COMP561: Computational Biology Methods & Research

RNA minimum free energy secondary structures

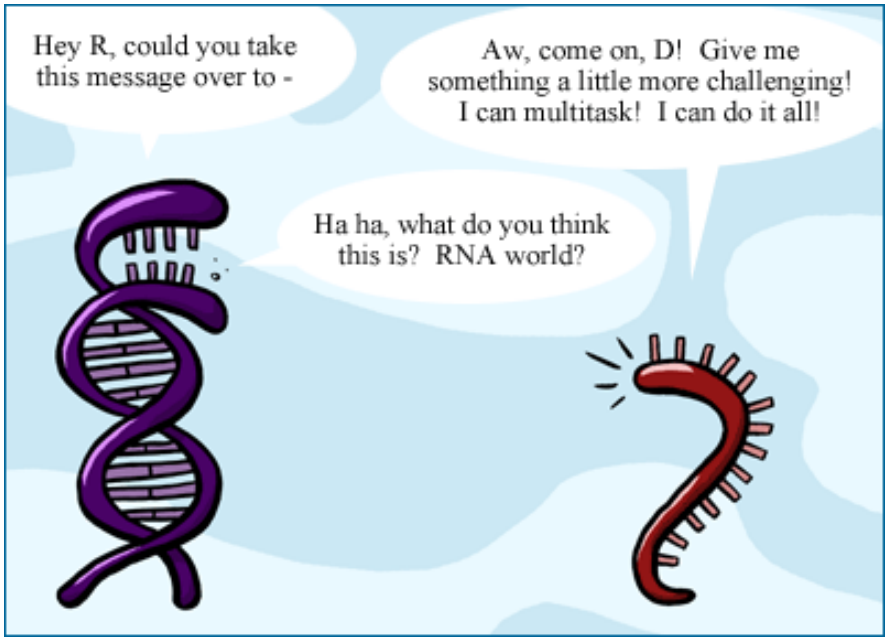
Jérôme Waldispühl
School of Computer Science, McGill

RNA world

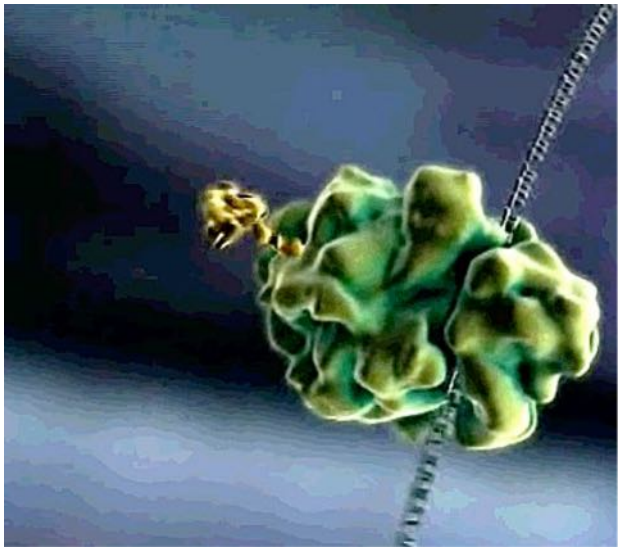
In prebiotic world, RNA thought to have filled two distinct roles:

1. an information carrying role because of RNA's ability (in principle) to self-replicate,
2. a catalytic role, because of RNA's ability to form complicated 3D shapes.

Over time, DNA replaced RNA in
Its first role, while proteins replaced
RNA in its second role.



RNA classification

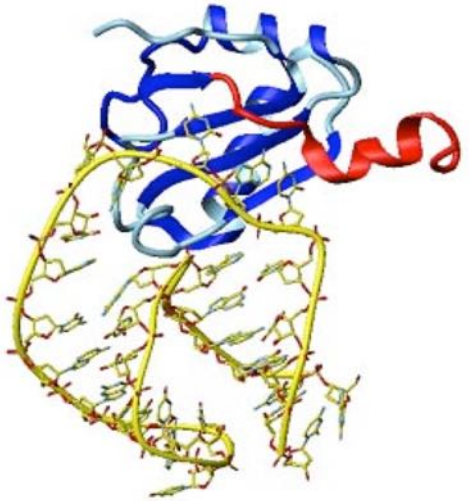


Messenger RNA:

- Carry genetic information,
- Structure less important.

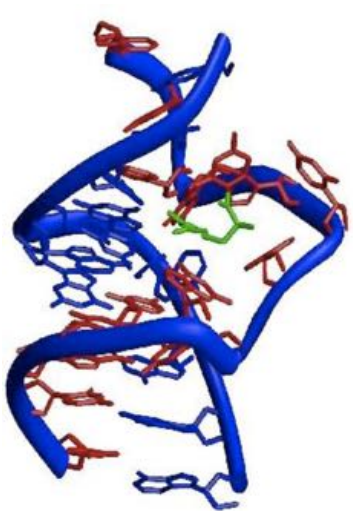
Non-coding RNA:

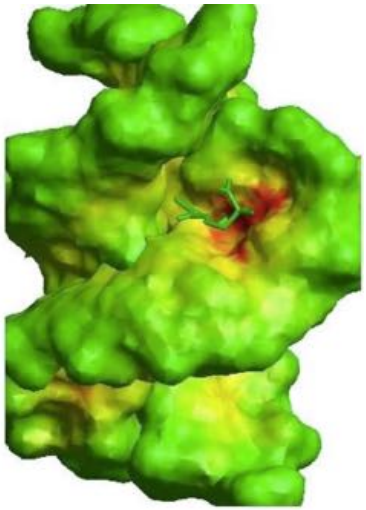
- Functional,
- Structure is important.

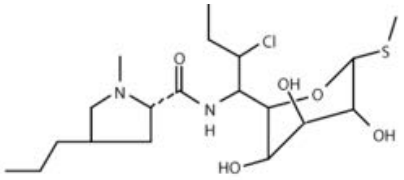


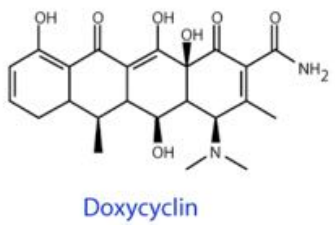
RNA structure and function

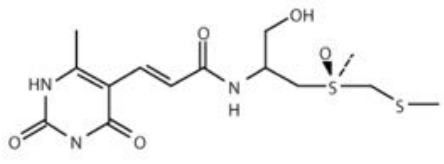
- RNAs have a 3D structure,
- This 3D structure allow complex functions,
- The variety of RNA structures allow the specific recognition of a wide range of ligands,
- Some molecules target these RNA structures (antibiotics, antiviruses):

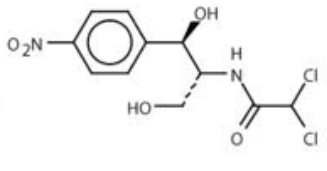




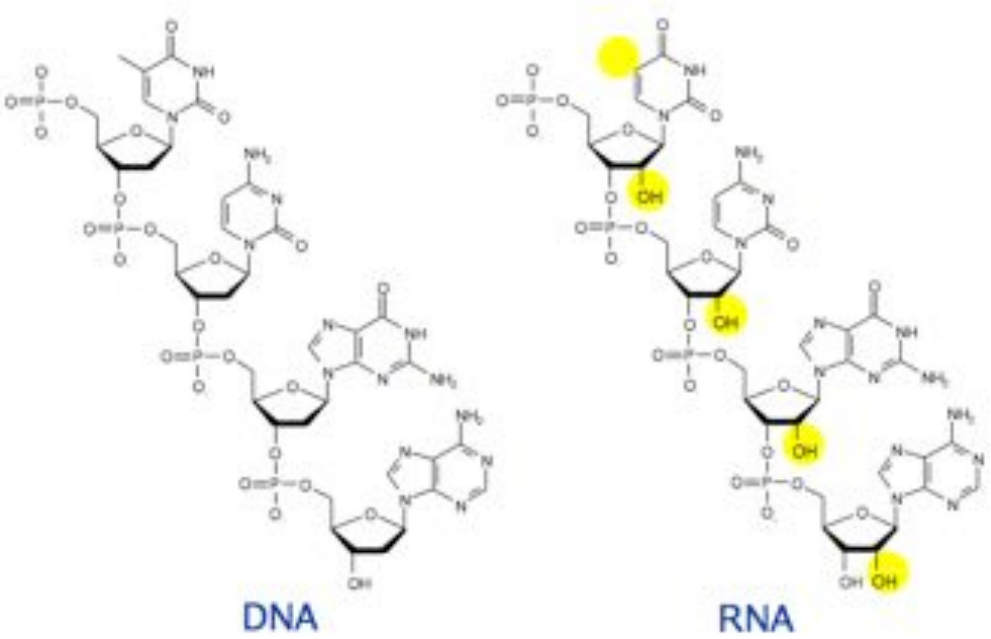








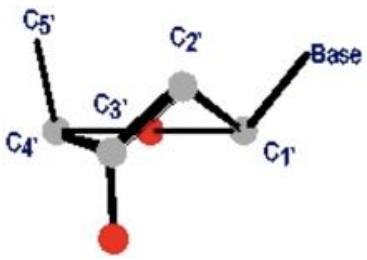
RNA vs DNA: Chemical nature



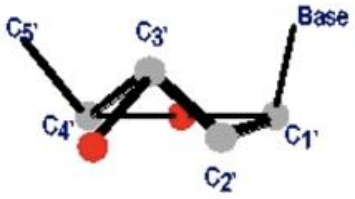
- 2’-OH group attached to sugar (instead of 2’-H): *more polar*
- Substitution of thymine by uracile = suppression of group 5-CH3

Small modifications => big effects

RNA vs DNA: Modification of the local and global geometry



C2’ endo

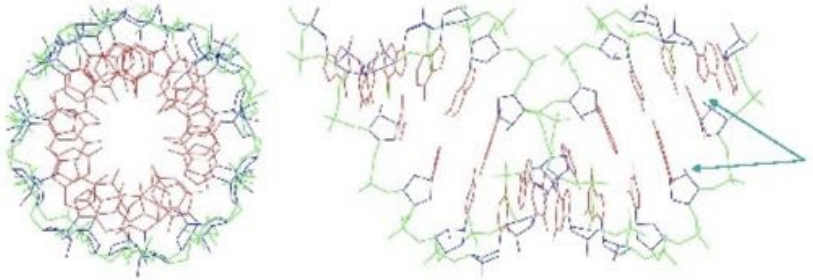


C3’ endo

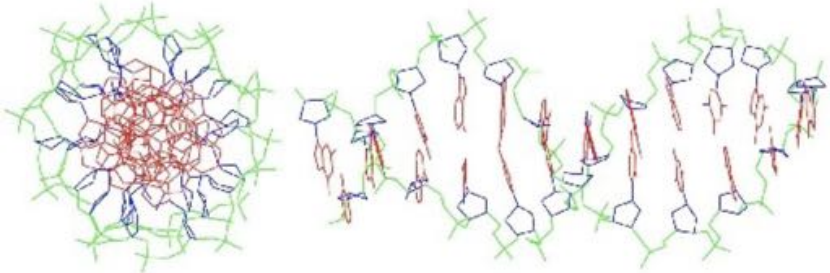
:Local conformation

Global conformation:

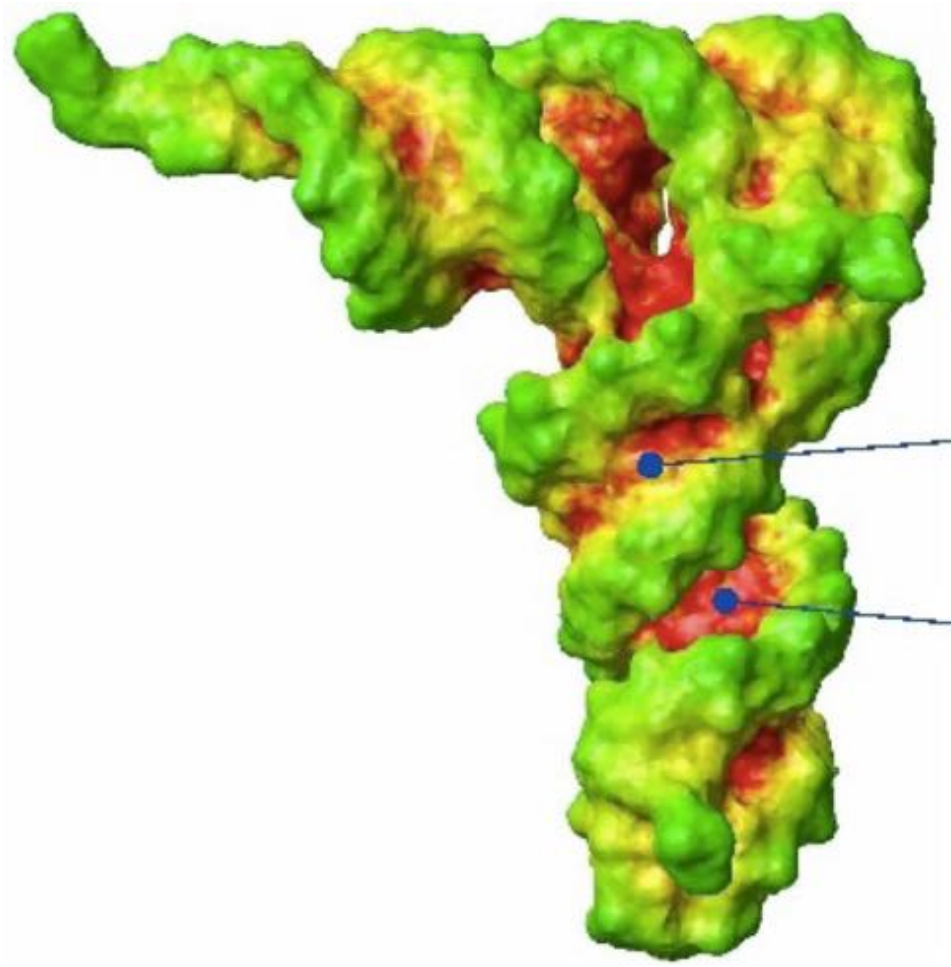
RNA favorite:



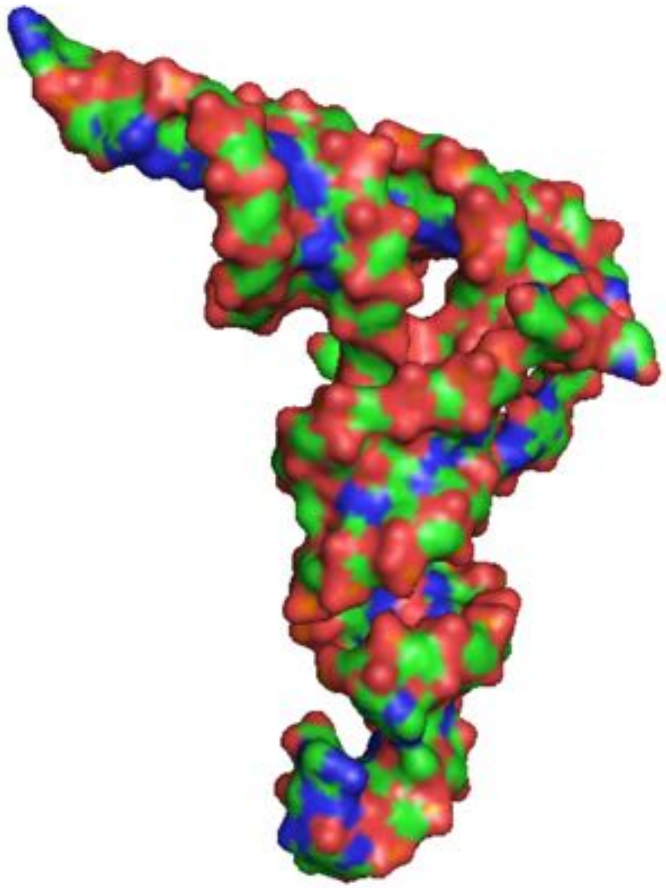
DNA favorite:



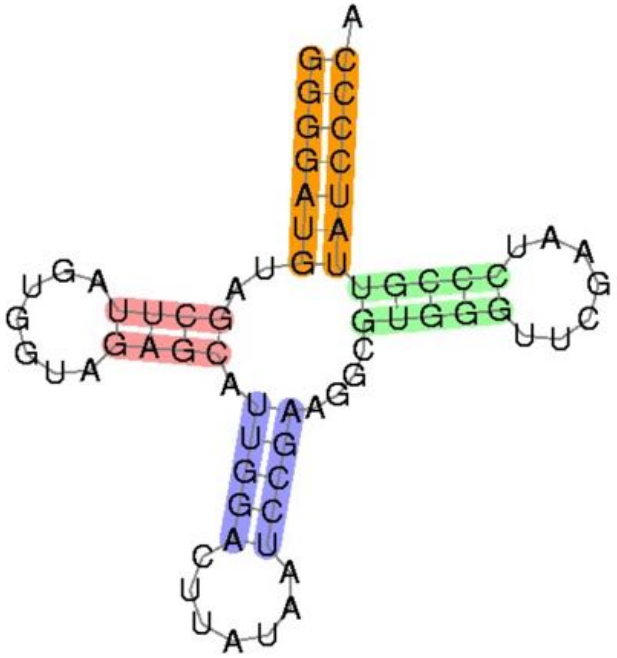
RNA vs DNA: Consequence of the modification of the geometry



RNA secondary structure



The **secondary structure** is the set of canonical base-pairs forming the scaffold of the 3D structure.



RNA secondary structure

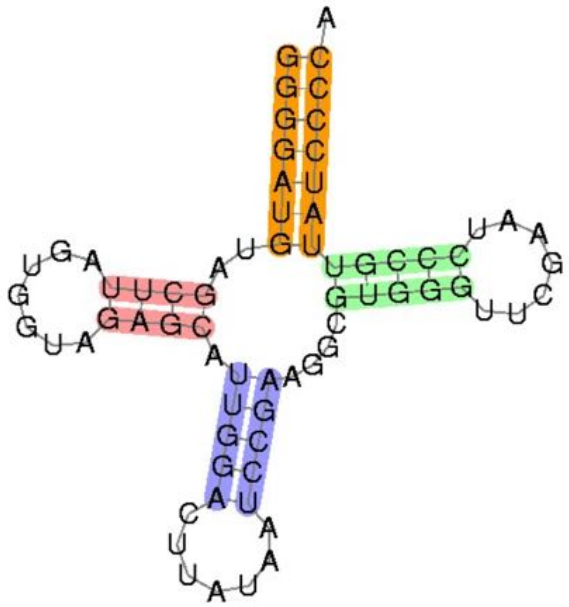
Central assumption: RNA secondary structure forms before the tertiary structure.

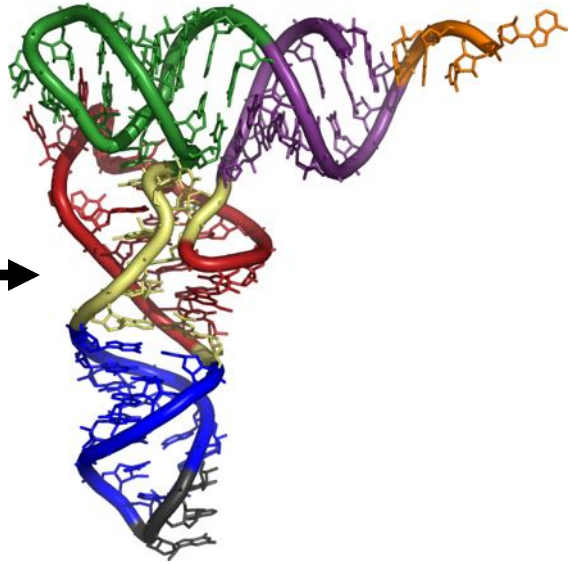
Primary structure

Secondary structure

Tertiary structure

cgcgggggttgatataatataaaaaataat
aaataataataataataattatcatcattt
ccgacccatattataataatacgggttgg
aaatatagatataatatttattatattgat
ataatacatatatataagttagaggaaat
gttgtttaaaggttaaactgttagattgca
aatctacacatttagagttcgattctcttc
atttcttatatatatactacccacgcg





This class: Secondary structure prediction using energy minimization principles

Principle of minimum energy

For a closed system with fixed entropy, the total energy is minimized at equilibrium.

Application to RNA folding:

- Closed system: Isolated RNA molecule
- Energy of system: Folding energy of the RNA
- State of the system: An RNA (secondary) structure

Definition

- Let $\omega \in \{A,C,G,U\}^*$ be a RNA sequence
- Let Δ be the ensemble of all secondary structures S compatible with ω.
- Let $E(S, \omega)$ be the free energy on ω folded in S.

Then, the minimum free energy (MFE) of ω is:

$$MFE(\omega) = \min_{S \in \Delta}(E(S,\omega))$$

And the minimum free energy secondary structure of ω is the structure S such that $E(S,\omega) = MFE(\omega)$.

Note: Here, we assume it exists an unique structure S satisfying the equation.

Free energy of a RNA secondary structure

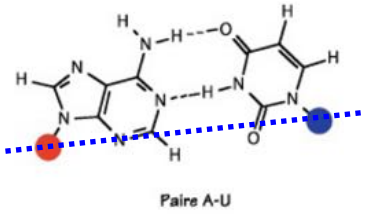
First approximation:

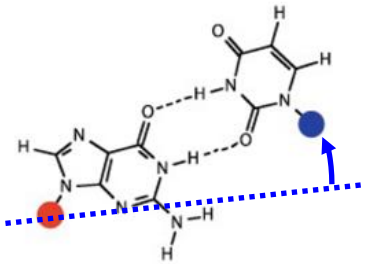
- Base pairs stabilize the RNA secondary structure
- Free energy ≡ number of base pairs

Second approximation:

- Base pairs have different energies
- Free energy: sum of all base pair energies

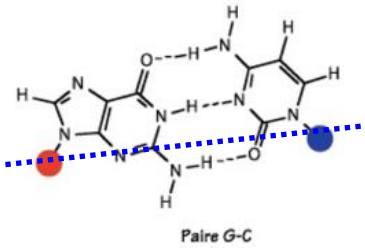
A-U or U-A : 2





G-U or U-G : 1

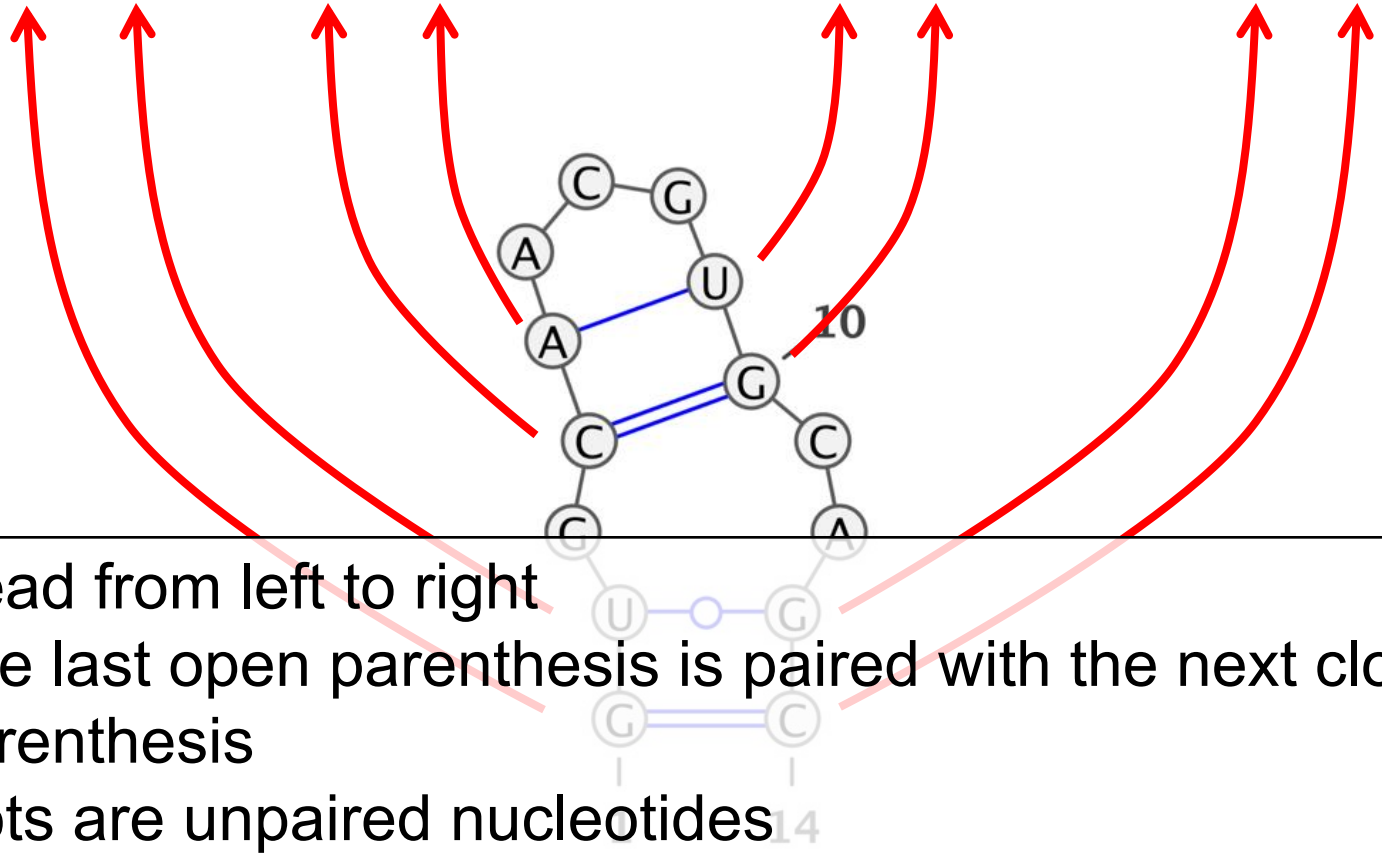
C-G or G-C : 3



Modeling RNA secondary structure

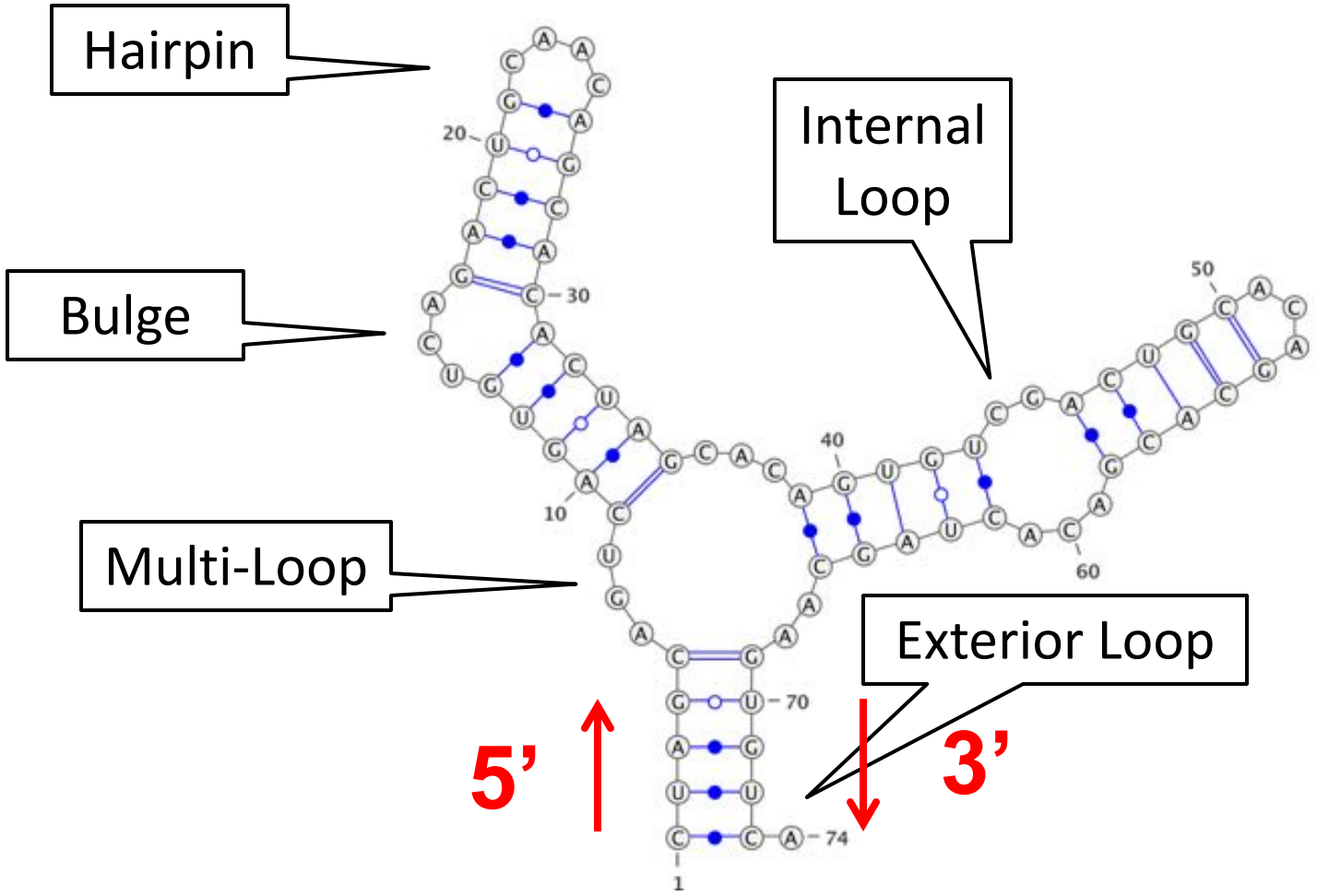
G U G C A A C G U G C A G C

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



- Read from left to right
- The last open parenthesis is paired with the next closed parenthesis
- Dots are unpaired nucleotides

RNA Nomenclature

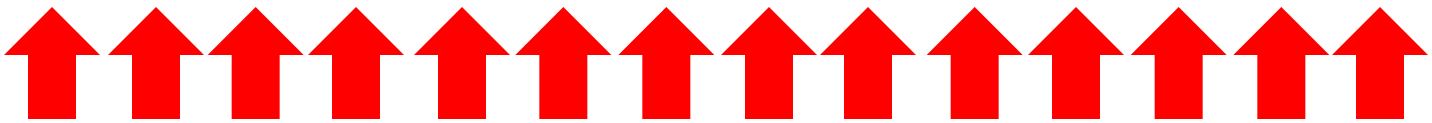


CUAGCAGUCAGUGUCAGACUGCAACAGCACACUAGCACAGUGUCGACUGCACAGCACGACACUAGCAAGUGUCA

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

String to contacts

```
1 2 3 4 5 6 7 8 9 0 1 2 3 4
G U G C A A C G U G C A G C
( ( . ( ( . . . ) ) . . ) )
```



1	2	4	5		

Contacts:
(5,9)
(4,10)
(2,13)
(1,14)

Notes:

- Assume no crossing interactions.
- Each it exist a closing parenthesis for each opening one.
- Base pair opens before closing.

Contacts to String

Principle: Look up at the rightmost position.

- If there is a base pair, print it and print recursively before and between this base pair.
- Otherwise, print unpaired and move one position left.

Contacts: (4,10), (5,8), (1,3).

```
1234567890
xxx(xxxxx)    f(1,10)
xxx(xxxx.)    f(1,3)+f(5,9)
xxx((xx).)    f(1,3)+f(5,8)
xxx((x.).)    f(1,3)+f(6,7)
xxx((..))     f(1,3)+f(6,6)
(x)(.(..))    f(1,3)
(.)(.(..))    f(2,2)
```

RNA secondary structure prediction using dynamic programming

Compute the secondary structure with the maximal number of canonical base pairs (Nussinov-Jacobson, 1980).

$$\delta(i,j) = \begin{cases} 1 & (i,j)\ is\ a\ valid\ base\ pair \\ -\infty & Otherwise \end{cases}$$

Algorithm (Nussinov-Jacobson):

$$M(i,j) = \begin{cases} 0 & if\ i \geq j - \theta \\ M(i,j-1) & No\ base\ pair\ at\ j \\ \max\limits_{i \leq k < j-\theta} \left(\delta(k,j) + M(i,k-1) + M(k+1,j-1) \right) & (k,j)\ is\ a\ base\ pair \end{cases}$$

Example

Ω=GCCAGU, θ=1

	0	1	2	3	4	5	6
0	M	G	C	C	A	G	U
1	G	0	0	1	1	1	2
2	C	–	0	0	0	1	1
3	C	–	–	0	0	1	1
4	A	–	–	–	0	0	1
5	G	–	–	–	–	0	0
6	U	–	–	–	–	–	0

M(1,6)

M(1,3) = max(M(1,2),δ(1,3)+M(2,2)) = max(0,1+0) = 1

M(1,4) = max(M(1,3), δ(1,4)+M(2,3), δ(2,4)+M(1,1)+M(3,3))

= max(1,0+0,0+0+0) = 1

Backtracking

M(1,|ω|) returns the maximal number of base pairs but not the structure.

How do we retrieve the secondary structure?

Backtracking!

Idea: Once we know the value of M(1,|ω|), we can trace the base pairs that were used to obtain it.

Example

Ω=GCCAGU, θ=1

M	G	C	C	A	G	U
G	0	0	1	1	1	2
C	–	0	0	0	1	1
C	–	–	0	0	1	1
A	–	–	–	0	0	1
G	–	–	–	–	0	0
U	–	–	–	–	–	0

$$M(1,6)=2=\begin{cases} M(1,5)=1 \\ \delta(1,6)+M(2,5)=1+1=2 \\ M(1,1)+\delta(2,6)+M(3,5)=0+0+1=1 \\ M(1,2)+\delta(3,6)+M(4,5)=0+0+0=0 \\ M(1,3)+\delta(4,6)+M(5,5)=1+1+0=2 \end{cases}$$

Example (option 1)

Ω=GCCAGU, θ=1

M	G	C	C	A	G	U
G	0	0	1	1	1	2
C	–	0	0	0	1	1
C	–	–	0	0	1	1
A	–	–	–	0	0	1
G	–	–	–	–	0	0
U	–	–	–	–	–	0

(? ? ? ?) Base pairs={(1,6)}

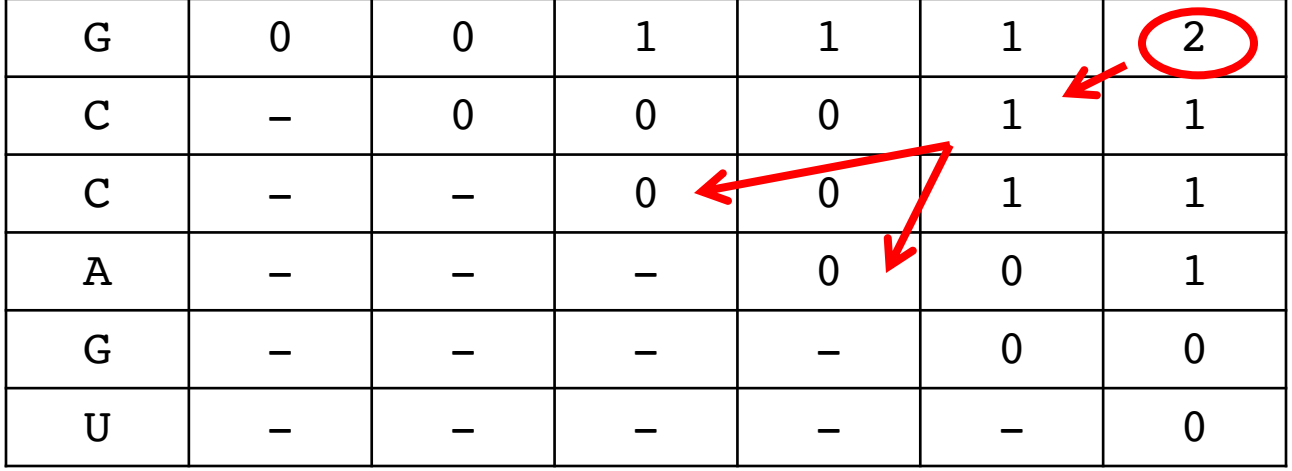
((? ?)) Base pairs={(1,6),(2,5)}

((. .)) Base pairs={(1,6),(2,5)}

Example (option 2)

Ω=GCCAGU, θ=1

M	G	C	C	A	G	U
G	0	0	1	1	1	2
C	–	0	0	0	1	1
C	–	–	0	0	1	1
A	–	–	–	0	0	1
G	–	–	–	–	0	0
U	–	–	–	–	–	0



(? ? ? ?) Base pairs={(1,6)}

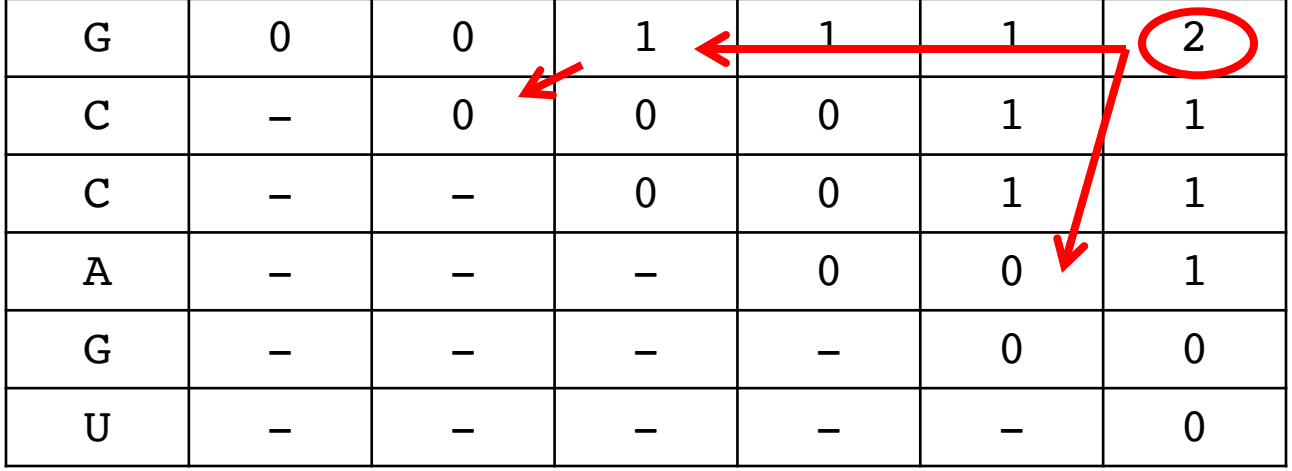
(? (?)) Base pairs={(1,6),(3,5)}

(. (.)) Base pairs={(1,6),(3,5)}

Example (option 3)

Ω=GCCAGU, θ=1

M	G	C	C	A	G	U
G	0	0	1	1	1	2
C	–	0	0	0	1	1
C	–	–	0	0	1	1
A	–	–	–	0	0	1
G	–	–	–	–	0	0
U	–	–	–	–	–	0



`???(?)` Base pairs={(4,6)}

`(?)(?)` Base pairs={(1,3),(4,6)}

`(.)(.)` Base pairs={(1,3),(4,6)}

RNA nearest neighbor energy model

Accuracy of the Nussinov-Jacobson model is moderate. We need a better model to weight the structures.

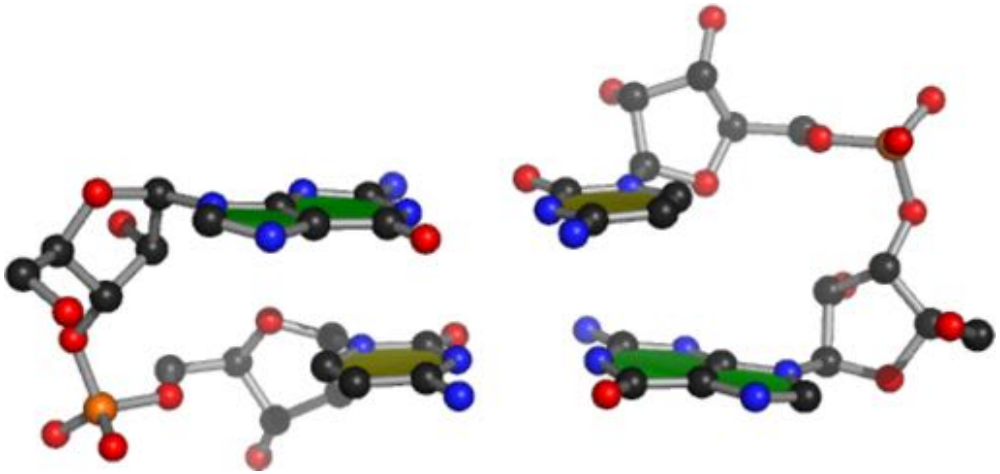
How?: Build an energy model from experimental measures (D. Turner).

But we need:

- *to define what are the important structural features that has to be evaluated.*
- *to keep the energy contribution local in order to allow a divide-and-conquer aproach (fast).*

Stacking base pairs

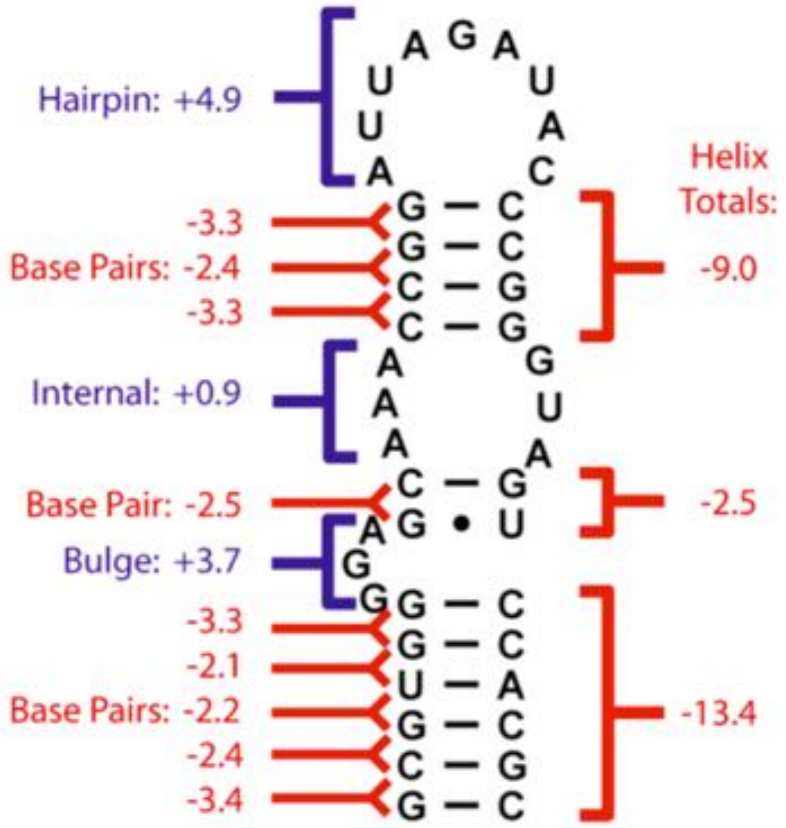
- Base stacking interactions between the pi orbitals of the bases' aromatic rings contribute to stability.
- GC stacking interactions with adjacent bases tend to be more favorable.



Note: Stacking energy are orientated.

$$\begin{matrix} 5' - CG - 3' \\ 3' - GC - 5' \end{matrix} \neq \begin{matrix} 5' - GC - 3' \\ 3' - CG - 5' \end{matrix}$$

Nearest Neighbor Energy Model



http://biomol.bme.utexas.edu

Zuker Algorithm

- Introduced by M. Zuker and P. Stiegler in 1981.
- Calculate the secondary structure with the MFE.
- Adaption of the Nussinov-Jacobson model to the thermodynamical nearest energy model.
- Algorithm originally implemented in the *mfold* software.
- Other popular implementation include:
  - *RNAfold* in the Vienna RNA Package
  - *RNAstructure*
  - *UNAfold* (*mfold* successor)

Want to know more?

Enroll COMP564 “Advanced Computational Biology Methods & Research” !!!

When? Winter 2019

Why? You liked COMP561 and want to know about bioinformatics.

What? We cover fundamental algorithms in computational structural & system biology.

jeromew@cs.mcgill.ca