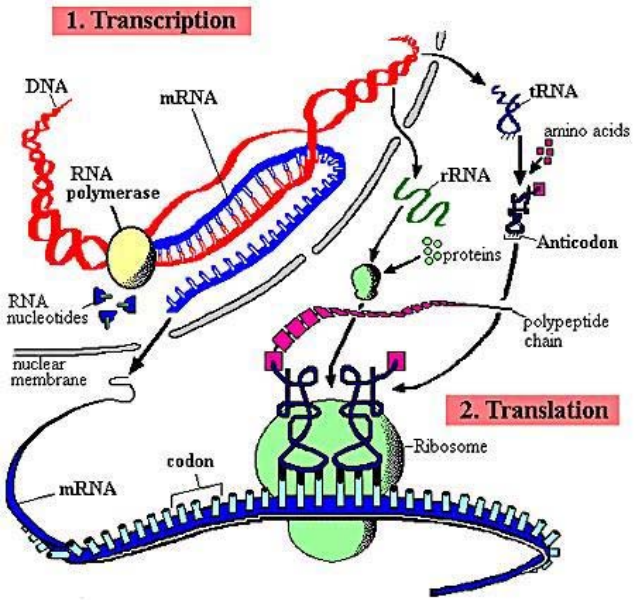
RNA Secondary Structure Prediction

Chin Lung Lu

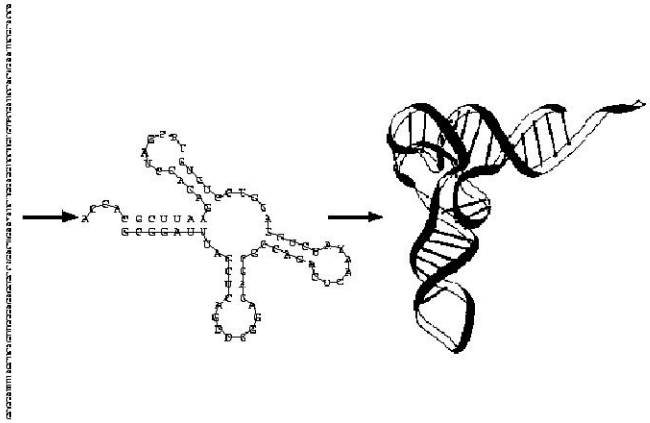
Computational Biology

Analyses and Applications of Sequences



RNA

- The function of an RNA (mRNA, tRNA, rRNA) is determined by its 3D structure.

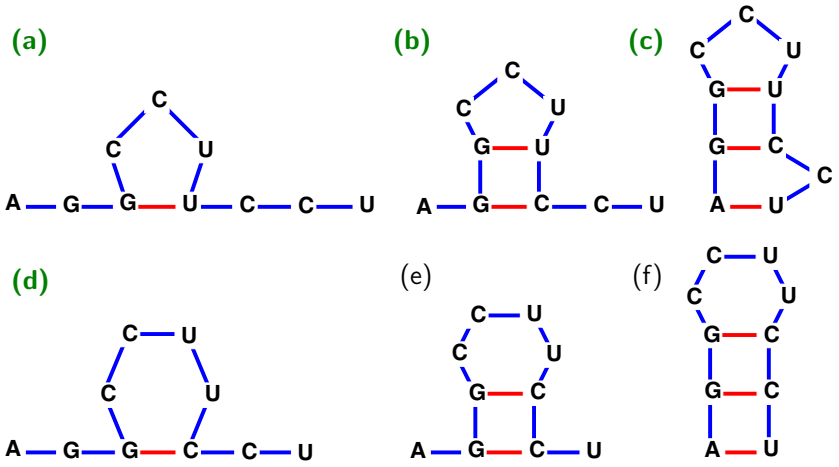


RNA secondary structure

- Primary structure of an RNA: a sequence of the bases A, G, C and U
- Due to hydrogen bonds, the bases of an RNA may form the base pair.
  1. Watson-Click base pairs:
     - G≡C: formed by a triple-hydrogen bond
     - A=U: formed by a double-hydrogen bond
  2. Wobble base pairs:
     - G−U: formed by a single hydrogen bond
- Secondary structure of an RNA: the Watson-Crick and wobble base pairs occurring in the RNA fold

RNA secondary structure

- **Example**: RNA = **A– G– G– C– C– U– U– C– C– U**



RNA secondary structure

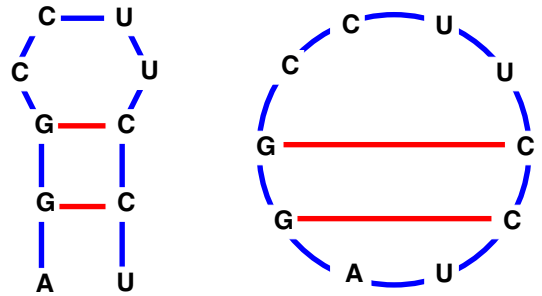
- What is the actual secondary structure of an RNA sequence?
- By the thermodynamic hypothesis, the actual secondary structure of an RNA sequence is the one with the minimum free energy.
  - The base pairs will increase the structural stability, but the unpaired bases will decrease the structural stability.
  - Watson-Crick base pairs are more stable than wobble base pairs.

Secondary structure of RNA

- RNA sequence: a string of n characters $R = r_1 r_2 \cdots r_n$, where $r_i \in \{A, C, G, U\}$
- Secondary structure of R: a set S of base pairs (r_i, r_j), where $1 \leq i < j \leq n$, such that
  1. $j - i > t$, t is a small positive constant (typically, $t = 3$ which means RNA does not fold too sharply on itself)
  2. If $(r_i, r_j), (r_k, r_l) \in S$ and $i \leq k$, then either
     - $i = k$ and $j = l$, $(r_i, r_j) = (r_k, r_l)$
     - $i < j < k < l$, (r_i, r_j) precedes (r_k, r_l), or
     - $i < k < l < j$, (r_i, r_j) includes (r_k, r_l)

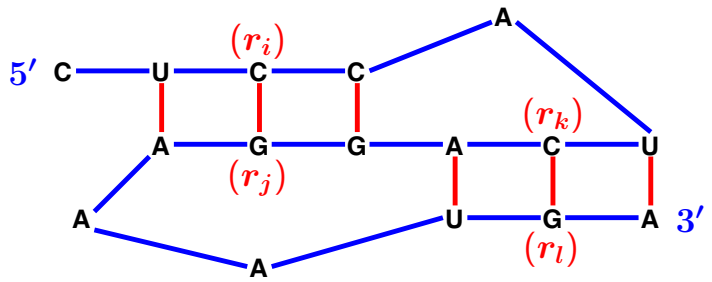
Outerplanar graph

- A secondary structure can be represented as an outerplanar graph with degree at most 3.
- Outerplanar graph: a graph in which all vertices are arranged on a circle and all edges lie inside the circle and do not intersect



Pseudoknot of an RNA

- Two base pairs (r_i, r_j) and (r_k, r_l) are called a pseudoknot if $i < k < j < l$.
- Pseudoknots do occur in RNA molecules, but their exclusion simplifies the problem.



Free energy

- How to calculate the free energy of a secondary structure S?
- Assign an energy to each base pair of S and then the free energy of S is the sum of the energies of all base pairs.
  - **Weighted version**: the energies for G≡C, A=U and G–U are different
  - **Unweighted version**: the energies of base pairs are all equal (maximum base pair matching problem: find a secondary structure with maximum number of base pairs)

RNA maximum base pair matching

- Given an RNA sequence $R = r_1 r_2 \cdots r_n$, find a secondary structure of R with maximum number of base pairs.
- $S_{i,j}$: the secondary structure of maximum number of base pairs on the substring $R_{i,j} = r_i r_{i+1} \cdots r_j$
- $\mathcal{M}_{i,j}$: the size of $S_{i,j}$ (i.e., $|S_{i,j}|$)
- $\mathcal{WW}$: {(A, U), (U, A), (G, C), (C, G), (G, U), (U, G)}

RNA maximum base pair matching

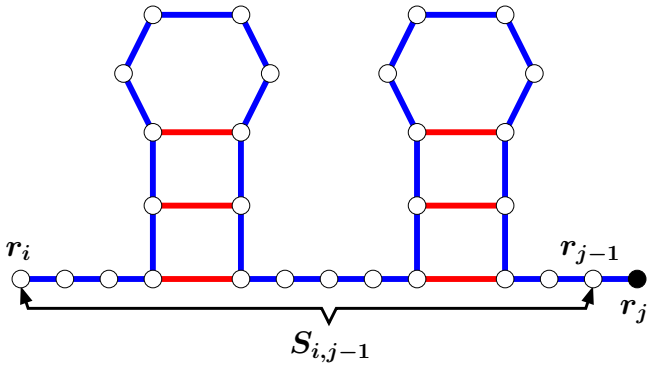
- $\rho(r_i, r_j)$: indicate whether any two bases r_i and r_j can be a legal base pair:

$$\rho(r_i, r_j) = \begin{cases} 1 & \text{if } (r_i, r_j) \in \mathcal{WW} \\ 0 & \text{otherwise} \end{cases}$$

- By definition, if $j - i \leq 3$, then r_i and r_j cannot be a base pair of $S_{i,j}$.
  - Hence, we let $\mathcal{M}_{i,j} = 0$ if $j - i \leq 3$.

Compute $\mathcal{M}_{i,j}$ for $j - i > 3$

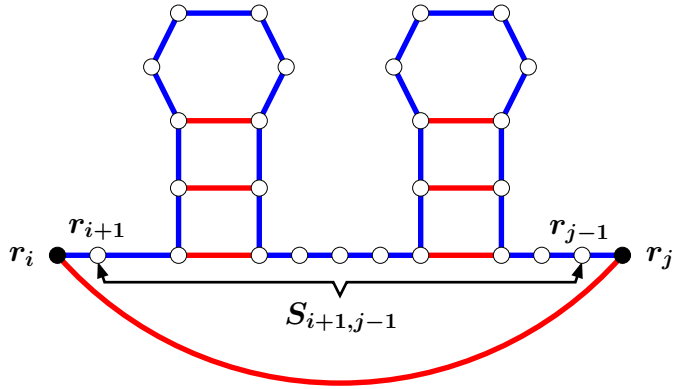
Case 1: In the optimal solution, r_j is not paired with any other base



$$\mathcal{M}_{i,j} = \mathcal{M}_{i,j-1}$$

Compute $\mathcal{M}_{i,j}$ for $j - i > 3$

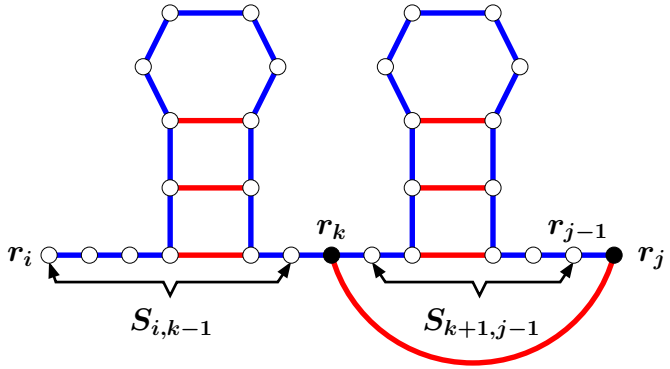
Case 2: In the optimal solution, r_j is paired with r_i and $\rho(r_i, r_j) = 1$



$$\mathcal{M}_{i,j} = 1 + \mathcal{M}_{i+1,j-1}$$

Compute $\mathcal{M}_{i,j}$ for $j - i > 3$

Case 3: In optimal solution, r_j is paired with some r_k, $i + 1 \leq k \leq j - 4$ and $\rho(r_k, r_j) = 1$



$$\mathcal{M}_{i,j} = \max_{i+1 \leq k \leq j-4} \{1 + \mathcal{M}_{i,k-1} + \mathcal{M}_{k+1,j-1}\}$$

RNA maximum base pair matching

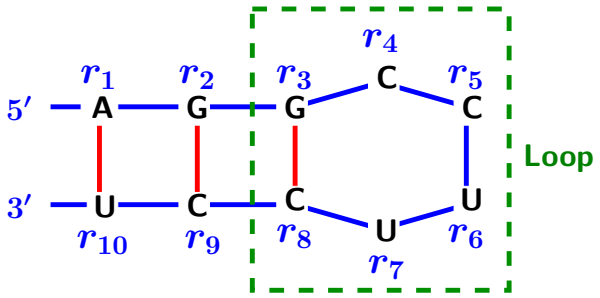
In summary, we have the following recursive formula to compute $\mathcal{M}_{i,j}$.

- If $j - i \leq 3$, then $\mathcal{M}_{i,j} = 0$.
- If $j - i > 3$, then $\mathcal{M}_{i,j} = \max\{C1, C2, C3\}$
  - $C1 = \mathcal{M}_{i,j-1}$
  - $C2 = (1 + \mathcal{M}_{i+1,j-1}) \times \rho(r_i, r_j)$
  - $C3 =$

$$\max_{i+1 \leq k \leq j-4} \left\{ \begin{array}{l} (1 + \mathcal{M}_{i,k-1} + \mathcal{M}_{k+1,j-1}) \times \\ \rho(r_k, r_j) \end{array} \right\}$$

Loops of RNA secondary structure

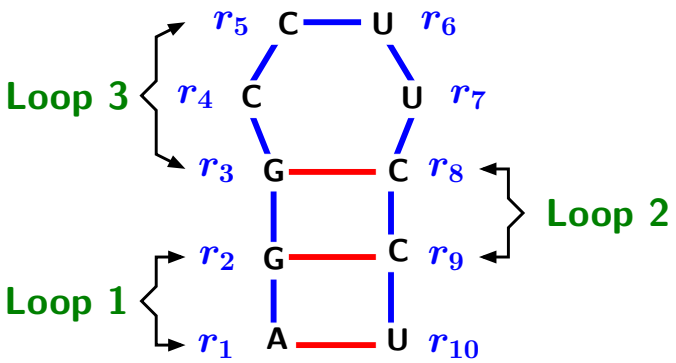
- Loop of a secondary structure: a substructure consisting of a base pair (r_i, r_j) and all bases accessible from (r_i, r_j)
  - r_p is accessible from (r_i, r_j): $i < p < j$ and no other $(r_{i'}, r_{j'})$ with $i < i' < p < j' < j$



Loops of RNA secondary structure

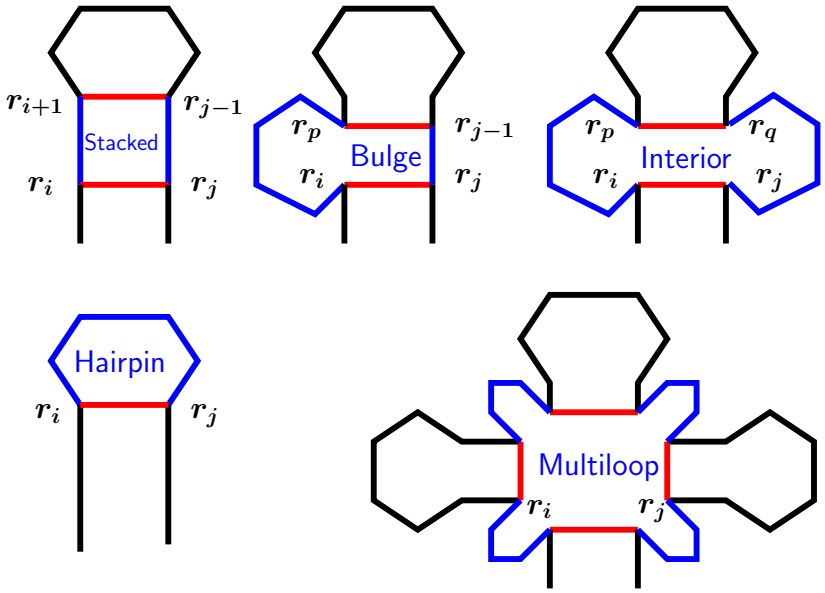
- Every base pair (r_i, r_j) corresponds to a loop:
  - Exterior (closing) base pair: (r_i, r_j)
  - Interior base pairs: those base pairs (r_p, r_q) accessible from (r_i, r_j) (i.e., both r_p and r_q are accessible from (r_i, r_j))
  - Size of a loop: the number of unpaired bases in the loop
  - Degree of a loop: the number of base pairs in the loop

Loop	Exterior BP	Interior BP	Size	Degree
1	(r_1, r_{10})	(r_2, r_9)	0	2
2	(r_2, r_9)	(r_3, r_8)	0	2
3	(r_3, r_8)	NO	4	1



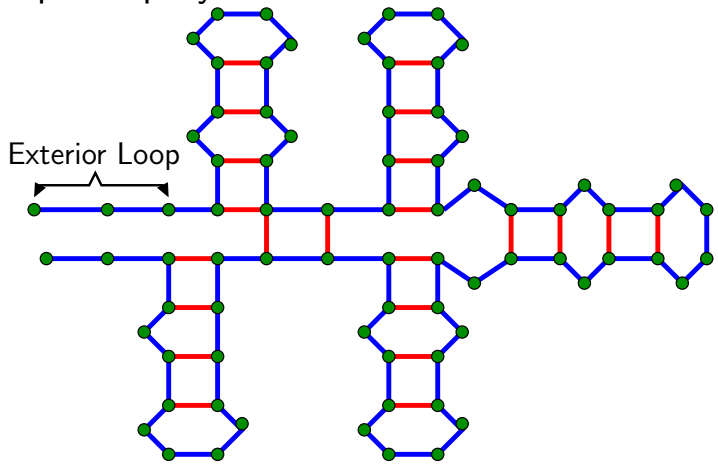
Various types of loops

- Hairpin loop: a loop of degree 1
- Stacked pair: a loop of degree 2 whose size is zero
- Bulge loop: a loop of degree 2 and non-zero size whose exterior and interior base pairs are adjacent
- Interior loop: a loop of degree 2 and non-zero size whose exterior and interior base pairs are not adjacent
- Multiloop: a loop of degree greater than 2
- Exterior loop: the collection of adjacent unpaired bases which are not accessible by any base pair



Unique loop decomposition

- Any secondary structure can be decomposed into loops uniquely.



Free energies of loops

- $\mathcal{H}(i,j)$: the free energy of a hairpin loop closed by base pair (i,j)
- $\mathcal{S}(i,j)$: the free energy of stacking base pair (i,j) with base pair $(i+1,j-1)$
- $\mathcal{BI}(i,j,p,q)$: the free energy of a bulge or interior loop with closing base pair (i,j) and interior base pair (p,q)
- $\mathcal{M}(i,j)$: the free energy of a multiloop with closing base pair (i,j), usually expressed as $a + b \times (deg - 1) + c \times size$, where a, b and c are constants

Loop free energy minimization

- Free energy of a secondary structure: the sum of the free energies of all loops in the secondary structure
- Loop dependent energy minimization problem: given an RNA sequence, find an optimal secondary structure (i.e., a secondary structure with the minimum free energy)
  - Solvable in $\mathcal{O}(n^4)$ time using the technique of dynamic programming
  - Can be further improved to $\mathcal{O}(n^3)$ time

Loop energy minimization

- $\rho(r_i, r_j)$: indicate whether any two bases r_i and r_j can be a legal base pair:

$$\rho(r_i, r_j) = \begin{cases} 1 & \text{if } (r_i, r_j) \in \mathcal{WW} \\ 0 & \text{otherwise} \end{cases}$$

- $S_{i,j}$: denote the optimal secondary structure of the substring $R_{i,j} = r_i r_{i+1} \cdots r_j$
- $\mathcal{E}_{i,j}$: denote the free energy of $S_{i,j}$

Computation of $\mathcal{E}_{i,j}$ ❶

Case 1: In the optimum solution, r_j is not paired with any other base. Then we have $\mathcal{E}_{i,j} = \mathcal{E}_{i,j-1}$.

Case 2: In the optimum solution, r_i is paired with r_j and $\rho(r_i, r_j) = 1$.

- Then there may be one or more loops between r_i and r_j.
- Let $L_{i,j}$ denote the structure with the minimum free energy in this case.
- Let $\mathcal{F}_{i,j}$ denote the free energy of $L_{i,j}$.
- Then we have $\mathcal{E}_{i,j} = \mathcal{F}_{i,j}$.

Computation of $\mathcal{E}_{i,j}$ ❷

Case 3: In the optimum solution, r_i is paired with some r_k, $i+1 \le k \le j-1$, and $\rho(r_k, r_j) = 1$.

- In this case, we can divide $R_{i,j}$ into two subsequences $R_{i,k-1}$ and $R_{k,j}$ such that

$$\mathcal{E}_{i,j} = \mathcal{E}_{i,k-1} + \mathcal{F}_{k,j}$$

- Since we want to find the k between $i+1$ and $j-1$ such that $\mathcal{E}_{i,j}$ is the minimum, we have

$$\mathcal{E}_{i,j} = \min_{i+1 \le k \le j-1} \{\mathcal{E}_{i,k-1} + \mathcal{F}_{k,j}\}$$

Computation of $\mathcal{E}_{i,j}$ ❸

In summary, we have the following recursive formula:

$$\mathcal{E}_{i,j} = \min \begin{cases} \mathcal{E}_{i,j-1} \\ \mathcal{F}_{i,j} \cdot \rho(r_i, r_j) \\ \min\limits_{i+1 \le k \le j-1} \{(\mathcal{E}_{i,k-1} + \mathcal{F}_{k,j}) \cdot \rho(r_k, r_j)\} \end{cases}$$

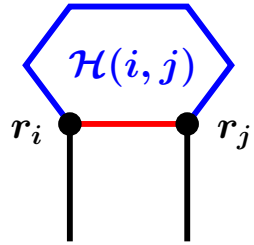
By definition, r_i and r_j cannot form a base pair if $j - i \le t = 3$ since $R_{i,j}$ does not fold itself too sharply. Hence, we have

$$\mathcal{E}_{i,j} = \mathcal{F}_{i,j} = +\infty \text{ if } j - i \le 3.$$

Computation of $\mathcal{F}_{i,j}$ ❶

Note that (r_i, r_j) is a base pair in $L_{i,j}$, (r_i, r_j) must be an exterior base pair of some one loop, say $\mathcal{L}$.

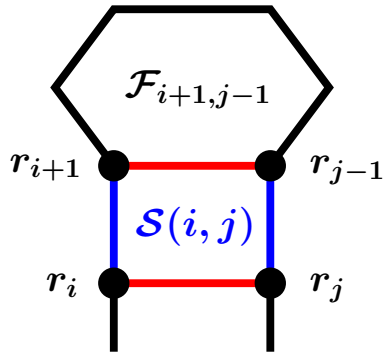
Case 1: $\mathcal{L}$ is a hairpin loop



$$\therefore \mathcal{F}_{i,j} = \mathcal{H}(i,j)$$

Computation of $\mathcal{F}_{i,j}$ ❷

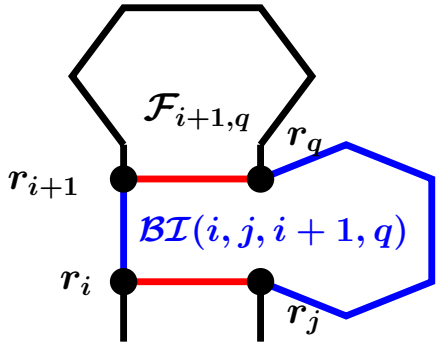
Case 2: $\mathcal{L}$ is a stacked loop closed by base pair (r_i, r_j)



$$\therefore \mathcal{F}_{i,j} = \mathcal{S}(i,j) + \mathcal{F}_{i+1,j-1}$$

Computation of $\mathcal{F}_{i,j}$ ❸

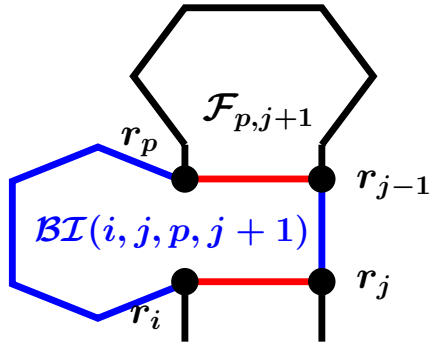
Case 3.1: $\mathcal{L}$ is a bugle loop with the interior base pair (r_{i+1}, r_q), where $i+2 \le q \le j-2$



$$\therefore \mathcal{F}_{i,j} = \min_{i+2 \le q \le j-2} \{\mathcal{BI}(i,j,i+1,q) + \mathcal{F}_{i+1,q}\}$$

Computation of $\mathcal{F}_{i,j}$ ❹

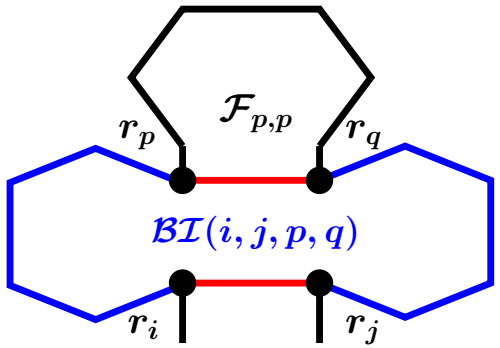
Case 3.2: $\mathcal{L}$ is a bugle loop with the interior base pair (r_p, r_{j-1}), where $i+1 \le p \le j-2$



$$\therefore \mathcal{F}_{i,j} = \min_{i+2 \le p \le j-2} \{\mathcal{BI}(i,j,p,j-1) + \mathcal{F}_{p,j+1}\}$$

Computation of $\mathcal{F}_{i,j}$ ❺

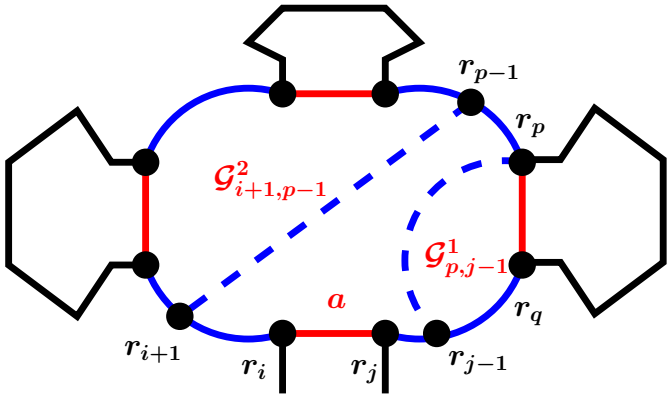
Case 4: $\mathcal{L}$ is an interior loop with the interior base pair (r_p, r_q), where $i+2 \le p < q \le j-2$



$$\therefore \mathcal{F}_{i,j} = \min_{i+2 \le p < q \le j-2} \{\mathcal{BI}(i,j,p,q) + \mathcal{F}_{p,q}\}$$

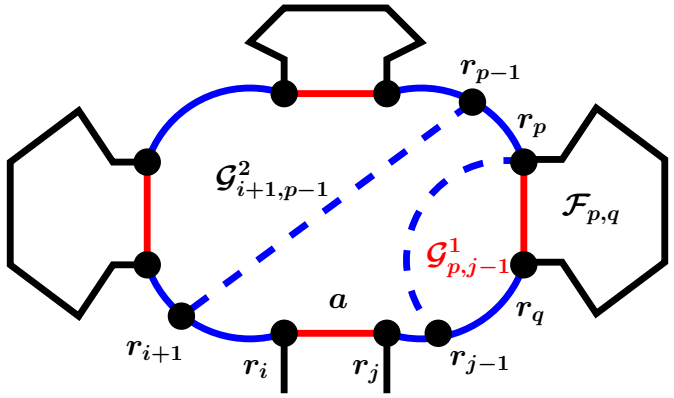
Computation of $\mathcal{F}_{i,j}$ ❻

Case 5: $\mathcal{L}$ is a multiloop, where suppose that (r_p, r_q) is the rightmost interior base pair of $\mathcal{L}$



$$\therefore \mathcal{F}_{i,j} = \min_{i-1 < p < j} \{a + \mathcal{G}^1_{p,j-1} + \mathcal{G}^2_{i+1,p-1}\}$$

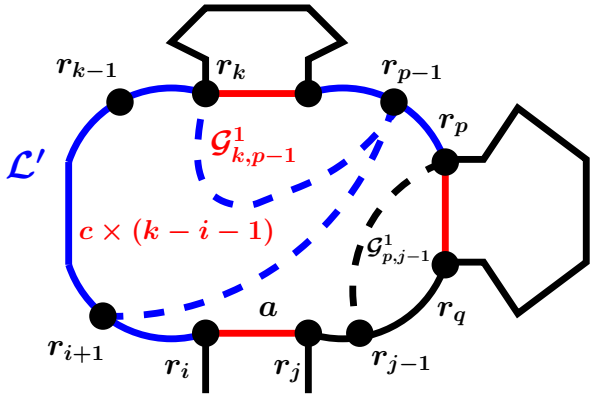
Computation of $\mathcal{G}^1_{p,j-1}$ ❼



$$\therefore \mathcal{G}^1_{p,j-1} = \min_{p < q < j-1} \{\mathcal{F}_{p,q} + b + c \times (j-q-1)\}$$

Computation of $\mathcal{G}^2_{i+1,p-1}$ ❽

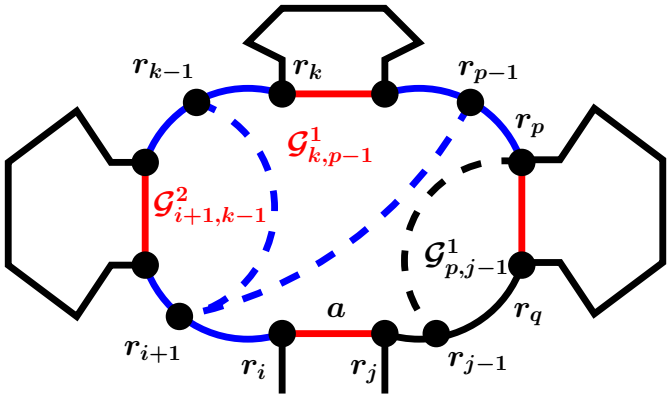
Case 2: Suppose that $\mathcal{L}'$ contains only one loop.



$$\therefore \mathcal{G}^2_{i+1,p-1} = \min_{i < k < p-1} \{\mathcal{G}^1_{k,p-1} + c \times (k-i-1)\}$$

Computation of $\mathcal{G}^2_{i+1,p-1}$ ⑨

Case 2: Suppose that $\mathcal{L}'$ contains two or more loops.



$$\therefore \mathcal{G}^2_{i+1,p-1} = \min_{i<k<p-1} \{\mathcal{G}^1_{k,p-1} + \mathcal{G}^2_{i+1,k-1}\}$$