RNA Folding: Beyond The Thermodynamic Hypothesis

Max Ward (supervised by Amitava Datta)

Computer Science and Software Engineering, University of Western Australia

Introduction & Motivation

Ribonucleic Acid (RNA) is a biologically active molecule with many poorly understood functions. For example, it is:

- Involved in regulation of DNA expression
- Implicated in developmental pathways
- A catalyst for many biological processes

Quick and accurate prediction of RNA structure is therefore essential. RNA folding algorithms are currently not able to reliably predict correct structures. These algorithms typically globally optimize some scoring function, usually thermodynamic stability. However, there is evidence that many RNAs fold into suboptimal states.

My Contribution

I hypothesized that local interactions are stronger than global interactions during RNA structure formation. To test this, a sliding window was used to generate locally optimal structures, then various algorithms were devised to merge these substructures. For some window sizes, the resulting complete structures were more accurate than those predicted by state of the art algorithms, which globally optimize. This constitutes strong support for my hypothesis. In addition, a new algorithm called '*ab*-splat', which was based on the computation of locally optimal windows, was introduced.

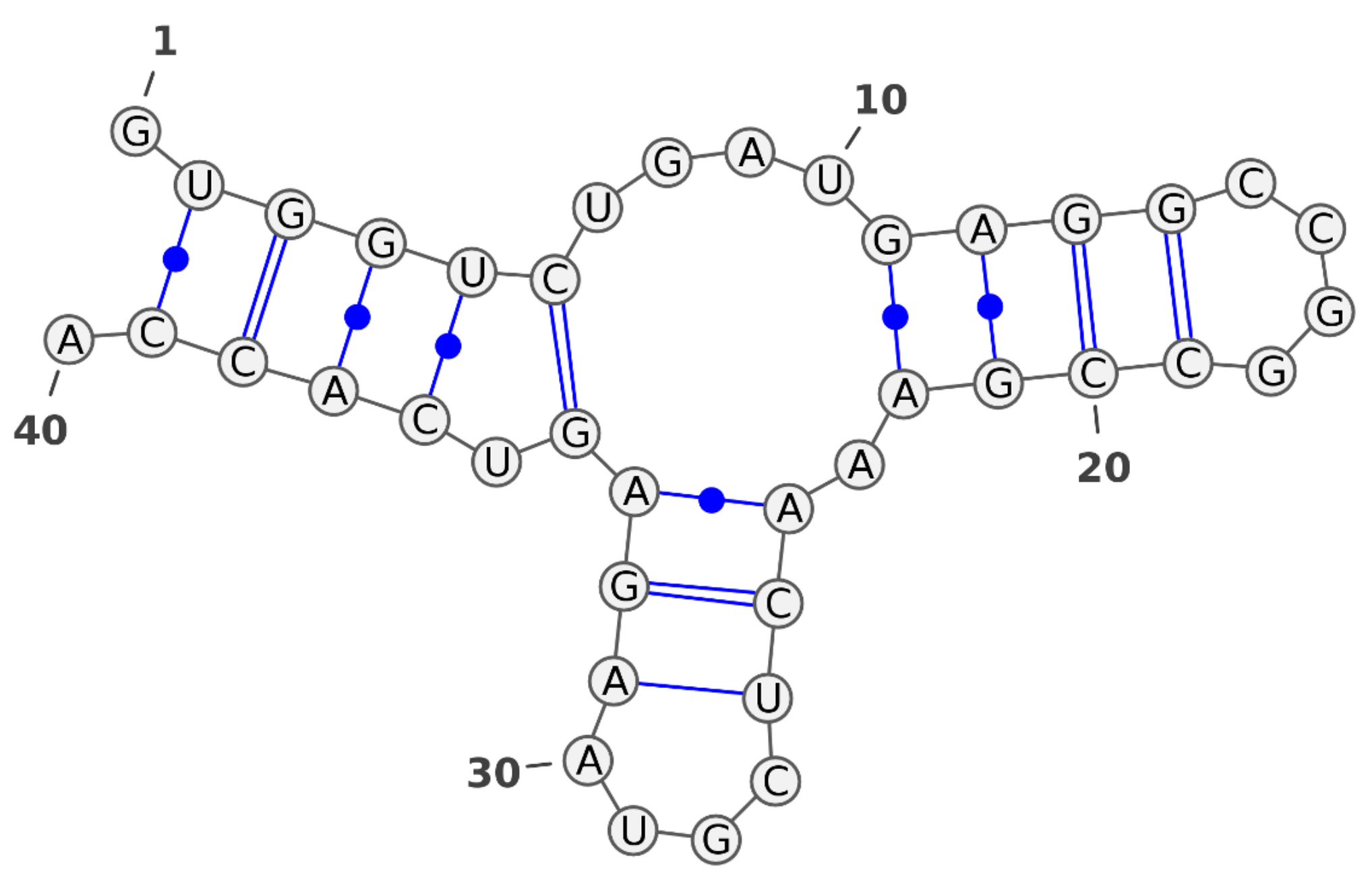

Figure 1: A RNA molecule

Sliding Windows

Hofacker, Priwitzer, and Stadler [1] devised an algorithm to compute the structure of consecutive windows over a RNA in $O(nL^2)$ time, where n is RNA length and L is window size. This is useful when searching DNA (which can be very large) for RNA structural motifs.

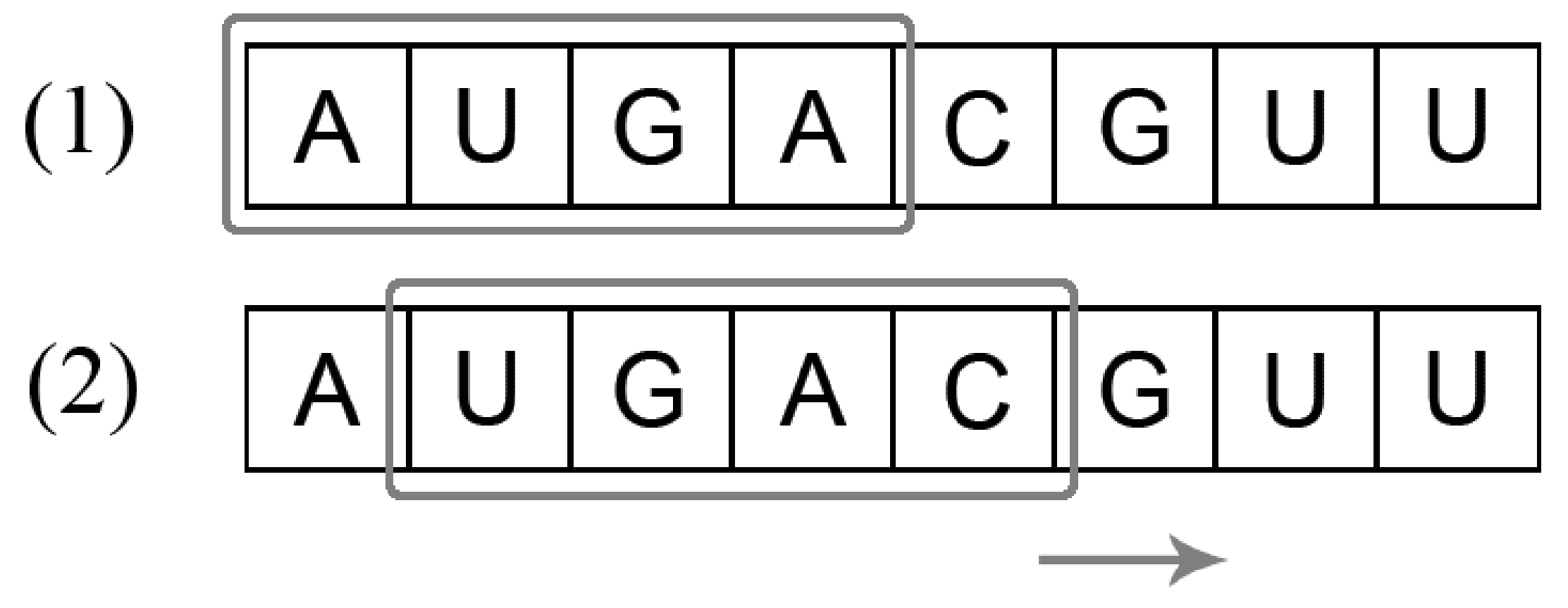

Figure 2: A sliding window over a RNA sequence

Weighted Activity Selection

The weighted activity selection problem is a classical optimization problem, and is defined as follows: Given a set of weighted intervals on a line, find a subset of intervals such that none overlap, and the sum of their weights is maximized. Such a subset can be found in $O(n \log n)$ time using dynamic programming.

Testing Local Interactions

To test if local interactions were stronger than global interactions, I examined window sizes 5 to 500. For every window size:

- The sliding window algorithm was run. The structures computed for every window position were saved in a set S.
- Every structure in the set S was assigned a score. This score was defined as the free energy of the structure. The lower the free energy of a structure, the more stable it is.
- Weighed activity selection was done on the set S to create a structure with minimum sum free energy.

This procedure was run on a large corpus of test RNA molecules whose structure had been experimentally determined. The single most accurate prediction was recorded.

Local Versus Global

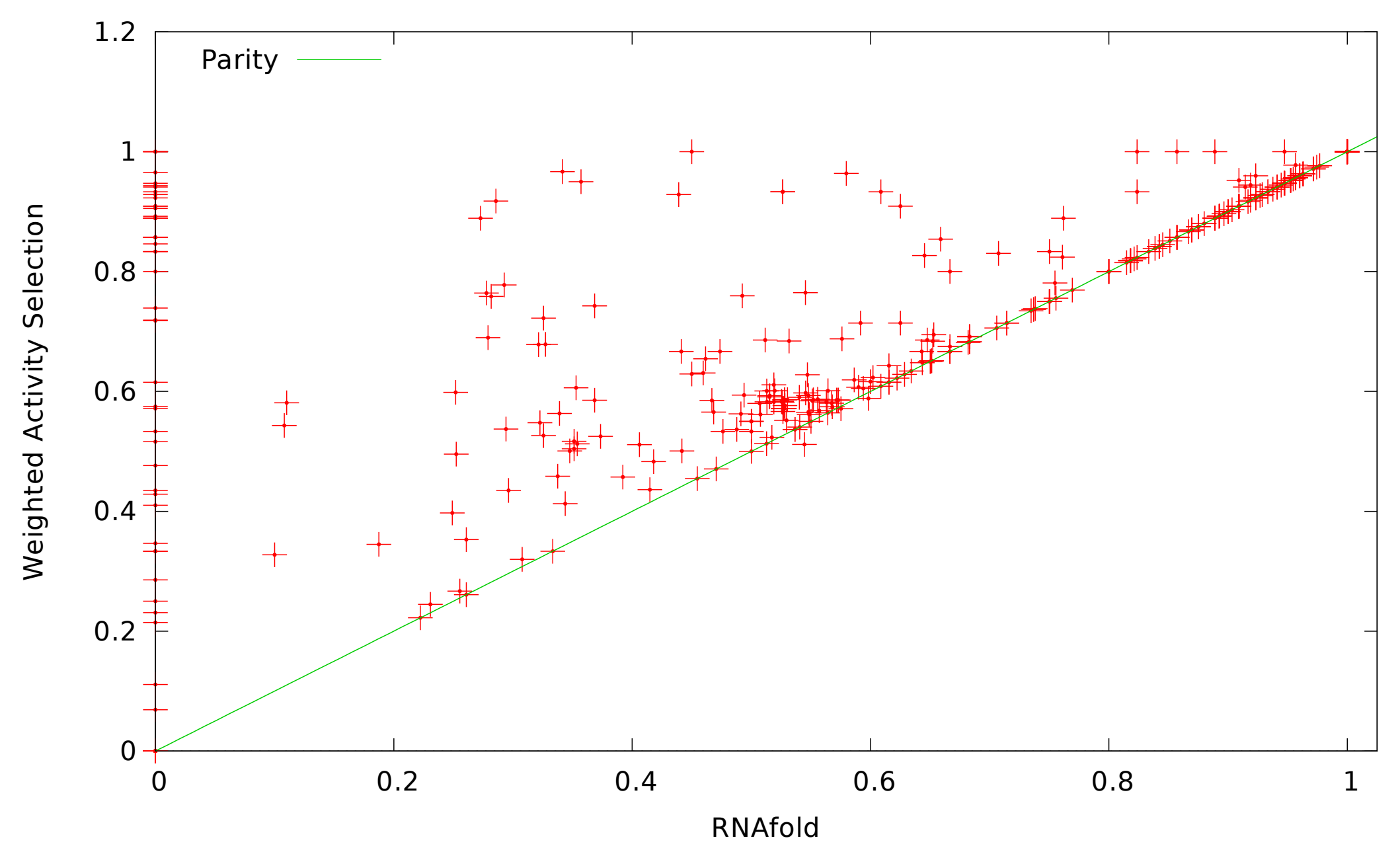

Figure 3: Scatter plot of the best locally optimized structure, versus the prediction of RNAfold

The Zuker algorithm, first described in 1981 [2], is the most widely used RNA folding algorithm today. Because it has been continually refined, it is considered state of the art. In addition, the Zuker algorithm finds a structure with minimum free energy, it globally optimizes a thermodynamic scoring function. RNAfold [3] is a modern implementation of the Zuker algorithm, and was chosen for comparison. As can be seen in Figure 3, the procedure I described earlier was consistently more accurate than RNAfold. This result was verified by a Wilcoxon signed-rank test ($p < 0.001$). This strongly supports my hypothesis. Because there exist window sizes for almost all tested RNA which are more accurate than the global optimum, local interactions appear to be more important than global interactions during RNA folding.

The *ab*-splat Algorithm

I also created the *ab*-splat algorithm, which is a complete structure prediction algorithm. It utilizes locally optimal sliding windows, and weighted activity selection.

```
Let $RNA$ be the primary RNA sequence being folded
Let $i = a$
Let $S = \{\}$
while $i \leq$ TRESHOLD($RNA$) do
    $S = S \cup$ COMPUTESLIDINGWINDOW($RNA$, $i$)
    $i = i \times b$
end while
return WEIGHTEDACTIVITYSELECTION($S$)
```

Testing *ab*-splat

Machine learning was used to find good values for a and b ($a = 24$, $b = 1.8$). In addition, an effective `Threshold` function ($9.5 \times \sqrt{RNA\_Length}$) was found through empirical testing. Using this `Threshold` function, the time complexity of *ab*-splat is $O(n^2)$, where n is RNA length. This is an order of magnitude faster than the Zuker algorithm, which requires $O(n^3)$ time. Again, RNAfold was used for comparison.

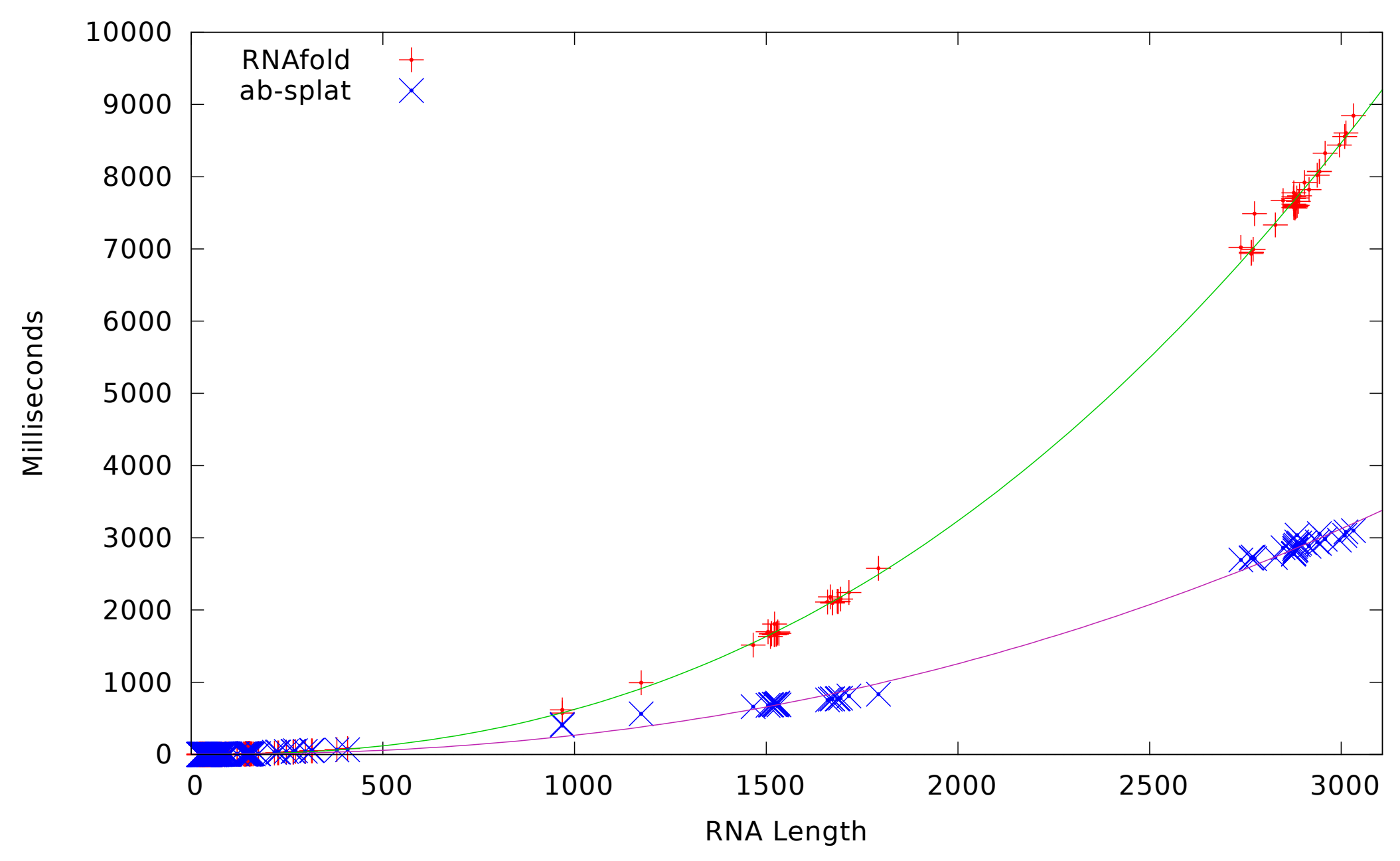

Figure 4: Time taken by RNAfold (red), and by *ab*-splat (blue)

A Wilcoxon signed-rank test showed that there was no statistically significant different in accuracy between RNAfold and *ab*-splat. This implies that both algorithms are equally accurate. However, the *ab*-splat algorithm was much faster in theory, and in practice (see Figure 4).

References

[1] Hofacker, I. L., Priwitzer, B., and Stadler, P. F. Prediction of locally stable rna secondary structures for genome-wide surveys. *Bioinformatics* **20**, 2 (2004), 186-190

[2] Zuker, M., and Stiegler, P. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Research* **9**, 1 (1981), 133-148.

[3] Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., Hofacker, I. L., et al. Viennarna package 2.0. *Algorithms for Molecular Biology* **6**, 1 (2011), 26.

THE UNIVERSITY OF
WESTERN AUSTRALIA