RNA Secondary Structure Prediction (continued)

Feb. 15, 2001
Notes: Jihong Ren

1. "Basic" energy functions:

- Stacked pairs

$$eS(i,j) = \begin{cases} -2 & \text{if } i,\ i+1 \in \{G,C\},\ j\ ,\ j-1 \text{ complement } i,\ i+1 \\ -1 & \text{if } i,\ i+1 \text{ not all } \{G,\ C\},\ j,\ j-1 \text{ complement } i,\ i+1 \\ +\infty & \text{otherwise} \end{cases}$$

- Hairpin

$$eH(I,\ j) = \begin{cases} +\infty \text{ if loop size } 0 \\ +3 \text{ if loop size } 1 \\ +2 \text{ if loop size } 2 \\ +1 \text{ if loop size } 3 \\ |\text{loop size}| \text{ otherwise} & //\text{loop size} = j - i - 1 \end{cases}$$

- Internal/Bulge loops

$eL(i,\ j\ ,\ i',\ j') = +1$

Here we only concern a bulge loop case.

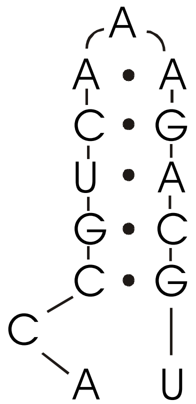
- Multiloop

$eM(i,\ j,\ i_1,\ j_1,\ \ldots\ ,\ i_k,\ j_k)$

Example:

Given above energy functions, and sequence 5'- ACCGUCAAAGACGU – 3', what is the total free energy of the sequence intuitively.

Intuitively, we can get the following secondary structure for above sequence.



Above structure contains : 1 hairpin with loop size 3, 3 stack pairs
Total energy = +1 + (-1) + (-1) + (-2) = -3.

Note: Minimum energy of secondary structure for a sequence is always <= 0 since a free structure without any base pairs has free energy 0.

2. Recursive Relations

- **W(j)**: minimum free energy of the strand $S_1 \dots S_j$

  $W(0) = 0$
  $W(j) = \min(W(j-1), \min_{1 \le i < j}(V(i,j) + W(i-1)))$, for j > 0

  First term W(j-1): j is not paired (external base, contribute nothing to free energy)
  Second term: The base S_j pairs with otherbase S_i in $S_1, S_2, \dots, S_{j-1}$, where i is chosen to minimize the resulting free energy. That energy is the sum of the energy V(i,j) of the compound structure closed by pair i·j, plus the energy W(i-1) of the remainder $S_1, S_2, \dots, S_{i-1}$. V(i, j) is defined to be the minimum energy of secondary structures on $S_i, \dots, S_j$, in which S_i is paired with S_j.

- **V(i,j)**: minimum energy of secondary structures on $S_i, \dots, S_j$, in which S_i is paired with S_j.

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

The terms in the second equation correspond to choosing the minimum free energy structure among the following possible solutions:

a. eH(i,j), i·j is the exterior pair in a hairpin loop.
b. i·j, and (i+1)·(j-1) forms stacked pair. V(i+1,j-1) is the energy of the compound structure closed by pair (i+1)·(j-1).
c. VBI(i, j), i·j is the exterior pair of a buldge or internal loop.
d. VM(i, j), i·j is the exterior pair of a multiloop.

- **VBI(i, j)**: minimum energy of secondary structures on $S_i, \dots, S_j$, in which i·j is the exterior pair of a buldge or internal loop.

$$VBI(i,j) = \min_{\substack{i',j' \\ i<i'<j'<j}} (\mathrm{eL}(i,j,i',j') + V(i',j'))$$

In this case, i·j is the exterior pair of a bulge or interior loop, and we must search all possible interior pairs i'·j' for the pair that results in the minimum free energy. For each such interior pair, the resulting free energy is sum of the energy of the bulge or internal loop eL(i, j, i', j'), plus the energy of the com-pound structure closed by i'·j'. It is easy to see that this search for the best interior pair is computationally intensive, simply because of the number of possibilities that must be considered. We will see later how to speed up this calculation, which is the new contribution of Lyngsø *et al.* [1].

- **VM(i,j)**: the free energy of the optimal structure for $S_i \ldots S_j$, assuming i·j closes a multibranched loop.

$$VM(i,j) = \min_{\substack{k, i_1, j_1, i_2, j_2, \ldots, i_k, j_k \\ i<i_1<j_1<i_2<j_2<\ldots<i_k<j_k<j \\ k\geq 2}} (\mathrm{eM}(i, j, i_1, j_1, i_2, j_2, \ldots, i_k, j_k) + \sum_{h=1}^{k} V(i_h, j_h))$$

In the same way that the recurrence for VM requires a search for the best structure among all the possible interior pairs, the calculation for VM is even more intensive, requiring a search for k interior pairs $i_h \cdot j_h$, each of which closes its own branch out of the multibranched loop and contributes free energy $V(i_h, j_h)$. A direct implementation of the calculation shown for is infeasibly slow.

- **WM(i, j)**: used to compute VM. WM(i,j) gives the free energy of an optimal structure for $S_i, \ldots, S_j$, assuming that S_i and S_j are on a multibranched loop.

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

The terms in the second equation correspond to the following possible solutions:

a. i·j forms a base pair and therefore defines one of the k branches, whose free energy is V(i,j).
b. S_i and S_j are not paired with each other, so the free energy is given by the minimum partition of the sequence into two contiguous subsequences.

Calculating VM then reduces to partitioning the loop into at least two pieces with the minimum total free energy:

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

3. Example:

S1 S2 S3 S4 S5
G G C C C

Ignore the multiloops, alg. Mantains 3 arrays: V(i,j), VBI(i,j), W(i).

V(I,J)	1	2	3	4	5
1	∞	∞	+3	+2	+1
2		∞	∞	+3	+2
3			∞	∞	∞
4				∞	∞
5					∞

VBI(I,J)	1	2	3	4	5
1	∞	∞	∞	∞	∞
2		∞	∞	∞	∞
3			∞	∞	∞
4				∞	∞
5					∞

	0	1	2	3	4	5
W(I)	0	0	0	0	0	0

4. Running time: (first ignoring multiloops)

The running time to fill in each of the complete tables (assuming the values on which it depends have already been computed and stored in their tables) is determined as follows:

- W: O(n^2). Each of n entries requires the computation of the min of O(n) terms.
- V: O(n^2). Each of O(n^2) entries requires the computation of the min of 4 terms.
- VBI: O(n^4). Each of O(n^2) entries requires the computation of the min of O(n2) terms.

For multiloops, if we assume that:

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

Where a,b and c are constants. a – cost for starting a multiloop, b – charge for 1 branch, c – charge associates with each unpaired bases on the loop.
Above assumption makes calculate VM reasonably efficiently although it has been suggested that it would be more accurate to approximate the free energy as a logarithmic function of the loop size.

There is a way to get the time complexity of calculating VBI to O(n^3). Please refer to notes from University of Washington.