CITS4008 Assignment 1

Max Ward

RNA and DNA

DNA is a double helix molecules 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 (U) 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 and peptide bond formation, and can also alter the regulation of genes [31]. Because of its inherently single stranded nature, RNA forms bonds with itself, folding into secondary and tertiary structures [4].

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 [4], but also because it will allow the detection and classification of unknown RNAs, enable prediction of novel RNA function, and assist the design of new RNA based drugs [3]. The secondary structure of RNA is also highly conserved during evolution, indicating its importance [7]. Secondary and tertiary structures can be treated hierarchically. As a result, it is possible to predict the secondary structure of an RNA without understanding the tertiary structure. My focus in this paper in on secondary structure prediction..

Nussinov's Algorithm

The first such algorithms were based on a relatively naive brute force approach in which all possible secondary structures were enumerated and the one with the most bonds was selected as the solution [18]. 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 [18]. In the late 1970s, when the first large RNA molecules were being

successfully sequenced, Nussinov et al. [18] 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 had to be entirely nested, an assumption that is generally true for naturally occurring RNA.

Because of its dynamic programming nature, this algorithm performs recursive decompositions of the RNA and builds larger structures out of repeated substructures. A natural representation of this is depicted in Figure 1.3. Part A of Figure 1.3 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.

$$
\begin{aligned}
M(i,j) &= \max_{A,B,C,D} \\
A &= M(i,j-1) \\
B &= M(i+1,j) \\
C &= M(i+1,j-1) + W(i,j) \\
D &= M(i,k) + M(k+1,j)\, \forall k : i < k < j
\end{aligned}
\tag{1}
$$

In the recurrence relation (show in Equation 1) the first two cases (A and B) find the score associated with not allowing i and j to bond. The 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 here means decomposition of the RNA into two separate structures between. This suggests a $O(N^3 3)$ worst case time complexity and a $O(N^2)$ space complexity, as an $O(N^2)$ state space (all combinations of i and j) is explored with a linear time recurrence relation

Circle Graphs and RNA

A circle graph represents the intersection of a set of chords contained inside a circle. A pair of chords (i, j), (k, l) can be described in three ways. The pair is said to overlap if i ¡= k ¡= j ¡= l. A chord (i, j) is said to contain (k, l) if i ¡ k ¡ l ¡ j. Finally, the pair is said to be disjoint if the pair does not overlap, and if neither chord contains the other. A circle graph is obtained by transforming every chord into a node. Edges between nodes indicate overlapping chords. Such a graph can also be represented as a set of intervals on a line. This is geometrically equivalent to cutting the circle, and laying the line out in one dimension. Intuitively one can see that the potential bonds of an RNA molecule form such a graph. Indeed, circle graphs were used to depict RNA is Nussinov's original paper [ref] (ref to figure).

An independent set of a graph is a set of vertices which share no edges. Concordantly, it is a selection of chords which do not overlap. A maximum independent set is an independent set of maximum size—it contains the greatest possible number of vertices. If vertices are assigned weights, a maximum-weight independent set can also be computed, which is an independent set comprising verticies with maximum

summed weight. Generally finding such sets is NP-Hard. However, in the case of circle graph, it can be solved in polynomial time. This is fortunate, as these algorithms have practical applications in register allocation, VLSI design, and bioinformatics [refs?].

O(md) and (m2)

Bonsma and Breuer found an O(nm) solution. They noted that in dense graphs, $m >= n^2$, this algorithm is more optimal than the best known algorithms which take $O = (n^4)$. Shortly thereafter [gregg] presented an algorithm that computed the maximum-weight independent set of a circle graph using only $O(m + n * mina^2, m)$ time. Here a denotes the independence number of the circle graph. The independence number of a graph is the cardinality of the largest independent set. In other words, it is the size of the maximum independent set. This is the best known algorithm to date.

Interestingly, the RNA folding algorithm first presented by Nussinov finds the maximum-weight independent set of a circle graph containing n endpoints in $O(m+n^3)$ where m is the total number of intervals or chords. In RNA there can never be two intervals with endpoints (i, j); a pair of endpoints uniquely identify a bond. However in the general case, there may be many bonds with shared endpoints (i, j), but with different weights. To solve the more general non-RNA case, we must first iterate over all the intervals, and for each (i, j) index store the maximum-weight interval. This takes O(m) time, as every interval is examined once, and the maximum-weight interval is stored in a 2D matrix whose indexes (i, j) correspond to endpoints. The standard implementation of Nussinovs algorithm can then be run with the W(i, j) function returning the value of the maximum-weight interval for endpoints (i, j). Finally we are left with an algorithm that takes $O(m + n^3)$ time, and uses $O(n^2)$ space.

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 [greggs]. And also that when $m >= n^2$ it is more efficient than that of Bonsma and Breuer. Furthermore it has better time complexity than the O(md) [2003] algorithm for any moderately dense graph, and is superior to the earlier algorithm ($O(m^2)$), despite being first published in 1979 and its successor (Zukers algorithm) used widely in bioinformatics to fold RNAs. This is particularly surprising, since the maximum-weight independent set problem has often been concomitant with the use of RNA folding algorithms.

References

[1] D. E. Knuth. *The TEX book.* Addison-Wesley, Reading, Massachusetts, 1984.

[2] L. Lamport. *LATEX : A Document Preparation System.* Addison-Wesley, Reading, Massachusetts, 1986.

[3] Ken Wessen, Preparing a thesis using LATEX , private communication, 1994.

[4] L. Lamport. Document Production: Visual or Logical, *Notices of the Amer. Maths. Soc.*, Vol. 34, 1987, pp. 621-624.