# Algorithms for RNA folding

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

May 7, 2014

**Abstract**

Analysis of thermodynamically based algorithms for RNA structure prediction

**Keywords:** Ribonucleic acid, structure, prediction, thermodynamic hypothesis.

**CR Classification:** J.3 Biology and genetics.

# 1 Introduction

## 1.1 Motivation

Ribonucleic acid (RNA) is at the core of many biological processes. Traditionally is has been described as the messenger molecule of DNA, faithfully carrying the code for protein 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 an 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 [5] reported that microRNAs affect the expression of genes by interfering with 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 [10]. Mattick [8] 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 important to be able to predict its structure. The purpose is this paper is to provide a survey of some widely used RNA structure prediction algorithms. In the interest of keeping this report succinct, I review only algorithms based on a conventional thermodynamically based model. Other algorithms often use machined learned, statistical parameters; these shall not be explored here. The Zuker algorithm was the first thermodynamic algorithm to achieve usable prediction accuracy, and it forms the basis for all the methods I shall thence discuss.

## 1.2 Relevant Algorithms

### 1.2.1 The Zuker Algorithm

In 1981 Zuker & Stiegler [13] described an algorithm that predicted RNA structure by minimizing molecular free energy. The lower free energy a molecule has, the more stable it is. This was done by introducing a number of thermodynamic rules for canonical substructures like hairpin loops, internal bulges, multiloops (also called bifurcation loops), unbonded bases, and stacked base pairs. For examples of these structures, refer to Figure 1. The Zuker algorithm uses a mutually recursive dynamic programming recurrence to achieve a relatively comprehensive scoring scheme. This is possible because RNA structures are nested, and hence substructures with minimum free energy need not be recomputed. It follows naturally that dynamic programming can then be used to find a global minimum. The thermodynamic scoring system is borrowed from the work of Studnicka et al. [12] who presented a complex but theoretically similar algorithm, albeit having much worse asymptotic and implementation complexity.

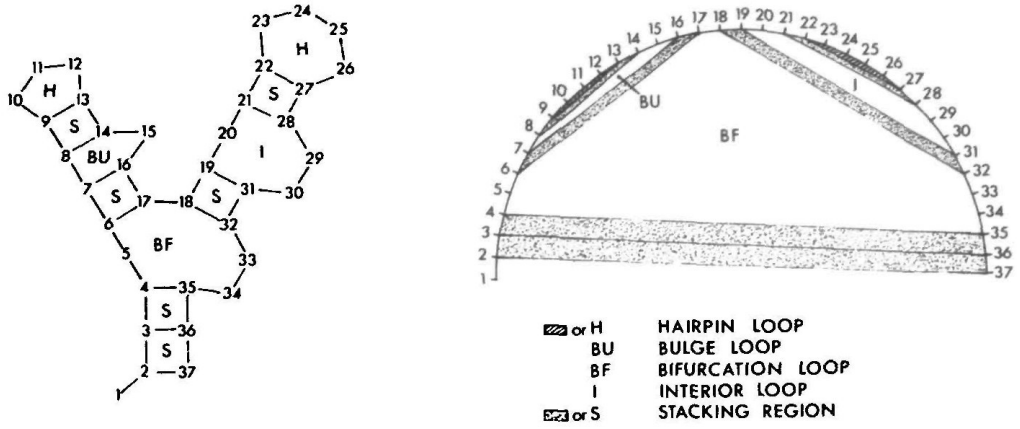


**Figure 1:** Diagram of substructures used in the Zuker algorithm. On the left is a diagram of a RNA secondary structure. On the right is the same structure laid out on a semi-circle. Bonds are represented as lines crossing the semi-circle. Taken from original publication [13].

Later, Matthews et al. [7] improved the standard thermodynamic model used by adding experimentally determined sequence dependencies. Then again in 2004, also by Matthews et al [6], this time by incorporating energy parameters for coaxial stacking of base pairs, and by refining previous parameters based on newer experiments. In both cases, the prediction accuracy of the Zuker algorithm was empirically improved. Under all of these incrementally improving thermodynamic models, the time complexity of the Zuker algorithm is $O(n^3)$, where $n$ is the number of nucleotides in the RNA. The space complexity for the Zuker and algorithm is $O(n^2)$, this is due to the memory used by dynamic programming tables. In short, it is able to efficiently predict secondary structures for RNAs with hundreds of bases, on modern hardware. Because of its efficiency, robustness, and extensibility, this method is, even today, still the most popular available. The most widely used packages for RNA secondary structure prediction all contain implementations of the Zuker algorithm [4, 11].

### 1.2.2 Maximum Expected Accuracy

The Maximum Expected Accuracy (MEA) technique for RNA prediction was first applied by Do, Woods & Batzoglou as part of the CONTRAfold algorithm [3]. The Zuker algorithm essentially finds the most likely structure under a thermodynamic model. An MEA prediction method is more compromising; it finds a structure containing the most probable set of bonds. This is done by computing the partition function for a RNA, as described by McCaskil [9]. The partition function computation is achieved by a modified version of the Zuker algorithm, and can be done in $O(n^3)$ time. It should be noted that the this computation uses the underlying thermodynamic model of the Zuker algorithm. This partition function can be processed to produce a two dimensional matrix of base pairing probabilities. Each pair of indexes in the matrix represents the probability that the bases corresponding to

those indexes will bond. Intuitively this represents the entire folding landscape of a RNA, rather than a single minimum free energy structure as computed by the Zuker algorithm. A MEA algorithm builds a structure by selecting bonds from this bond probability matrix. However, if the sum probability is naively maximized, it tends to produce structures with too many bonds. As such, a MEA algorithm must find a compromise between the number of bonds selected, and the total probability of the computed structure. Usually this is achieved by introducing some cut-off value, or, in the case of CONTRAfold, a scaling factor.

#### 1.2.3 Cotranscriptional folding

Explain in high level terms the three algorithms tested

# 2 Materials and Methods

## 2.1 Environment and Software

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 (gcc) 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 Vienna RNA Suite (version 2.1.7, released April 13th 2014) was used as the reference implementation of the MEA algorithm and the Zuker algorithm. The RNAfold module in particular can be used to fold structures using the Zuker algorithm, or to produce a structure with maximum expected accuracy. The entire Vienna RNA 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 (ref). 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 (also called 'gretl') was used for all statistical tests, and to produce the graphics found in Section 3.

## 2.2 Data Set

The RNA secondary 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 secondary 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 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

| Algorithm | Mean | Median | Std. Dev. |
|---|---|---|---|
| RNAfold | 0.574834 | 0.608696 | 0.341977 |
| MEA | 0.564584 | 0.615385 | 0.346070 |
| CoFold | 0.580802 | 0.608696 | 0.337973 |

**Table 1:** Summary statistics for the F-scores of RNAfold, MEA, and CoFold.

selected set. Hereafter, I shall refer to this collection of RNA secondary structures as the 'testing set'. The testing set contained 392 different RNA molecules ranging in length from 20 to 3032 nucleotides.

## 2.3 Accuracy

Well established procedures have been developed over the history of RNA structure prediction for comparing accuracy. Usually accuracy is determined by comparing predicted structures to known structures. True Positives ($TP$) is defined as the number of base pairs which appear in both the predicted structure and the actual structure. False Positives ($FP$) is the number of predicted base pairs not in the true structure [4]. Similarly, False Negatives ($FN$) is defined as the number of base pairings in the reference structure but not present in the predicted structure [4]. Sensitivity, also called the True Positive Rate ($TPR$), can be defined using the previously introduced values. I have given a mathematical definition of the $TPR$ in Equation 1.

$$\frac{TP}{TP + FN} \tag{1}$$

Precision, also known as Positive Predictive Value ($PPV$), can also be calculated using these values (see Equation 2).

$$\frac{TP}{TP + FP} \tag{2}$$

For the results presented in this report, the F-score was used as a measure of accuracy. F-score is defined as the harmonic mean of sensitivity and precision. It provides a good balance between these two metrics.

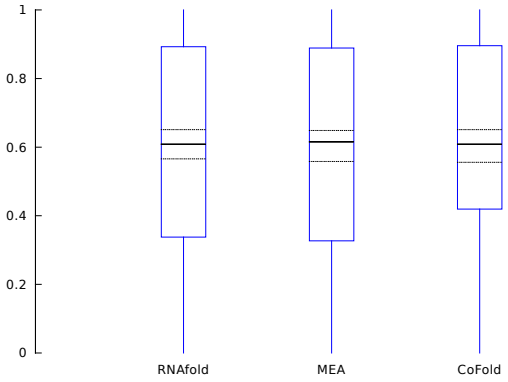


**Figure 2:** Box plots depicting the spread of F-scores for RNAfold, MEA, and CoFold.

| Algorithms | $z$-value | P-value |
|---|---|---|
| MEA & CoFold | -0.225675 | 0.410727 |
| RNAfold & CoFold | -0.85571 | 0.196079 |
| RNAfold & MEA | -0.790628 | 0.214581 |

**Table 2:** The results of Wilcoxon Signed-rank tests between RNAfold, MEA, and CoFold. The $z$-value indicates the magnitude of difference between F-scores. The P-value indicates statistical significance.

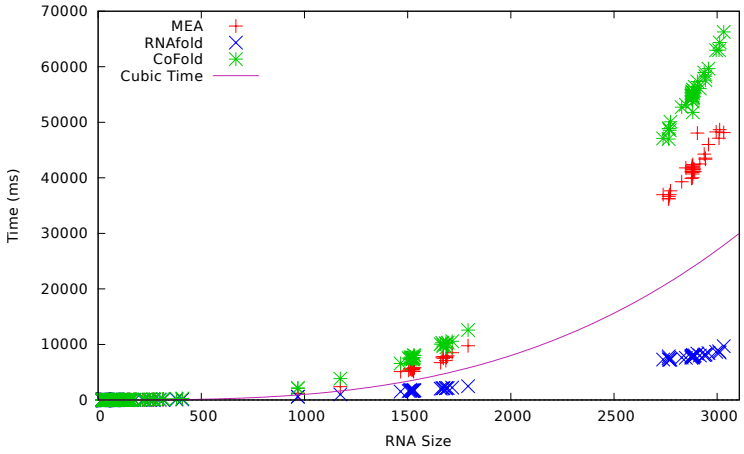


**Figure 3:** Here the recorded run time for RNAfold, MEA, and CoFold is depicted as a scatter plot. The time used by all algorithms increases with increasing RNA length. The length of an RNA is defined as the number of nucleotides it compromises. A purple line representing a generic cubic curve is included for comparison. This line has been scaled by a constant so that it remains visible on the diagram.

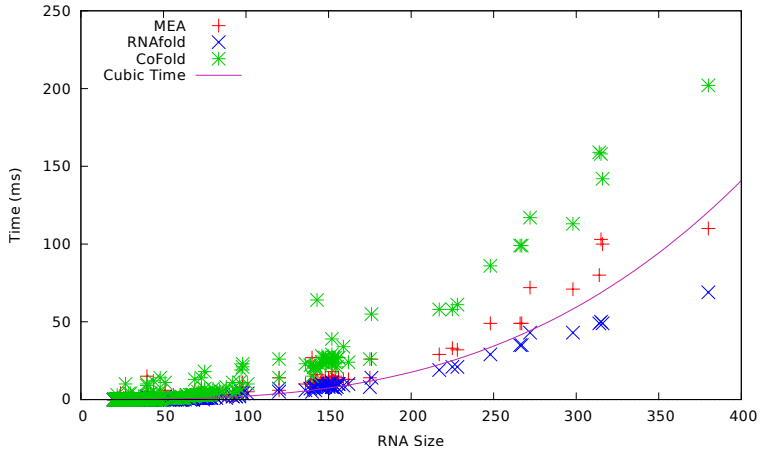


**Figure 4:** Here the recorded run time for only smaller RNAs ($< 400$ nucleotides) is depicted as a scatter plot. A purple line representing a generic cubic curve is included for comparison. This line has been scaled by a constant so that it remains visible on the diagram.

# 3 Results

## 3.1 Accuracy

### 3.1.1 Small RNA

### 3.1.2 Moderate RNA

### 3.1.3 Large RNA

## 3.2 Time

# 4 Discussion and Conclusions

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

[4] 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.

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

[6] 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.

[7] 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.

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

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

[10] 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.

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

[12] 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.

[13] 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.