The RNA Bond Problem

Max Ward

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

June 1, 2014

Abstract

Ribonucleic Acid (RNA) is an important biological molecule with myriad functions. For example, it drives developmental processes, regulates the expression of genes. Because of this, algorithms for predicting RNA structures have been proposed. 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) performs myriad important biological functions. Some recently discovered and notable examples include its ability to regulate the expression of certain genes [3]. As a result of this, it has been implicated as the driving molecule behind developmental processes [3]. Because function is determined by structure, algorithms for predicting the structure of RNA molecules have been under active investigation since the early 1970s. This problem is formally called the RNA folding problem. I describe a simplified version of this which I call the RNA bond problem. Furthermore, I introduce the Nussinov algorithm, which solves the RNA bond problem. Finally, I show that the RNA bond problem is actually equivalent to finding the maximum weight independent set of a circle graph. The implications of this are then discussed.

2 The RNA Bond Problem

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, 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.

2.1 The Nussinov Algorithm

In 1978 Nussinov et al. [4] 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. Nussinov et al.'s intuition was that every bond increases the stability of a RNA's structure, so a structure with maximum bonds should be maximally 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 in Equation 1.

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

(1)

This function can solve the RNA bond problem, $M(0, \texttt{RNA\_Length} - 1)$. 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.

3 Maximum Weight Independent Sets and Circle Graphs

An independent set of a graph is any collection of vertices such that none share an edge. The maximum weight independent set of a graph requires that vertices have weights. It is defined as any independent set such that the sum weight of vertices is maximal. A circle graph is a special kind of graph derived from a circle diagram. A circle diagram is easy to visualize. Imagine a circle and a collection chords crossing the inside edge of the circle. This is a circle diagram. A circle graph is derived from a circle diagram using a simple procedure. Every chord is represented by a vertex. Two vertices share an edge if and only if the chords they represent intersect in the circle diagram. This is best explained using a diagram, as such, I have included Figure **??**. Finding the maximum weight independent set of a circle graph is equivalent to finding a set of non-intersecting chords with maximum sum weight in the circle diagram.

Finding the maximum weight independent set of a circle graph has several practically important applications. For example, it is used in register allocation [2], VLSI design [1], and bioinformatics [5].

3.1 Graphs and RNA

I now wish to point out that the RNA bond problem is isomorphic to finding the maximum weight independent set of a circle graph. This is the core observation of this report. If we imagine a RNA joined end to end in a circle, and all of the possible valid bonds represented as chords across the circle, the solution to the RNA bond problem is any subset of chords that do not intersect, and which has maximum sum weight. This is, by definition, also a maximum weight independent set. As such, the Nussinov algorithm finds the maximum weight independent set of a circle graph.

Borrowing terminology from literature on graph theory, I shall refer to the number of distinct points on a circle graph as n, and the number of chords as m. In other words, there are m chords, and there are n distinct points where a chord or chords touch the interior edge of the circle diagram. Thus, we say that the Nussinov algorithm takes $O(n^3)$ time. The Nussinov algorithm was published in 1978. The most optimal algorithm for finding the maximum weight independent set of a circle graph at the time required $O(m^3)$ time, and was discovered in 1978. This is inferior to the Nussinov algorithm, as m is bounded from below by $n/2$. It wasn't until 2003 that a clearly better algorithm was found. This algorithm required $O(md)$ time, and was discovered by Valiente [6]. Note that d represents the density of the graph, but the precise complexity analysis is not important for this discussion. Despite the Nussinov algorithm being an asymptotically superior algorithm for 25 years, it was never used outside of RNA research.

4 Conclusions

I have briefly introduced the RNA folding problem. Following this, a precise definition of a simpler problem, the RNA bond problem, was described. The Nussinov algorithm is an relatively simple and reasonably efficient solution to this problem. In addition, I have showed that the RNA bond problem, and finding the maximum weight independent set of a circle graph, is equivalent. Thus the Nussinov algorithm solves both problems. In addition, it is more optimal than the other algorithms that solve the same problem at the time of its publication. This has not been recognized until now. I submit that this illustrates the danger of over specialization within scientific communities. I also conjecture that this situation would have been avoided with a more holistic approach to research. All knowledge is valuable; insular research has no place in the advancement of science

References

[1] Jason Cong and CL Liu. Over-the-cell channel routing. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 9(4):408–418, 1990.

[2] Dominique de Werra, Ch Eisenbeis, Sylvain Lelait, and Bruno Marmol. On a graph-theoretical model for cyclic register allocation. *Discrete Applied Mathematics*, 93(2):191–203, 1999.

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

[4] Ruth Nussinov, George Pieczenik, Jerrold R Griggs, and Daniel J Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied mathematics*, 35(1):68–82, 1978.

[5] Krister M Swenson, Yokuki To, Jijun Tang, and Bernard ME Moret. Maximum independent sets of commuting and noninterfering inversions. *BMC bioinformatics*, 10(Suppl 1):S6, 2009.

[6] Gabriel Valiente. A new simple algorithm for the maximum-weight independent set problem on circle graphs. In *Algorithms and Computation*, pages 129–137. Springer, 2003.