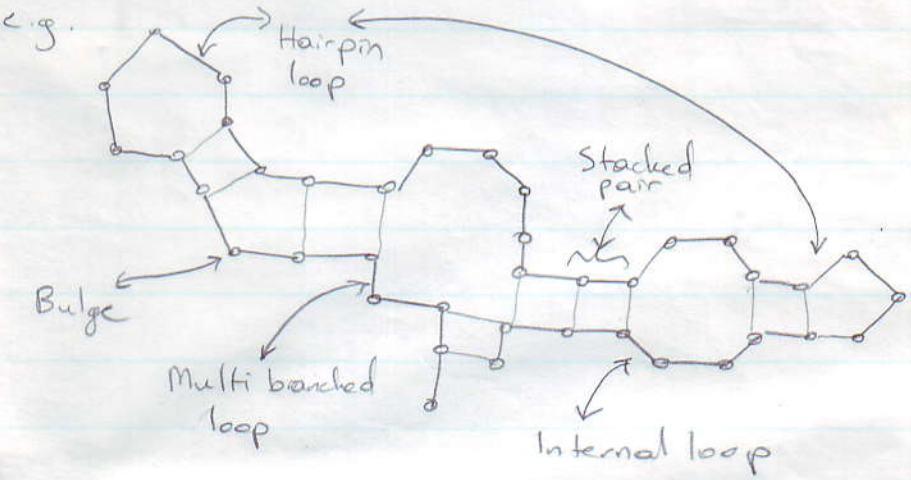
RNA Secondary Structure

RNA is single stranded and folds into a functional shape when bases pair up. This is called the secondary structure.

After folding some bases pair up and others remain free forming loops. Bonded bases provide negative free energy and stabilize RNA, unbonded ones provide positive free energy and destabilize RNA.

Secondary structure determines the functionality of RNA. The RNA folding problem is to compute a folding which minimizes the free energy. (Lyngsø et al)



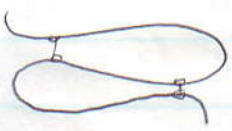
Def: Harrpin: Contains one set of paired bases
Internal loop: Contains two sets of paired bases
Bulge: Internal loop with two of the paired bases adjacent to each other
Stacked pair: Loop formed by $i \cdot j$ and $(i+1) \cdot (j-1)$
Multibranched loop: A loop with > two paired base sets
External base: A base not contained in a loop.

Notions & Definitions:

If $S = s_1 \dots s_n$ is an RNA sequence & $1 \leq i < j \leq n$ then $i \cdot j$ is the pairing of s_i with s_j.

Def: A secondary structure of S is a set of base pairs such that each base is paired at most once.
More precisely $\forall i \cdot j$ and $i' \cdot j'$ $\quad i = i' \Leftrightarrow j = j'$.

Assumption. No pseudo knots allowed:



Justification: pseudoknots occur rarely and without them the problem is simpler.

Def: A pseudo knot is a pair of base pairs $i \cdot j$ & $i' \cdot j'$ s.t. $i < i' < j < j'$

Def: One of the pairs is closest to the ends of RNA strand: exterior / closing base pair - maximum $|i-j|$ for any $i \cdot j$
All other pairs are interior base pairs.

Free Energy Functions

Assumptions: Free energy of RNA is sum of free energy of loops each of which are independent from others.

Functions:

1) $eS(i,j)$: free energy between a stacked pair: $i \cdot j$ and $(i+1) \cdot (j-1)$.

This depends on s_i, s_j and s_{i+1}, s_{j-1}. These are stabilizing and if both pairs are complementary (A-U & C-G) or wobble-pairs (G-U) the energy will be negative.

2) $eH(i,j)$: This is free energy in a loop closed by $i \cdot j$. Depends on length of the hairpin, values of s_i, s_j, and unpaired bases next to s_i, s_j.

3) $eL(i,j,i',j')$: Free energy of an internal loop or bulge with exterior pair $i \cdot j$, and interior pair $i' \cdot j'$. Depends on the values of $s_i, s_j, s_{i'}, s_{j'}$ and unpaired bases next to these as well as the size of the loop.

4) $eM(i,j, i_1, j_1 \cdots i_k, j_k)$: The energy of multi branched loop closed by $i \cdot j$ with interior pairs $i_1, j_1, \ldots i_k, j_k$. Not well understood.

Dynamic Programming:
We describe five tables:

$W(j)$: free energy of optimal structure of first j bases $s_1 \dots s_j$. Answer: $W(n)$

$V(i,j)$: free energy of opt. structure for $s_i \dots s_j$. assuming $i \cdot j$ forms a paired set of bases.

$VBI(i,j)$: free energy for $s_i \dots s_j$ assuming $i \cdot j$ closes a bulge or internal loop.

$VM(i,j)$: Free energy of $s_i \dots s_j$ assuming $i \cdot j$ closes a multi-branched loop.

$WM(i,j)$: used to compute $VM(i,j)$.

Recursively each one of them can be defined as follows:

$W(j)$: $W(0) = 0$

$$W(j) = \min \begin{cases} W(j), \\ \min_{1 \le i < j} (V(i,j) + W(i-1)) \end{cases}$$

assumption: external bases do not contribute to overall energy.

$V(i,j)$: $V(i,j) =$

$$\begin{cases} +\infty & \text{for } i \ge j \\ \min(eH(i,j)), \\ eS(i,j) + V(i+1, j-1), \\ VBI(i,j) \\ VM(i,j) & \text{for } i < j \end{cases}$$

$$VBI(i,j) = \min_{\substack{i',j' \text{ s.t.} \\ i<i'<j'<j}} \{ eL(i,j,i',j') + V(i',j') \}$$

The following seems to require too much time but can be improved by a simple trick.

$$VM(i,j) = \min_{\substack{k, i_1, i_2 \cdots i_k j_1 \cdots j_k \\ \text{s.t. } i<i_1<j_1<i_2 \cdots <j_k<j}} \{ eM(i,j,i_1,j_1,\cdots i_k,j_k) + \sum_{h=1}^{k} V(i_h,j_h) \}$$

The speedup is achieved through the assumption that

$$eM(i,j,i_1,j_1 \cdots i_k,j_k) \simeq a + bk + c\left((i_1 - i - 1) + (j - j_k - 1) + \sum_{h=1}^{k-1} (i_{h+1} - j_h - 1)\right)$$

This ignores the content of the loop.

Define: $WM(i,i) = c$

$$WM(i,j) = \min \begin{cases} V(i,j) + b \\ \min_{i<h\leq h} (WM(i,h-1) + WM(h,j)) \end{cases} \quad \text{for } i<j$$

Thus $$VM(i,j) = \min_{i+1<h\leq j-1} (WM(i+1,h-1) + WM(h,j-1) + a)$$

Here $WM(i,j)$ is the free energy of an optimal structure for $s_i \cdots s_j$ assuming $s_i \cdots s_j$ is on a multi-branch loop.

Running Time

W : $O(n^2)$

V : $O(n^2)$

VBI : $O(n^4)$ → bottleneck

WM : $O(n^3)$

VM : $O(n^3)$