

Lecture 20

RNA Secondary Structure Prediction

March 7, 1996

Lecturer: Gary Stormo

Notes: David Adams

Topic: Minimum Energy Modeling of RNA structures

1.0 Obtaining RNA structure from sequence.

There are two basic methods of obtaining the secondary, and possibly tertiary structure of RNA from its sequence. The first is the use of minimum energy modeling. Minimum energy modeling has the advantage that it can be used on single sequences. Its disadvantage is that it does not resolve all types of structure well. The second method is that of comparative modeling. Comparative modeling may have more resolving power for some types of structure, but it relies on the presence of multiple, aligned sequences.

RNA is single stranded, and thus can anneal to itself in the same manner that complementary DNA strands anneal to each other. This "self-annealing" creates complicated structures. It is thought that, in biological systems, RNA has a preferred structure. In fact, in some cases the structure of an RNA molecule has demonstrable functional significance. A striking example of an RNA with structurally imparted function is a group of RNAs that have enzymatic activity—they can act as catalysts in biochemical reactions in a manner similar to that of a protein enzyme. RNA base pairing follows DNA base pairing rules with a few exceptions. As RNA contains the molecule uridine in places where DNA would contain thymidine, the following base pairings occur:

DNA base pairs:	G::C	RNA base pairs:	G::C
	A::T		A::U
			G::U <--rare, specific to RNA

RNA structure is difficult to determine by any method. Most of the methods discussed in this lecture will deal with the determination of secondary structure. Deducing tertiary structure (complicated three dimensional structure) from sequence adds additional difficulties to computational models.

Finally, there are non-computational methods for studying RNA structure. These include biochemical experiments which cut only at single-stranded RNA, NMR and crystallography experiments which use electromagnetic imaging techniques, and genetics which makes use of mutations in experimental organisms.

2.0 Representation of RNA structure.

[I will name pages of the notes in the order that they occurred in the handout, calling the top page "1," its back side "2," etc. -ed. note]


```

AGCGUCGUCGUGUCGCGUUUGCCGU  j pairs with i (A-U)
|-----|
i                               |j
                               j-1

```

$\max(M(i, k-1) + M(k, j)) = j$ is paired with something in k
 k is segment between i and j , this step is the one
 requiring the pseudoknot constraint.

```

UGCUCGUCGAGUCGCGUUUGCCGU  j pairs with k (A-U)
|-----|-----|
i           k           |j
                   j-1

```

This has a time complexity of $O(n^3)$.

The problem with such an algorithm is that the energy in RNA structures does not come from bp formation, but from bp stacking. This process is in turn dependent on the local orientation of nucleotides in a base paired section of the RNA molecule.

Base pair stacking energies can be categorized into groups of similar RNA structures. On page 5 of the lecture notes, on the right hand side, several examples of such structures are illustrated. Notice that a negative score is equivalent to a lower energy cost, e.g. loops with many unpaired bases are costly and have a high positive value. Areas of base-pairing subtract from the total score. A special case is the bulge loop:

```

G-C
. . negative score for interaction of G-C to G-C
G-C
. . negative score for interaction of G-C to U-A
U-A
/ .
A . sum of scores for A (+) and U-A to U-A interaction (-)
\ .
U-A
. . negative score for interaction of U-A to A-U
A-U

```

Bulge loops have a positive score associated with unpaired bases, and a negative score associated with the paired bases.

Another special case is the M-loop (see item A3, Figure 2, right side of handout page 6).

A paper by Zuker, et al. (1981) (see references in handout), presents a dynamic programming algorithm which takes stacking energies into account.

This algorithm, which is outlined on page 6 of the notes, makes use of two "half-matrices."

Above the diagonal, $V(i,j)$ with $i \leq j$ is the energy of the best structure in which i is paired with j . Note that $V(i,j)$ equals infinity if i cannot pair with j . There are five possibilities at each iterative step.

close H-loop

Looping structure at the terminus of a series of paired bases.

close B-loop

Bulge loop, as above.

close I-loop

Looping structure within a series of paired bases.

close M-loop

As above.

stack on S_{i+1}, S_{j-1}

Add new pair of bases to existing series of paired bases

Below the diagonal, $W(i,j)$ is the minimum energy for $S_{i,j}$. $W(i,j) \leq 0$. There are three iterative possibilities here.

$V(i, j)$

$W(i, j) = \min(W(i+1, j), W(i, j-1))$

i or j unpaired

$W(i, j) = \min_k (W(i, k) + W(k+1, j))$

i or j paired elsewhere

This algorithm also has a $O(n^3)$ time complexity.

Reliability:

The mathematical solution is not always the optimal one. This is due, at least, to the following:

1. Energy rules are approximate. (+/- 5% to 10%)
2. Pseudoknots and multifurcated loops are disallowed.

A variety of constraints may be added to improve the performance of the Zuker algorithm. A few examples are a priori knowledge:

1. that S_k is single stranded

$$V(i,k) = V(k,i) = \text{infinity for } 1 \leq i \leq k \text{ and } k \leq j \leq N$$

2. that sequence 3' of S_k is cleavable

(means that S_k and S_{k+1} can't both be paired)

3. that S_i and S_j are paired.

can set a bias toward pair formation

3. that S_k is paired to something

The algorithm can be used to look at sub-optimal structures. Page 8 of the lecture notes demonstrates a plot of $V(i,j)$ to $V'(j,i)$ which gives the energy of the best structure with i,j over the whole sequence (1 ... N). By restricting the output and by using a dot plot, all pairing structures within a given percentage of the minimum can be examined. In particular, such methods allow a search for regions with few acceptable alternative structures.

COMPARATIVE METHODS.

Until now we have been discussing methods using one sequence. Comparative methods draw conclusions from two or more aligned sequences. The central idea of comparative methods is to look for compensating complementary changes in two regions of an RNA sequence. For instance, if changes at site 1 correlate strongly with complementary changes at site 2, then sites 1 and 2 are candidates for a secondary structure base pair. Comparative methods are better at finding pseudoknots than thermodynamic methods.

Regions of highly conserved sequence, with or without compensating complementary changes can be tested genetically. If a genetic mutation in the RNA sequence disrupts normal activity, it is a candidate for participation in secondary structure.

Methods for automating and quantifying the correlation among RNA sites is discussed on page 10 and 11 of the lecture notes. Frequencies are determined for how often each type of base occurs at each site. Sites which show conservation are used to form a hypothesis. Observed and expected data are compared for a significant difference.