Lecture 7: December 5, 2006

Lecturer: Michal Ziv-Ukelson *Scribe: Erez Katzenelson and Ofer Lavi*

7.1 RNA Basics

7.1.1 Introduction - RNA Base Pairing

The RNA (RiboNucleic Acid) is a biological polymer similar to DNA, except for three main differences: (a) its sugar-phosphate backbone contains ribose rather than deoxyribose; (b) RNA is usually found as a single strand rather than double strand; (c) it contains the base *Uracil(U)* rather than *Thymine(T)*, in addition to *Adenine (A), Cytosine (C)* and *Guanine (G)*. As for DNA bases, RNA bases can also bond and form pairs. The canonical pairs are **A-U, G-C, G-U**, where A-U, G-U are based on two hydrogen bonds and G-C is based on three hydrogen bonds (as seen in Figure 7.1). For simplicity, we will only discuss A-U, G-C pairing, since G-U is highly unstable (referred to as a *"wobble"* pair) and thus quite rare.

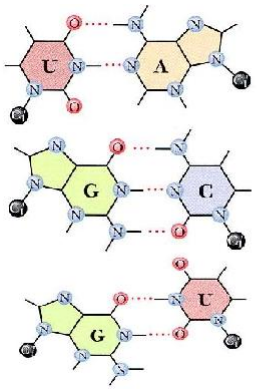

Figure 7.1: The three possible base pairs form two or three hydrogen bonds

7.1.2 Roles of RNA

There are several types of RNA, each with its own structural and functional characteristics:

- *Messenger RNA* (mRNA): Encodes for the primary sequence of a protein through the genetic code.

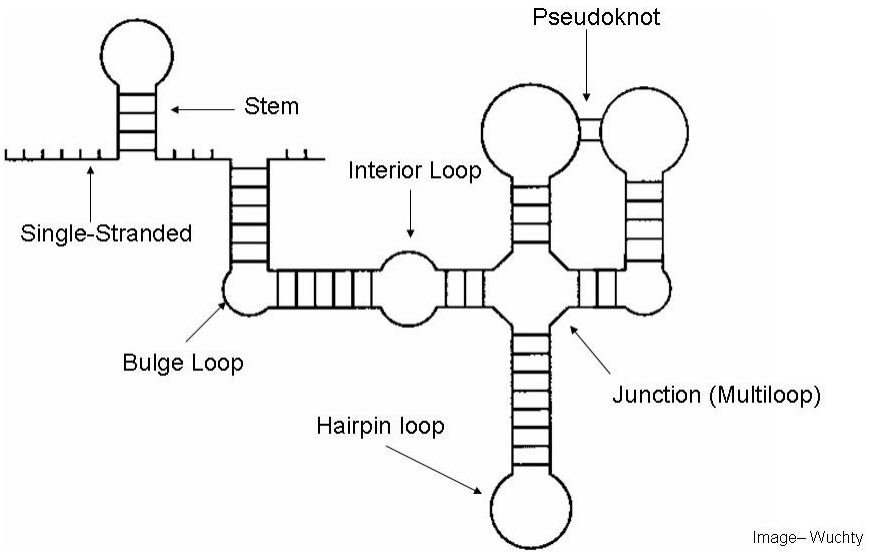

Figure 7.2: Typical motifs of RNA secondary structure (taken from *Wuchty*)

- Transfer RNA (tRNA): Binds an amino acid with its anti-sense codon in order to perform translation at the ribosome.
- Ribosomal RNA (rRNA): Binds with proteins to form the ribosome - a complex which translates an mRNA strand to the suitable protein.
- Small Nuclear RNA (snRNA): Short RNA sequences which perform "maintenance" on RNA (such as splicing and regulation of transcription factors) within the nucleus.
- MicroRNA (miRNA): Short RNAs which can inhibit the translation of mRNA or increase its RNA structure degradation.

7.1.3 RNA Structure

We consider three levels of description for the RNA structure degradation:

- Primary structure - the sequence of RNA bases
- Secondary structure - a two dimensional folding containing an annotation of which base pairs are formed.
- Tertiary structure - a three dimensional folding containing a base sequence with base pair annotation and we describe the spatial location of every atom

The method will focus on predicting the secondary structure for a given primary structure.

The secondary structure consists of the following base-pairing patterns: *Single Strand (unpaired bases), stem, hairpin loop, bulge loop, interior loop, junction (multiloop) and pseudoknot*, as shown in Figure 7.2. It is easy to see that if we know the primary structure and the stem positions, we can determine all the other characteristics of the secondary structure, except for the pseudoknots. Fortunately, the common belief is that pseudoknots contribute very little to the energy balance of the RNA molecule. For that reason, and for another computational reason (which will be discussed later in 7.3.1), it is a common practice to ignore their locations when predicting RNA folding.

7.2 RNA secondary structure prediction

7.2.1 Formal definition of RNA secondary structure

In the most general definition, given a single stranded RNA sequence of length N, $R = r_1, ..., r_N$, a secondary structure of R is defined to be a set S of disjoint base pair indices, (i, j) s.t. $1 \leq i < j \leq N$ and r_i and r_j form a valid base pair.

This definition can be further restricted so that adjacent bases cannot be paired by adding the following restriction over all pairs (i, j): $j - i > 3$

The biochemical justification for this restriction is that the minimal hairpin loop size is empirically known to be greater than 3.

7.2.2 Nested Edges Graph representation

The secondary structure can be described as a graph $G = (V, E)$ where V is the set of bases in R and E is defined as $E = \{(v_i, v_j)|i = j - 1\} \bigcup \{(i, j) \in S\}$ (see Figure 7.3).

Recall that S is the set of base pair indices defined in 7.2.1

This defines a graph with two types of edges: edges between subsequent bases, and edges between base pairs defined by the secondary structure.

If we restrict the secondary structure, so that it cannot contain Pseudoknots, the graph representation of the secondary structure would be a Nested Edges Graph, formally:

$$\forall (v_i, v_j), (v_k, v_l) \in E, k \in [i, j] \Leftrightarrow l \in [i, j]$$

7.2.3 Optimal RNA secondary structure

The optimal secondary structure of RNA is commonly viewed as the most energetically stable RNA secondary structure. Thus, in order to predict the optimal structure, we first need to define a stability score for a given secondary structure.

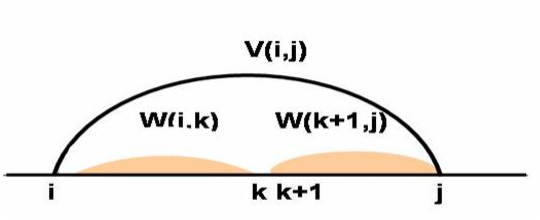

Figure 7.3: The base pair matching problem represented as a Nested Edges Graph Problem - *image by Sean Eddy*

Intuitively, a greater number of valid pairs (A-U and C-G) corresponds to a more stable secondary structure. This definition is not complete, and there are other methods that give more refined predictions: predictions that allow also the loose pair G-U, and others that also take into account different weights for different motifs in the secondary structure (see [7])

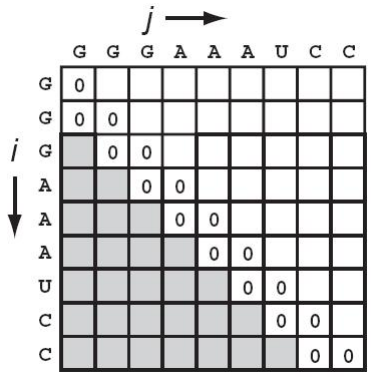

Figure 7.4: Matrix initialization for the DP algorithm on RNA sequence *GGGAAAUCC*

7.3 Algorithm for maximizing the number of base pairs

The problem of finding the optimal base pair matching of a given primary RNA sequence is an NPC problem, since the number of possible pairs is exponential regarding to the sequence length (as explained in [3]). Therefore, even though the base pair maximization is not truly optimal, it has a major advantage, as it suits a well established algorithmic approach for predicting the RNA secondary structure, using sequence alignment.

The algorithm we describe is a dynamic programming algorithm, that dynamically builds the optimal secondary structure of all subsequences of the given RNA ([6]).

$S(i,j)$ is defined to be the number of base pairs in the folding of the subsequence $r_i...r_j$ of R which results in the highest number of base pairs.

Base condition:

$S(i,j) = 0$ for each i,j s.t. $i = j, i = j+1$ (see Figure 7.4)

The four recurrence options:

1. Adding another base pair: r_i, r_j are a valid base pair, and we add it to the best folding of subsequence $r_{i+1}...r_{j-1}$ (see Figure 7.5)
2. Adding another base at the beginning of the sequence $r_{i+1}...r_j$ (see Figure 7.6)
3. Adding another base to the end of the sequence $r_i...r_{j-1}$ (see Figure 7.7)
4. Finding the best cut point k of the sequence $r_i...r_j$ (a bifurcation) (see Figure 7.8)

Recurrence relation:

$$S(i,j) = max \begin{cases} S(i+1,j-1)+1 & \text{(if i,j is a base pair)} \\ S(i+1,j) & \\ S(i,j-1) & \\ max_{i<k<j} S(i,k) + S(k+1,j) & \end{cases}$$

Thus, we can compute an optimal folding for $r_i...r_j$ if we already have all optimal foldings for all subsequences of $r_i...r_j$ (see Figure 7.9)

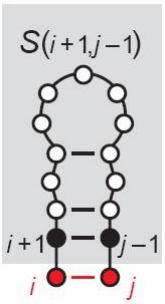

Figure 7.5: Adding another base pair: r_i, r_j are a valid base pair, and we add it to the best folding of subsequence $r_{i+1}...r_{j-1}$.

7.3.1 The Pseudoknots pitfall

A most important notion, which allows us to use the dynamic programming algorithm is that each score $S(i,j)$ is independent of the overall structure, and relies only on the local

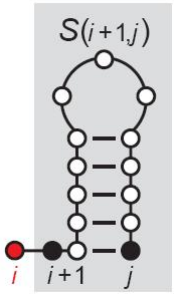

Figure 7.6: Adding another base to at the beginning of the sequence $r_{i+1}...r_j$ (if r_i and r_j are not a valid base pair) *images - Sean Eddy.*

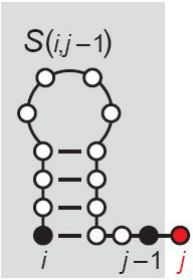

Figure 7.7: Adding another base to to the end of the sequence $r_{i+1}...r_j$ (if r_i and r_j are not a valid base pair).

structure of $r_i...r_j$. However, we know that it is possible to have Pseudoknots, which are base pairs between distant hairpin loop motifs. Two distant hairpin loops are found in the middle of two different subsequences, and although not changing the local score of the two subsequences, a pseudoknot increases the overall score of the fold. The trivial solution takes us back to enumerating all potential structures, which is exponential with the number of bases.

The base pair maximization dynamic programming algorithm avoids this pitfall by simply not allowing pseudoknots. This has been found to be a reasonable assumption[2] due to two reasons:

1. Close pseudoknots contribute very little to the overall stability of the fold.

2. Distant pseudoknots are very rare.

This approach is implied in both [3] and [4] for obtaining the optimal base pair matching of RNA sequences. For other purposes, such as three-dimensional modeling of the RNA structure, Pseudoknots cannot be ignored, since they actually affect the spatial position of atoms.

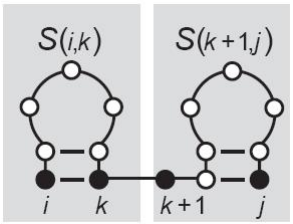

Figure 7.8: Finding the best cut point k of the sequence $r_i ... r_j$ (a bifurcation).

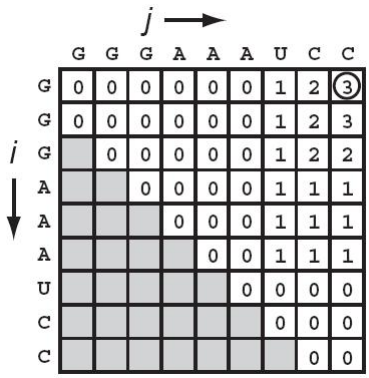

Figure 7.9: Final matrix. Secondary structure score in the upper-right corner.

7.3.2 Time and space complexity of the algorithm

The algorithm requires the computation of a two dimensional matrix of size $N \times N$. At each recurrence step, calculating the first three options take a constant time, but going over at most N values of k at each cell of the matrix takes $O(N)$. Total time complexity is then $O(N^3)$. Space complexity, storing the matrix is $O(N^2)$

7.4 Criticism on the base pair maximization algorithm

The first two drawbacks of this algorithm are:

1. It ignores Pseudoknots.

2. The score incompletely reflects the stability model of the secondary structure obtained.

As the first drawback can be excused from a biochemical perspective, the second was addressed and amended by other algorithms that model the stability of the secondary structure better. The basic DP (Dynamic Programming) algorithm[9] only takes the number of base pairs into account, while other algorithms like mFOLD[8] and Vienna[5] are based on a scoring system that takes the different motifs into account.

The score of each motif is drawn from laboratory experiments[2] that measured the stability of the different motifs.

Still, the dogma "more stable → more likely" holds, but the definition of "more stable" is improved. Also, like in the basic algorithm, these algorithms run under the assumption that the stability of a given subsequence, is independent from the rest of the sequence, and only depends on the subsequence itself and its subsequences. This allows using DP, albeit with more sophisticated recurrence rules.

However, not that in reality the equation "more stable=more likely" does not always hold. In fact, algorithms based on thermodynamics predict only 50-70% of the base pairs correctly. In addition, a large number of RNA structures lie within 5-10% of the predicted global energy minimum[1]. There is still a lot of room for research regarding the reasons for getting a specific fold for a given RNA sequence, and for algorithms for prediction of the right secondary structure.

Bibliography

[1] Sean R Eddy. How do rna folding algorithms work? *Nature Publishing Group*, 2004.

[2] David H. Mathews et al. Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *Journal of Molecular Biology*, pages 911–940, 1999.

[3] R. Durbin et al. *Biological sequence analysis*. Cambridge University Press, 1998.

[4] Dan Gusfield. *Algorithms on strings, trees and sequences*. Canbridge University Press, 1997.

[5] Ivo L. Hofacker. Vienna rna secondary structure server. *Nucleic Acid Research*, 31:3429–3431, 2003.

[6] Ruth Nussinov and Ann B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded rna. *PNAS*, 1980.

[7] Tinoco and Bustamante. How rna folds. *Journal of Molecular Biology*, 1999.

[8] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acid Research*, 31:3406–3415, 2003.

[9] Michael Zuker and Patrick Stiegler. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acid Resaerch*, 9:133–148, 1981.