

RNA FOLDING

WITH ENERGY MINIMIZATION

Presented by
Aasminpreet Singh Kainth

Shape and Structure of Molecules

The primary molecules of biological interest (DNA, RNA, and proteins) all have fundamentally linear structures.

However, the biological *function* of these molecules depends upon how they interact with other molecules.

These interactions depend heavily upon the shape of the molecules involved.

Experimental methods for determining structures, such as x-ray crystallography, are slow, expensive, tricky, and work only in certain environmental conditions.

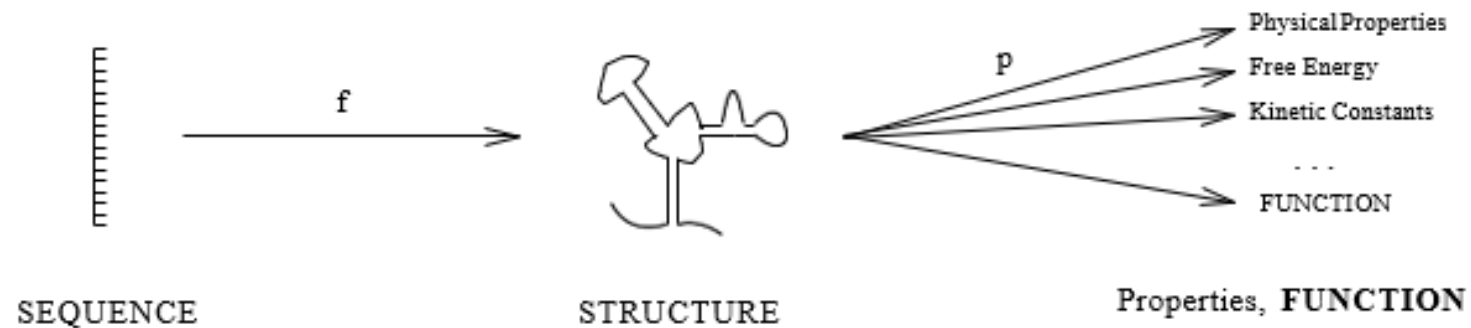
All this combines to make computational prediction of structures an important and messy problem.

Why look at Structure?

Basic paradigm of structural biology:

Sequence → **Structure** → **Function**

Structure based methods can succeed where sequence analysis fails



Factors determining Molecular shape/structure

Molecular shape/structure is determined by many factors:

- Molecular bonds

- Electrostatic forces

- The size/shape of molecular subunits

- Hydrophobicity of the associated bases

- The current environmental conditions

Levels of Structure

Primary structure

Secondary structure

Ternary structure

Quandary structure

Interest in these different levels?

The interest in these different levels of abstraction is

- (1) that sometimes it is much easier to get accurate predictions at lower levels
- (2) accurate lower-level knowledge may be more useful than less-precise higher-level knowledge.

DNA folding

DNA molecules usually come in the form of double-stranded molecules, whose secondary structure is the famous 'double helix' discovered by Watson and Crick.

The tertiary structure of DNA is the shape of the chromosomes it folds into.

RNA Folding

RNA molecules, are usually **single strands** which **fold back** onto themselves into predefined 3D shapes or structures.

The **folding problem seeks the structure or shape** of a given sequence.

The **shape** of certain RNAs plays a major role in **determining its interaction** with other molecules, for example tRNAs.

Since RNA is single-stranded, its component bases tend to bond with other bases.

RNA (Ribonucleic acid)

RNA molecules are an **important component of biological substance**.

They are the **carriers of genetic information** between DNA molecules and proteins.

They also play an **important role in many biological processes**, such as catalysis, protein synthesis, immunity, development and many other important biological processes.

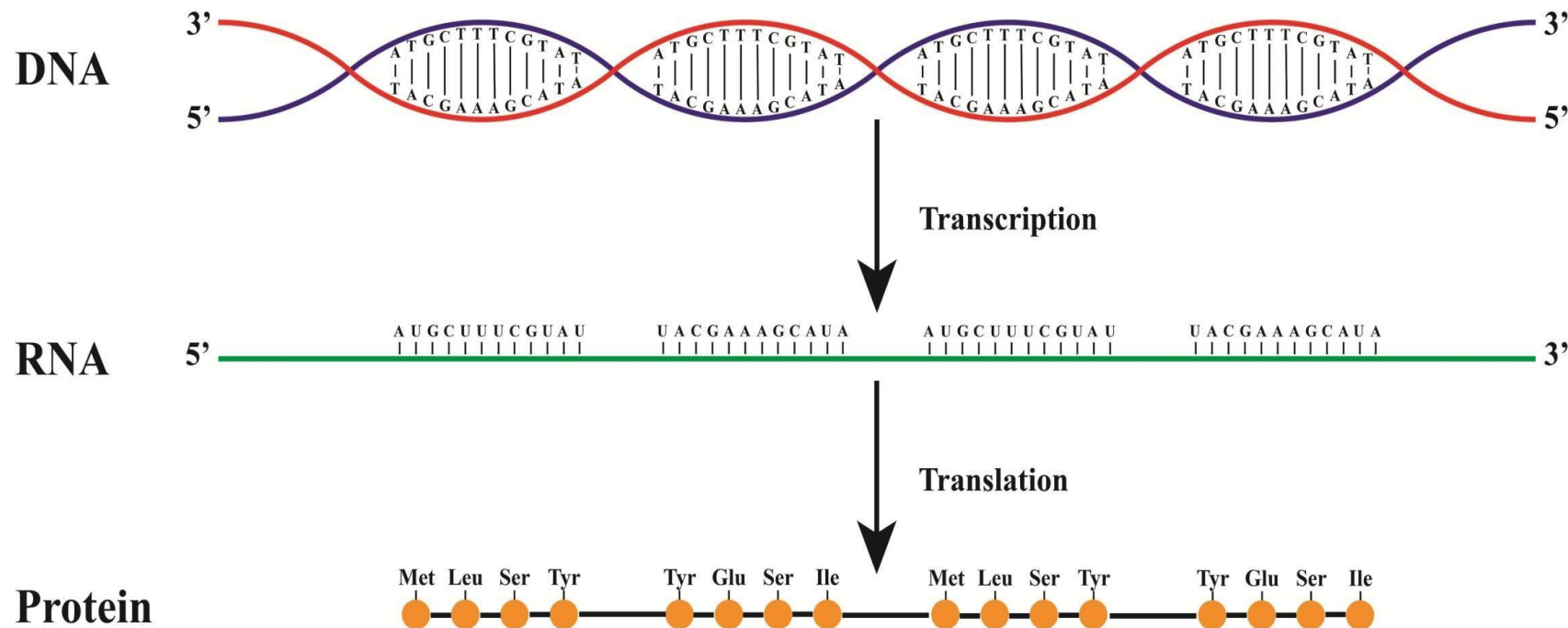
From DNA to Protein

Protein synthesis occurs with the help of three RNA types.

Messenger RNA (**mRNA**)

Transfer RNA (**tRNA**)

Ribosomal RNAs (**rRNA**)



What other does RNA do?

It is widely **believed** that **RNA molecules are the closest thing** to the molecules from which **life originally evolved**.

RNA molecules can **perform** the function of **coding for proteins** (i.e. information storage) usually associated with DNA.

RNA molecules can have **enzymatic functions** usually associated with proteins.

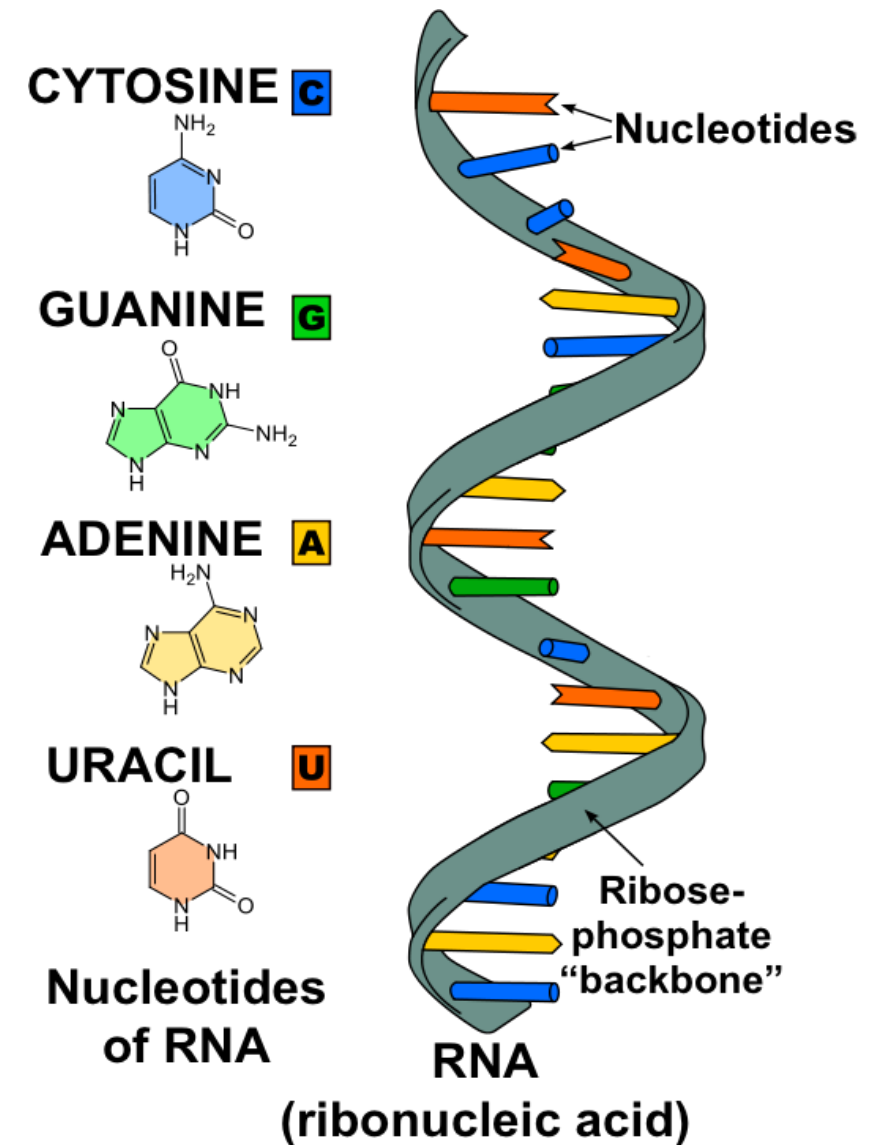
Structure of RNA

RNA typically is a single-stranded biopolymer

RNA consists of Ribose Nucleotides which are **nitrogenous bases** appended to a ribose sugar.

These Ribose Nucleotides are attached by phosphodiester bonds, forming strands of varying lengths.

The nitrogenous bases in RNA are adenine, guanine, cytosine and uracil, which replaces thymine in DNA.



Features of RNA

RNA: polymer composed of a combination of four nucleotides

adenine (A); cytosine (C); guanine (G); uracil (U)

G-C and A-U form complementary hydrogen bonded base pairs (canonical Watson-Crick)

G-C base pairs being more stable (3 hydrogen bonds) A-U base pairs less stable (2 bonds)

non-canonical pairs can occur in RNA -- most common is G-U

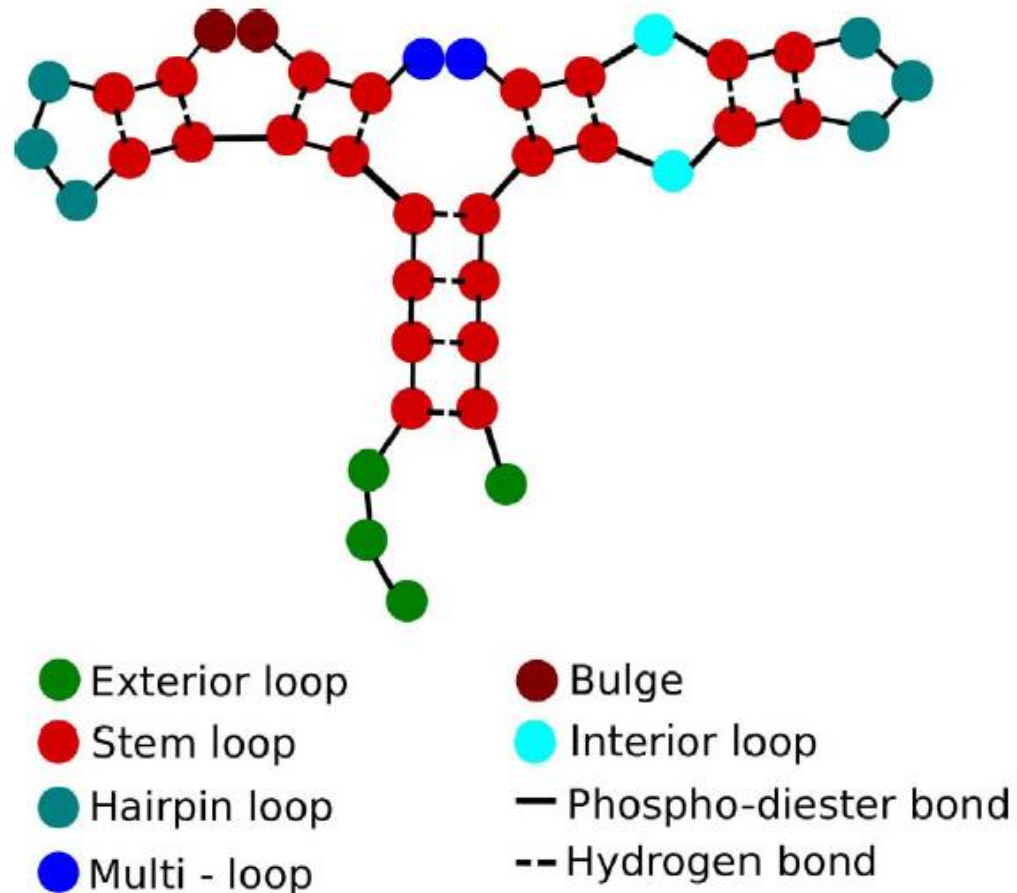
Understanding the RNA Secondary Structure

A **secondary structure** is a list of base pairs (i, j) on a sequence x , with (rules for normal SS)

1. Any nucleotide can form at most one pair
2. No pseudo-knots: No pairs (i, j) and (k, l) with $i < k < j < l$
3. If (i, j) is a pair then $x_i x_j \in \{GC, CG, AU, UA, GU, UG\}$
4. If (i, j) is a base pair, then $j - i > 3$.

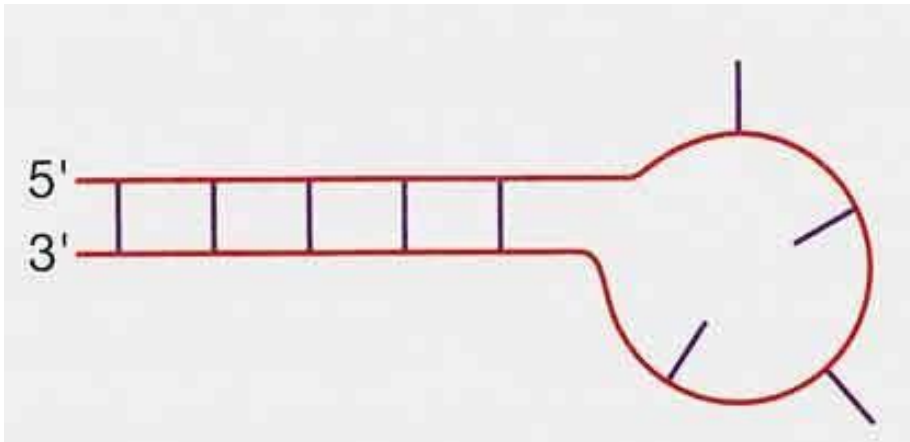
Understanding the RNA Secondary Structure

Decomposition of an RNA secondary structure into **structural elements**.



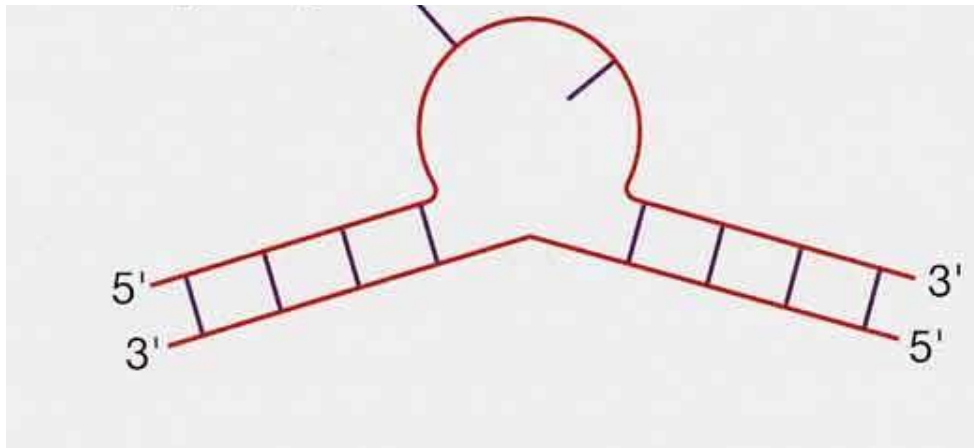
Loop Decomposition

Secondary structures can be uniquely decomposed into loops.

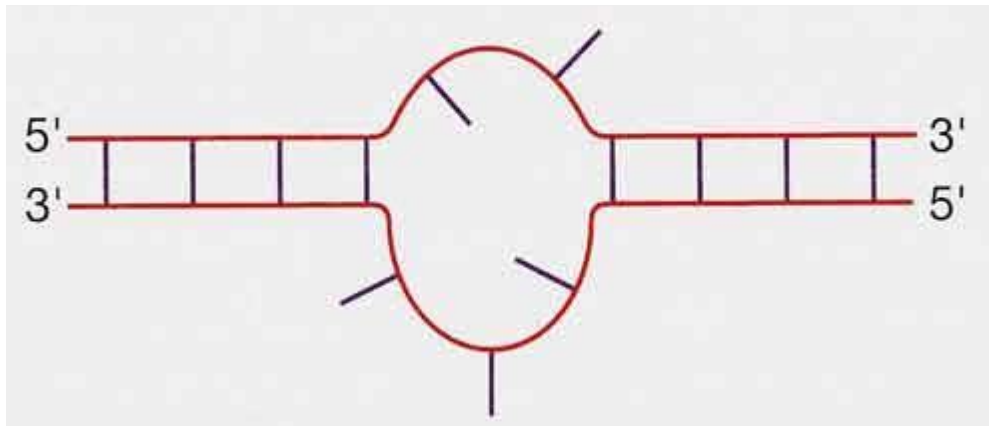


Stem Loop (Hairpins)

Loop Decomposition

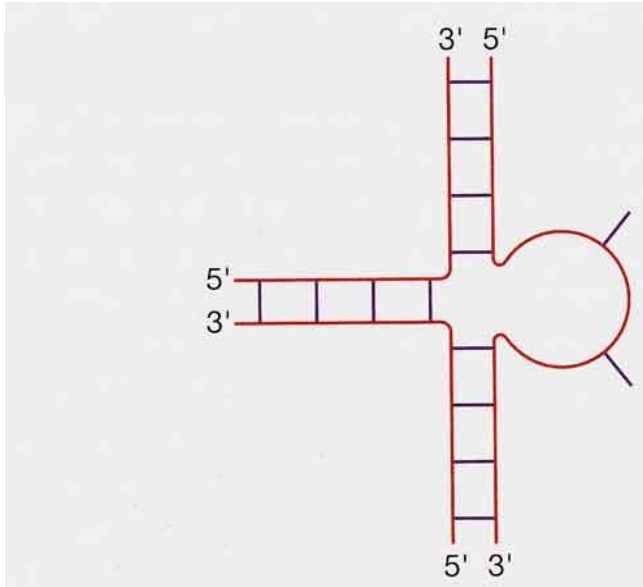


Bulge Loops

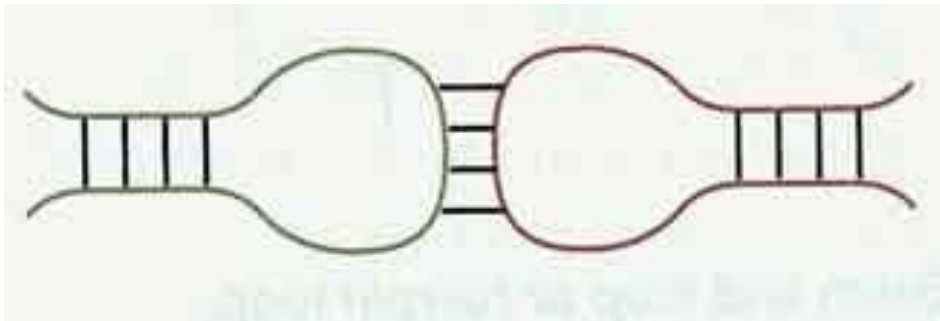


Interior Loops

Loop Decomposition

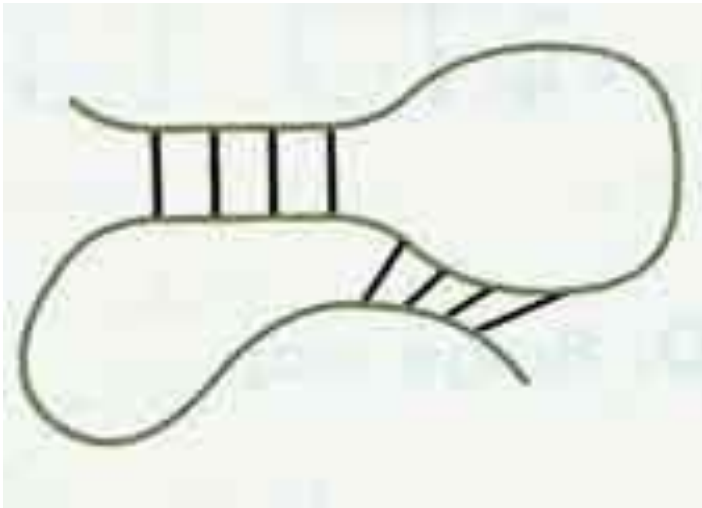


Junctions (Multiloops)

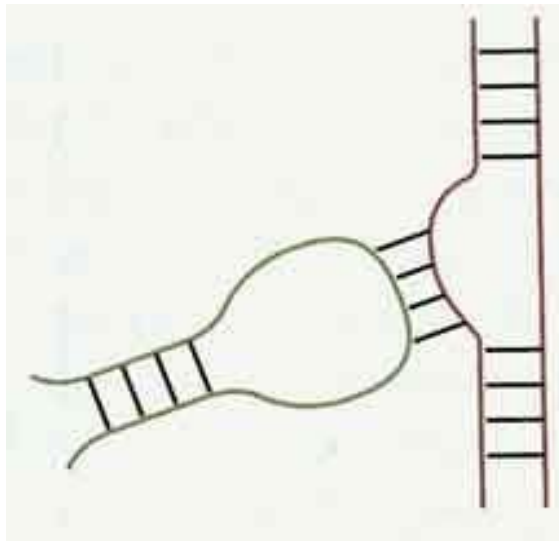


Kissing Hairpins

Loop Decomposition



Pseudoknots



Hairpin-Bulge Interactions

Pseudo knots

Excluding pseudo knots makes life easy

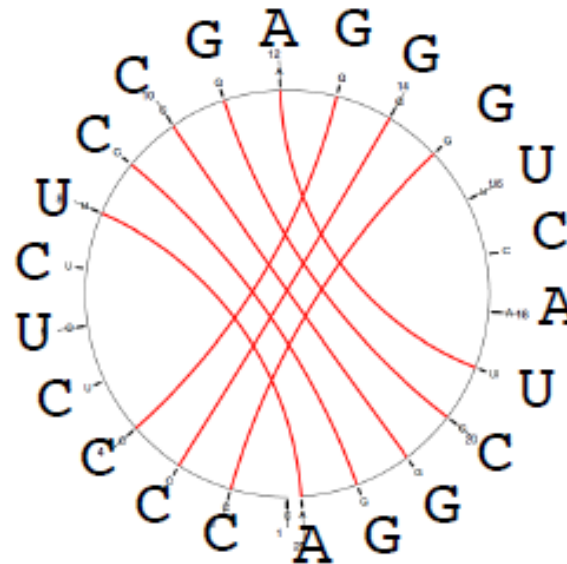
On the other hand:

Pseudo knots can have important function

```
          A-C
3' - A-G-G-C-U /   U
      U-C-C-G-A-G-G-G
          U       C-C-C - 5'
          C--U--C/
```

Circle plot=====>

<=== Example of a simple pseudoknot.

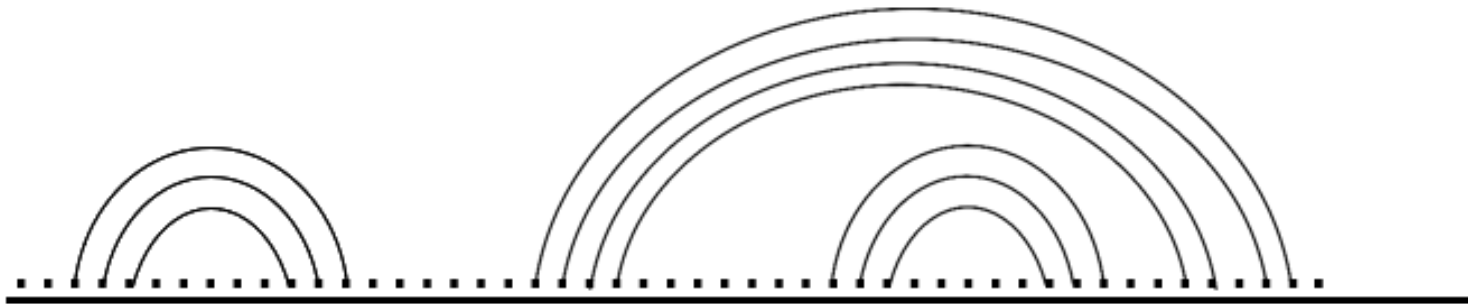


RNA Secondary Structure Notation

Parentheses notation

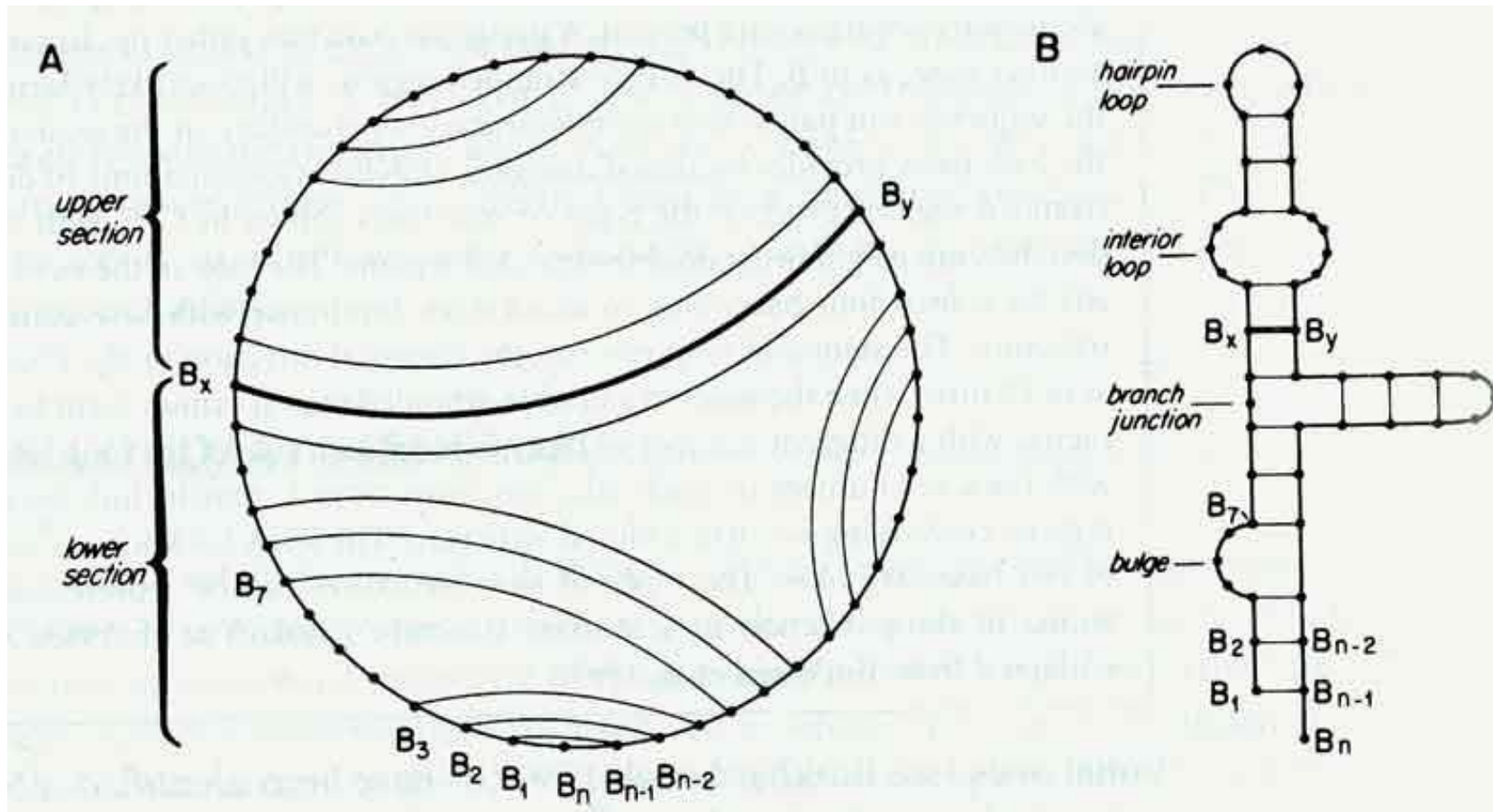
..(((.....))).....((((.....))).))...

Arc ('rainbow') notation



RNA Secondary Structure Notation

Circle Plot



RNA structure prediction methods

Dot Plot Analysis

Base-Pair Maximization

Free Energy Methods

Covariance Models

Algorithms for Secondary Structure Prediction

RNA structures can be predicted by Dynamic programming algorithms in many variants.

- Minimum free energy structure (Zuker & Stiegler '81)
- Optimal and certain suboptimal structures (Zuker '89)
- All structures within an energy range (Wuchty et al. '99)
- Partition function and base pair probabilities (McCaskill '90)
- Stochastic suboptimals (Ding & Lawrence '01)
- Maximum expected accuracy structures (Do et al '06)
- Consensus structure prediction from alignment (Knudsen & Hein '99, Hofacker et al. '02)
- Minimum free energy with pseudo-knots (Rivas & Eddy '99)
- Extended secondary structures with non-canonical pairs (Parisien & Major '08, Höner et al '11)

Algorithms for Secondary Structure Prediction

The dynamic programming approach to RNA secondary structure prediction relies on the fact that structures can be **recursively decomposed into smaller components** with independent energy contributions.

In each of the decomposition steps only a single loop (or stacking of two consecutive base pairs) needs to be evaluated.

Algorithms for Secondary Structure Prediction

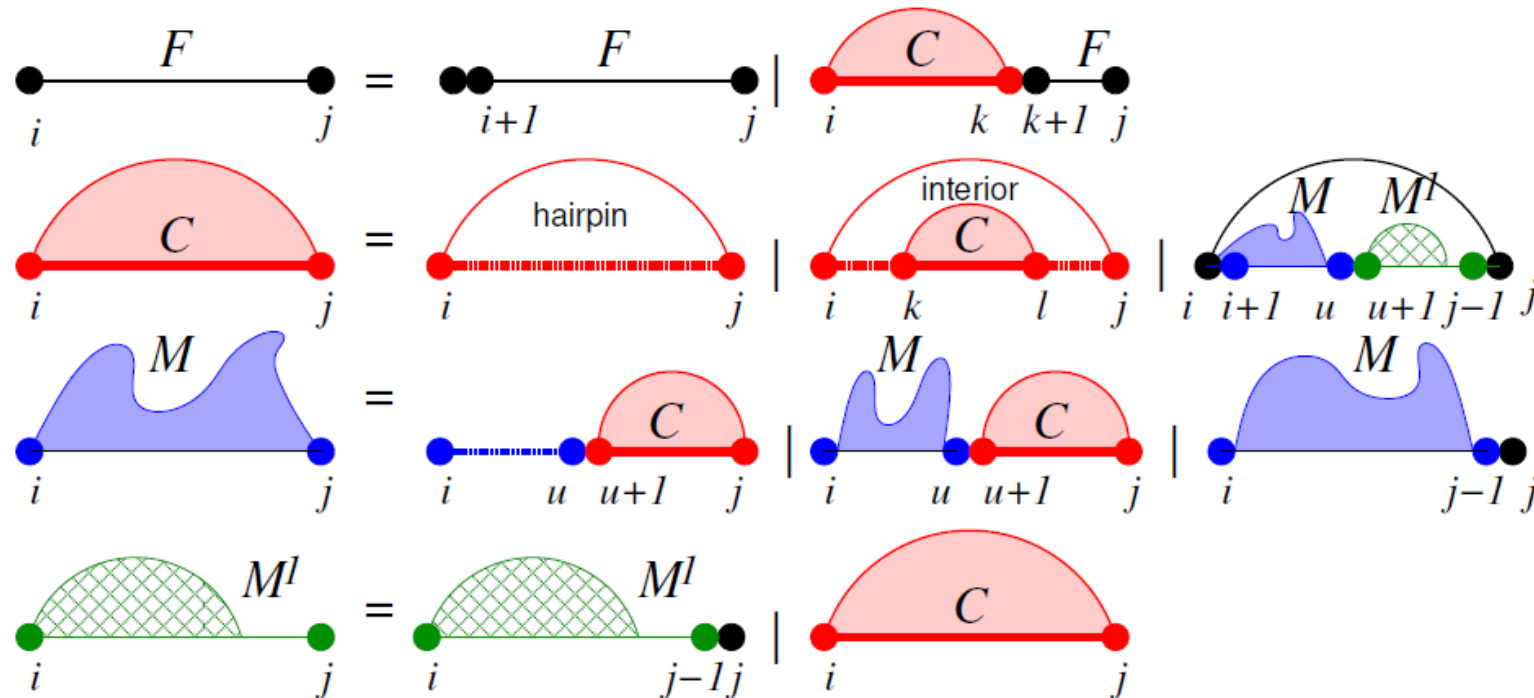


Fig. The classical recursions of the standard model of RNA folding (drawing from Lorenz et al. (2011)). The hieroglyphic symbols denote different types of RNA secondary structures: F is an arbitrary secondary structures, C a structure enclosed by a base pair, and M and M^l denote components of multibranch loops.

Base-Pair Maximization

Find structure with the most base pairs

Efficient dynamic programming approach introduced by Ruth Nussinov.

Nussinov Algorithm

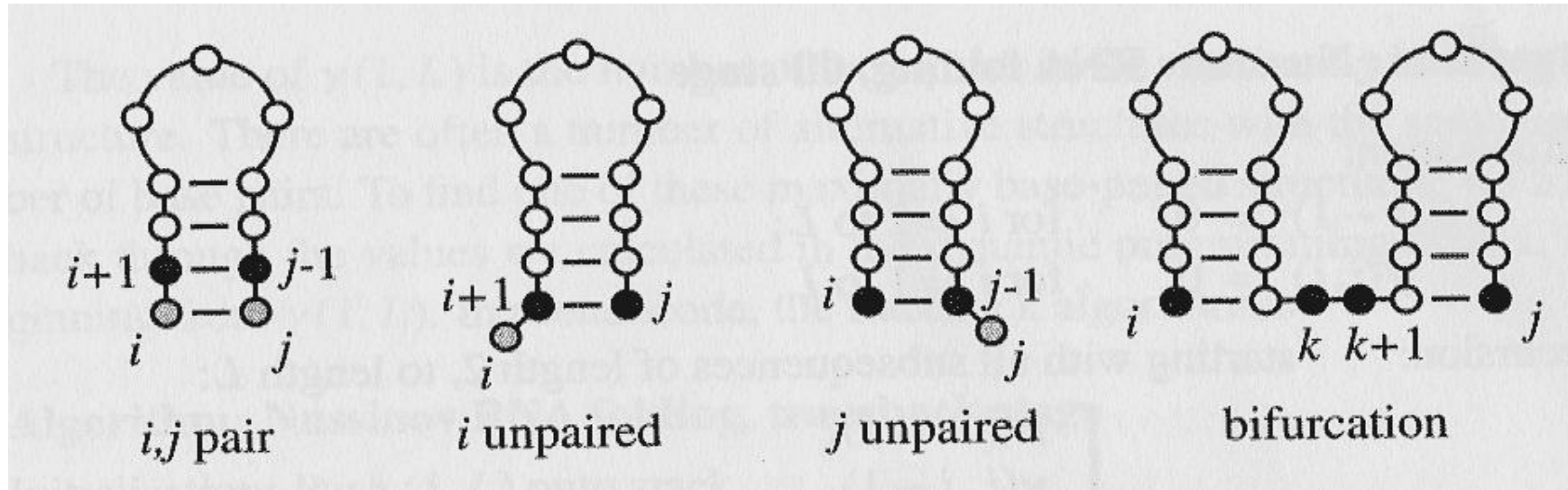
It computes the highest number of nucleotide coupling with 2 structure.

This algorithm was proposed in 1978.

Nussinov Algorithm

Nussinov Algorithm is a recursive algorithm which calculates the best structure for small subsequence, and works its way outward to larger and larger subsequence.

Four ways to get the best structure between position i and j from the structures of the smaller sub sequences.



Nussinov Folding Algorithm Matrix

Suppose we have a RNA sequence of GGCAAUGC.

Create a Matrix.

Label as i and j as shown here.

A diagram showing a grid of 10 rows and 10 columns. The top row is green. The first column is light blue. The cell at row 1, column 1 is labeled $i \ j$.

Nussinov Folding Algorithm Matrix

Fill RNA Sequence GGCAAUGC in the second row and column

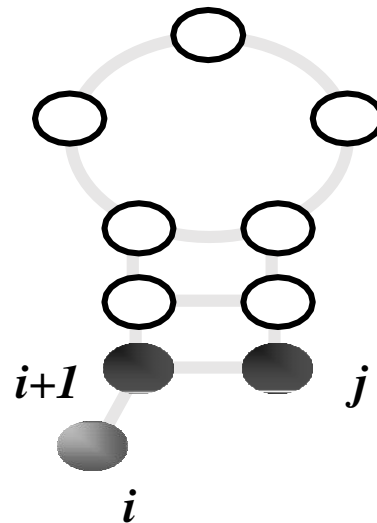
[illegible]

Nussinov Folding Algorithm

The score $S(i, j)$ is the maximum of the following 4 possibilities:

1. $S(i + 1, j) - i$ unpaired—

Take lower element.



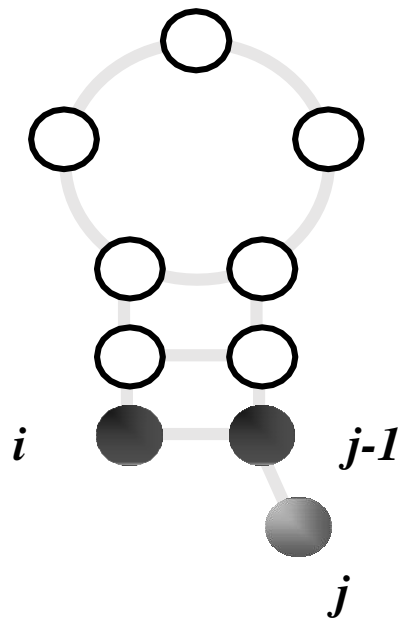
	1	2	3	4
1		A	G	C
2	A	0	i	
3	G		0	
4	C			0

Nussinov Folding Algorithm

The score $S(i, j)$ is the maximum of the following 4 possibilities.

2. $S(i, j - 1) - j$ unpaired–

Take left element.



	1	2	3	4
1		A	G	C
2	A	0	ⁱ ←	
3	G		0	
4	C			0


Nussinov Folding Algorithm

The score $S(i, j)$ is the maximum of the following 4 possibilities.

3. $S(i+1, j-1) + e(i, j-i)$ i, j pair—

$e(i, j)$ – Energy of pairing

Take the diagonally left Lower Down.

	1	2	3	4
1		A	G	C
2	A	0		
3	G		0	
4	C			0

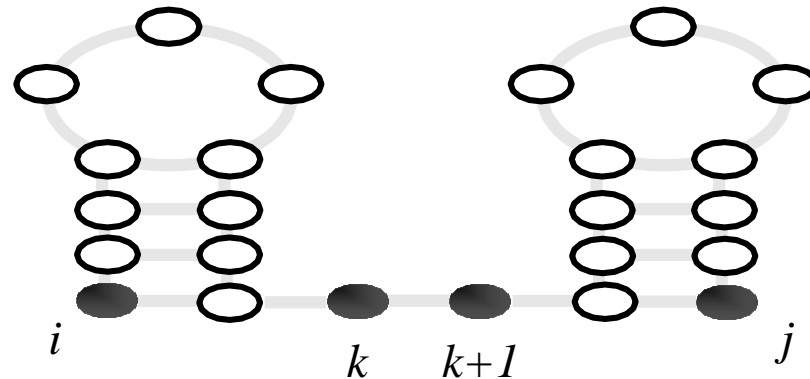
Nussinov Folding Algorithm

The score $S(i, j)$ is the maximum of the following 4 possibilities.

4. Bifurcation

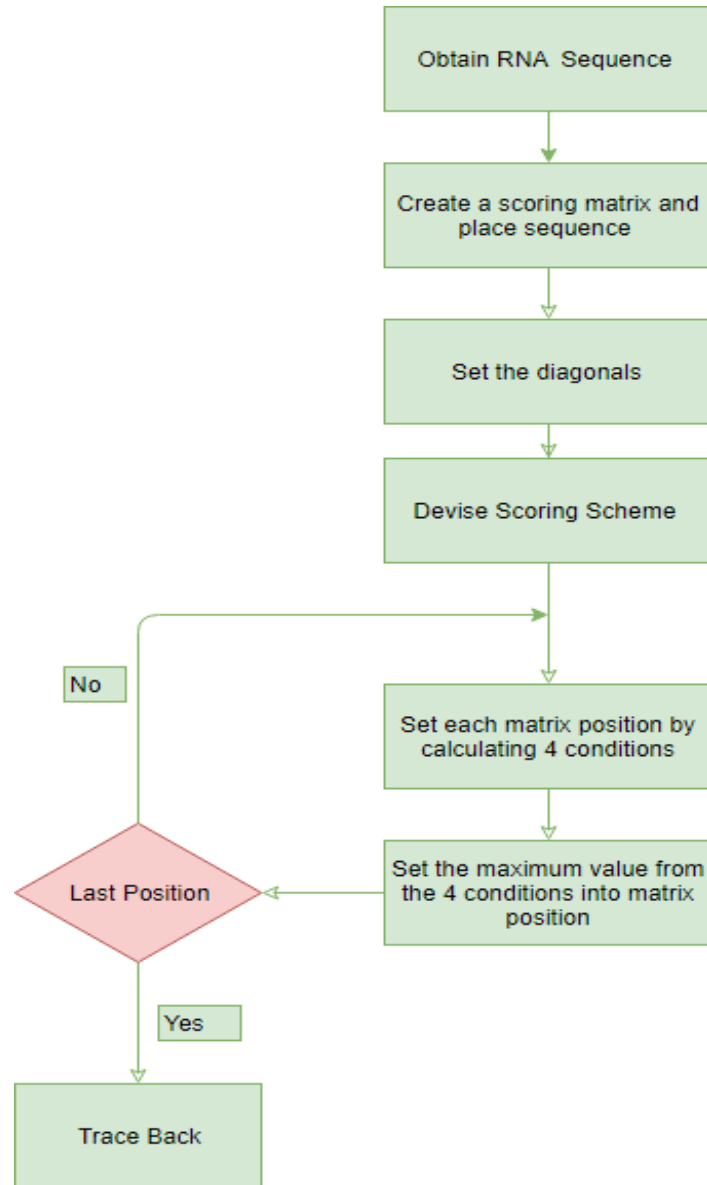
$$\text{Max } i < k < j \{S(i, k) + S(k+1, j)\} - \text{bifurcation} -$$

Take Left row, Bottom Column.



	1	2	3	4
1		A	G	C
2	A	0	i	
3	G		0	
4	C			0

Flow Chat



Nussinov Folding Algorithm

Algorithm

- Input: Sequence $x = (x_1, x_2, \dots, x_L)$
- Output: Maximal number $S(i, j)$ of base pairs for (x_i, \dots, x_j) .
- Initialization:

$$\begin{array}{ll} S(i, i) = 0 & \text{for } i = 1 \text{ to } L. \\ S(i, i-1) = 0 & \text{for } i = 2 \text{ to } L; \end{array}$$

for $n = 2$ to L do
 for $j = n$ to L do
 Set $i = j - n + 1$

$$S(i, j) = \max \left\{ \begin{array}{l} S(i+1, j) \\ S(i, j-1) \\ S(i+1, j-1) + e(i, j) \\ \mathbf{Max}_{i < k < j} \{ S(i, k) + S(k+1, j) \} \end{array} \right.$$

Return $S(1, L)$

Initialization:

$$S(i, i) = 0$$

for $I = 2$ *to* L .

[illegible]

Example

Initialization:

$S(i, i) = 0$ $S(i, i-1) = 0$ *for* $I = 2$ *to* L .
for $I = 2$ *to* L ;

	A	C	U	G
A	0	0	1	0
C	0	0	0	1
U	1	0	0	1
G	0	1	1	0

$j \longrightarrow$

		1	2	3	4	5	6	7	8	9
$i \downarrow$		G	G	G	A	A	A	U	C	C
	1	G	0							
	2	G	0	0						
	3	G		0	0					
	4	A			0	0				
	5	A				0	0			
	6	A					0	0		
	7	U						0	0	
	8	C							0	0
	9	C								0

Recursive Relation

For all subsequences from length 2 to length L:

$$S(i, j) = \max \left\{ \begin{array}{l} S(i + 1, j) \\ S(i, j - 1) \\ S(i + 1, j - 1) + e(i, j) \\ \mathbf{Max}_{i \leq k \leq j} \{ \mathcal{S}(i, k) + \mathcal{S}(k + 1, j) \} \end{array} \right.$$

[illegible]

Example

$j \longrightarrow$

		1	2	3	4	5	6	7	8	9
$i \downarrow$		G	G	G	A	A	A	U	C	C
	1 G	0	0	0						
	2 G	0	0	0	0					
	3 G		0	0	0	0				
	4 A			0	0	0	0			
	5 A				0	0	0	1		
	6 A					0	0	1	1	
	7 U						0	0	0	0
	8 C							0	0	0
	9 C								0	0

$$S(i, j) = \max \begin{cases} S(i + 1, j) \\ S(i, j - 1) \\ S(i + 1, j - 1) + e(i, j) \\ \mathbf{Max}_{i < k < j} \{S(i, k) + s(k + 1, j)\} \end{cases}$$

Example

$j \longrightarrow$

$i \downarrow$

		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
1	G	0	0	0	0					
2	G	0	0	0	0	0				
3	G		0	0	0	0	0			
4	A			0	0	0	0			
5	A				0	0	0	1	1	
6	A					0	0	1	1	1
7	U						0	0	0	0
8	C							0	0	0
9	C								0	0

$$S(4,7) = \max \begin{cases} S(5,7) \\ S(4,6) \\ S(5,6) + e(4,7) \\ \mathbf{Max}_{4 \leq k < 7} \{S(4,k) + s(k+1,7)\} \end{cases}$$

Example

$j \longrightarrow$

		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
$i \downarrow$	1	G	0	0	0					
	2	G	0	0	0	0				
	3	G		0	0	0	0			
	4	A			0	0	0			
	5	A				0	0	1	1	
	6	A				0	0	1	1	1
	7	U					0	0	0	0
	8	C						0	0	0
	9	C							0	0

$$S(4,7) = \max \left\{ \begin{array}{l} S(5,7) \leftarrow \\ S(4,6) \\ S(5,6) + e(4,7) \\ \text{Max}_{4 \leq k < 7} \{S(4,k) + s(k+1,7)\} \end{array} \right.$$

Example

$j \longrightarrow$

		1	2	3	4	5	6	7	8	9
$i \downarrow$		G	G	G	A	A	A	U	C	C
	1 G	0	0	0	0					
	2 G	0	0	0	0	0				
	3 G		0	0	0	0	0			
	4 A			0	0	0	0			
	5 A				0	0	0	1	1	
	6 A					0	0	1	1	1
	7 U						0	0	0	0
	8 C							0	0	0
	9 C								0	0

$$S(4,7) = \max \begin{cases} S(5,7) \\ S(4,6) \leftarrow \\ S(5,6) + e(4,7) \\ \text{Max}_{4 < k < 7} \{S(4,k) + s(k+1), 7\} \end{cases}$$

Example

$j \longrightarrow$

		1	2	3	4	5	6	7	8	9
$i \downarrow$		G	G	G	A	A	A	U	C	C
	1 G	0	0	0	0					
	2 G	0	0	0	0	0				
	3 G		0	0	0	0	0			
	4 A			0	0	0	0			
	5 A				0	0	0	1	1	
	6 A					0	0	1	1	1
	7 U						0	0	0	0
	8 C							0	0	0
	9 C								0	0

$$S(4,7) = \max \left\{ \begin{array}{l} S(5,7) \\ S(4,6) \\ S(5,6) + e(4,7) \leftarrow \\ \text{Max}_{4 < k < 7} \{S(4, k) + s(k+1, 7)\} \end{array} \right.$$

Example

→

		1	2	3	4	5	6	7	8	9	j
i		G	G	G	A	A	A	U	C	C	
	1 G	0	0	0	0						
	2 G	0	0	0	0	0					
	3 G		0	0	0	0	0				
	4 A			0	0	0	0				
	5 A				0	0	0	1	1		
	6 A					0	0	1	1	1	
	7 U						0	0	0	0	
	8 C							0	0	0	
	9 C								0	0	

↓

$$S(4,7) = \max \begin{cases} S(5,7) \\ S(4,6) \\ S(5,6) + e(4,7) \\ \mathbf{Max_{4 \leq k \leq 7} \{S(4,k) + s(k+1), 7\}} \end{cases}$$

↗

Example

$j \longrightarrow$

		1	2	3	4	5	6	7	8	9
$i \downarrow$		G	G	G	A	A	A	U	C	C
	1 G	0	0	0	0					
	2 G	0	0	0	0	0				
	3 G		0	0	0	0	0			
	4 A			0	0	0	0	1		
	5 A				0	0	0	1	1	
	6 A					0	0	1	1	1
	7 U						0	0	0	0
	8 C							0	0	0
	9 C								0	0

$$S(4,7) = \max \begin{cases} S(5,7) \\ S(4,6) \\ S(5,6) + e(4,7) \\ \text{Max}_{4 < k < 7} \{S(4,k) + s(k+1), 7\} \end{cases}$$

$j \longrightarrow$

		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
$i \downarrow$	1	G	0	0	0	0	0	1	2	3
	2	G	0	0	0	0	0	1	2	3
	3	G		0	0	0	0	1	2	2
	4	A			0	0	0	1	1	1
	5	A				0	0	1	1	1
	6	A					0	1	1	1
	7	U						0	0	0
	8	C							0	0
	9	C								0

Traceback

```
if  $i < j$  then
  if  $S(i, j) = S(i + 1, j)$  then
    traceback( $i + 1, j$ )
  else if  $S(i, j) = S(i, j - 1)$  then
    traceback( $i, j - 1$ )
  else if  $S(i, j) = S(i + 1, j - 1) + w(i, j)$ 
  then
    print base pair ( $i, j$ )
    traceback( $i + 1, j - 1$ )
  else for  $k = i + 1$  to  $j - 1$  do
    if  $S(i, j) = S(i, k) + S(k + 1, j)$  then
      traceback( $i, k$ )
      traceback( $k + 1, j$ )
    break
end
```

PAIRS

STACK
(1,9)

CURRENT

$j \longrightarrow$

		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
1	G	0	0	0	0	0	0	1	2	3
2	G	0	0	0	0	0	0	1	2	3
3	G		0	0	0	0	0	1	2	2
4	A			0	0	0	0	1	1	1
5	A				0	0	0	1	1	1
6	A					0	0	1	1	1
7	U						0	0	0	0
8	C							0	0	0
9	C								0	0

$i \downarrow$

Traceback

if $i < j$ then

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

$\text{traceback}(i + 1, j)$

else if $S(i, j) = S(i, j - 1)$ then

$\text{traceback}(i, j - 1)$

else if $S(i, j) = S(i + 1, j - 1) + w(i, j)$

then

 print base pair (i, j)

$\text{traceback}(i + 1, j - 1)$

else for $k = i + 1$ to $j - 1$ do

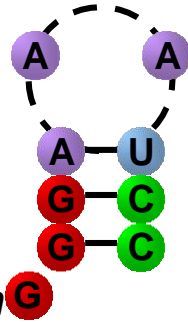
 if $S(i, j) = S(i, k) + S(k + 1, j)$ then

$\text{traceback}(i, k)$

$\text{traceback}(k + 1, j)$

 break

end



PAIRS

(2,9)

(3,8)

(4,7)

STACK

-

CURRENT

(6,6)

		$j \longrightarrow$								
		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
$i \downarrow$	1	G	0	0	0	0	0	1	2	3
	2	G	0	0	0	0	0	1	2	3
	3	G		0	0	0	0	1	2	2
	4	A			0	0	0	1	1	1
	5	A				0	0	1	1	1
	6	A					0	1	1	1
	7	U						0	0	0
	8	C							0	0
	9	C								0

Traceback

if $i < j$ then

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

$$\text{traceback}(i + 1, j)$$

else if $S(i, j) = S(i, j - 1)$ then

$$\text{traceback}(i, j - 1)$$

else if $S(i, j) = S(i + 1, j - 1) + w(i, j)$

then

```
print base pair (i, j)
```

$$\text{traceback}(i + 1, j - 1)$$

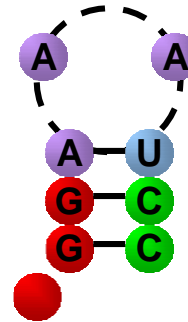
```
else for  $k = i + 1$  to  $j - 1$  do
```

if $S(i, j) = S(i, k) + S(k + 1, j)$ then

$$\text{traceback}(i, k)$$
$$traceback(k + 1, j)$$

break

end



$j \longrightarrow$

		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
	1 G	0	0	0	0	0	0	1	2	3
	2 G	0	0	0	0	0	0	1	2	3
	3 G		0	0	0	0	0	1	2	2
	4 A			0	0	0	0	1	1	1
	5 A				0	0	0	1	1	1
	6 A					0	0	1	1	1
	7 U						0	0	0	0
	8 C							0	0	0
	9 C								0	0

i
↓

Base Pair Maximization Algorithm Issues

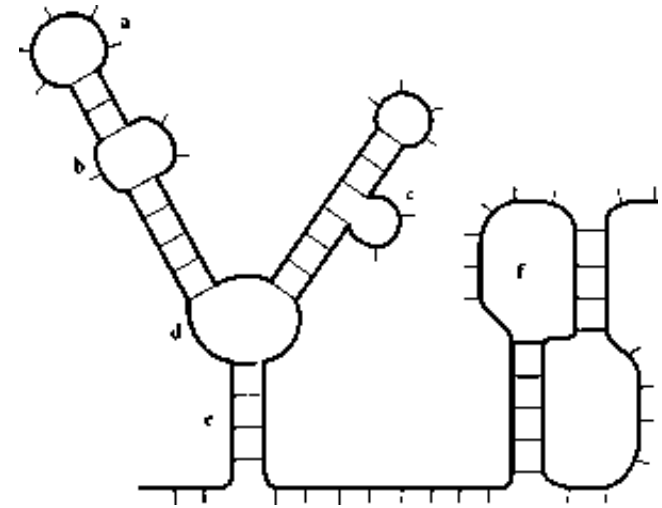
What is computational complexity of algorithm?

(for sequence of length N)

Answer: Memory - $O(N^2)$ Time - $O(N^3)$

Can it handle pseudoknots?

Answer: No. Pseudoknots invalidate recursion for $S(i,j)$



Evaluation of Maximizing Base pairs

Simplistic approach

Does not give accurate structure predictions, as

- no stacking of base pairs considered

- loop sizes not distinguished

- no special scoring of multi-loops

Misses:

- nearest neighbour interactions

- stacking interactions

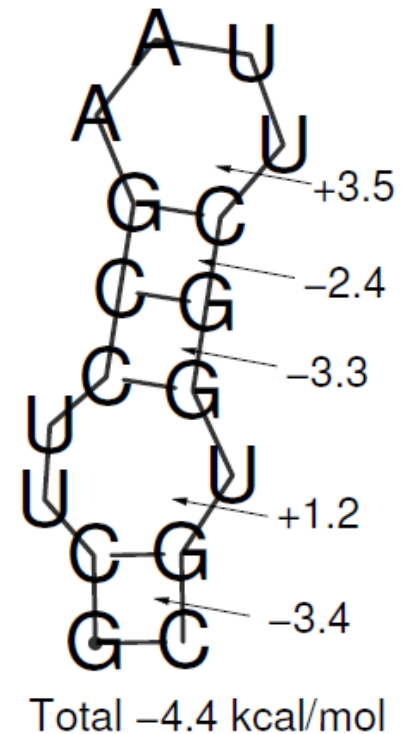
- loop length preferences

Nearest Neighbor Energies

Free energy of a structure approximated as the sum of loop energies

Loop energies depend on loop type and size

Free energies are dependent on temperature and ionic conditions



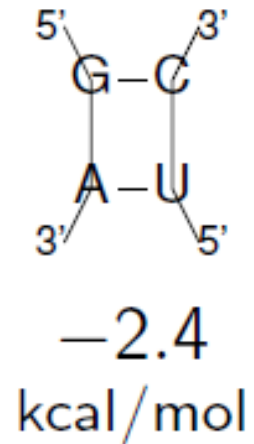
Stacked Pairs

Major source of stabilizing energy

all 21 combinations measured, accuracy at least 0.1 kcal/mol

include the hydrogen bonding energy of pair formation

energies of G·U pairs depend on context, i.e. violate the nearest-neighbor model



		$(i + 1, j - 1)$					
		CG	GC	GU	UG	AU	UA
(i, j)	CG	-2.4	-3.3	-2.1	-1.4	-2.1	-2.1
	GC	-3.3	-3.4	-2.5	-1.5	-2.2	-2.4
	GU	-2.1	-2.5	1.3	-0.5	-1.4	-1.3
	UG	-1.4	-1.5	-0.5	0.3	-0.6	-1.0
	AU	-2.1	-2.2	-1.4	-0.6	-1.1	-0.9
	UA	-2.1	-2.4	-1.3	-1.0	-0.9	-1.3

For comparison: Thermal energy $RT \approx 0.6$ kcal/mol at 37C.

Energy Minimization and the Zuker folding algorithm

RNA folding algorithm is dictated by **biophysics** rather than by counting and maximizing the number of base pairs.

The most sophisticated secondary structure prediction method for single RNAs is the **ZUKER algorithm**, an energy minimization algorithm which assumes that the correct structure is the one with the lowest equilibrium free energy (ΔG)

Stabilized Energy

Destabilized Energy

Sum of energy

Energy Minimization Methods

Energy minimization algorithm predicts the correct secondary structure by minimizing the free energy (ΔG)

Gibbs Free Energy G of a system:

$$\mathbf{G = H - TS}$$

$$\Delta G = 0; \quad \Delta G > 0; \quad \Delta G < 0$$

G calculated as sum of individual contributions of:

- loops

- base pairs

- secondary structure elements

Energies of stems calculated as stacking contributions between neighboring base pairs

Zuker's Algorithm

The minimum energy structure can be calculated recursively by a dynamic programming algorithm.

The principal difference is that because of stacking parameters, two matrices (called V and W) are kept instead of one.

$W(i, j)$ is the energy of the best structure on i, j .

$V(i, j)$ is the energy of the best structure on i, j given i, j are paired.

Assumptions

In predicting minimum energy of RNA secondary structure, several simplifying assumptions are made

1. The most likely structure is identical to the energetically preferable structure
2. Nearest-neighbor energy calculations give reliable estimates of an experimentally achievable energy measurements
3. Usually we can neglect pseudoknots

Energy of Secondary Structure Elements

- Energy contributions of the various structure elements:
 - $eH(i, j)$
 - $eS(i, j)$
 - $eL(i, j, l', j')$
 - $eM(i, j, i_1, j_1, \dots, i_k, j_k)$

General multi loop contribution will be too expensive in prediction:

We need to use a simplified contribution scheme.

- multiloop $eM(i, j, k, k') = a + bk + ck'$

a, b, c = weights

a = energy contribution for closing of loop

k = number of inner base pairs

k' = number of unpaired bases within loop

$W(i)$ -Zuker's Algorithm for RNA Energy Minimization

energy of an optimal structure of subsequence 1 through i :

For an RNA sequence S , define the Zuker-matrix W as a matrix of entries W_{ij} for $1 \leq i \leq j \leq n$ by

$$W(i) = \min \begin{cases} W(i-1) & \text{---- } j \text{ Unpaired} \\ \min_{i < j \leq i} \{ W(j-1) + V(j, i) \} & \text{---- } j \text{ paired} \end{cases}$$

$V(i,j)$ - Zuker's Algorithm for RNA Energy Minimization

- energy of an optimal structure of subsequence i through j closed by $i \bullet j$:
- For an RNA sequence S , define the Zuker-matrix V as a matrix of entries V_{ij} for $1 \leq i \leq j \leq n$ by

$$V(i,j) = \min \begin{cases} eH(i,j) & \text{---- hairpin loop} \\ eS(i,j) + V(i+1,j-1) & \text{---- stacking loop} \\ VBI(i,j) & \text{---- interior loop/bulge} \\ VM(i,j) & \text{---- multi-loop} \end{cases}$$

$eH(i,j)$

- energy of hairpin loop closed by $i \bullet j$

computed with:

$$\delta\delta G = 1.75 \times RT \times \ln(l_s),$$

Loop Energy Table

DESTABILIZING ENERGIES BY SIZE OF LOOP			
SIZE	INTERNAL	BULGE	HAIRPIN

1	.	3.8	.
2	.	2.8	.
3	.	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

$eS(i,j)$

- energy of **stacking base pair** $i \bullet j$ with $i+1 \bullet j-1$

		5'	→	3'	
		CX			
		GY			
		3'	←	5'	
	Y:	A	C	G	U

X:	A		.	.	-2.1
	C		.	-3.3	.
	G		-2.4	.	-1.4
	U		-2.1	-2.1	.

$VBI(i,j)$

- energy of an optimal structure of the subsequence from i through j , where $i \bullet j$ closes a **bulge or an internal loop**

$$VBI(i, j) = \min_{\substack{i < i' < j' < j \\ i' - i + j - j' > 2}} \{eL(i, j, i', j') + V(i', j')\}$$

$$eL(i,j,i',j')$$

- energy of a **bulge** or **internal loop** with exterior base pair $i \bullet j$ and interior base pair $i' \bullet j'$

		5'	-->	3'	
			X		
			C	A	
			G	U	
			YA		
		3'	<--	5'	
Y:	A	C	G	U	

A		3.2	3.0	2.4	4.8
C		3.1	3.0	4.8	3.0
G		2.5	4.8	1.6	4.8
U		4.8	4.8	4.8	4.8

$VM(i,j)$

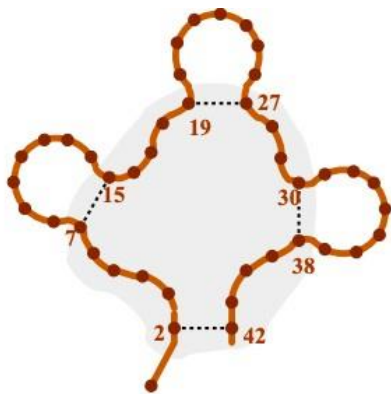
- energy of an optimal structure of the subsequence from i through j , where $i \bullet j$ closes a **multibranched loop**

$$VM(i,j) = \min_{\substack{i < i_1 < j_1 < \dots \\ < i_k < j_k < j}} \{eM(i,j,i_1,j_1,\dots,i_k,j_k) + \sum_{l=1}^k V(i_l,j_l)\}$$

Simplified Multi-loop Energy — Example

- In general: multi-loop energy depends on everything: inner base pairs $(i_1, j_1) \dots (i_k, j_k)$, closing base pair (i, j) , and sequence.

Example:



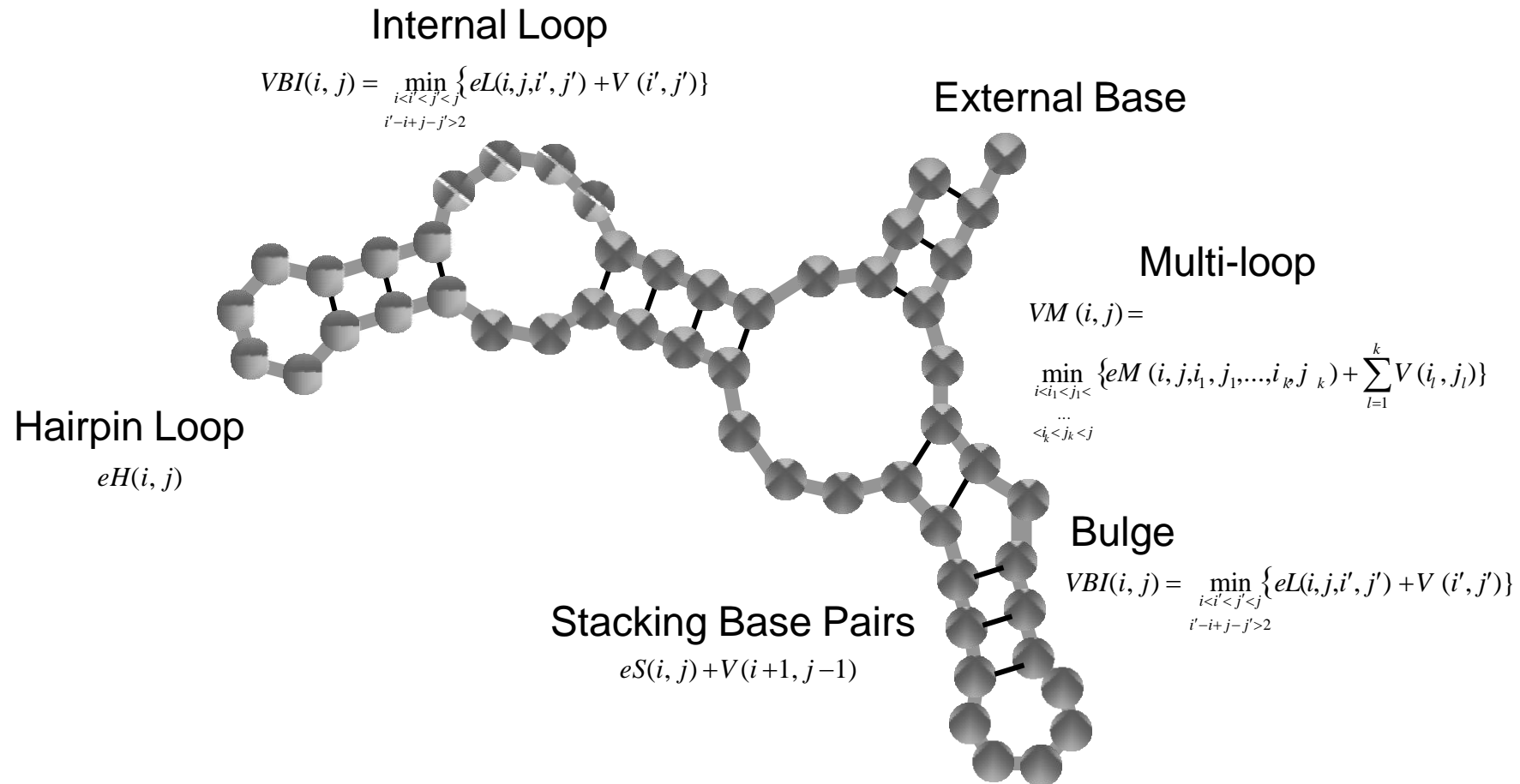
general: $eM(2, 42, 7, 15, 19, 27, 30, 38)$

simplified: $eM(2, 42, k, k') = a + bk + ck'$, where

$k = 3$: inner base pairs within loop

$k' = 12$: unpaired bases within multi-loop

Assembling the Pieces



Zuker-Algorithm: Summary

3 matrices:

W — minimal energy of **general substructure** $i \dots j$

V — minimal energy of **closed substructure** $i \dots j$

WM — minimal energy of true part of a **multi-loop** $i \dots j$

Recursions equations

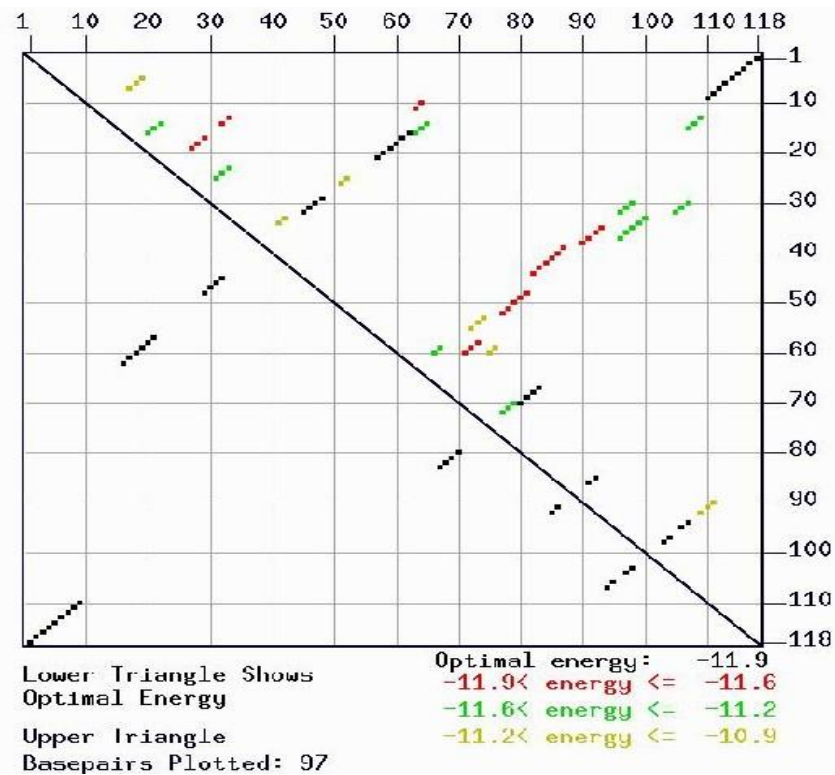
$$WM(i, j) = \min \left\{ \begin{array}{l} V(i, j) + b \\ WM(i, j-1) + c \\ WM(i+1, j) + c \\ \min_{i < k \leq j} \{ WM(i, k-1) + WM(k, j) \} \end{array} \right\}$$

Zuker Algorithm Implementation (MFOLD)

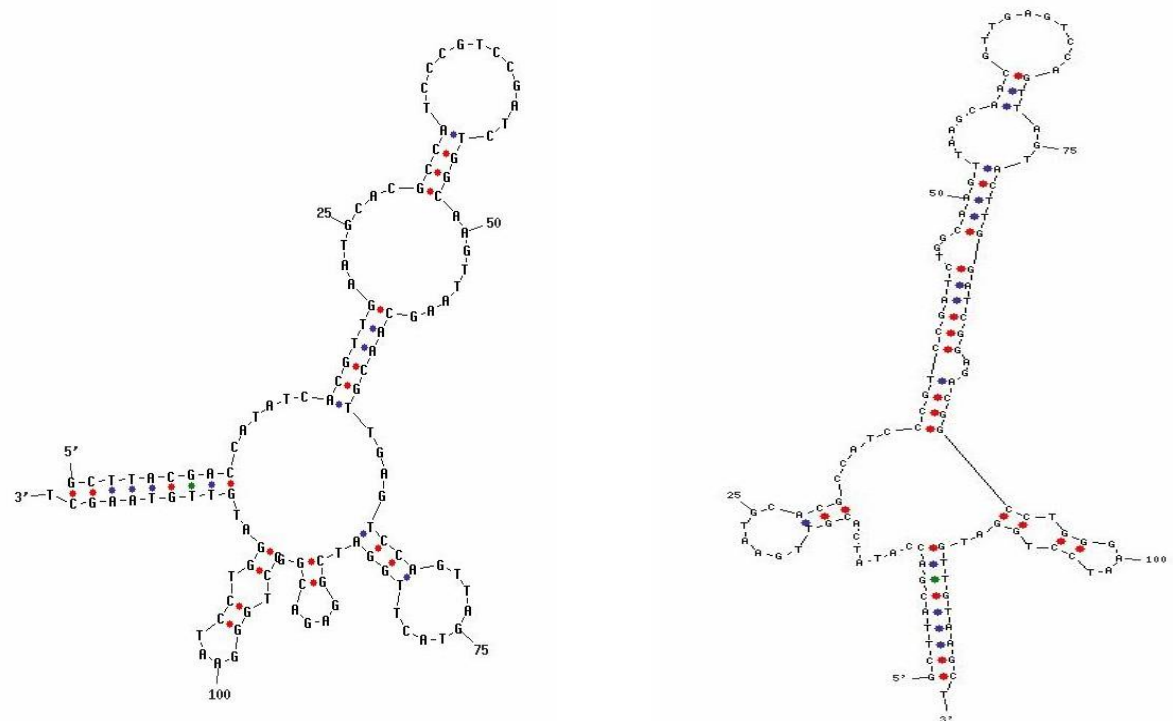
Example Sequence:

```
GCTTACGACCATATCACGTTGAATGCACGCCATCCCGTCCGATCTGGCAAGTTAAGCAAC  
GTTGAGTCCAGTTAGTACTTGGATCGGAGACGGCCTGGGAATCCTGGATGTTGTAAGCT
```

MFOLD Energy Dot Plot:



Resulted Optimal Structures:



CDPfold – New method

New Method of RNA Secondary Structure Prediction

It combines a convolutional neural network and dynamic programming as well as a sequence alignment method.

Proposed by Zhang et. al in 2019

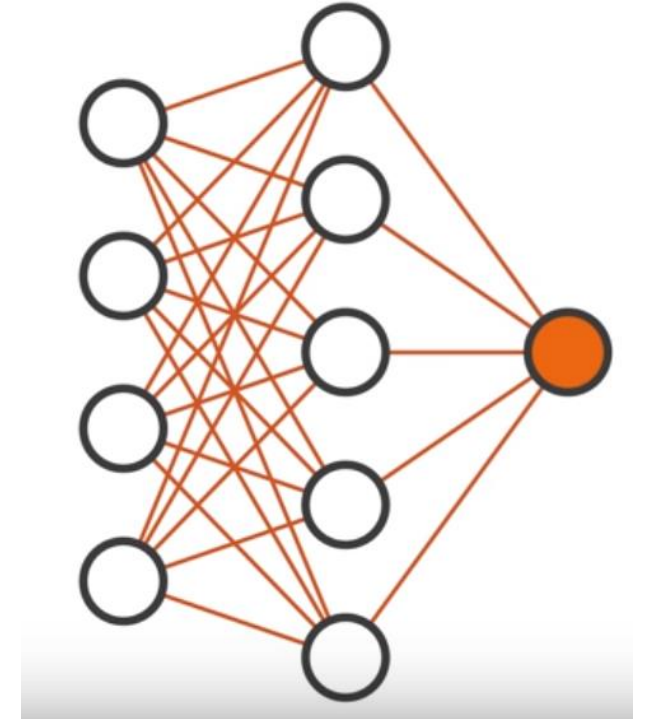
What is a convolutional neural network?

A convolutional neural network (CNN) is a type of neural network

Neural Network

A regular neural network has an **input layer**, **hidden layers** and an **output layer**.

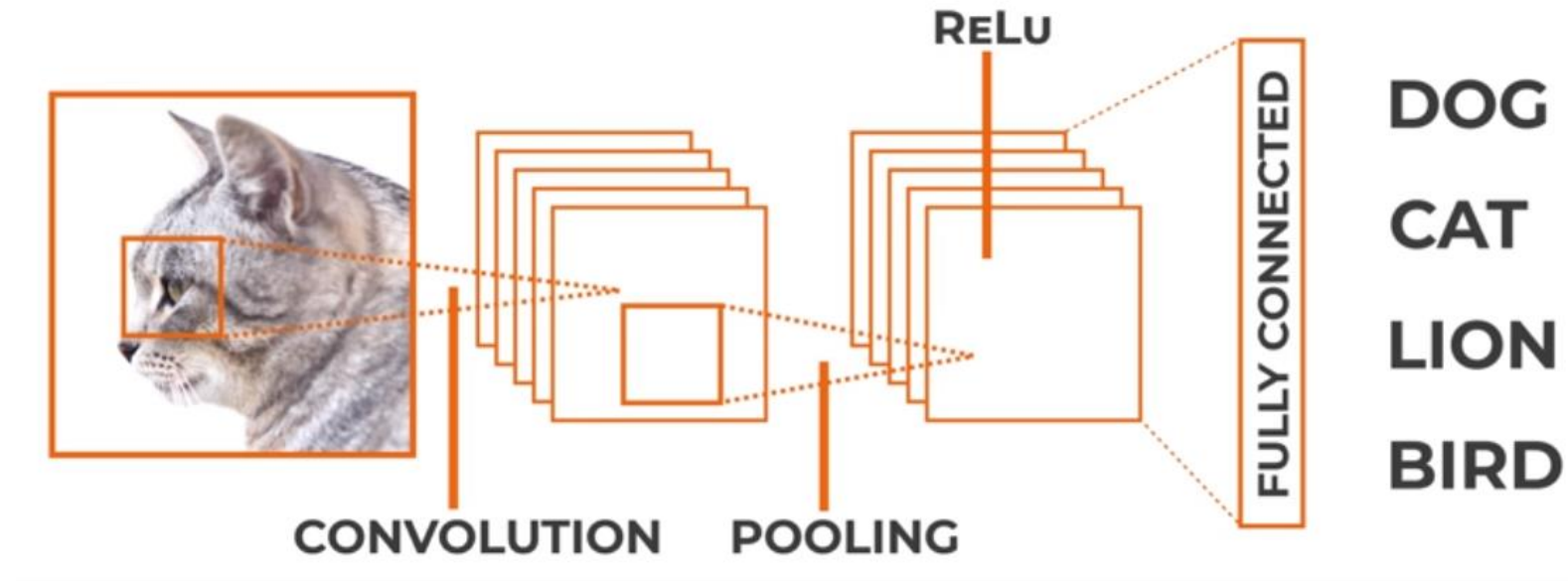
Each of these layers contains neurons that are connected to neurons in the previous layer, and each neuron has its own weight.



What's inside a convolutional neural network?

The word 'convolutional' refers to the filtering process.

Like a normal neural network, a convolutional neural network is made up of multiple layers.



The **ReLU layer (rectified linear unit layer)** acts as an activation function, ensuring non-linearity as the data moves through each layer in the network.

Fully connected layer

What's inside a convolutional neural network?

Convolutional layer

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

What's inside a convolutional neural network?

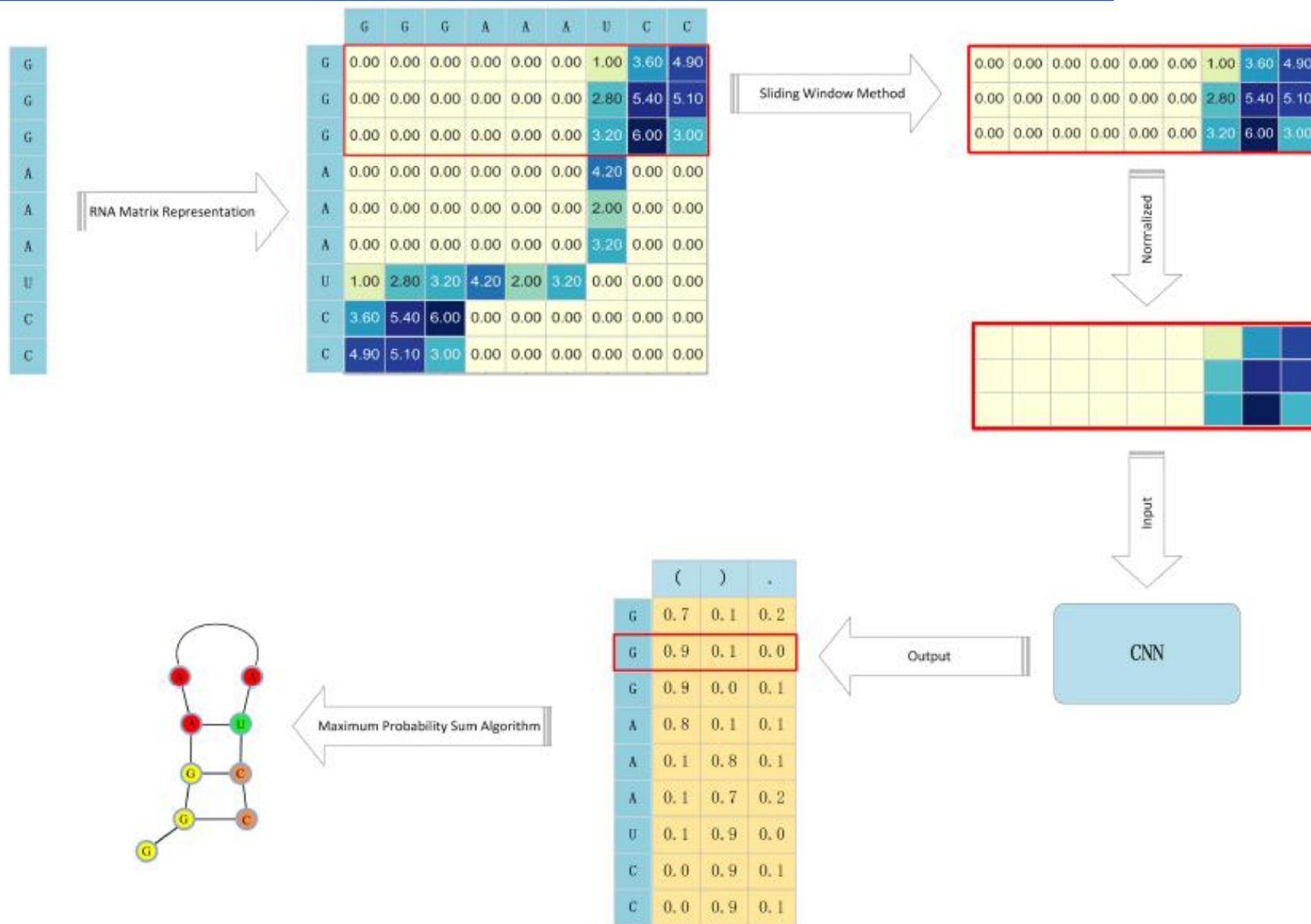
The **pooling layer** – this down samples or reduces the sample size of a particular feature map. This makes processing much faster.

There are two ways of doing this

- **max pooling**
- **average pooling**

These steps amount to feature extraction.

CDPfold – The process and Analysis



RESULTS (from article)

For the prediction of an RNA secondary structure obtained by the CDPfold, results are obtained by two indicators, sensitivity and specificity.

1. **Sensitivity** (recall-rate in machine learning)
2. **Specificity** (precision-rate in machine learning)

The **F-score** can be used to measure the precision and recall.

$$\text{F-score} = \frac{2 \times \text{Sensitivity} \times \text{Specificity}}{\text{Sensitivity} + \text{Specificity}}$$

RESULTS

Table shows the accuracy of the designed algorithm compared with other algorithms on the 5sRNA dataset.

TABLE | Comparison of algorithms in 5sRna.

Software	5sRNA		
	Sensitivity	Specificity	<i>F</i> -score
mfold	0.693	0.704	0.698
RNAfold	0.694	0.704	0.699
cofold	0.585	0.591	0.588
Sfold	0.703	0.733	0.718
CDPfold	0.932	0.916	0.924

RESULTS

Using the F-score, they get the predicted effect of the designed generic model on the three types of RNA datasets.

They used the same test data to perform experiments under other published algorithms.

TABLE | Comparison of three types of RNA based on their prediction accuracy.

Software	5sRNA	tRNA	srpRNA
Mfold	0.698	0.631	0.566
RNAfold	0.699	0.632	0.577
CDPfold	0.911	0.905	0.823

Discussion

Experimentally, the method has had good performance in predicting the accuracy of a RNA secondary structure.

Although CDPfold has achieved good results in RNA secondary structure prediction, some problems encountered during the experimental process.

The results predicted by the CDPfold method still need to be further **corrected** in the results predicted by the convolutional neural network.

Discussion

In the current prediction of the RNA secondary structure, the prediction of pseudoknots is still a difficult point.

The RNA structure representation method used in this paper uses the dot bracket representation. However, the dot bracket representation does not reflect the false knots present in the RNA structure.

The prediction of the secondary structure of longer RNA sequences is not reflected. This is because the current experimental methods are not perfect enough.

References

- [1] Recent advances in RNA folding; Jörg Fallmann, Sebastian Willb, Jan Engelhardt, Björn Grüning, Rolf Backofen, Peter F. Stadler
- [2] Improved predictions of secondary structures for RNA; JOHN A. JAEGER*, DOUGLAS H. TURNER*, AND MICHAEL ZUKER*
- [3] Fast algorithm for predicting the secondary structure of single-stranded RNA (computer program/polynucleotide/RNA folding); RUTH NUSSINOV* AND ANN B. JACOBSON
- [4] A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming; Hao Zhang¹, Chunhe Zhang¹, Zhi Li², Cong Li¹, Xu Wei¹, Borui Zhang³ and Yuanning Liu¹
- [5] Hogeweg P Hesper, B 1984 Energy directed folding of RNA sequences Nucleic Acids Research, 12 1 Part 1 67 74
doi 10.1093/nar/12.1.67
- [6] Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization, D Mathews Rna 2004
- [7] Haoyue Fu, Lianping Yang, Xiangde Zhang, "An RNA secondary structure prediction method based on minimum and suboptimal free energy structures", Journal of Theoretical Biology, vol. 380, pp. 473, 2015
- [8] M. E. Nebel and A. Scheid, "Analysis of the Free Energy in a Stochastic RNA Secondary Structure Model," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 6, pp. 1468 1482, Nov. Dec. 2011

Thank You.