CITS4008 Assignment 2

Max Ward

April 3, 2014

Abstract

RNA is the messenger molecule for DNA. It also plays a vital role in cellular metabolism. As a result, it is valuable to predict the secondary structure of RNA molecules. I have found that the RNA folding problem is equivalent to finding the maximum-weight independent set of a circle graph. In addition the Nussinov algorithm, which is used to fold RNA, can be used to solve the maximum-weight independent set problem, and is competitive with similar algorithms.

RNA and DNA

DNA is a double helix molecule that codes for proteins used by cells [1]. This code, which can be thought of as the 'digital' representation for the 'analogue' protein used by our cells, must be carried to ribosomes which translate it into protein [1]. This is a task carried out by Ribonucleic Acid (RNA). RNA is much like DNA as its chemical structure is similarly composed of Guanine, Adenine, and Cytosine. The fundamental difference, however, is that it is single stranded in structure, and has Uracil in place of Thymine [1].

Recent research has found myriad functions for RNA other than being the messenger molecule for DNA. For example, RNA can act as a catalyst for RNA splicing, peptide bond formation, and can alter the regulation of genes [19]. Because of its inherently single stranded nature, RNA forms bonds with itself, folding into secondary and tertiary structures [6].

It is axiomatic that chemical structure is tantamount to biological function, and RNA is no exception. For this reason, there has and continues to be an intense interest in predicting the secondary and tertiary structure of RNA molecules. This is in part because it will elucidate the underlying principles of RNA structure formation and function [6], but also because it will allow the detection and classification of unknown RNAs, and assist the design of new RNA based drugs [4]. The secondary structure of RNA is also highly conserved during evolution, indicating its importance [9]. Secondary and tertiary structures can be treated hierarchically. As a result, it is possible to predict the secondary structure of RNA without understanding the tertiary structure [17]. This report focuses on secondary structure prediction.

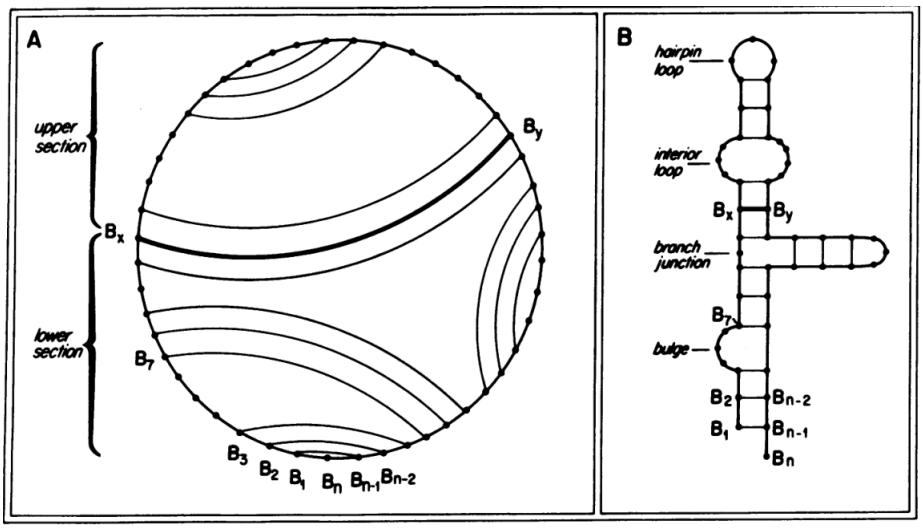

Figure 1: RNA secondary structure as described in the Nussinov algorithm. Taken from a publication by Nussinov & Jacobson [13].

The Nussinov Algorithm

The first RNA secondary structure prediction algorithms were relatively naive brute-force searches in which all possible secondary structures were enumerated and the one with the most bonds was selected as the solution [14]. While being very simplistic, these initial techniques introduce an important assumption: RNA molecules will form energetically stable secondary structures. Maximising bonds is a crude but nonetheless accurate measure of energetic stability, as every bond increases the stability of a structure [14]. In the late 1970s, when the first large RNA molecules were being successfully sequenced, Nussinov et al. [14] introduced an algorithm based on loop matching for bonding pairs. Their algorithm was designed to find a single structure with the maximal number of bonds using dynamic programming. This is possible only with the restriction that all bonding pairs are nested, an assumption that is generally true for naturally occurring RNA.

Because of its dynamic programming nature, this algorithm performs recursive decompositions of an RNA, and builds larger structures out of repeated substructures. A natural representation of this is depicted in Figure 1. Part A of Figure 1 shows bonds as arcs across a circular graph. In it, we see the nested nature of the structures being explored by the Nussinov algorithm. Part B shows how these structures translate to actual RNAs, and how these appear in vivo.

$$M(i,j) = \max\{A, B, C, D\} \tag{1}$$
$$A = M(i, j-1)$$
$$B = M(i+1, j)$$
$$C = M(i+1, j-1) + W(i,j)$$
$$D = \max\{M(i,k) + M(k+1,j)\} \; when \; i < k < j$$

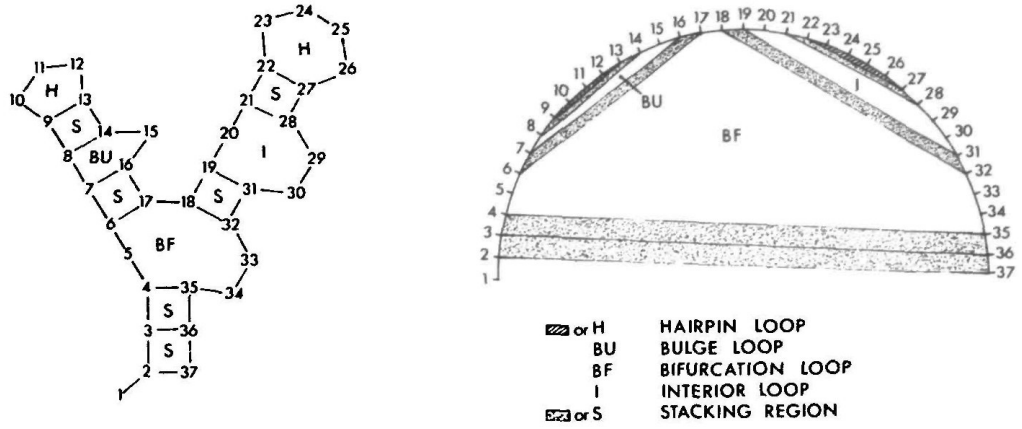

Figure 2: Diagram of faces used in the Zuker algorithm. Taken from original publication [**?**].

In the recurrence relation show in Equation 1, the first two cases (A and B) find the score associated with not allowing the positions i and j to bond. The case C conversely determines the score given that positions i and j are bonded. The final case D computes the score associated with a bifurcation. A bifurcation here means decomposition of the RNA into two separate structures between. This 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.

The Zuker Algorithm

Soon after the work of Nussinov & Jacobson, Zuker & Stiegler [**?**] described an altered version of the same algorithm which, instead of maximising base pair weights, minimized the free energy of the secondary structure. This was done by introducing a number of thermodynamic rules for canonical structures like hairpin loops, internal bulges, multiloops, unbonded base pairs, and stacked base pairs. The algorithm is similar to the Nussinov algorithm but adds another mutually recursive dynamic programming recurrence to inject a complex and relatively comprehensive energy system. The original energy system used is borrowed from the work of Studnicka et al. [**?**] who presented a complex but theoretically similar algorithm, albeit with much worse asymptotic and implementation complexities.

First I shall introduce some useful terminology which should clarify aspects of Zuker & Stieglers algorithm. The bases of an RNA molecule can be thought of as vertices in a planar graph. Edges between such vertices are then represented as chords on a semicircular diagram (Figures 1 and 2). These chords are not allowed to touch. A chord is admissible if it represents a chemically valid bond, and an admissible structure is a structure whose graph contains only admissible bonds. Thence, one can define a face of such a graph as any planar region bounded on all sides—either

by chords, or the edge of the graph. The folding algorithm of Zuker & Stiegler considers such faces as the basic contributing factor to a molecule's stability, unlike the algorithm of Nussinov & Jacobson which considers only individual bonds.

Let $E(F)$ represent the energy of a face F; impossible structures are given an energy value of infinity, for example hairpin loops with less than three bases in the intervening loop region. In addition let $V(i,j)$ be defined as the minimum free energy of all structures in which bases i and j are bonded, and let $W(i,j)$ represent the minimum free energy of all structures contained within bases i and j inclusive. Note that for $W(i,j)$ there may or may not be a bond between bases i and j. Also, if i and j cannot bond then $V(i,j) = \infty$. Finally note that $FH(i,j)$ represents a hairpin loop structure from i to j, and that $FL(i,j,i',j')$ is defined as the region bounded by the bonds i,j and i',j'. Examples of these decompositions are shown diagrammatically in the right half of Figure 2. The labelled regions show faces in a semicircular graph representing a strand of RNA. In the accompanying left half of the figure, the same RNA structure is shown as it would appear in a real RNA rather than in a purely diagrammatic depiction.

$$\begin{gathered} V(i,j) = \min\{E1, E2, E3\} \\ E1 = E(FH(i,j)) \\ E2 = \min\{E(FL(i,j,i',j')) + V(i',j')\} \ where\, i < i' < j' < j \\ E3 = \min\{W(i+1,i') + W(i'+1,j1)\} \ where\, i+1 < i' < j-2 \end{gathered} \tag{2}$$

As shown by the definition provided in Equation 2, $V(i,j)$ is computed by minimizing three cases. The first case considers the bond between i and j closing off a hairpin loop (H in Figure 2). The second accounts for situations in which i and j are bonded. This can result in a bulge (BU in Figure 2), internal loop (I in Figure 2), or the continuation of a stacking region with the interior bond i',j' (S in Figure 2). The third and final case considers bifurcations (BF in Figure 2).

$$\begin{gathered} W(i,j) = \min\{W(i+1,j), W(i,j-1), V(i,j), E4\} \\ E4 = \min\{W(i,i') + W(i'+1,j)\} \ where\, i < i' < j-1 \end{gathered} \tag{3}$$

Equation 3 is the recurrence for $W(i,j)$ as described by Zuker & Stiegler. Again there are three cases. The first two cases $W(i+1,j)$ and $W(i,j-1)$ should be thought of as a single case which consider situations in which there is no bond between i and j. This is similar to cases A and B from the Nussinov algorithm (Equation 1). The final case considers taking the bond from i to j. This final case allows for bifurcations in which two bonding pairs split the structure into two sections. The final minimum free energy of the best structure is defined by $W(1,n)$, where n is the length of the RNA molecule. It should be noted that the free energy for small molecules (fewer than 6 nucleotides in length) can easily be precomputed, and forms the base case of the given recurrence relations. Because of its efficiency ($O(N^3)$ time and $O(N^2)$ space), 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 [11, **?**].

Stochastic Context Free Grammars

Conclusion

It should be noted that the bound of $O(m+n^3)$ means that, when $a = n$ and $a^2 < m$, this algorithm is equivalent to that of Nash & Gregg. Also, given $m \geq n^2$, it is more efficient than that of Bonsma & Breuer. Furthermore it has better time complexity than Valentino's algorithm for any moderately dense graph. It is also superior to earlier algorithms which take $O(m^3)$ time. Despite this, the Nussinov algorithm was first published in 1978, and it's successor is currently widely used in bioinformatics to predict RNA secondary structures [11]. This is particularly surprising, since the maximum-weight independent set problem has often been concomitant with the use of RNA folding algorithms [15, 2].

Li, Ranka and Sahni [10] implemented the Nussinov algorithm on the GPU. This was reportedly up to 1582 times faster than the single threaded version. They also outlined efficient multi-core implementations for the CPU, based on the same model.

For 25 year the Nussinov algorithm was the most optimal algorithm for finding the maximum-weight independent set of a circle graph. Before 2013, it was still the most optimal algorithm in some cases. In addition, excellent parallel implementations of the Nussinov algorithm are available. Despite this, it was never recognised as finding, or used for finding, the maximum-weight independent set for a circle graph. This starkly illustrates the problems inherent to cliquing within science. I 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] B Alberts, D Bray, K Hopkin, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. Essential cell biology; garland science: New york, 2009. pages 427–452.

[2] Michaël Bon and Henri Orland. Tt2ne: a novel algorithm to predict rna secondary structures with pseudoknots. *Nucleic acids research*, 39(14):e93–e93, 2011.

[3] Paul Bonsma and Felix Breuer. Counting hexagonal patches and independent sets in circle graphs. *Algorithmica*, 63(3):645–671, 2012.

[4] Anne Condon. Problems on rna secondary structure prediction and design. In *Automata, Languages and Programming*, pages 22–32. Springer, 2003.

[5] 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.

[6] Graeme L Conn and David E Draper. Rna structure. *Current opinion in structural biology*, 8(3):278–285, 1998.

[7] 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.

[8] Fanica Gavril. Algorithms for a maximum clique and a maximum independent set of a circle graph. *Networks*, 3(3):261–273, 1973.

[9] Ivo L Hofacker. Rna consensus structure prediction with rnaalifold. In *Comparative Genomics*, pages 527–543. Springer, 2008.

[10] Junjie Li, Sanjay Ranka, and Sartaj Sahni. Multicore and gpu algorithms for nussinov rna folding. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on*, pages 1–2. IEEE, 2013.

[11] 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.

[12] Nicholas Nash and David Gregg. New algorithms for maximum independent sets of circle graphs. 2013.

[13] Ruth Nussinov and Ann B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded rna. *Proceedings of the National Academy of Sciences*, 77(11):6309–6313, 1980.

[14] 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.

[15] Jana Sperschneider and Amitava Datta. Knotseeker: Heuristic pseudoknot detection in long rna sequences. *RNA*, 14(4):630–640, 2008.

[16] 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.

[17] Ignacio Tinoco Jr and Carlos Bustamante. How rna folds. *Journal of molecular biology*, 293(2):271–281, 1999.

[18] 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.

[19] Zhenjiang Xu, Anthony Almudevar, and David H Mathews. Statistical evaluation of improvement in rna secondary structure prediction. *Nucleic acids research*, 40(4):e26–e26, 2012.