a free-to-use, curated collection of RNA structures taken from various publicly available databases and publications. A subset of RNA structural data was extracted from the database. This subset contained only RNA structures that were marked as having been verified using X-ray crystallography, or nuclear magnetic resonance imaging. It also comprises only whole RNAs; none of the RNAs used were fragments or subsequences of larger RNA molecules. Finally, no duplicates were allowed in the selected set. Hereafter, I shall refer to this collection of RNA structures as the 'testing set'. The testing set contained 392 different RNA molecules ranging in length from 20 to 3032 nucleotides.

3 Conclusions

I have made the conjecture that newer algorithms should be more accurate at predicting RNA structures. The algorithms I have tested, in decreasing order of age, were the Zuker algorithm, MEA based algorithms, and CoFold. I ran several statistical tests to compare the F-scores of these algorithms. In every test, no statistically significant difference was found. There are only two reasonable conclusions, that the testing set was insufficiently large, or that no algorithm is significantly more accurate than the others. The testing set was large (392 RNAs), and varied, with many different classes of RNAs represented. As such, I submit that there is little difference in the predictive power of the algorithms tested. Unfortunately I was only able to test a very sparse collection of RNA prediction algorithms. Having said this, the Zuker algorithm and MEA based techniques are widely used. Additionally CoFold is a new algorithm, and represents state-of-the-art. Hence, I also submit that my findings are relevant to both research and practice. Nonetheless, further investigations should use a larger collection of algorithms, and implementations. Stochastic context free grammar based methods in particular show promise. Furthermore, if a larger testing set is available, it should be used.

A secondary goal of this investigation was to test the time requirements of the selected algorithms. While MEA and CoFold appear to have much larger constant factors compared to RNAfold, they all have a curve that is visibly cubic. This supports claims of asymptotic complexity. The Zuker algorithm has had decades of optimization, and the RNAfold implementation in particular is well optimized. As such, this may be partly an optimization discrepancy. A more detailed investigation might compare naive and optimized implementations of RNA folding algorithms.

I had supposed that one would see a progression of accuracy with newer algorithms. This was not the case. Despite this, increasingly large constant factors are evident in newer algorithms. These findings constitute compelling evidence that little progress is being made in algorithmic RNA prediction. Unfortunately, it is increasing clear that RNA has a complex and fundamental role in biology. For the time being, the most practical algorithm appears to be the Zuker algorithm—as implemented in RNAfold. It provides the best balance between predictive power and run-time efficiency. However, better algorithms are an excellent direction for future research, as better predictive power will be invaluable for understanding RNA.

References

[1] Paulo P Amaral, Marcel E Dinger, Tim R Mercer, and John S Mattick. The eukaryotic genome as an rna machine. *Science*, 319(5871):1787–1789, 2008.

[2] Mirela Andronescu, Vera Bereg, Holger H Hoos, and Anne Condon. Rna strand: the rna secondary structure and statistical analysis database. *BMC bioinformatics*, 9(1):340, 2008.

[3] Giovanni Baiocchi and Walter Distaso. Gretl: Econometric software for the gnu generation. *Journal of Applied Econometrics*, 18(1):105–110, 2003.

[4] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

[5] Fred Russell Kramer and Donald R Mills. Secondary structure formation during rna synthesis. *Nucleic acids research*, 9(19):5109–5124, 1981.

[6] Ronny Lorenz. Viennarna package. `http://www.tbi.univie.ac.at/~ronny/RNA/index.html`. Accessed: 2014-05-12.

[7] Ronny Lorenz, Stephan HF Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, Ivo L Hofacker, et al. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[8] Eugene V Makeyev and Tom Maniatis. Multilevel regulation of gene expression by micrornas. *Science*, 319(5871):1789–1790, 2008.

[9] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–7292, 2004.

[10] David H Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *Journal of molecular biology*, 288(5):911–940, 1999.

[11] John S Mattick. A new paradigm for developmental biology. *Journal of Experimental Biology*, 210(9):1526–1547, 2007.

[12] John S McCaskill. The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

[13] Steven R Morgan and Paul G Higgs. Evidence for kinetic effects in the folding of large rna molecules. *The Journal of chemical physics*, 105(16):7152–7157, 1996.

[14] Joan U Pontius, James C Mullikin, Douglas R Smith, Kerstin Lindblad-Toh, Sante Gnerre, Michele Clamp, Jean Chang, Robert Stephens, Beena Neelam, Natalia Volfovsky, et al. Initial sequence and comparative analysis of the cat genome. *Genome research*, 17(11):1675–1689, 2007.

[15] Jeff Proctor. Cofold webserver. `http://www.e-rna.org/cofold/download.cgi`. Accessed: 2014-05-12.

[16] Jeff R Proctor and Irmtraud M Meyer. Cofold: an rna secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic acids research*, 41(9):e102–e102, 2013.

[17] Jessica S Reuter and David H Mathews. Rnastructure: software for rna secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):129, 2010.

[18] Gary M Studnicka, Georgia M Rahn, Ian W Cummings, and Winston A Salser. Computer method for predicting the secondary structure of single-stranded rna. *Nucleic acids research*, 5(9):3365–3388, 1978.

[19] Michael Zuker and Patrick Stiegler. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, 1981.