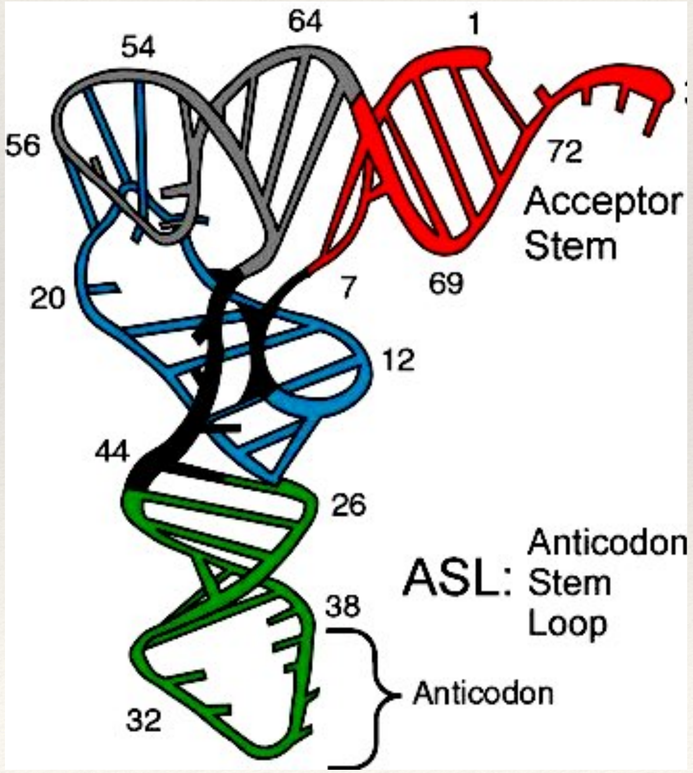
RNA Structure Prediction

Hierarchical organization of RNA molecules

Primary structure:

5' ACCACCUGCUGA 3'

Tertiary structure:



Secondary Structure

Hierarchical organization of RNA molecules

Primary structure:

5’ to 3’ list of covalently linked nucleotides, named by the attached base

Secondary Structure

List of base pairs, denoted by i•j for a pairing between the i-th and j-th Nucleotides, r_i and r_j, where $i<j$ by convention.
Pairing mostly occur as A•U and G•C (Watson Crick), and G •U (wobble)
By definition, base pairs in secondary structure are nested: if i is paired with j, Then i+1 can only be paired with k such that $i+2<k<j$.
Helices are inferred when two or more base pairs occur adjacent to one another

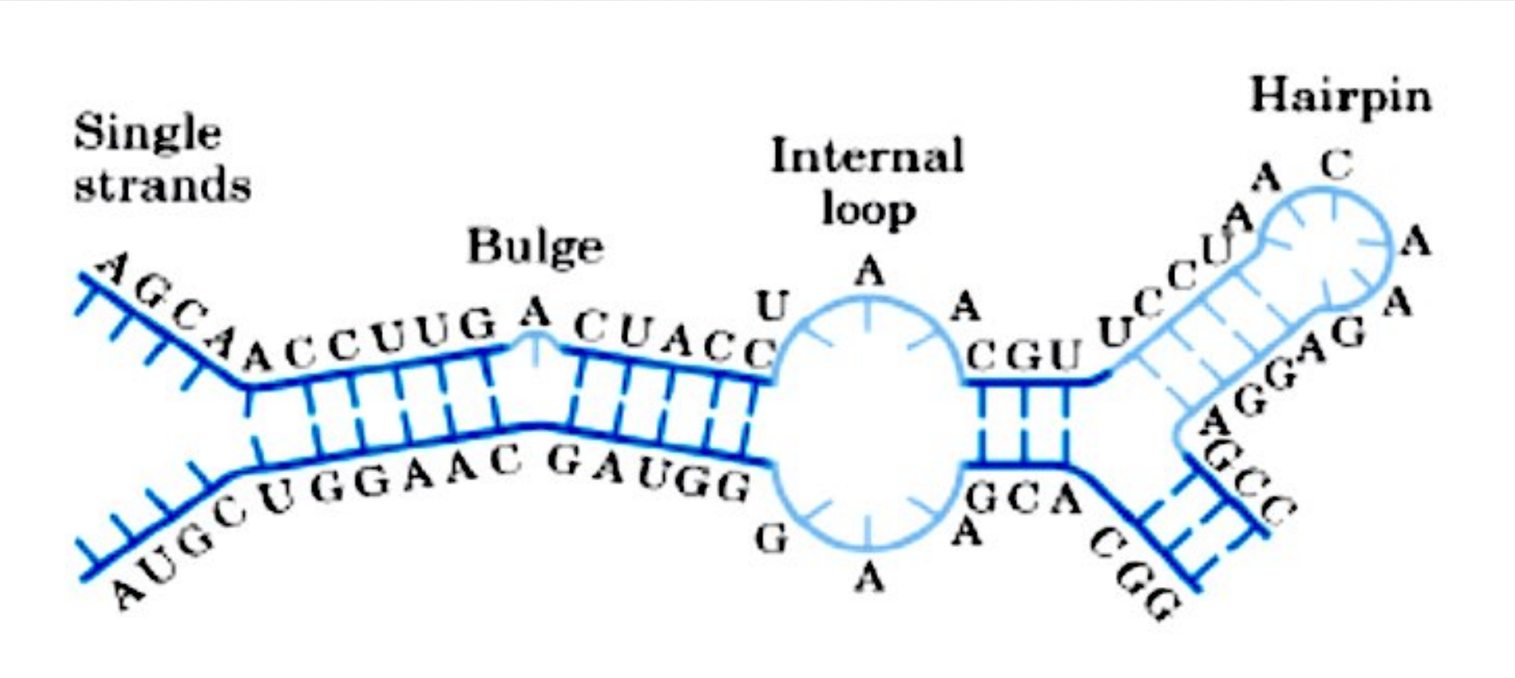
Tertiary structure:

List of interactions between secondary structures

RNA secondary structures

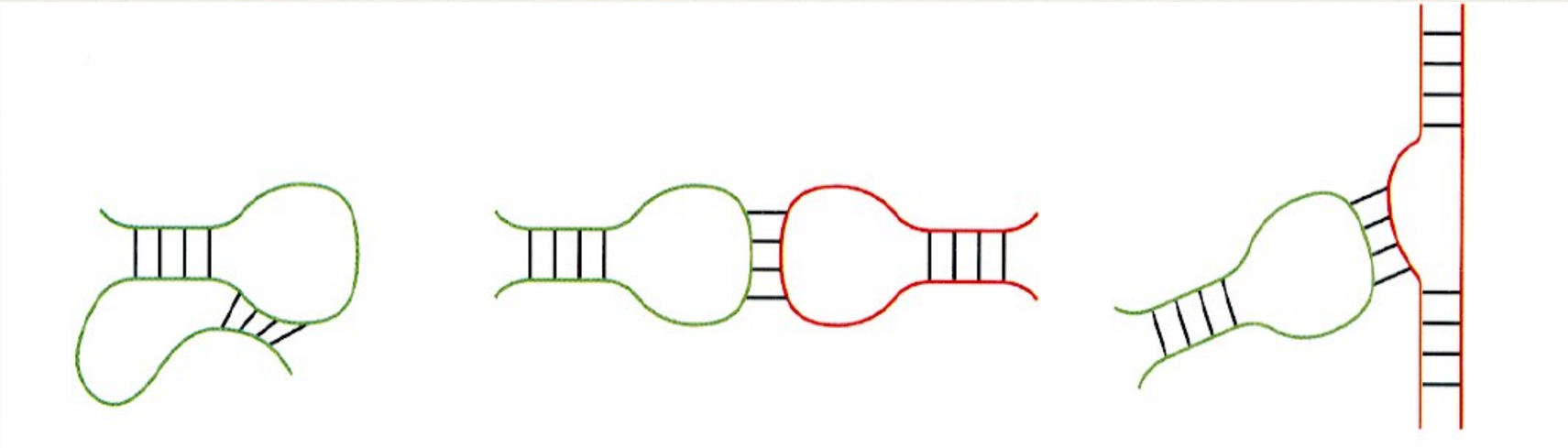
Single stranded bases within a stem are called a bulge of bulge loop if the single stranded bases are on only one side of the stem.

If single stranded bases interrupt both sides of a stem, they are called an internal (interior) loop.



RNA “tertiary interactions”

In addition to secondary structural interactions in RNA, there are also tertiary interactions, including: (A) pseudoknots, (B) kissing hairpins and (C) hairpin-bulge contact.



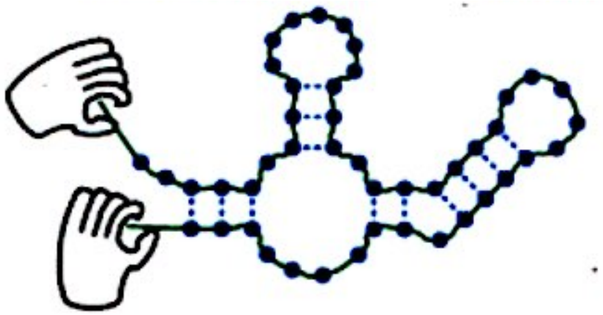
Pseudoknot **Kissing hairpins** **Hairpin-bulge**

RNA secondary structure representation

- Grammatically correct string of parentheses

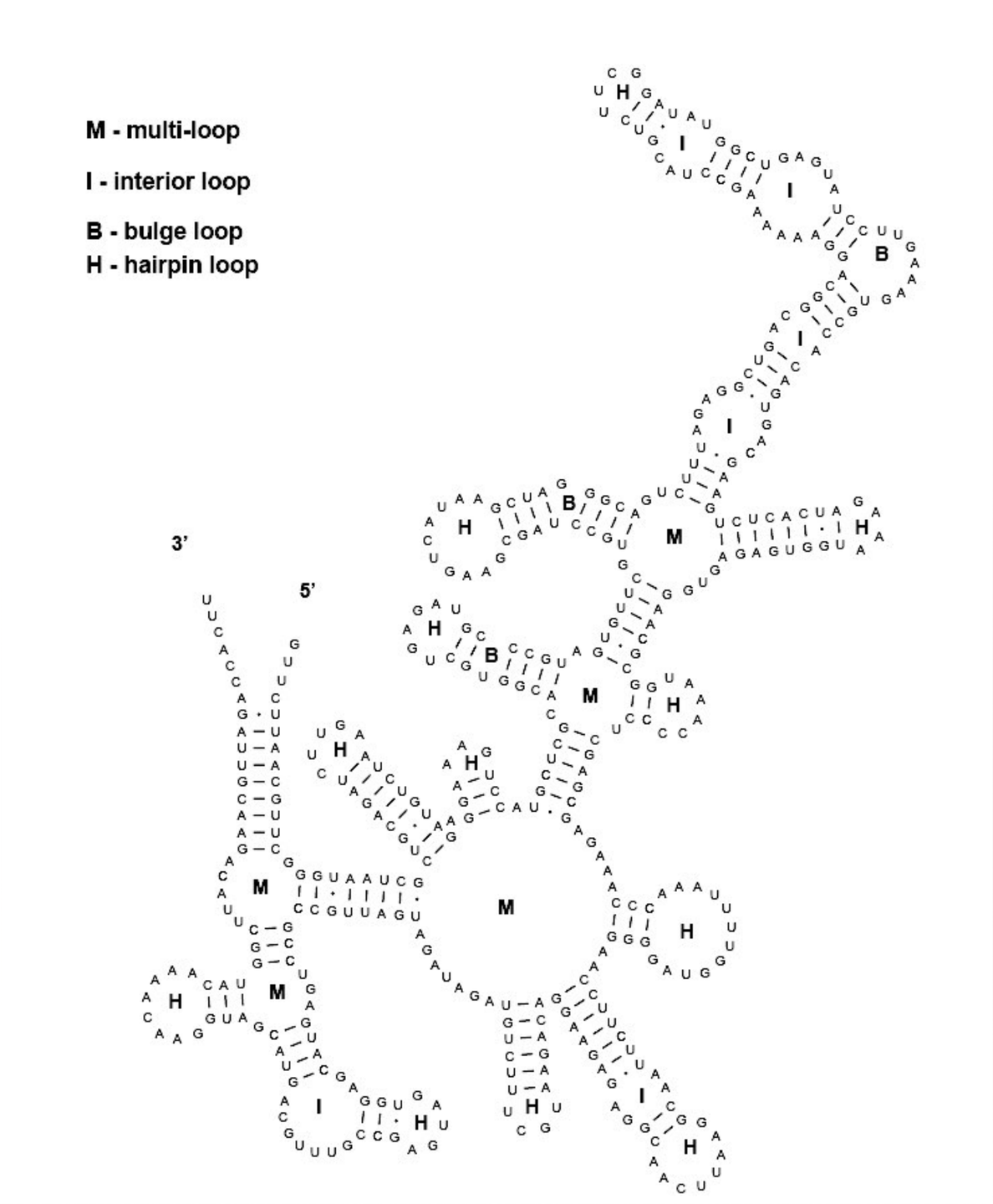
```
..(((.(((.....))).((((((....)))).)).....)))
AGCUACGGAGCGAUCUCCGAGCUUUCGAGAAAGCCUCUAUUAGC
```

- Planar graph



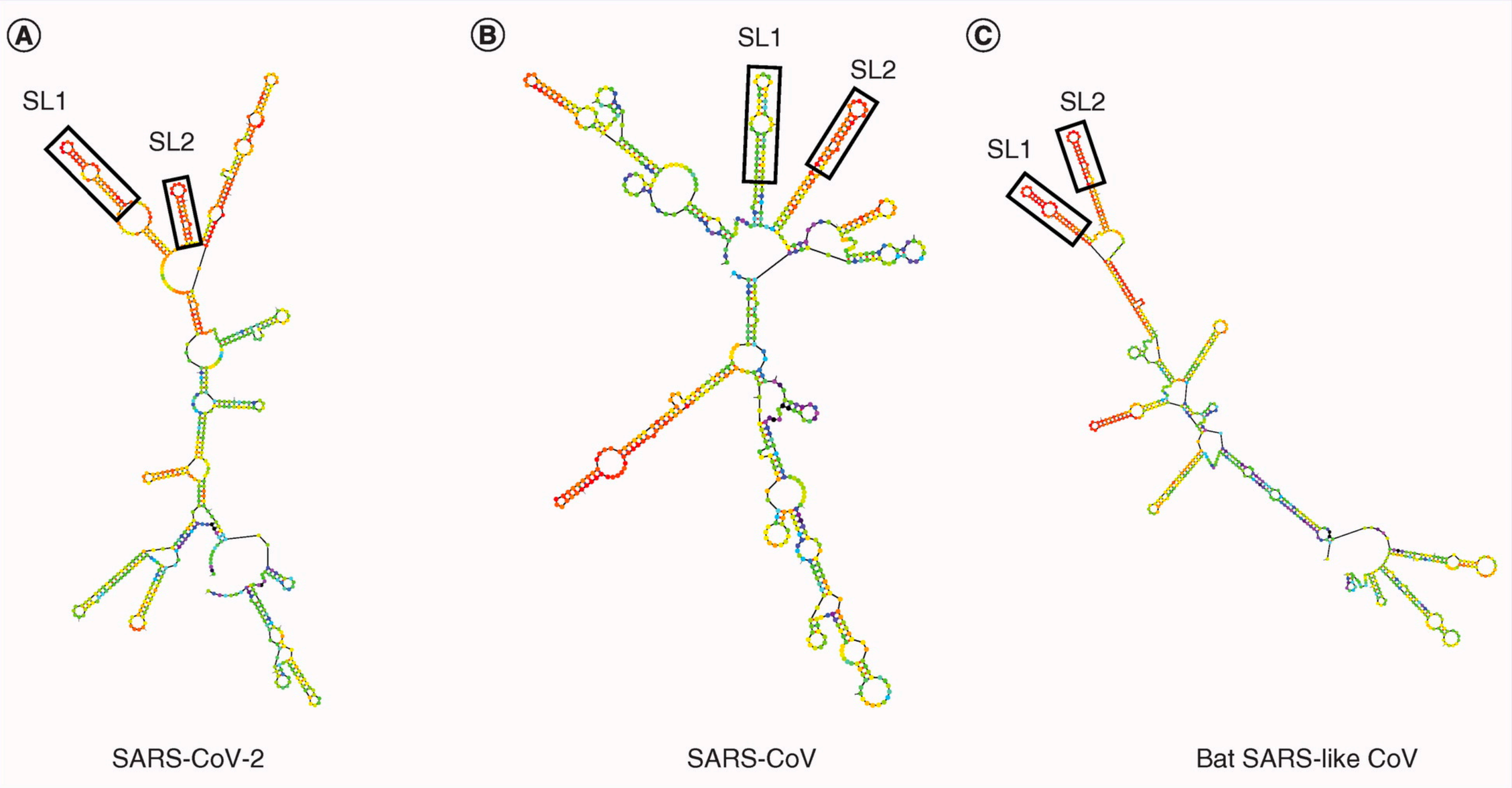
- Arch diagram

- Mountain diagram



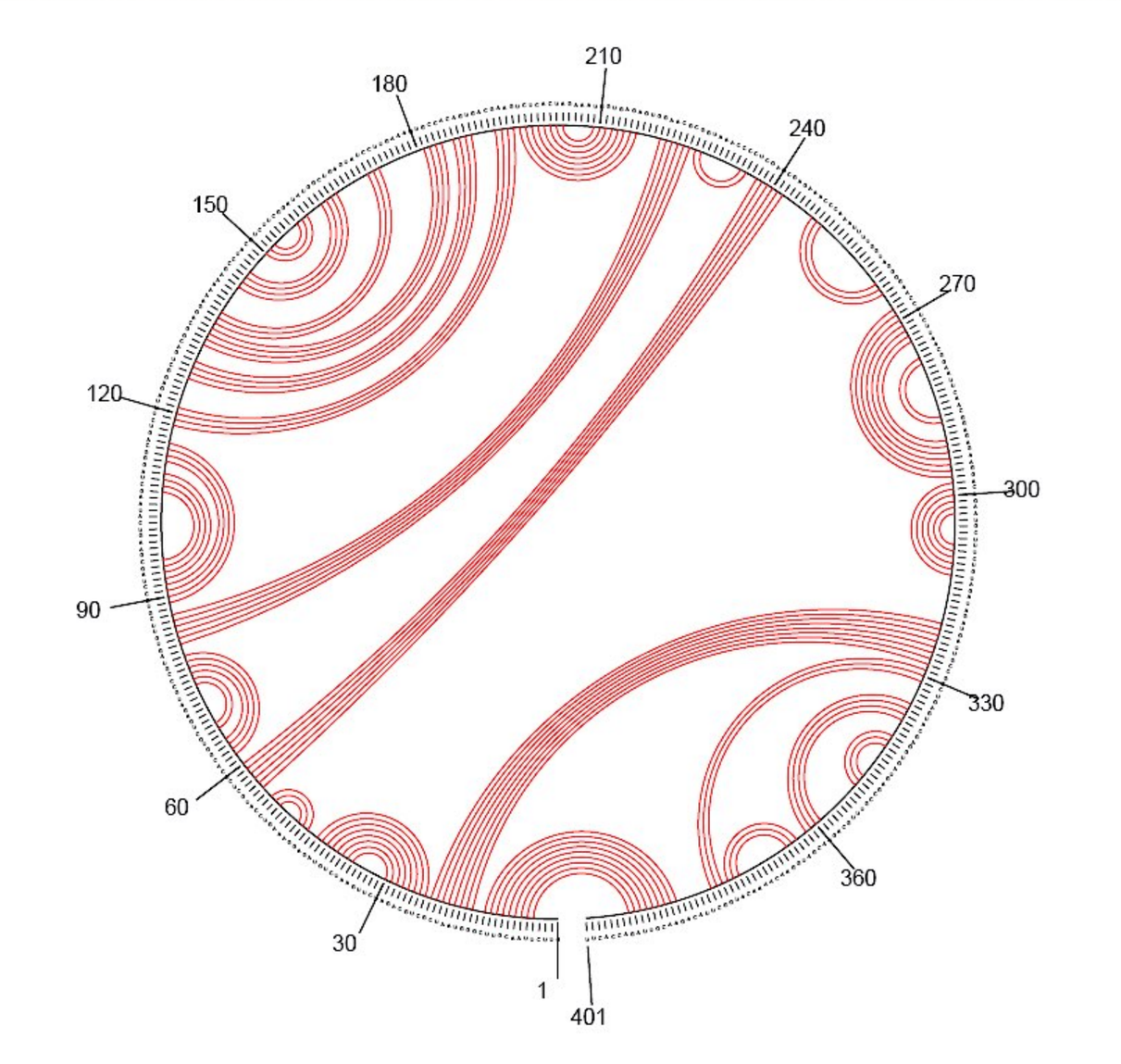
Predicted secondary structure for Bacillus Subtilis RNase P RNA

(from Zuker)


A
SL1
SL2
SARS-CoV-2
B
SL1
SL2
SARS-CoV
C
SL2
SL1
Bat SARS-like CoV

RNA secondary structure representation

Circular representation:

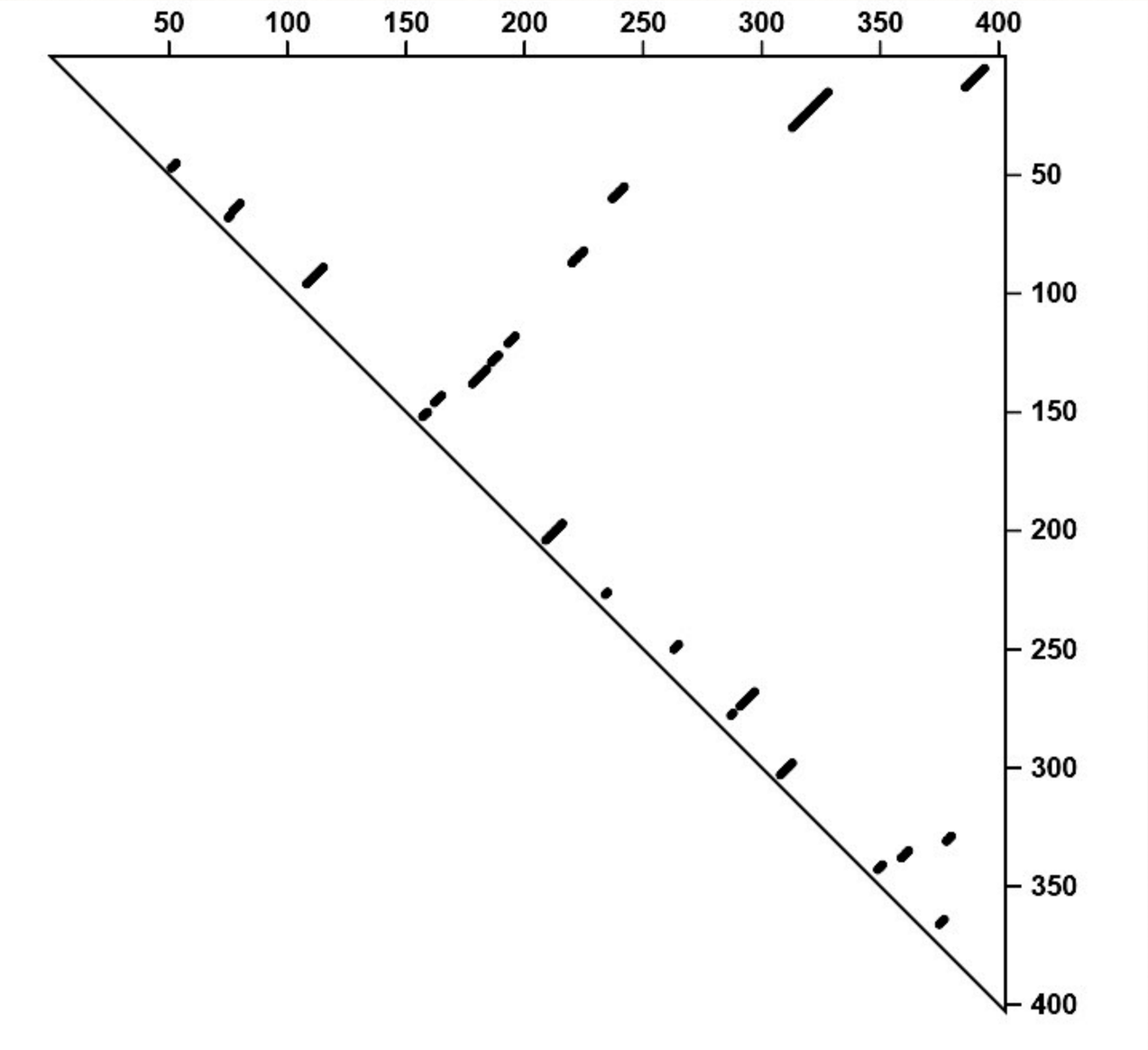


Bacillus Subtilis RNase P RNA

RNA secondary structure representation

DotPlot representation of the same Bacillus Subtilis RNA folding:

A dot is placed to represent a base pair

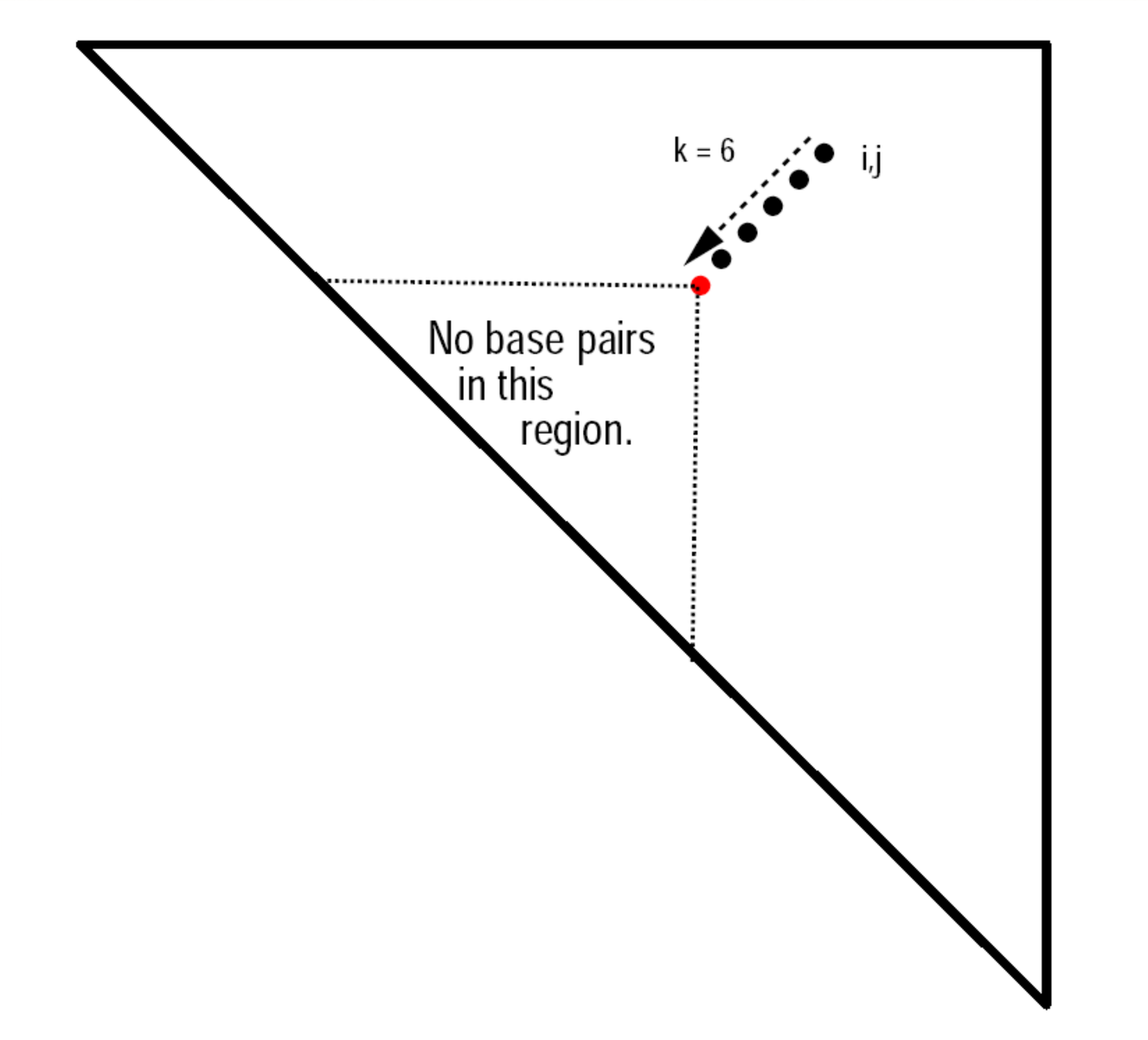


Understanding RNA structure dot plot

Simple stem loop:

-Single helix closed by the base pair i•j.
The other base pairs are (i+1) •(j-1) … (i+5) •(j-5) (6 total)

-The last base pair, shown in red, closes a hairpin loop.
If k •l closes a hairpin loop, there can be no base pair k' •l' such that k<k'<l'<l



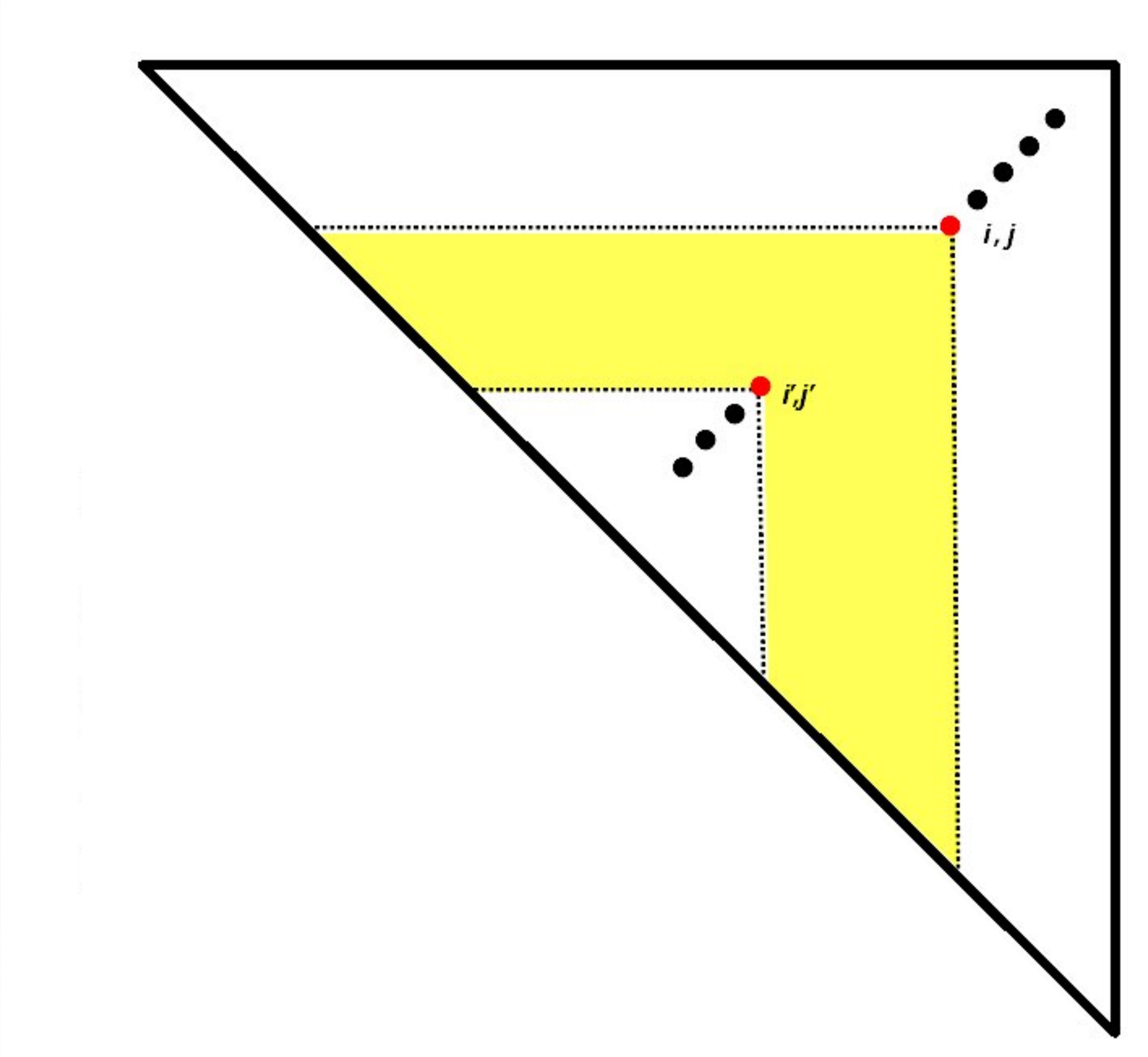
Understanding RNA structure dot plot

Interior loop (or bulge):

i•j and i' •j' close an interior loop if i<i'<j'<j and max{i'-i, j-j'}>1.

It is a bulge loop if min{i'-i,j-j'}=1.

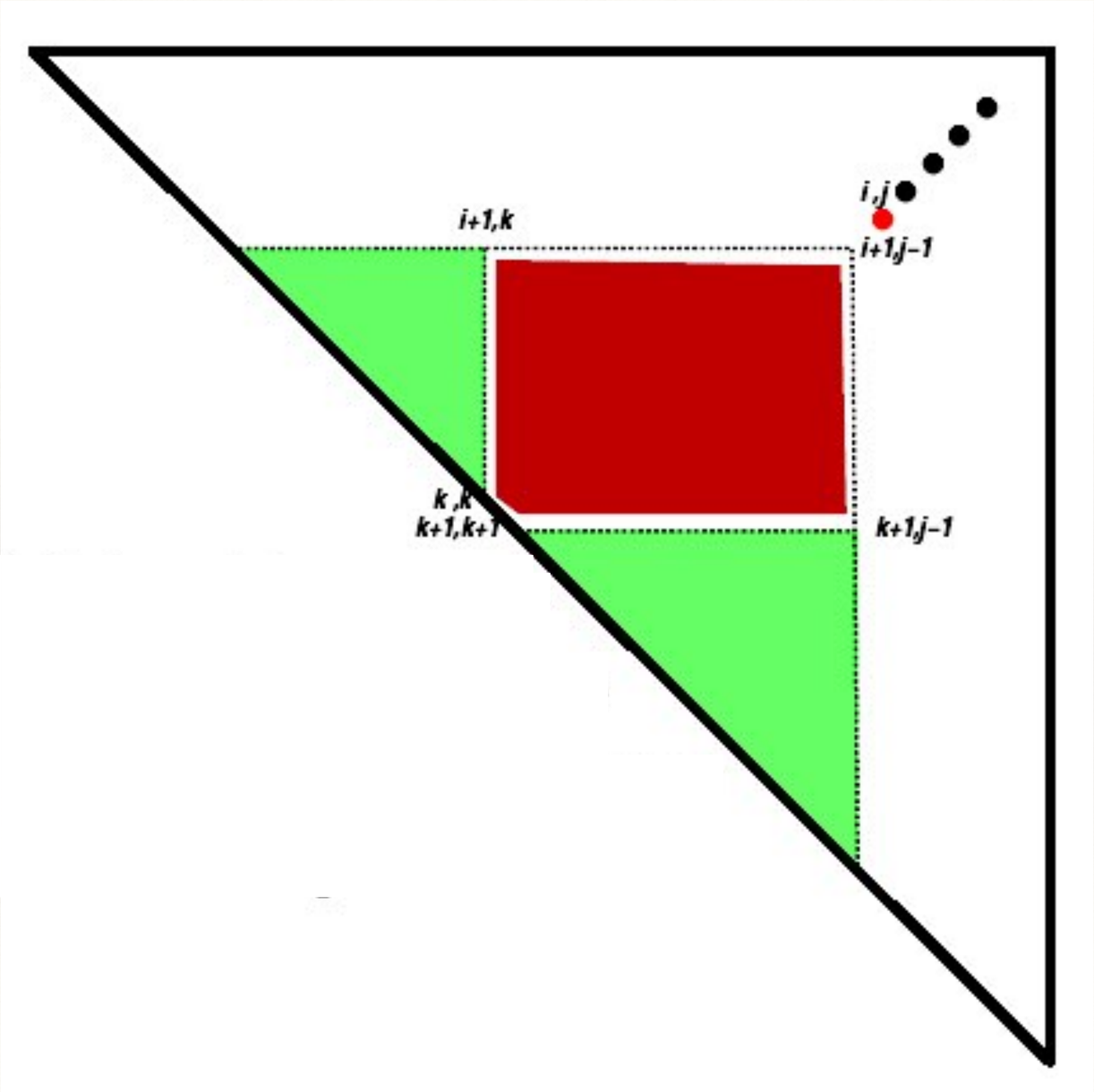
The yellow area is empty of base pair.



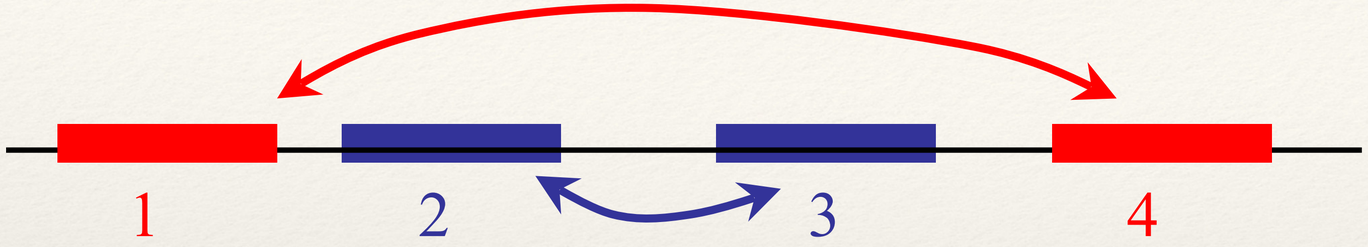
Understanding RNA structure dot plot

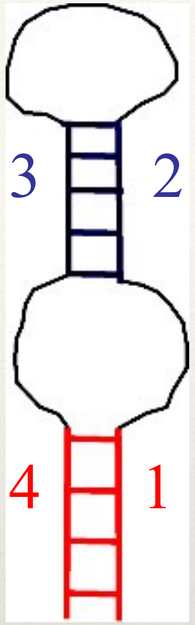
Multi-branch loop:

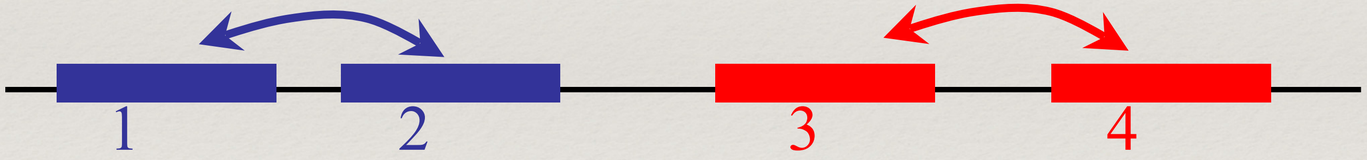
Base pair i•j closes a multi-branch if and only if there is a k, i<k<j such that both regions shaded In green contain base pairs, and the other shaded region is empty

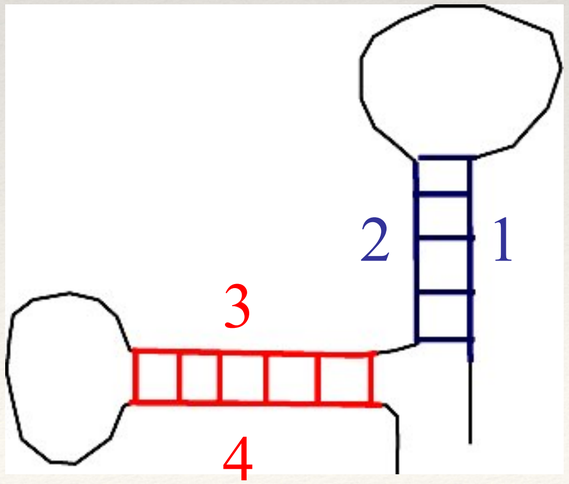


Only three ways to pair four segments…

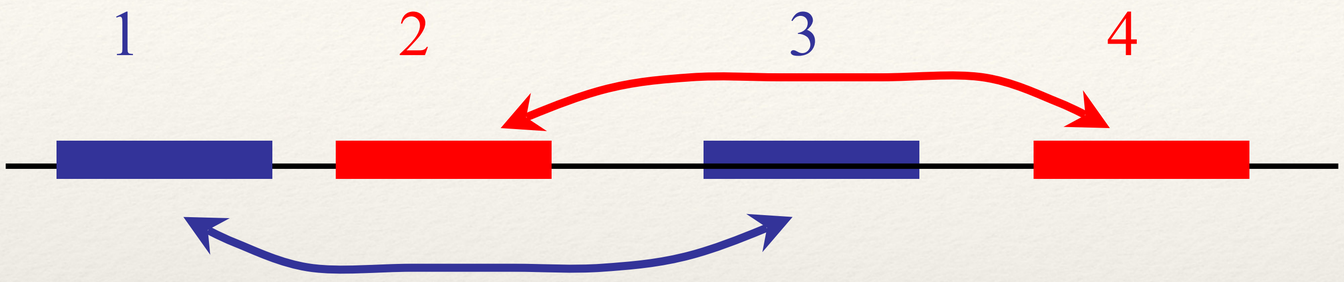




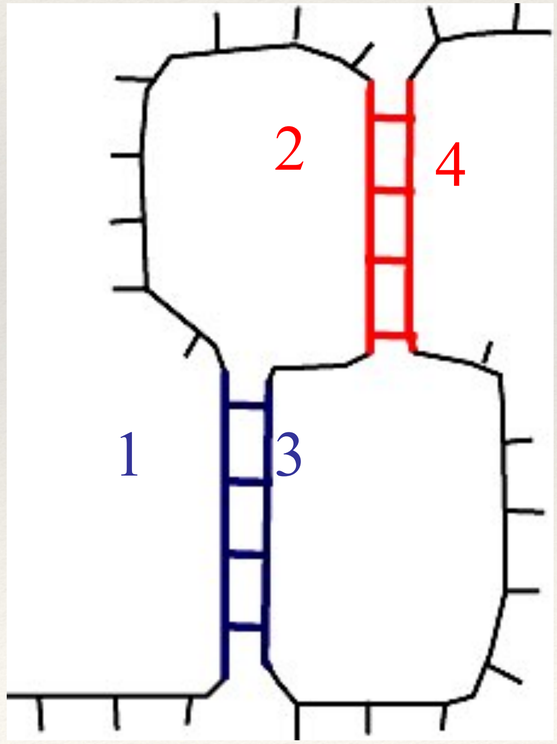




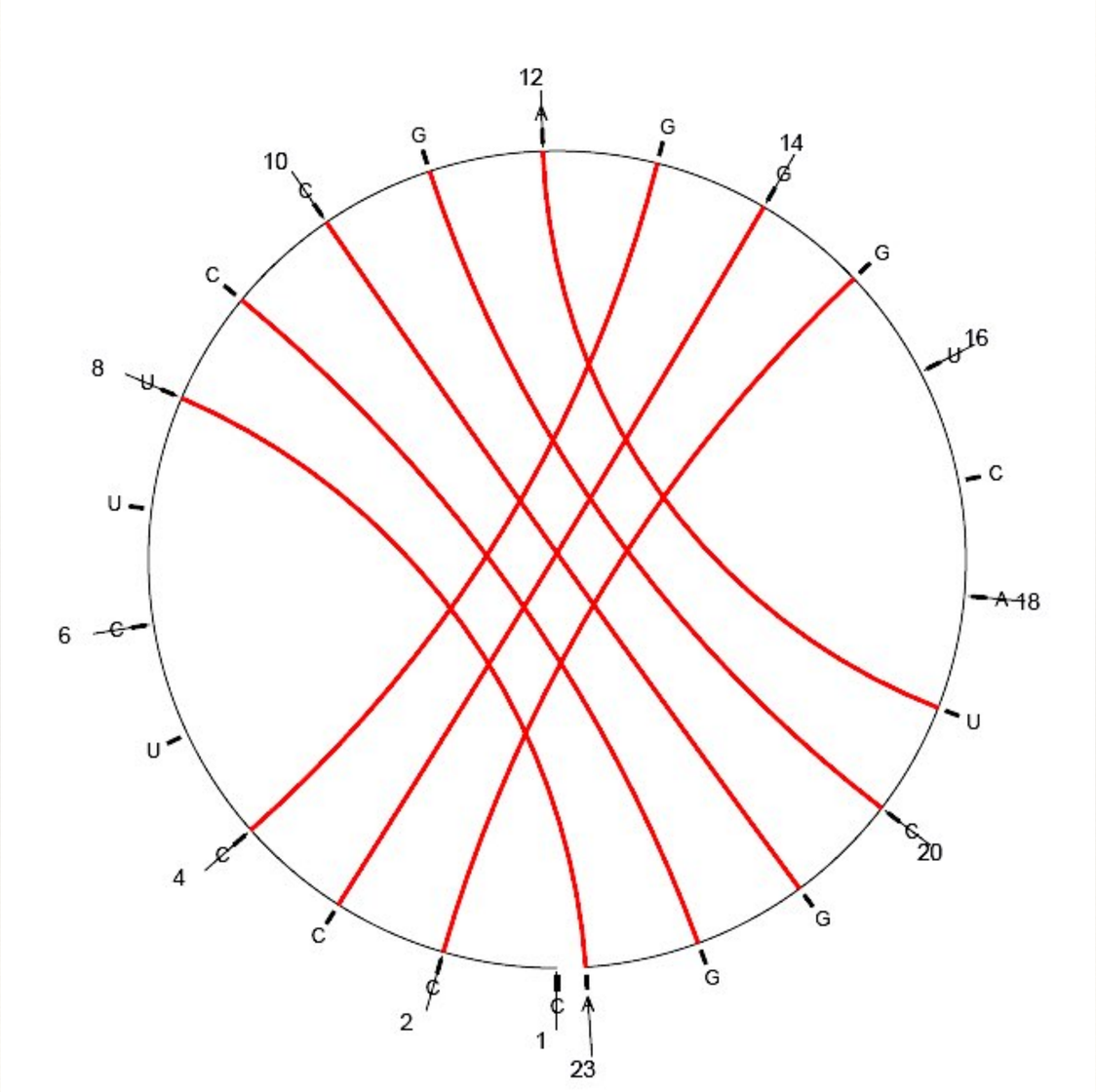
Only three ways to pair four segments…



Pseudo-knot : usually not considered a secondary structure…as it is difficult to predict !!



Circular representation of a pseudo-knot



RNA secondary structure definition

An RNA sequence is represented as:

$R = r_1, r_2, r_3, \ldots, r_n$ (r_i is the i-th nucleotide).

Each r_i belongs to the set {A, C, G, U}.

A secondary structure on R is a set S of ordered pairs, written as i•j, $1 \leq i < j \leq n$, satisfying:

1. $j - i > 3$ (exclude "close" base pairs)

2. if i •j and k •l are 2 base pairs, with $i \leq k$, then either:
   (a) $i = k$ and $j = l$ (same base pair)
   (b) $i < j < k < l$ (i •j precedes k •l)
   (c) $i < k < l < j$ (i •j includes k •l)

RNA Secondary Structure Prediction

Two primary methods for RNA secondary structure prediction:

-Co-variation analysis (comparative sequence analysis)

Takes into account conserved patterns of basepairs during evolution (2 or more sequences)

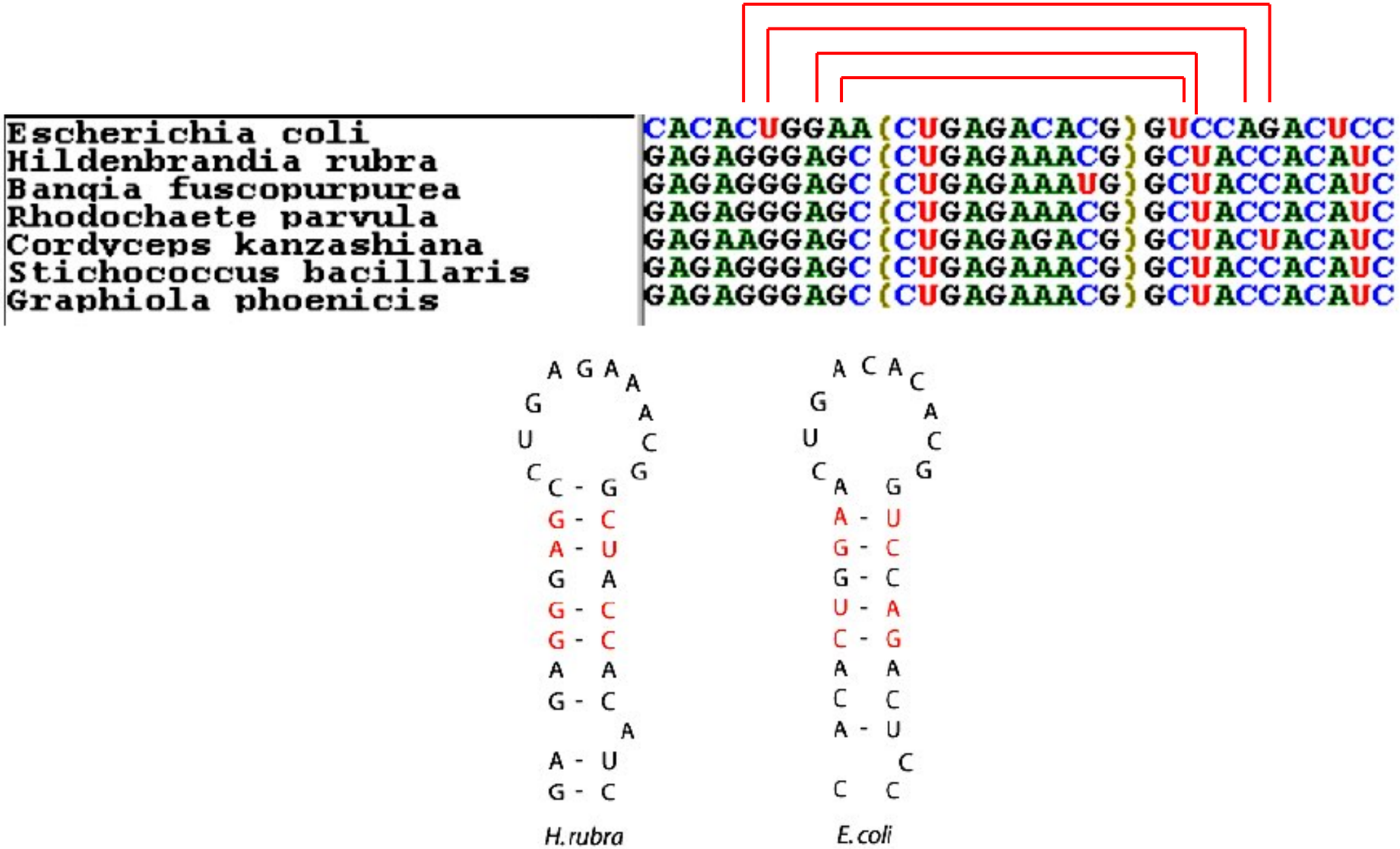
-Minimum free-energy method

Determines structure of complementary regions that are energetically stable

Comparative Sequence Analysis

- Molecules with similar functions and different nucleotide sequences will form similar structures
- Correctly identifies high percentage of secondary structure pairings and a smaller number of tertiary interactions
- Primarily a manual method

Co-variation



Quantitative Measure of Co-variation

Mutual Information Content:

$$H(i,j) = \sum_{N_1,N_2 \in \{A,C,G,U\}} f_{i,j}(N_1,N_2) \log_2 \frac{f_{i,j}(N_1,N_2)}{f_i(N_1) f_j(N_2)}$$

f_{ij}(N1,N2) : joint frequency of the 2 nucleotides, N_1 from the i-th column, and N_2 from the j-th column

f_i(N) : frequency in the i-th column of the nucleic acid N

How well does it work ?

Table 1

Summary of the evolution of the Noller-Woese-Gutell 16S and 23S rRNA structure models from the first to the most recent covariation-based structure models (adapted from Table 3a,b in [23]).

Model	16S rRNA		23S rRNA	
Date	1980	1999	1981	1999
1. Approximate number of complete sequences	2	7000	2	1050
2. Percentage of 1999 sequences*	0.03	100	0.2	100
3. Number of bp proposed correctly*	284	478	676	870
4. Number of bp proposed incorrectly*	69	0	102	0
5. Total bp in model (3 + 4)	353	478	778	870
6. Percentage of bp in model present in the current model (3 / X)*[1]	59.4	100	77.7	100
7. Accuracy of proposed bp (3 / 5)	80.5	100	86.9	100
8. Number of bp in current model missing from this model (X – 3)*[1]	194	0	194	0
9. Number of tertiary bp proposed correctly*	4	40	4	65
10. Percentage of tertiary bp proposed correctly*	10.0	100	6.2	100
11. Number of base triples proposed correctly*	0	6	0	7
12. Percentage of base triples proposed correctly*	0	100	0	100

*Comparisons are made against the current (1999) models. [1]X = 478 for 16S rRNA; X= 870 for 23S rRNA. bp, base pairs.

Gutell, Lee, Cannone, COSB, 2002, 12:301

Computing RNA secondary structure

- ***Working hypothesis:***

The native secondary structure of a RNA molecule is the one with the minimum free energy

- ***Restrictions:***
  - ***No knots***
  - ***No close base pairs***
  - ***Base pairs: A-U, C-G and G-U***

Computing RNA secondary structure

- ***Tinoco-Uhlenbeck postulate:***
  - ***Assumption:*** ***the free energy of each base pair is independent of all the other pairs and the loop structures***
  - ***Consequence:*** ***the total free energy of an RNA is the sum of all of the base pair free energies***

Independent Base Pairs Approach

- Use solution for smaller strings to find solutions for larger strings
- This is precisely the basic principle behind dynamic programming algorithms!

RNA folding: Dynamic Programming

Notation:

- $e(r_i,r_j)$: free energy of a base pair joining r_i and r_j

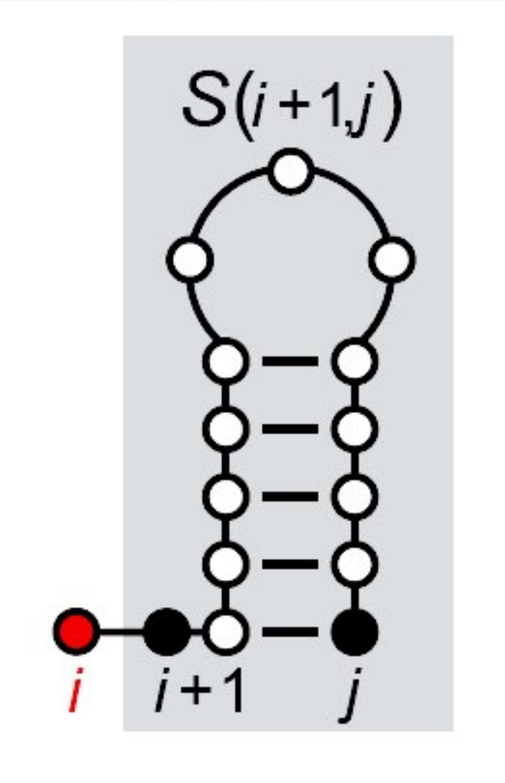
- B_{ij} : secondary structure of the RNA strand from base r_i to base r_j. Its energy is $E(B_{ij})$

- $S(i,j)$: optimal free energy associated with segment $r_i \ldots r_j$

$$S(i,j) = \max E(B_{ij})$$

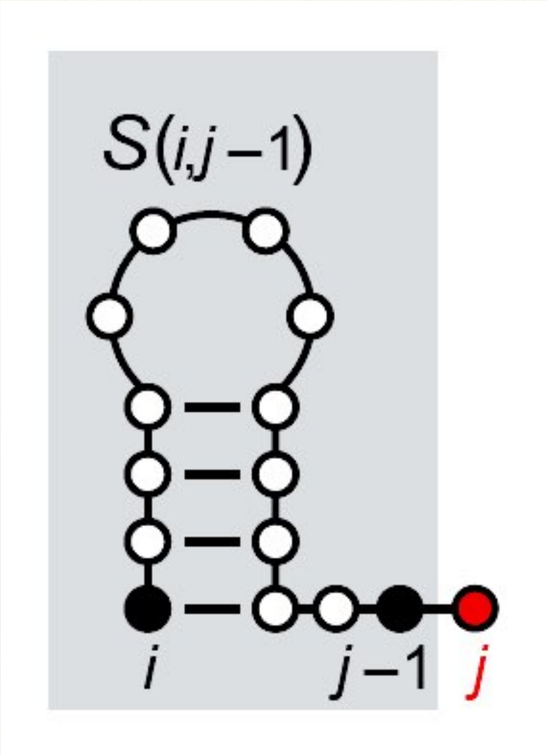
RNA folding: Dynamic Programming

There are only four possible ways that a secondary structure of nested base pair can be constructed on a RNA strand from position i to j:



1. i is unpaired, added on to a structure for i+1…j

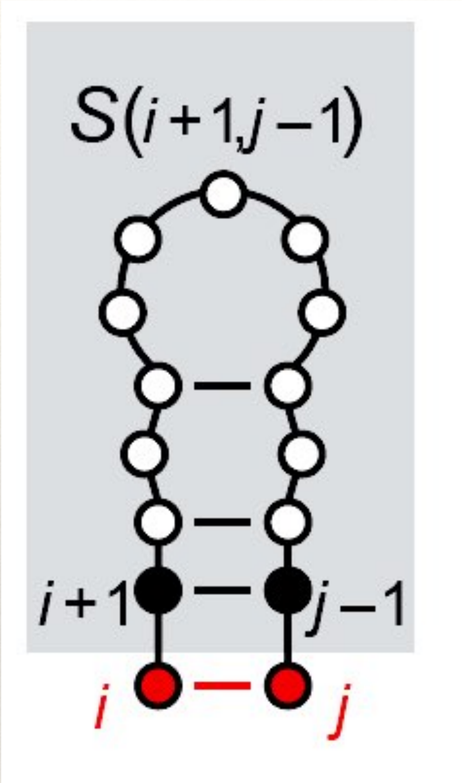
$S(i,j) = S(i+1,j)$



2. j is unpaired, added on to a structure for i…j-1

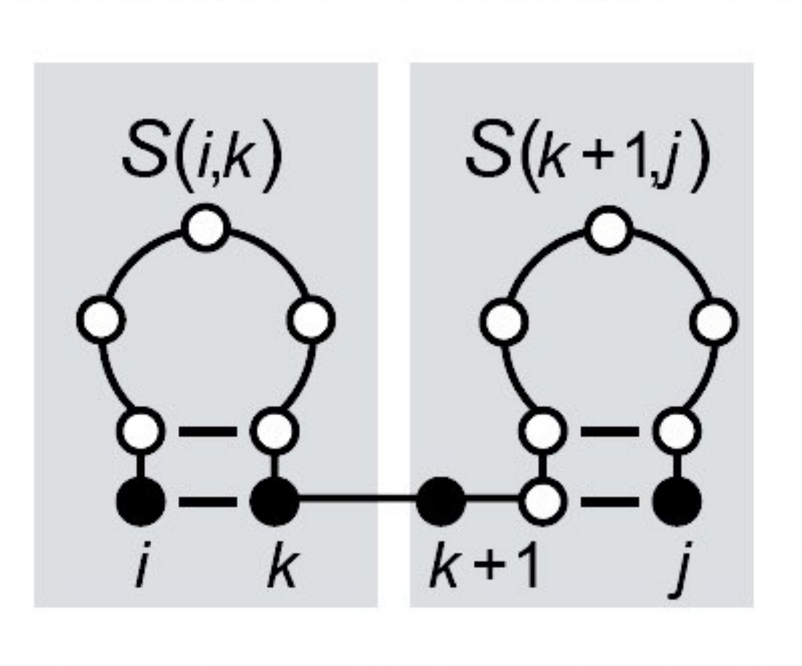
$S(i,j) = S(i,j-1)$

RNA folding: Dynamic Programming



3. i j paired, added on to a structure for i+1…j-1

$S(i,j) = S(i+1,j-1)+e(r_i,r_j)$



4. i j paired, but not to each other; the structure for i…j adds together structures for 2 sub regions, i…k and k+1…j

$S(i,j) = \max \{S(i,k)+S(k+1,j)\}$

RNA folding: Dynamic Programming

Since there are only four cases, the optimal score S(i,j) is just the maximum of the four possibilities:

$$S(i,j) = \max \begin{cases} S(i+1,j) & r_i \text{ unpaired} \\ S(i,j-1) & r_j \text{ unpaired} \\ S(i+1,j-1) + e(r_i, r_j) & i, j \text{ base pair} \\ \max\limits_{i<k<j} \{S(i,k) + S(k+1,j)\} & i, j \text{ paired, but not to each other} \end{cases}$$

To compute this efficiently, we need to make sure that the scores for the smaller sub-regions have already been calculated

Dynamic Programming !!

RNA folding: Dynamic Programming

Notes:

$S(i,j) = 0$ if $j-i < 4$: do not allow "close" base pairs

Reasonable values of e are -3, -2, and -1 kcal/mole for GC, AU and GU, respectively. In the DP procedure, we use 3, 2, 1 (or replace max with min)

Build upper triangular part of DP matrix:

- *start with diagonal – all 0*
- *works outward on larger and larger regions*
- *ends with $S(1,n)$*

Traceback starts with $S(1,n)$, and finds optimal path that lead there.

Initialisation:

No close basepairs

j →

i ↓

	A	U	A	C	C	C	U	G	U	G	G	U	A	U
A	0	0	0	0										
U		0	0	0	0									
A			0	0	0	0								
C				0	0	0	0							
C					0	0	0	0						
C						0	0	0	0					
U							0	0	0	0				
G								0	0	0	0			
U									0	0	0	0		
G										0	0	0	0	
G											0	0	0	0
U												0	0	0
A													0	0
U														0

Propagation:

C5….U9 :

C5 unpaired:
S(6,9) = 0

U9 unpaired:
S(5,8)=0

C5-U9 paired
S(6,8) +e(C,U)=0

C5 paired, U9 paired:
S(5,6)+S(7,9)=0
S(5,7)+S(8,9)=0

j →

	A	U	A	C	C	C	U	G	U	G	G	U	A	U
A	0	0	0	0	0									
U		0	0	0	0	0								
A			0	0	0	0	2							
C				0	0	0	0	3						
C					0	0	0	0	0					
C						0	0	0	**0**	3				
U							0	0	0	0	1			
G								0	0	0	0	1		
U									0	0	0	0	2	
G										0	0	0	0	1
G											0	0	0	0
U												0	0	0
A													0	0
U														0

Propagation:

C5….G11 :

C5 unpaired:
S(6,11) = 3

G11 unpaired:
S(5,10)=3

C5-G11 paired
S(6,10)+e(C,G)=6

C5 paired, G11 paired
S(5,6)+S(7,11)=1
S(5,7)+S(8,11)=0
S(5,8)+S(9,11)=0
S(5,9)+S(10,11)=0

j →

	A	U	A	C	C	C	U	G	U	G	G	U	A	U
A	0	0	0	0	0	0	2							
U		0	0	0	0	0	2	3						
A			0	0	0	0	2	3	5					
C				0	0	0	0	3	3	3				
C					0	0	0	0	0	3	6			
C						0	0	0	0	3	3			
U							0	0	0	0	1	1		
G								0	0	0	0	1	2	
U									0	0	0	0	2	2
G										0	0	0	0	1
G											0	0	0	0
U												0	0	0
A													0	0
U														0

Propagation:

j →

i ↓

	A	U	A	C	C	C	U	G	U	G	G	U	A	U
A	0	0	0	0	0	0	2	3	5	6	6	8	10	12
U		0	0	0	0	0	2	3	5	6	6	8	10	10
A			0	0	0	0	2	3	5	5	6	8	8	8
C				0	0	0	0	3	3	3	6	6	6	6
C					0	0	0	0	0	3	6	6	6	6
C						0	0	0	0	3	3	3	3	3
U							0	0	0	0	1	1	3	3
G								0	0	0	0	1	2	2
U									0	0	0	0	2	2
G										0	0	0	0	1
G											0	0	0	0
U												0	0	0
A													0	0
U														0

Traceback:

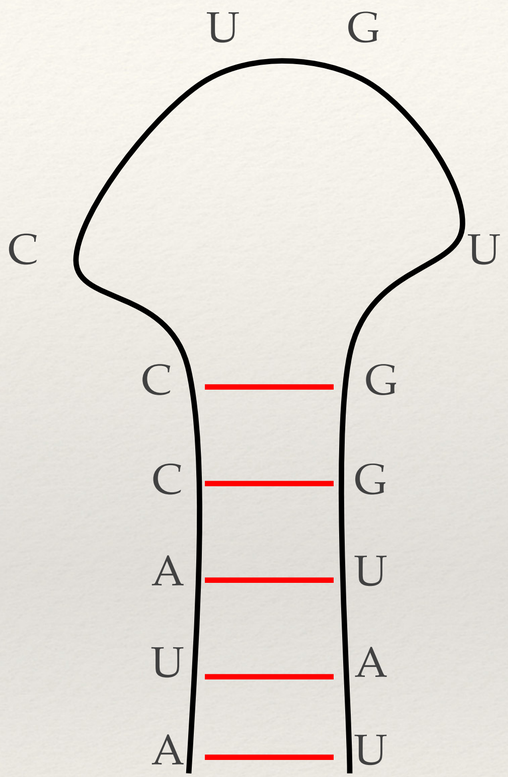
j →

i ↓

		A	U	A	C	C	C	U	G	U	G	G	U	A	U
	A	0	0	0	0	0	0	2	3	5	6	6	8	10	12
	U		0	0	0	0	0	2	3	5	6	6	8	10	10
	A			0	0	0	0	2	3	5	5	6	8	8	8
	C				0	0	0	0	3	3	3	6	6	6	6
	C					0	0	0	0	0	3	6	6	6	6
	C						0	0	0	0	3	3	3	3	3
	U							0	0	0	0	1	1	3	3
	G								0	0	0	0	1	2	2
	U									0	0	0	0	2	2
	G										0	0	0	0	1
	G											0	0	0	0
	U												0	0	0
	A													0	0
	U														0

FINAL PREDICTION

AUACCCUGUGGUAU



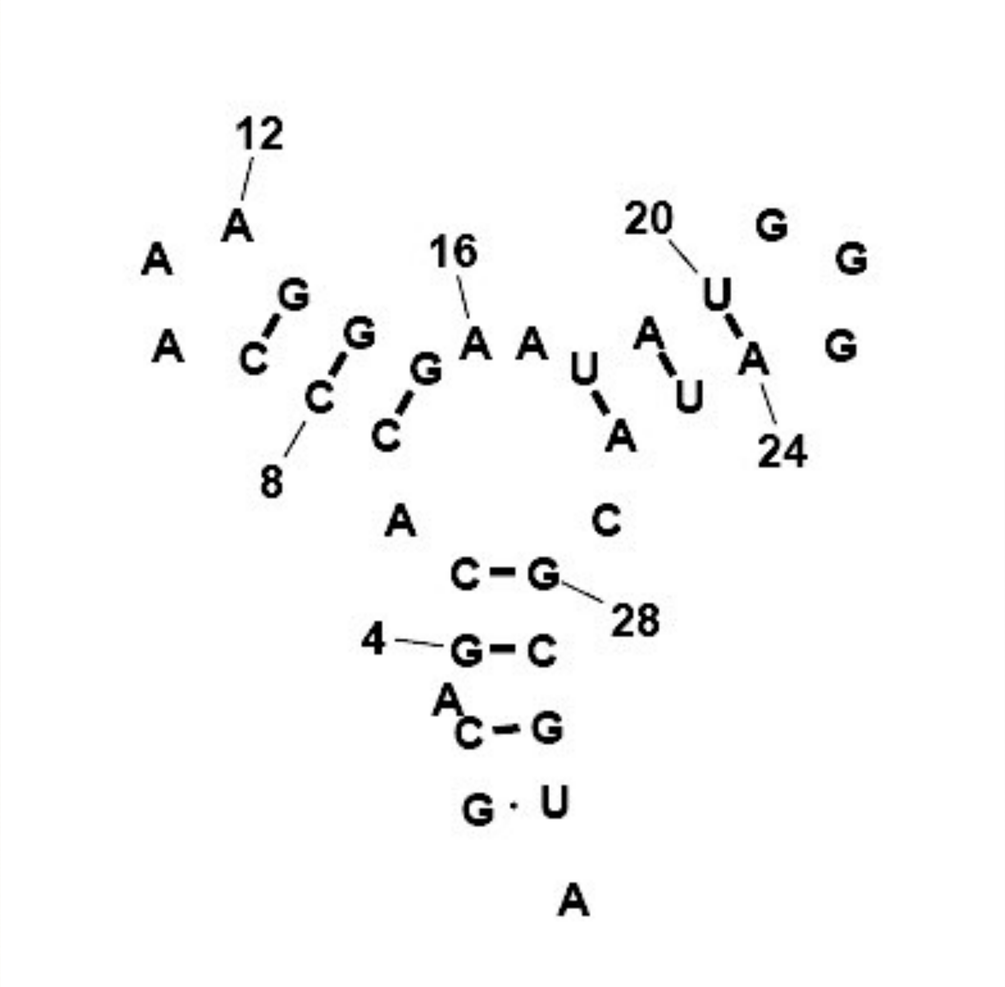
Total free energy: -12 kcal/mol

Try it yourself!!

Sequence:

GCAGCACCCAAAGGGAAUAUGGGAUACGCGUA

One possible solution:



Some notes

- Computational complexity: N^3

- Does not work with pseudo-knot (would invalidate DP algorithm)

- Methods that include pseudo knots:
  Rivas and Eddy, JMB 285, 2053 (1999)
  Orland and Zee, Nucl. Phys. B 620, 456 (2002)
  These methods are at least N^6

Some notes (2)

- The scoring scheme is too simplistic!
- Needs to take into account the cost of loops (both internal and in hairpins), of bulges, ….

Example: 2x2 interior loops in RNA closed by a GC and a CG base pair:

```
                    5' ------> 3'
                     G \/ \_/ C
                     C /\  |  G
                    3' <------ 5'
```

	Y:	A	A	A	C	C	C	G	G	U	U	U
		A	C	G	A	C	U	A	U	C	G	U
	AA	1.5	1.2	-0.5	1.2	1.8	0.80	0.10	-0.7	1.9	-0.3	1.5
	AC	1.2	0.9	-0.8	0.9	0.9	0.00	-0.20	-2.0	1.0	-1.6	0.2
	AG	0.1	-0.1	-1.9	-0.2	0.9	-0.10	-1.30	-1.3	0.9	-0.9	0.9
	CA	1.2	1.0	-0.8	0.9	1.0	0.00	-0.10	-1.9	1.0	-1.5	0.2
	CC	1.8	1.0	0.2	0.9	1.0	0.00	0.90	-0.9	1.0	-0.5	0.2
X	CU	1.9	1.0	0.3	1.0	1.0	0.00	0.90	-0.9	1.1	-0.5	0.3
	GA	-0.5	-0.8	-2.6	-0.8	0.2	-0.80	-1.90	-1.9	0.3	-1.5	0.3
	GG	1.1	0.9	-0.9	0.8	1.5	0.50	-0.20	-1.0	1.5	-0.6	1.1
	GU	-0.3	-1.5	-1.5	-1.6	-0.5	-1.50	-0.90	-4.5	-0.5	-4.1	-0.5
	UC	0.8	0.0	-0.8	0.0	0.0	-1.00	-0.10	-1.9	0.0	-1.5	-0.7
	UG	-0.7	-1.9	-1.9	-2.0	-0.9	-1.90	-1.30	-4.9	-0.9	-4.5	-0.9
	UU	1.5	0.2	0.3	0.2	0.2	-0.70	0.90	-0.9	0.3	-0.5	-0.5

Destabilizing energies of loops

Size	Internal	Bulge	Hairpin
1	NA	3.8	NA
2	NA	2.8	NA
3	NA	3.2	5.6
4	1.7	3.6	5.5
5	1.8	4.0	5.6
6	2.0	4.4	5.3
7	2.2	4.6	5.8
8	2.3	4.7	5.4
30	3.7	6.1	7.7

Prediction Programs

- MFOLD (Zuker) (web server)
- http://www.unafold.org/mfold/applications/rna-folding-form.php

- Genebee (both comparative + energy model) (web server)
  http://www.genebee.msu.su/services/rna2_reduced.html

- Vienna RNA package
  http://www.tbi.univie.ac.at/~ivo/RNA/

- Mc-Sym (Computer Science approach)
- https://major.iric.ca/MC-Sym/

- RNAFold
- http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi

How well do they perform?

- Current RNA folding programs get about 60% of base pairs correct, on average: useful, but not yet good.
- The problem is the scoring system: thermodynamic model is accurate within 5-10%, and many alternative structures are within 10%.
- Possible solution: combination of thermodynamic score with comparative sequence information

Useful web sites on RNA

- *Comparative RNA web site*
  https://crw-site.chemistry.gatech.edu

- *RNA structure database*
  http://ndbserver.rutgers.edu/ (nucleic acid database)

- *RNA structure classification*
  http://scor.berkeley.edu/

- *RNA visualisation*
  http://ndbserver.rutgers.edu/ndbmodule/services/download/rnaview.html
  http://x3dna.org