CITS4008 Assignment 2: Beyond the Thermodynamic Hypothesis

Max Ward

April 17, 2014

Abstract

The algorithmic prediction of RNA molecules dates back to the late 1970s. Despite the fields age, little progress has been made. There are two core approach for in silico RNA folding: ad-hoc dynamic programming algorithms, and stochastic context free grammar parsing methods. Both converge on the same accuracy upper limit, despite often using wildly different scoring schemes. In this report I outline these core algorithmic approaches for folding RNA. Furthermore, I highlight the lack of progress in prediction accuracy. Finally I attempt to explain why the predictive power of these algorithms is stunted, and suggest a direction for future research.

Contents

List of Figures

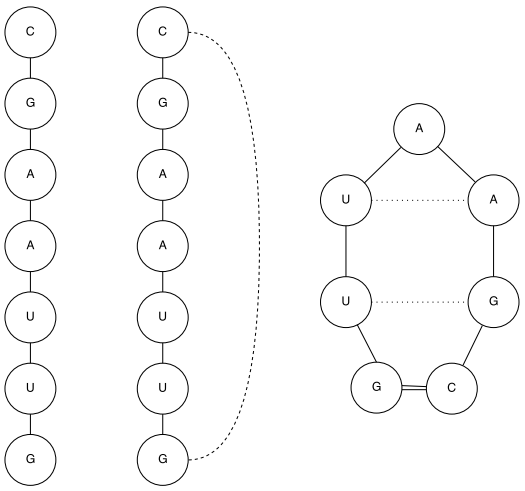

Figure 1: A simple example of how a RNA sequence might fold. Dotted lines represent potential bonds between nucleotides, double lines represent actual bonds. The progression of folding follows from left to right.

1 RNA and Aeroplanes

Ribonucleic Acid (RNA) molecules are very much like paper aeroplanes. In some ways, they are simpler. To understand this assertion the reader must unfortunately also understand the chemistry of RNA. Luckily, this chemistry is so straightforward that I shall now describe a working model of it.

As the name would imply, RNA is chemically similar to Deoxyribonucleic Acid (DNA). However, unlike DNA, RNA is single stranded. This means that it is made of a single sequence of connected nucleotides. These nucleotides are like links in a chain. Each such nucleotide is attached to it's direct neighbours in the sequence. Much like an actual chain, a complete RNA sequence is flexible, and can move about freely. While the nucleotides I have mentioned are chemically ingenious, we need not know the details. All that needs to be said here is that there are four types: Adenine (often abbreviated to A), Guanine (G), Cytosine (C), and Uracil (U). Adenine readily forms chemical bonds with Uracil, and vice-versa. Similarly, Guanine and Cytosine have a propensity to bond. Additionally, Guanine and Uracil sometimes form weak bonds. In the parlance of RNA research, these types of bonds are called 'A-U', 'G-C', and 'G-U' respectively.

So what about the paper aeroplanes? When constructing such a machine, one starts with a blank sheet of paper. One then folds this paper repeatedly upon itself. This is precisely what happens when RNA molecules form bonds. Say a G nucleotide sees an attractive looking C nucleotide and a chemical bond forms between them. This would cause the entire chain to fold on itself so that the G and C nucleotides come into close proximity. After the bond has formed, we are left with a loop containing the nucleotides that were between the G and the C in our theoretical RNA (see Figure 1). The folding process may then repeat itself for nucleotides that are contained inside this loop region. Hence we say that the bonds of RNA molecules

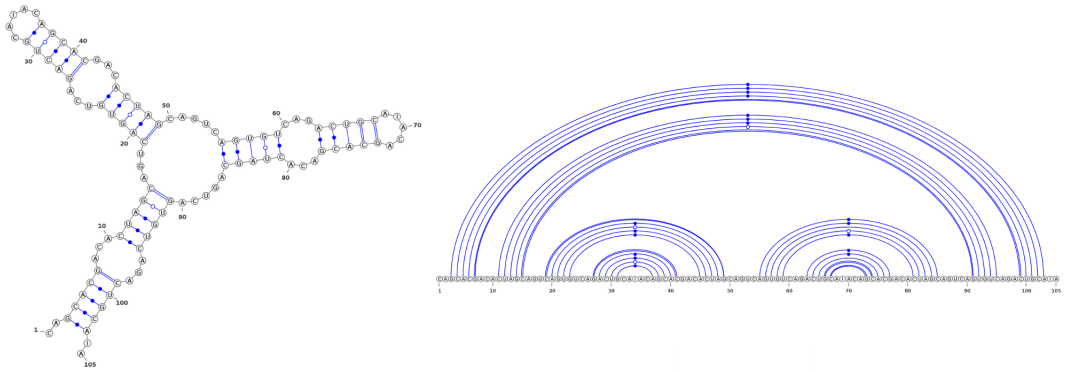

Figure 2: How a real RNA might fold. Depicted on the left is the final folded state. Depicted on the right is an arc diagram. In the arc diagram, the RNA has been laid out on a line. The blue arcs represent bonds between nucleotides in the RNA. The arc diagram corresponds to the structure on the left.

are nested. Folds form within folds, and therefore the structure of RNA is recursive. To clarify this concept, I shall recapitulate the paper aeroplane analogy. Paper aeroplanes are made from finite, two dimensional rectangles that fold repeatedly in three dimensions. In contrast, RNA molecules are finite, one dimensional line segments that fold repeatedly in two dimensions.

1.1 Why Bother?

RNA molecules fold into interesting structures, but why does this matter? RNA is actually a biologically active molecule. Recent research has found myriad functions for RNA. For example, RNA can act as a catalyst for mRNA splicing, peptide bond formation, and can alter the regulation of genes [14]. 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 structure of RNA molecules in silico. Furthermore, it will allow the detection and classification of unknown RNAs, and assist the design of new RNA based drugs [2]. The structure of RNA is also highly conserved during evolution, indicating its importance [4].

I endeavour to give the reader a brisk but incisive review of RNA secondary structure prediction algorithms in this report. For the sake of succinctness, my focus shall be methods able to predict RNA structures ex nihilo—that is with no information other than the RNA sequence itself. Additionally, I have omitted algorithms capable of predicting structures called pseudoknots. This is because such structures are uncommon, and because prediction of RNA structures with pseudoknots is an NP-complete problem [7]. The reader should also note that the kinds of structures I have here described are called ‘secondary structures’. The primary structure of RNA is simply the chain of nucleotides. Thus, as I have explained, the secondary structure of RNA is its two dimensional shape after folding. This shape is what is computed by ‘secondary structure prediction’ algorithms.

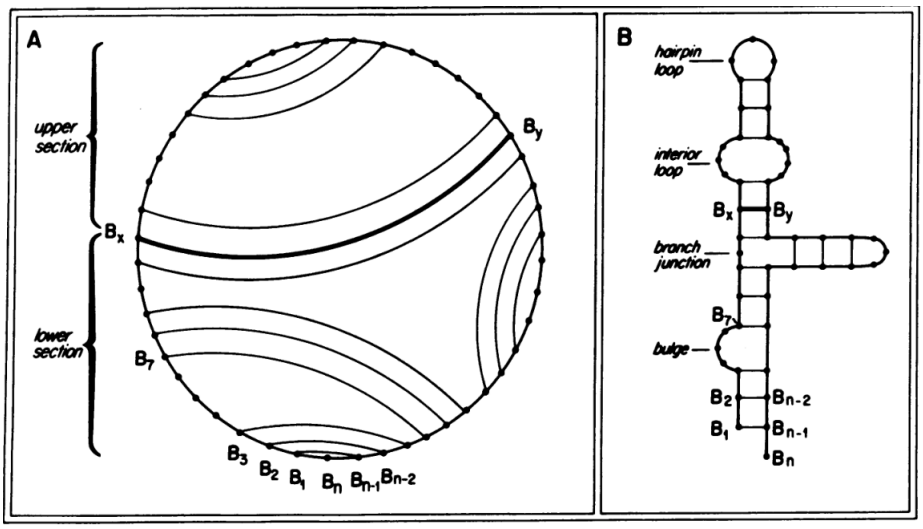

Figure 3: RNA secondary structure as described in the Nussinov algorithm. Taken from a publication by Nussinov & Jacobson [8].

2 Dynamic Programming

2.1 The Nussinov Algorithm

According to the 'Thermodynamic Hypothesis', biologically active molecules should form structures that have minimum free energy and thus maximum stability [1]. For RNA molecules, counting bonds is a crude but nonetheless accurate measure of energetic stability, as every bond increases the stability of a structure [9]. In the late 1970s, when the first large RNA molecules were being successfully sequenced, Nussinov et al. [9] introduced an algorithm to find a single structure with maximal number of bonds using dynamic programming. This is possible only with the restriction that all bonding pairs are nested, and hence form no pseudoknots.

Because of its dynamic programming nature, this algorithm performs recursive decompositions of a RNA and builds larger structures out of repeated substructures. A natural representation of this is depicted in Figure 3. Part A of Figure 3 shows bonds as arcs across a circular graph. One can also 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 they appear in vivo.

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

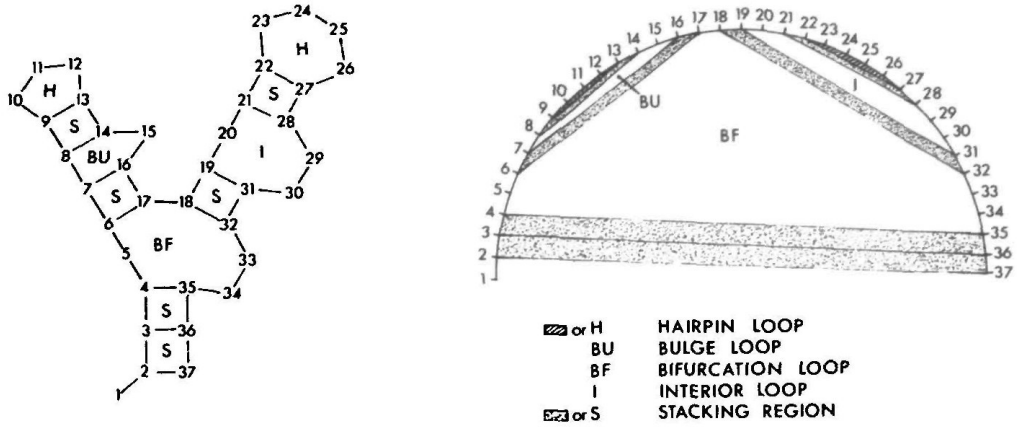

Figure 4: Diagram of faces used in the Zuker algorithm. Taken from original publication [15].

Equation 1 describes the recurrence relation for the Nussinov algorithm. 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 scores associated with bifurcations. A bifurcation here means decomposition of the RNA into two separate structures.

2.2 The Zuker Algorithm

Soon after the work of Nussinov & Jacobson, Zuker & Stiegler [15] described an altered version of the same algorithm which, instead of maximising bond weights, minimized free energy. This was done by introducing a number of thermodynamic rules for canonical structures like hairpin loops, internal bulges, multiloops, unbonded bases, and stacked base pairs. The algorithm is similar to the Nussinov algorithm, but requires another mutually recursive dynamic programming recurrence to inject a complex and relatively comprehensive thermodynamic scoring system.

First I shall introduce some useful terminology which should clarify aspects of Zuker & Stieglers algorithm. The bases of a RNA molecule can be thought of as vertices on a planar graph. Edges between such vertices are then represented as chords on a semicircular diagram (see Figures 3 and 4). 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. 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; inadmissible structures are given an energy value of infinity. In addition let $V(i, j)$ be defined as the minimum free energy of all structures between i and j, 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 $W(i,j)$ the bases at i and j need not be bonded. Alternatively if i and j cannot bond, then $V(i,j) = \infty$. Finally note that $FH(i,j)$ represents a hairpin loop 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 4. In the accompanying left half the same RNA structure is shown as it would appear in vivo.

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

As defined 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 4). The second accounts for situations in which i and j are bonded, resulting in a bulge (BU in Figure 4), internal loop (I in Figure 4), or the continuation of a stacking region with the interior bond i',j' (S in Figure 4). The third and final case considers bifurcations (BF in Figure 4).

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

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 $W(i,j-1)$ are conceptually a single scenario 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 third case considers taking the bond from i to j. The final case allows for bifurcations. The 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 [6, 10].

3 Stochastic Context Free Grammars

Later, in the early 90s, Hidden Markov models and Stochastic Context Free Grammars (SCFGs) were being used to model RNA folding. Sakakibara et al. [13] used SCFGs to accurately fold tRNAs, a family of RNAs with notoriously difficult to predict secondary structures. They did this by defining a formal grammar that parses secondary structure elements into a primary sequence, with probabilities assigned to the production rules. These probabilities can be trained using actual RNAs, yielding

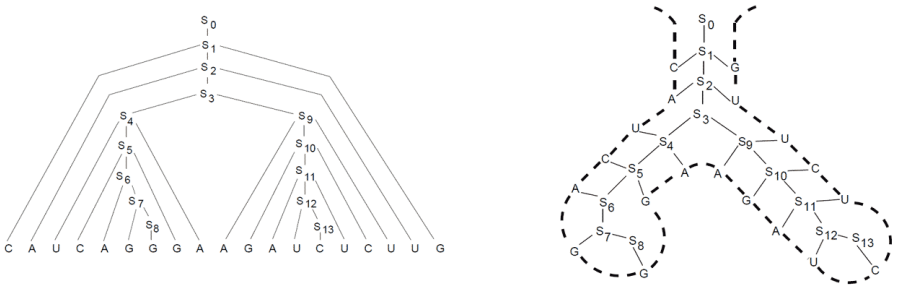

Figure 5: Left shows the parse tree for a context free grammar. Right shows how this tree relates to RNA secondary structure. Taken from original publication [13].

a model capable of parsing novel RNA sequences. Later methods were similar, but work for general classes of RNAs [3, 5].

In 2012 Rivas, Lang & Eddy [12] presented a computation tool called TORNADO which can parse various RNA grammars. It supports the typical grammars used in SCFG approaches, the grammar implicit in Zuker's algorithm upon which the thermodynamic model is based, and many more complex grammars. They used this meta-algorithm to compare the current state of the art models. Because TORNADO is essentially a super-SCFG parser, all of the supported grammars can have their parameters changed arbitrarily. Hence, the experimentally determined thermodynamic model, which the Zuker algorithm uses, can be applied to other grammars, and complex, machined learned parameters can be applied to the Zuker grammar. Rivas, Lang & Eddy found that the best machine learned models are comparable to the typical thermodynamic model in accuracy, however they often suffer from overfitting.

3.1 Unification of Techniques

In a subsequent publication, Rivas [11] unified pseudoknot-free RNA folding algorithms. Her core observation is that all such prediction algorithms contain the same four key components: an architecture, or the production rules of a grammar; a scoring scheme, or how scores are assigned to these production rules; and the parametrization of the scoring scheme, or the specific values assigned to it. These three features are referred to by Rivas as the 'model'. The fourth and final feature is the folding algorithm used to find the best structure given the model. Here Rivas notes that the two dominant folding algorithms are interchangeable. The Cocke-Younger-Kasami (CYK) algorithm used to parse SCFGs and algorithms based on the work of Zuker & Stiegler are isomorphic for the purpose of parsing RNA grammars. Rivas additionally notes that all scoring schemes and parametrizations appear to hit an accuracy upper limit, and that complex, machine learned models are only slightly more accurate than thermodynamic models. In fact, relatively basic grammars with hundreds of parameters seem to perform almost equivalently to those with tens of thousands.

4 Conclusions

Current state of the art algorithms fold RNA in a way that globally maximises their score according to some model. The Zuker algorithm, for example, finds the global minimum free energy configuration. SCFGs globally maximise the perceived probability of a parse tree. This bias is largely due to the Thermodynamic Hypothesis. Anfinsen [1] presented this hypothesis as the underlying principle behind the formation of biologically active proteins. He held that protein fold into the minimum Gibbs free energy conformation in their typical biological environment; environment being defined as the molecules' physiological state: pH, temperature, and ion concentration. Furthermore, through natural selection, molecules that are most likely to fold into the correct shape have evolved, and the atomic interactions of such molecules fully determine their final state. This insight has been invaluable for folding proteins and RNAs in silico. Despite this, it has recently become clear that methods for the prediction of RNA secondary structures have hit an upper limit in accuracy.

Perhaps the Thermodynamic Hypothesis is insufficiently powerful to explain in vivo RNA secondary structures. As Rivas [11] noted, there appears to be a ceiling on the accuracy of our current algorithms—all of which are based on the Thermodynamic Hypothesis. Finding a way past this accuracy barrier will be important for future advancements, as it is increasingly obvious that RNAs are essential biological agents. RNA prediction is a liminal field, and in dire need of a way beyond the Thermodynamic Hypothesis.

References

[1] CB Anfinsen. Principles that govern the protein folding chains. *Science*, 181:233–230, 1973.

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

[3] Robin D Dowell and Sean R Eddy. Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction. *BMC bioinformatics*, 5(1):71, 2004.

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

[5] Bjarne Knudsen and Jotun Hein. Pfold: Rna secondary structure prediction using stochastic context-free grammars. *Nucleic acids research*, 31(13):3423–3428, 2003.

[6] 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.

[7] Rune B Lyngsø and Christian NS Pedersen. Rna pseudoknot prediction in energy-based models. *Journal of computational biology*, 7(3-4):409–427, 2000.

[8] 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.

[9] 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.

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

[11] Elena Rivas. The four ingredients of single-sequence rna secondary structure prediction. a unifying perspective. *RNA biology*, 10(7):1185, 2013.

[12] Elena Rivas, Raymond Lang, and Sean R Eddy. A range of complex probabilistic models for rna secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, 18(2):193–212, 2012.

[13] Yasubumi Sakakibara, Michael Brown, Rebecca C Underwood, I Saira Mian, and David Haussler. Stochastic context-free grammars for modeling rna. In *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*, volume 5, pages 284–293. IEEE, 1994.

[14] 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.

[15] 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.