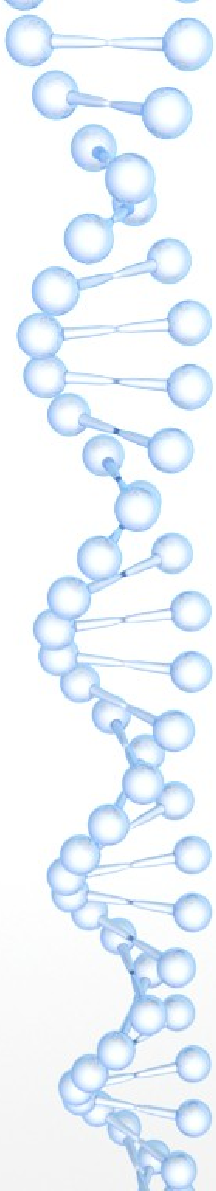
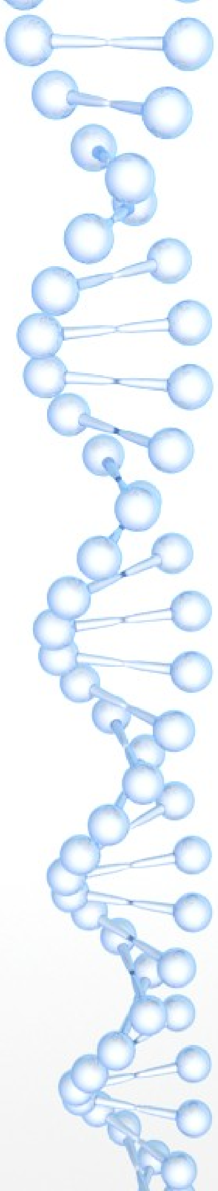


RNA folding problem

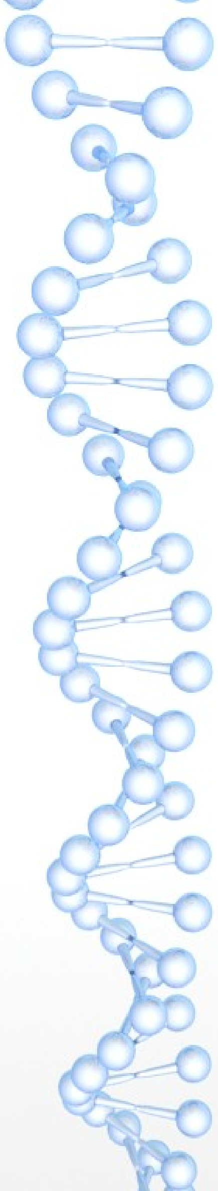
Course: Optimization Models and Algorithms for Data Science

Lorenzo Angelo Giovanni Gini   #457499



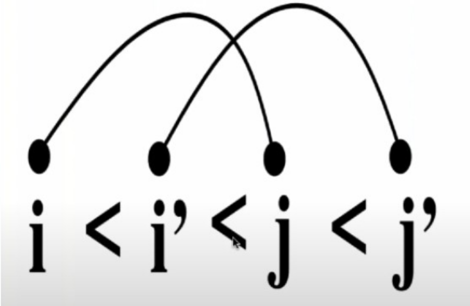
The problem

- The issue is to predict the RNA 2-D folding shape given the sequence of nucleots
- This is a very important problem in biology becouse the structure define the fisic and chemistry proprieties of the RNA molecules
- It was already solved with algorithms using dynamic programing but formulate as a (ILP) problem it’s esier to model (but it’s a bit computational slower than the one wrote in dynamic programming)



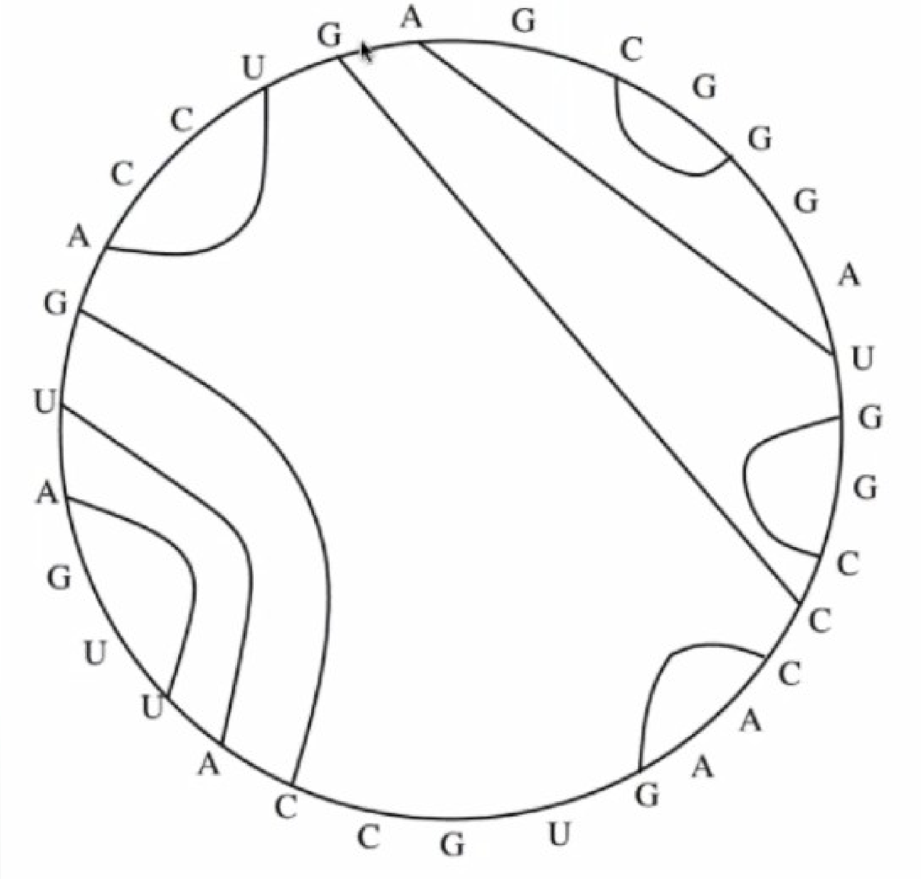
How does it form a RNA molecule?

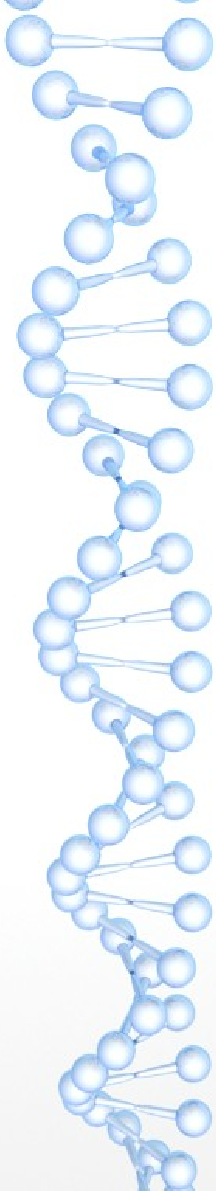
- RNA is composed by 4 different characters colled nucleots: A, C, G, U
- They are divided in 2 complamentary groups: {A, U}, {C, G}
- A nucleot of a complamentary group tend to bound to the other nucleot of its group
- RNA bond do not cross each others



Nested pairs

Two nucleots bond to each other that respect the parameters explained before are colled **nested pair**


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



First model

Colled S the sequence of nucleots; an (ILP) problem need:

- **Variable**: P(i,j) a binary variable that is set to one if and only if the nucleot i is bond to the nucleot j with a nested pair
- **Constraints**: 3 constraints are needed:

1) A nucleot has to bond to at most one other nucleot:

$$\sum_{i>j} P(i,j) + \sum_{i<j} P(i,j) \leq 1 \quad \forall\, j \in S$$

First Model

- **Constraints:**

2) $P(i,j)$ could be one if and only if i and j are complementary

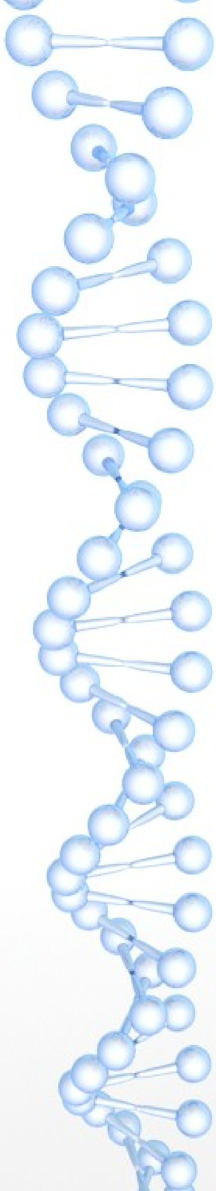
3) Are allowed only nested pairs:

$$P(i,j) + P(k,h) \leq 1 \qquad \forall\, i < k < j < h \in S$$

- **Objective function:**

$$\max \sum_{i<j} P(i,j)$$

that maximaze the bonds so the molecule is more stable



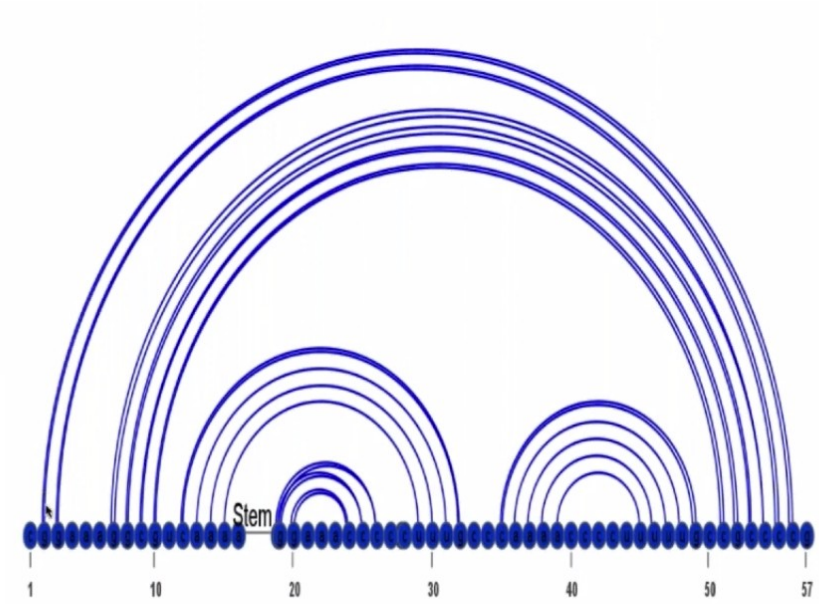
Some adjustments

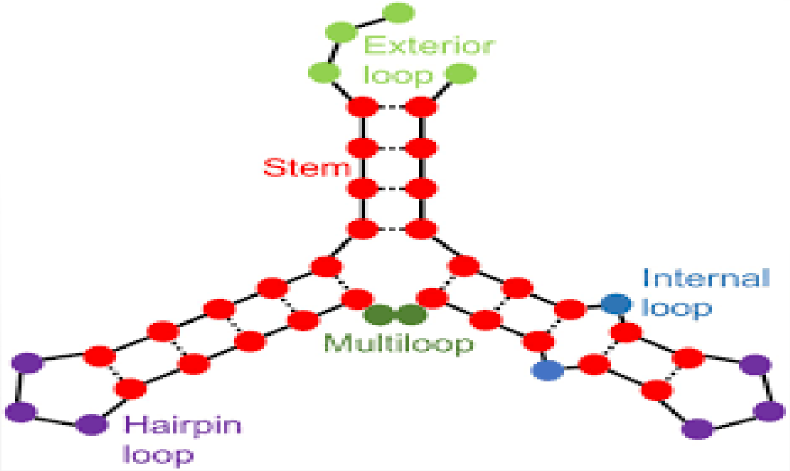
Some adjustments that can be done are:

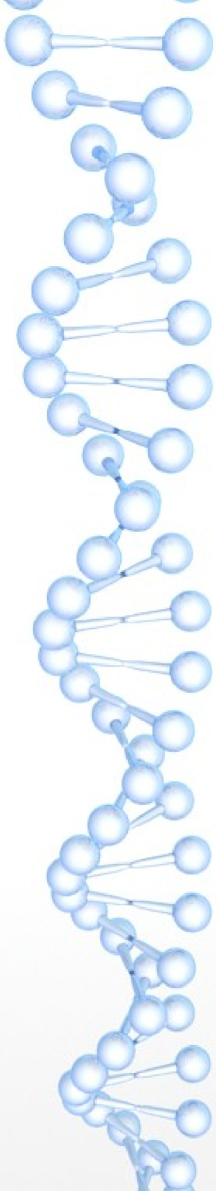
- Set P(i,j) = 0 if i and j are consecutive
- Permitt some bounds between nucleot that are not complementary
- Help the making of consecutive nested pairs colled **stack quartet**. The consecutive stack quartet are colled **stack**

Why the last adjustment?

More consecutive nested pairs, so a stack, form a more stable strucure in the molecule becouse they create a **steam**







Second Model

- **Variables**: P(i,j); Q(i,j). Q(i,j) is a new binary variable that is set to 1 if and only if i and j are part of a stack quarted
- **Weight matrix**: it’s buid a matrix that in position i,j got the weight of the bond between i and j
- **Constraints**:

1) Avoiding bond between cosecutive nucleots:

$$P(i,i+1) = 0 \quad \forall\, i \in S$$

$$P(i,i-1) = 0 \quad \forall\, i \in S$$

Second Model

- **Constraints**:

2) Constraint on Q variable:

$$P(i,j) + P(i+1,j-1) - Q(i,j) \le 1 \quad \forall\, i < j$$

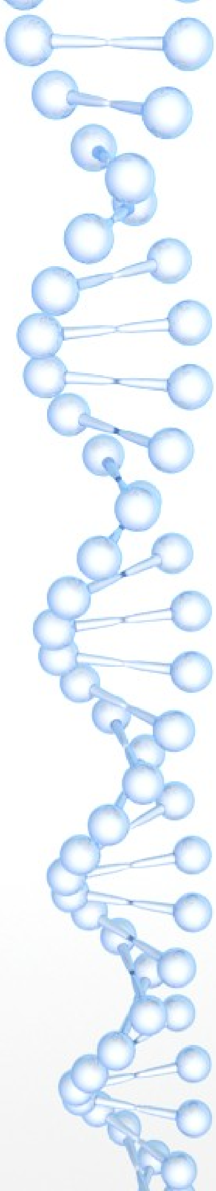
$$2 \ast Q(i,j) - P(i,j) - P(i+1,j-1) \le 0 \quad \forall\, i < j$$

Note: this two constraints are one the opposite implication of the other

- **Objective function**:

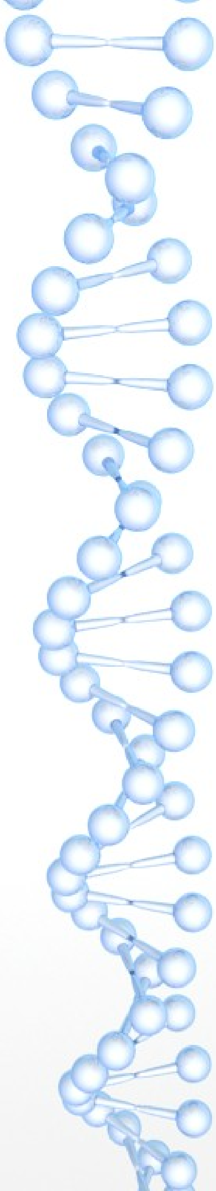
1) $\max \sum_{i<j} [W(i,j) \ast P(i,j) + Q(i,j)]$

2) $\max \sum_{i<j} Q(i,j)$



Extensions

- In the specilized biology libraries there is more or over 200 parameters to set for each different situations
- One common situation is try to create longer stem
- Another changing could be permit some cross pairs for maximazing the number of stack so the lenght of the stem



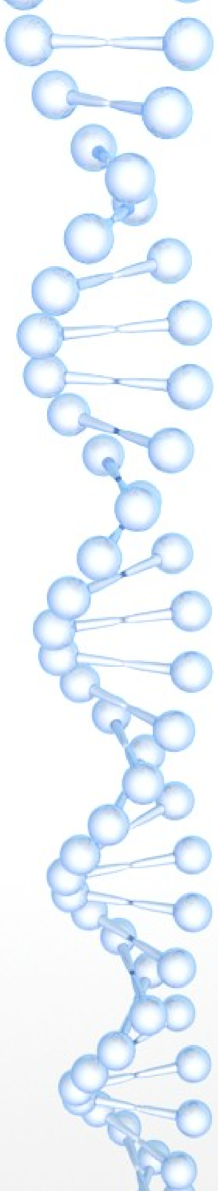
Last Model

This model wont minimizing the number of the stacks

- **Variables**: P(i,j), Q(i,j), F(i,j). Where F(i,j) is a new binary variable that is set to 1 if and only if i,j is the firts nested pair of a stack
- **Constraint**: introducing the new constraints for F(i,j):

$Q(i,j) - Q(i-1,j+1) - F(i,j) \leq 0 \quad \forall\, i < j$

$2 \ast F(i,j) - Q(i,j) + Q(i-1,j+1) \leq 1 \quad \forall\, i < j$



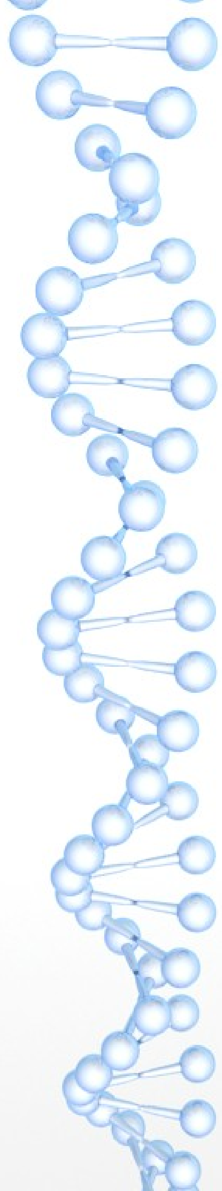
Last Model

- **Objective function:**

1) $\max \sum_{i<j} [W(i,j) \ast P(i,j) + Q(i,j) - F(i,j)]$

2) $\max \sum_{i<j} [Q(i,j) – F(i,j)]$

in this way the the objective function is penalized if there is a lot of short stack insted of a few longer



Thank you for the attention

Reference:

- Chapter 6,Gusfield, D. (2019). *Integer Linear Programming in Computational and Systems Biology: An Entry-Level Text and Course.* Cambridge: Cambridge University Press. doi:10.1017/9781108377737
- On-line surces for the biological facts