Algorithms for RNA structure Prediction

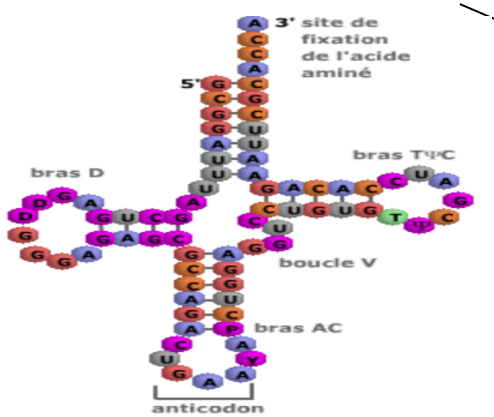
Fariza TAHI

IBISC laboratory
Université d'Evry-Val d'Essonne/Genopole

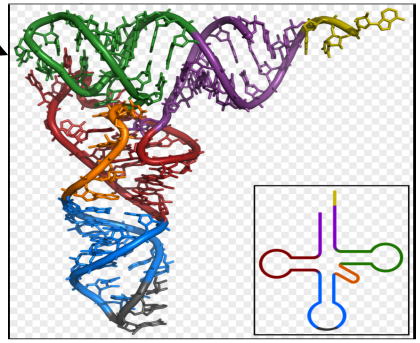
RNA Structures

primary, secondary and tertiary structures

```
...GUCGACUAGCU
AGGCUGGAUGUUAG
GGCUCUCUACACCU
CUAGCGUAGCUAGC
UACAAACUUUAAAA
AAGGGGGCGU...
```

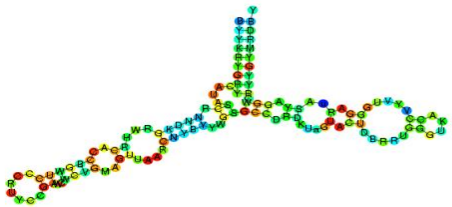


http://tp-svt.pagesperso-orange.fr/arn_fichiers/codants.htm

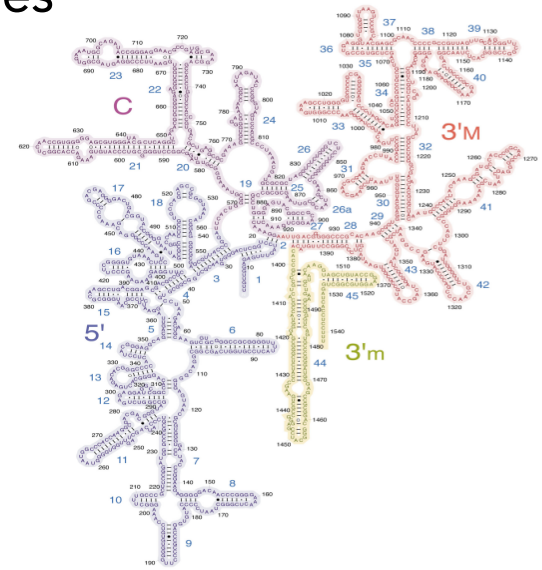


http://fr.wikipedia.org/wiki/Acide_ribonucléique_de_transfert

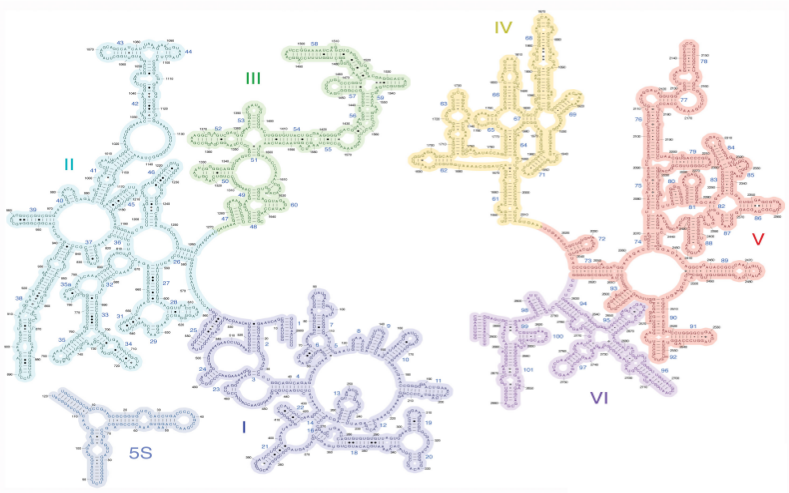
RNA secondary structures



5S rRNA

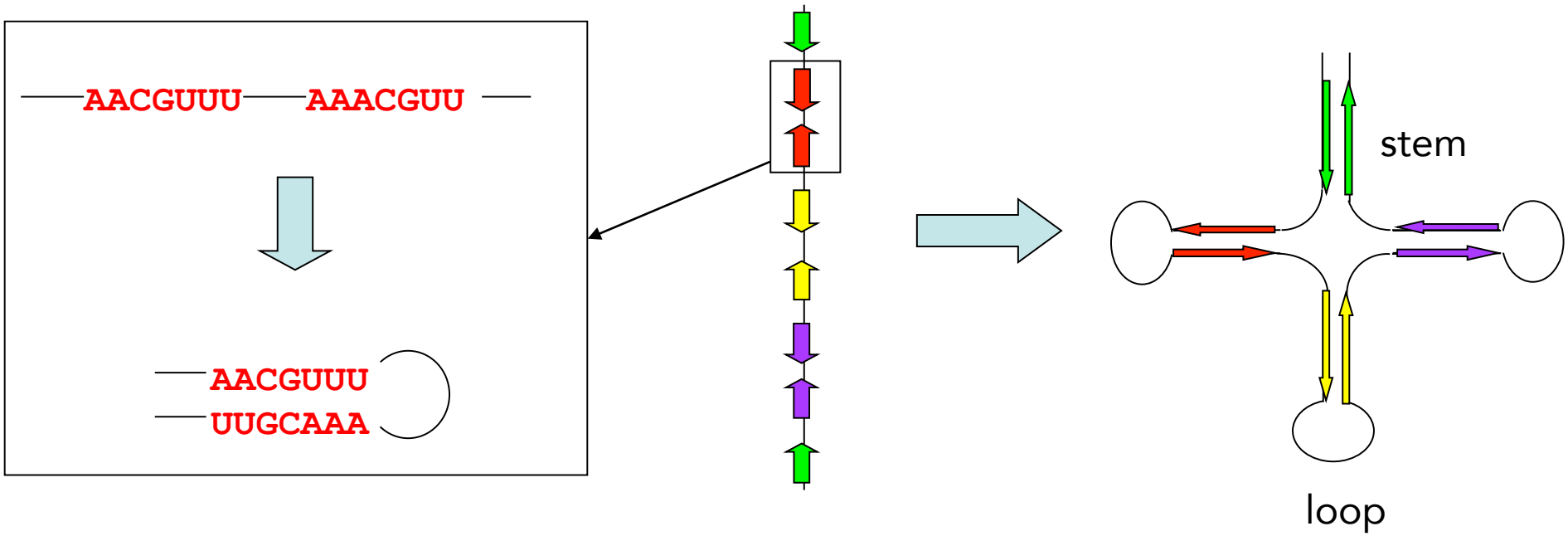


16S rRNA



23S rRNA

RNA secondary structure prediction

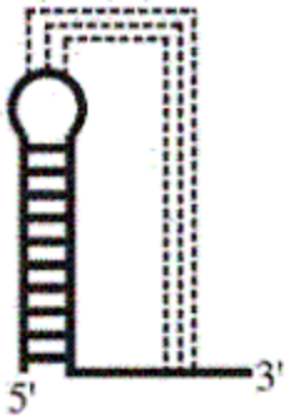


- Two main approaches
  - Thermodynamic approach → Structure of minimal energy
  - Comparative approach → Conserved structure between species

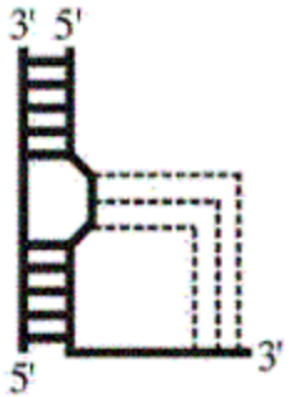
- Several algorithms have been developed
  - High complexities (≥ $O(n^3)$)

Pseudoknots

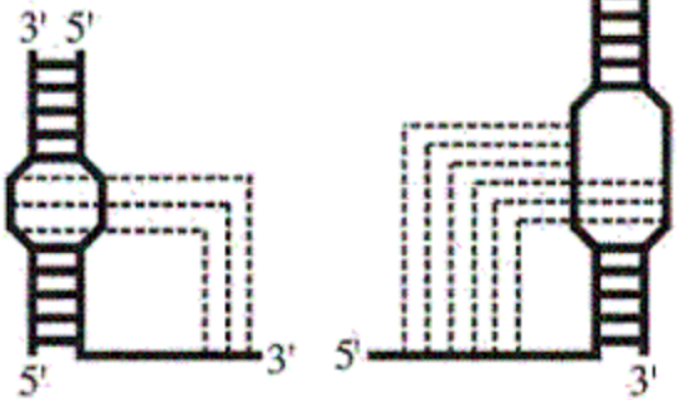
- Particular Motifs of the secondary structure
- Lost of the planar conformation of the secondary structure



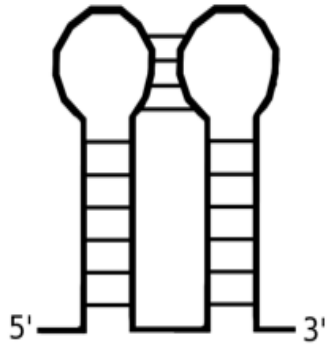
H-type pseudoknot



B-type pseudoknot



I-type pseudoknots



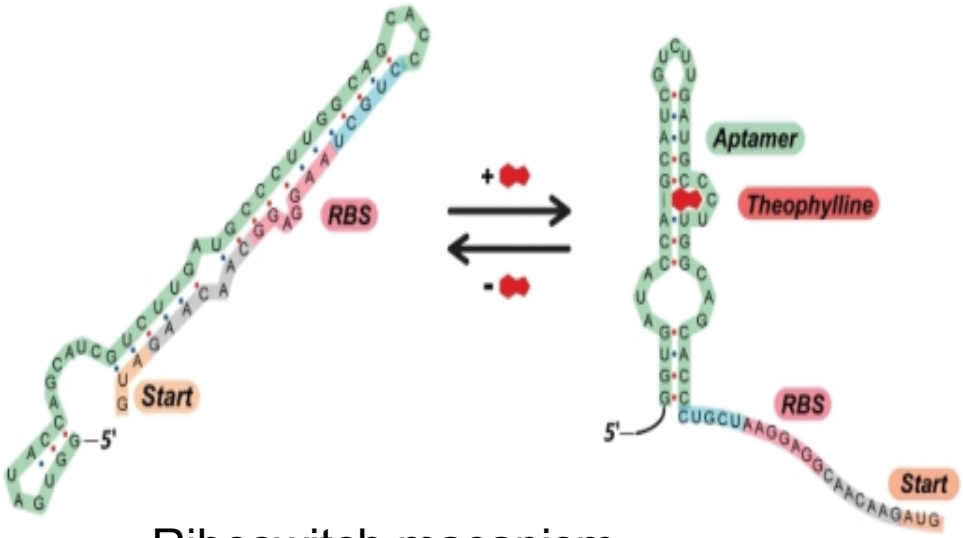
HHH (kissing hairpin)

→ Existing algorithms do not predict all types of pseudoknots

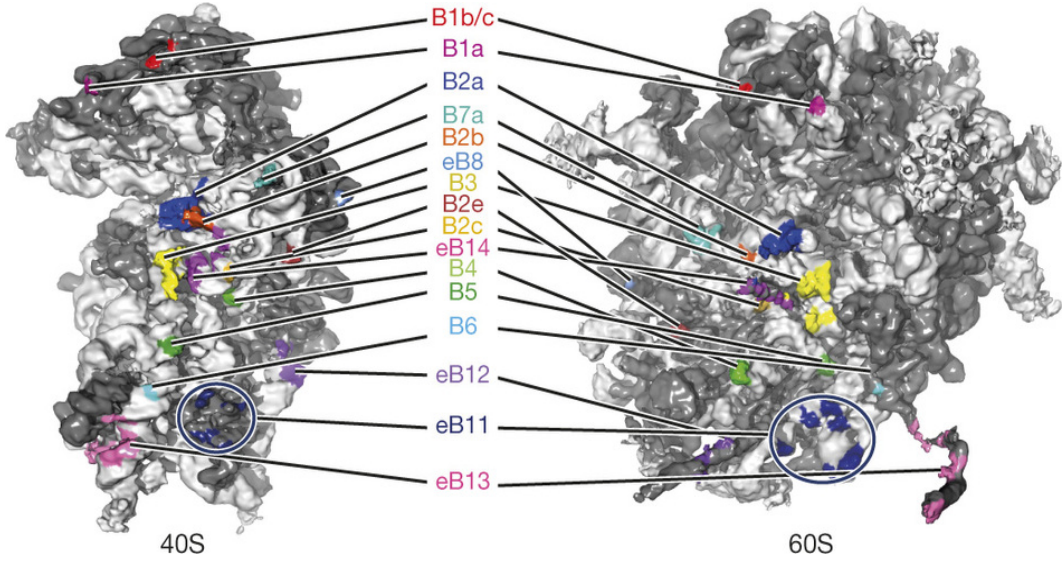
→ High complexities

RNA secondary structure prediction

- A given model can only approach the real structures
- RNA structures are not stable



Riboswitch macanism
(Seeliger and al. Plos one, 2012)



Structure of the human 80S ribosome
(Khatter et al., Nature 2015)

➡ Prediction of several solutions

➡ Combination of models

Algorithms based on Comparative approach

Phd thesis of Stéfan Engelen

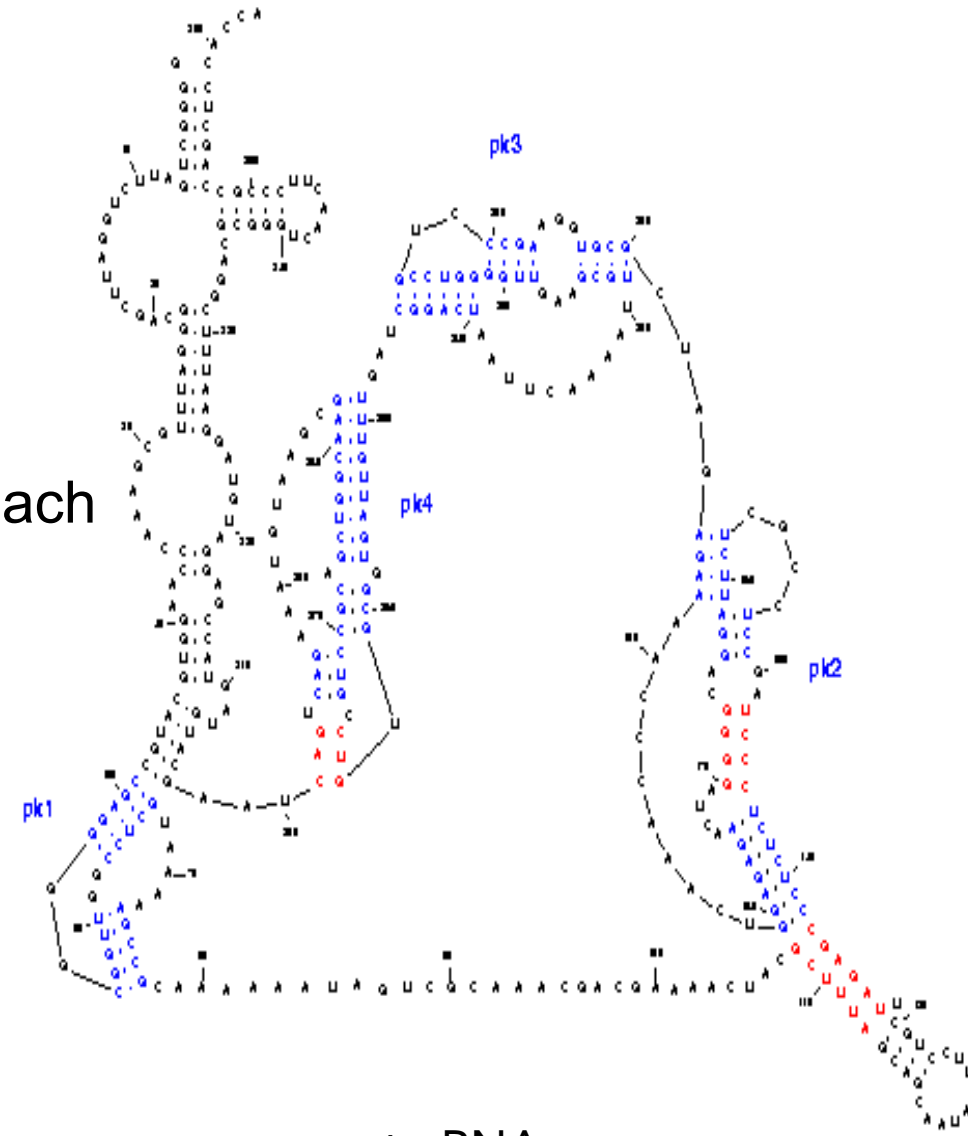
Tfold algorithm

Approach

- Comparative approach
- Thermodynamic criteria
- Helices instead of pairings
- “Divide and conquer” algorithmic approach
- Returns several alternative structures
- Predicts all types of pseudoknots

→ Time complexity of $O(n^2)$

(Engelen and Tahi, NAR 2010)



tmRNA

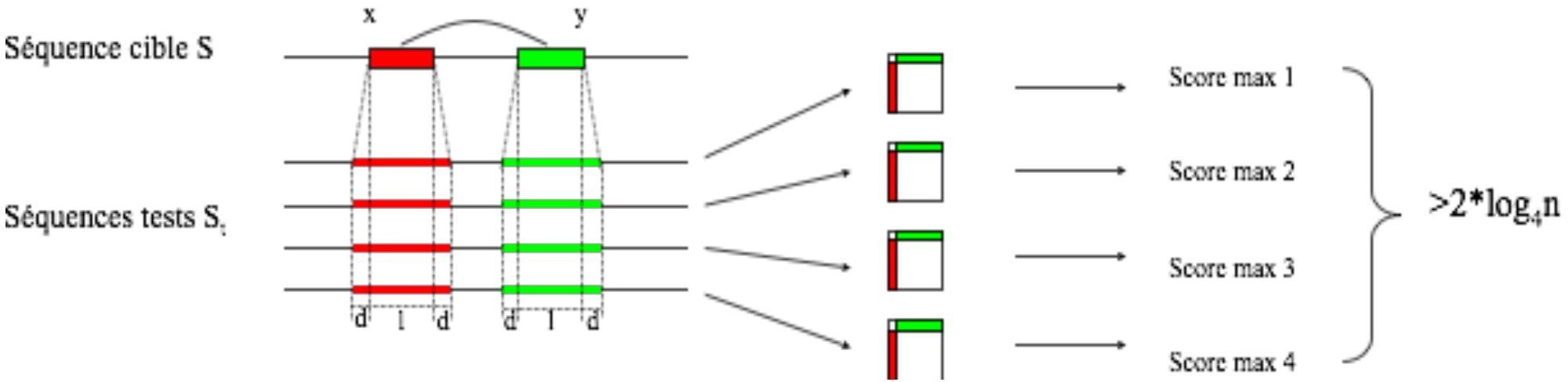
Tfold algorithm: search for helices

- Search for helices on the target sequences
- Verify their conservation in the homologous sequences

	1	2	3	4	5	6	7	8	9	10	11	12	13
	C	G	G	C	A	U	U	C	G	G	C	C	U
13 U	0	2	2	0	2	0	0	0	**2**	2	0	0	0
12 C	0	3	5	0	0	0	0	0	3	5	0	0	
11 C	0	3	**6**	0	0	0	0	0	3	**6**	0		
10 G	3	0	0	**9**	0	1	1	3	0	0			
9 G	3	0	0	3	0	1	3	**4**	0				
8 C	0	**6**	3	0	0	0	0	0					
7 U	0	2	**8**	**0**	2	0	0						
6 U	0	2	2	0	2	0							
5 A	0	0	**0**	0	0								
4 C	0	3	3	0									
3 G	3	0	0										
2 G	3	0											
1 C	0												

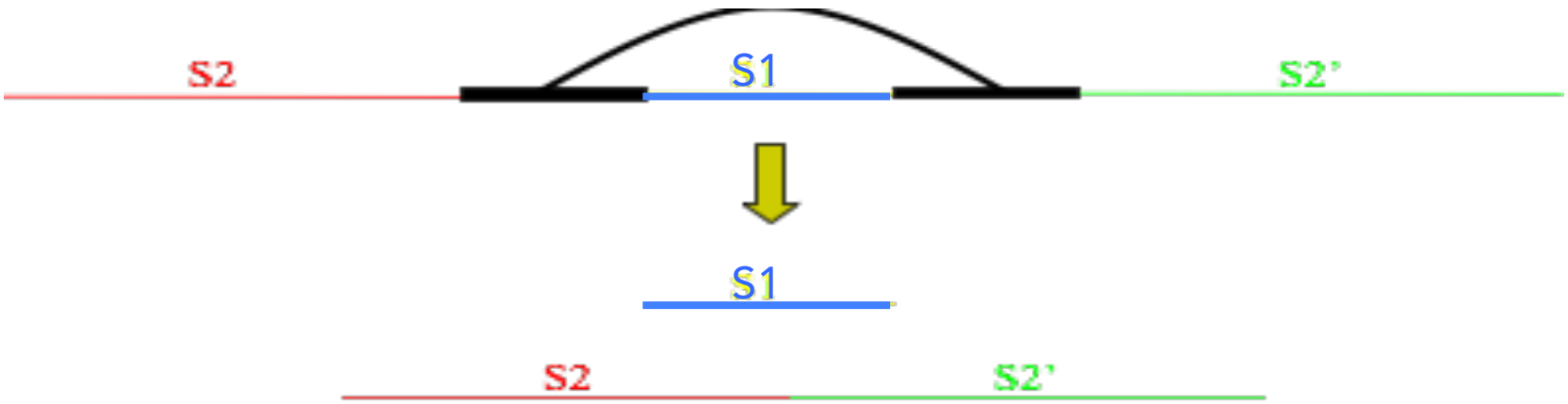
(2,12,3)
(1,9,3)
(8,13,2)

```
Case(9,13):B5'GU⇨0+2=2
Case(8,9):C↔G⇨1+3=4
Case(3,7):R5'GU⇨6+2=8
Case(4,7):C↔U⇨0
Case(3,5):H5'GA⇨Case(2,6)=1+1
```



→ Time complexity of $O(n^2)$

Tfold algorithm: Divide and conquer approach



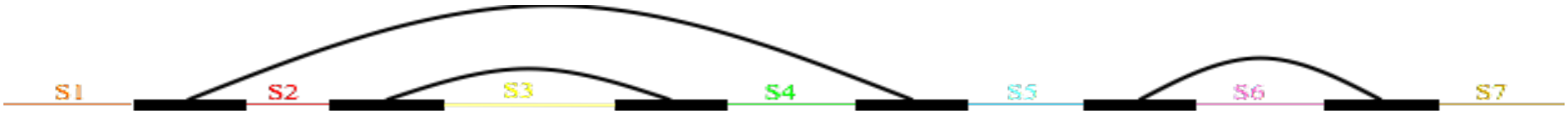
- Nested or disjoint helices

- If there are not compatible helices with same score

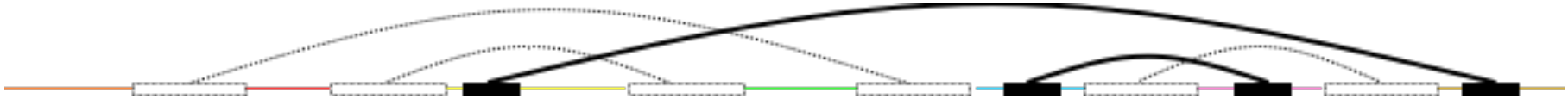
  → alternative structures

Tfold algorithm: Prediction of pseudoknots

- Let a sequence S, we find a list L1 of compatible helices



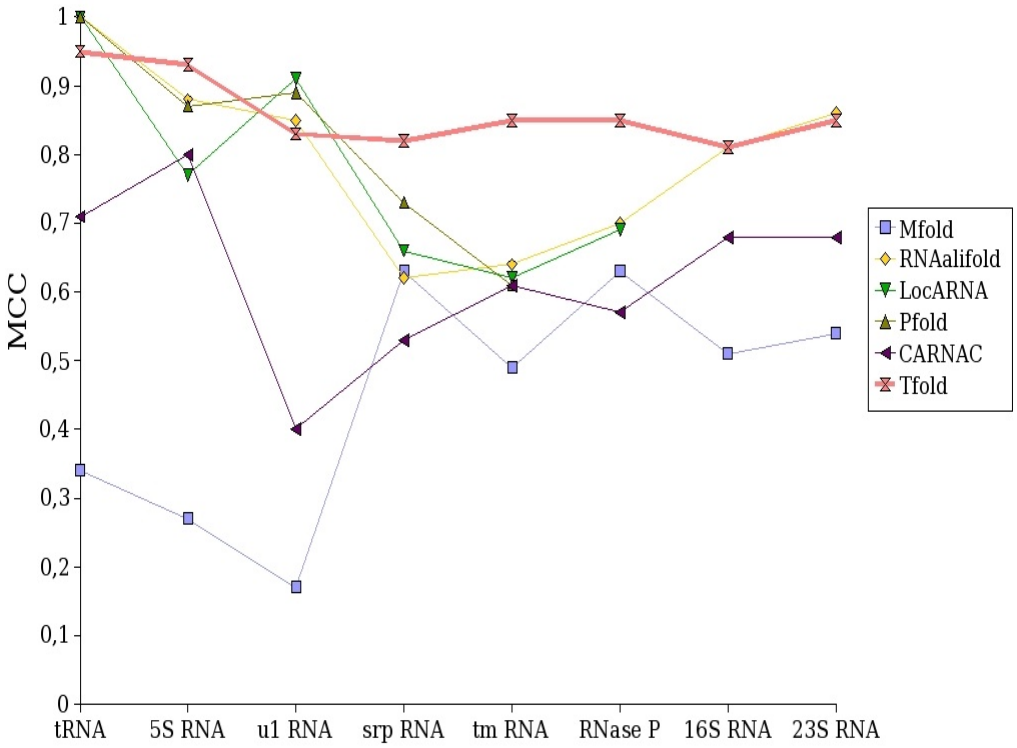
- The algorithm is applied on S without L1 helices, allowing to find anot her list L2 of compatible helices

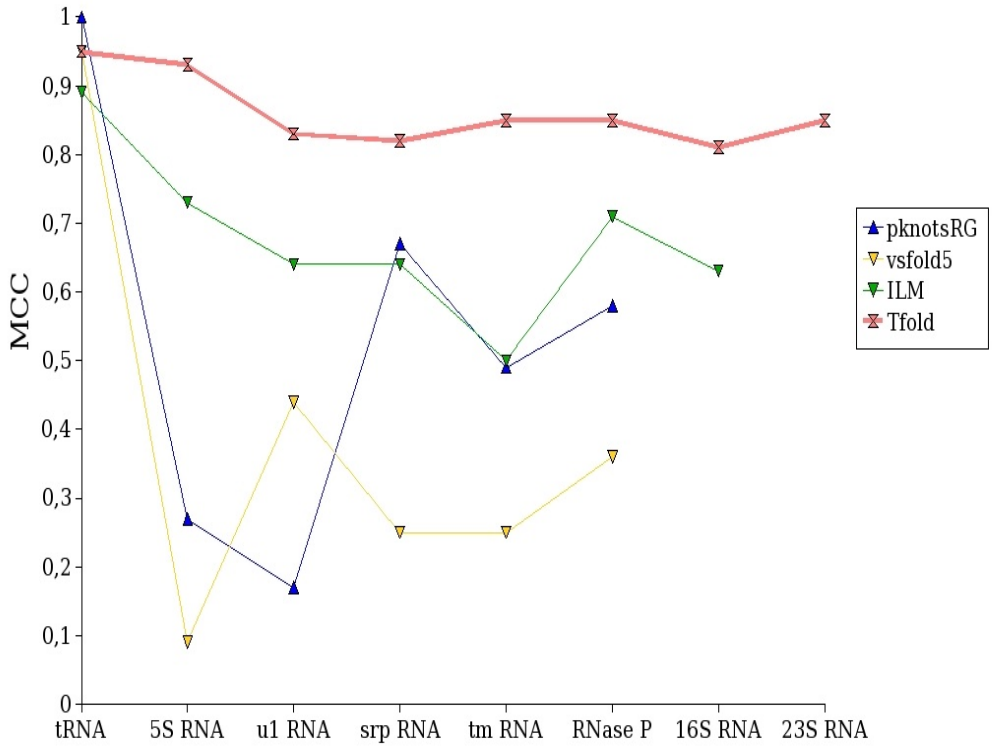


- Each helix of L2 forms a pseudoknot with a helix of L1

→ Prediction of all types of pseudoknots

Tfold results





MCC : coefficient of correlation between the predicted structure and the structure of reference

→ Sensibility, selectivity and MCC ≥ 80%

→ Homogeneous results for any considered RNA

Algorithms based on bi-objective integer programming approach

Phd thesis of Audrey Legendre

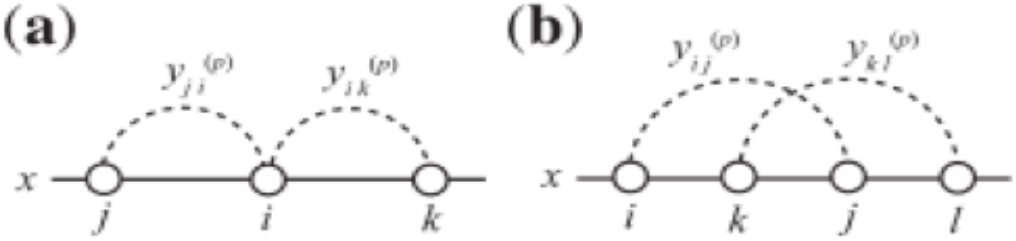
Co-supervision with Eric Angel

Biokop algorithm

- Combining two (thermodynamic approach) models
  - Maximum Expected Accuracy (MEA)
  - Minimum Free Energy (MFE)

  → bi-criteria optimization problem

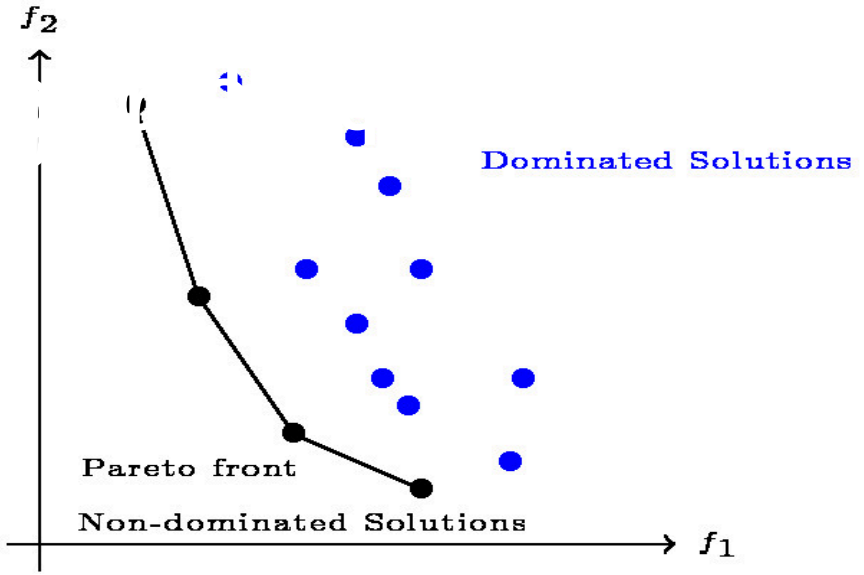
- Bi-objective integer programming approach
  - Objective function to minimize or maximize (MEA and MFE)
  - Subject to Constraints



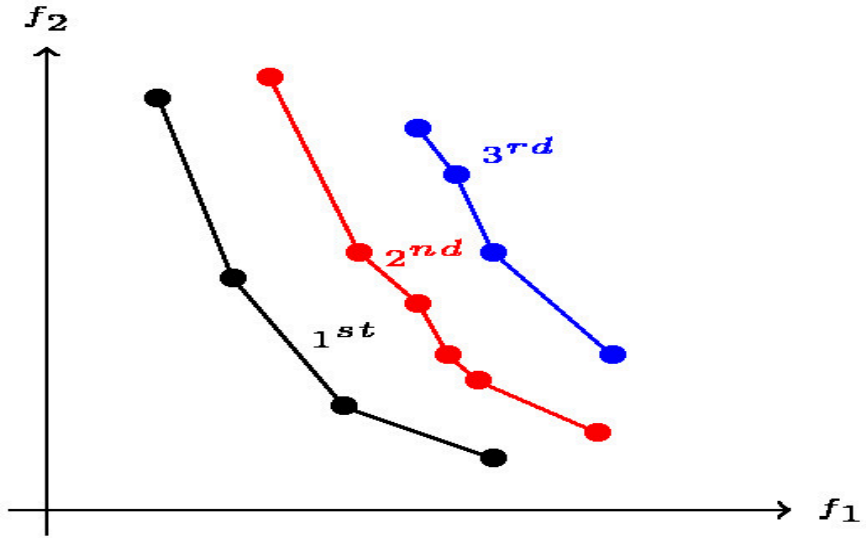
→ flexible
→ allow to model all kinds of pseudoknots

Biokop algorithm

- Bi-objective integer program (BOIP)
- Pareto set (set of solutions)



- Generic method to find the k-best Pareto sets for any BOIP
- Application to prediction of secondary structure of RNA with pseudoknots



→ Biokop Software

(Legendre et al., BMC Bioinfo, in Revision)

Biokop algorithm: Results

- 198 RNAs (21 to 128 nt) from Pseudobase++ (Taufer et al., 2009)
- Prediction of 30 solutions per RNA

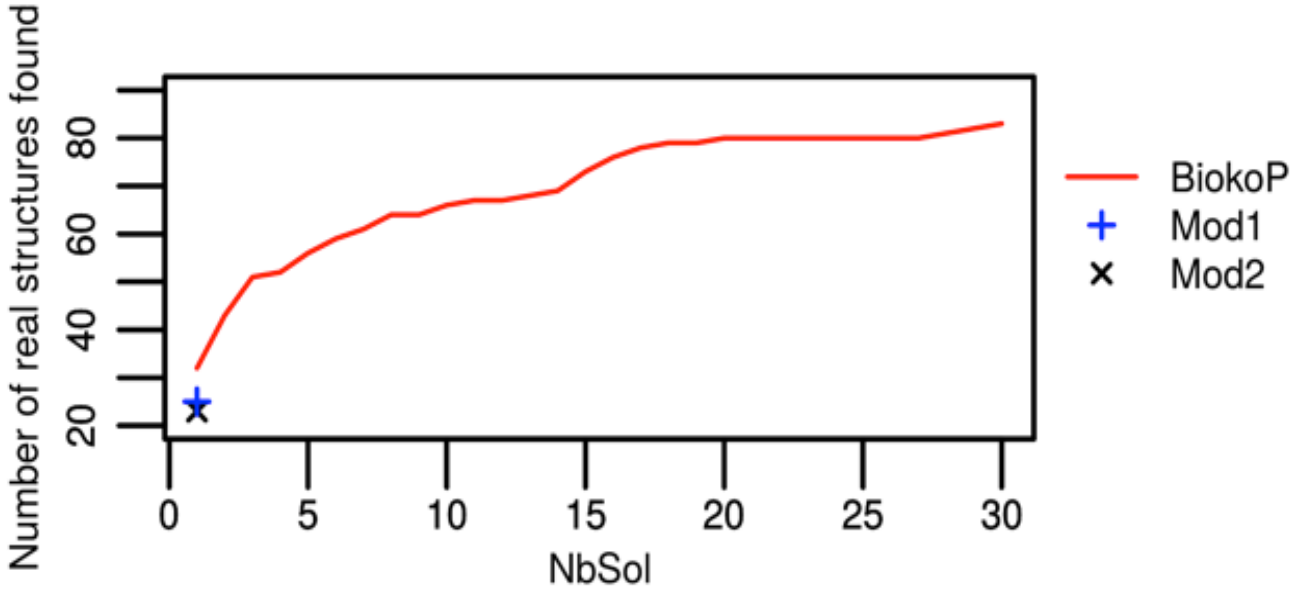
k-best Pareto set, $k =$	1	2	3	4	5	6	7	8	9	10	Total
Number of real solutions	45	15	13	7	1	1	0	0	1	1	84

- Mod1: MFE

Based on (Poolsap et al., 2009)

- Mod2: MEA

Based on (Sato et al., 2011)

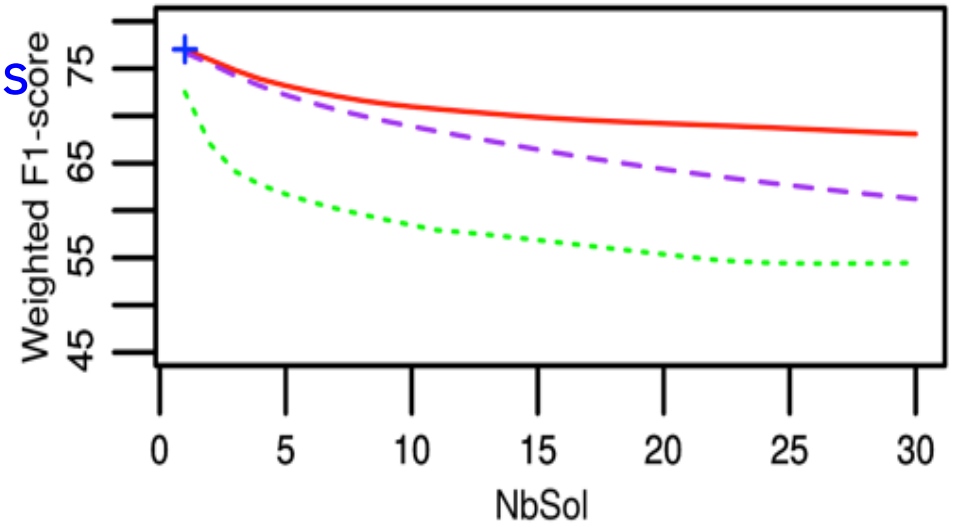


All solutions of Mod1 and Mod2 are returned by BiokoP as optimal solutions

Biokop algorithm: Results

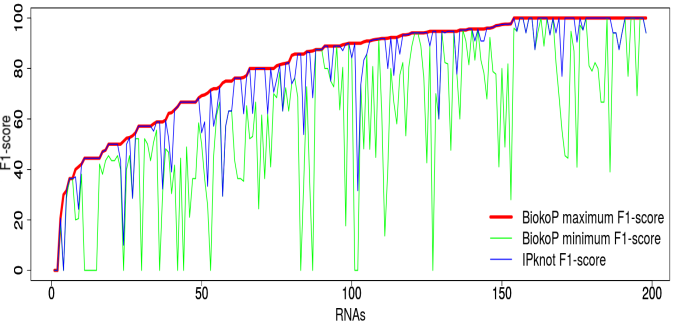
Weighted mean F1-scores

- pKiss
  (Janssen and Giegerich, 2014)
- McGenus (Bon et al., 2012)
- IPknot: (Sato et al., 2011)

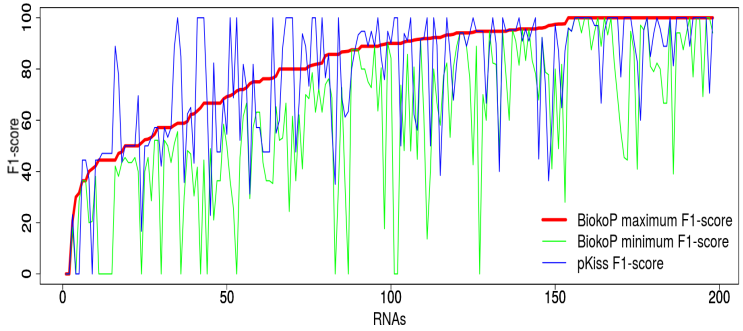


- BiokoP: several optimal solutions
- Other tools: one optimal solution

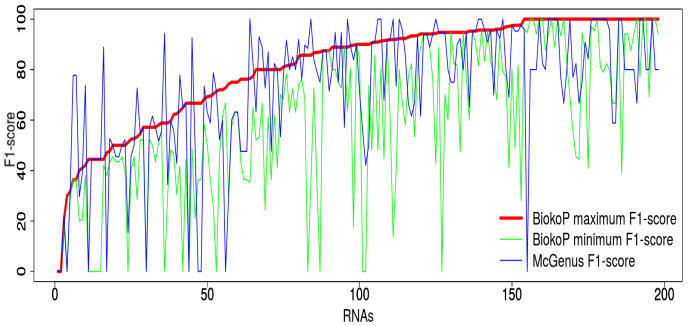
F1-scores of optimal solutions for each RNA



With IPknot



With pKiss



With McGenus

(http://EvryRNA.ibisc.univ-evry.fr)

Home
Publications
About us
SSOMncRNA
BiokoP
IpiRId
piRPred
miRBoost
ncRNAclassifier
miRNAFold
Tfold
SSCA
P-DCFold
RNA-SC
miRBoost+Plant

EvryRNA

Welcome to EvryRNA Platform

Contact:
fariza.tahi@ibisc.univ-evry.fr

EvryRNA platform is a web server providing various algorithms and bioinformatics tools developed in the laboratory IBISC of UEVE/Genopole, and dedicated to the prediction and the analysis of non-coding RNAs (ncRNAs). These RNAs are regulators of gene expression control and genome stability. They are involved in different biological processes, and some of them, including microRNAs, are known to be involved in many diseases such as cancer and neurodegenerative diseases. Their study provides insight into how living organisms function, including differentiation and cell proliferation, but also to consider new therapeutic approaches for genetic diseases and cancer.

EvryRNA includes many bioinformatics software for RNA secondary structure prediction and identification and prediction of ncRNAs including small RNAs (microRNAs, piARNs, etc.) in large-scale genomic sequences. These software correspond to different algorithms based on pattern matching and sequence algorithmic approaches, and machine learning approaches. They have the particularity, compared to the state of art, speed, enabling large-scale analyses, in addition to the effectiveness of predictions.

Home
Publications
About us

BiokoP
Home
Prediction

IpiRId
piRPred
miRBoost
ncRNAclassifier
miRNAFold
Tfold

BiokoP - Bi-objective programming pseudoknot Prediction

PREDICTION

Upload your own Fasta file or submit an RNA sequence directly:

Allowed nucleotides are A, U, T, G, C, a, u, t, g, c.

Size limitation: 130 nucleotides.

To test BiokoP with the example sequence, just press the Predict button.

FASTA FILE: Browse... No file selected.

SEQUENCE: AAGCCUUUUGGAUCGAAC

NUMBER OF PARETO SETS: 1

Greater the number of Pareto sets is, greater the number of predicted structures is.

FORNA VIEWER: Yes

Allows Forna Viewer secondary structure visualization tool.

IBISC EvryRNA Tanuki Login

PREDICTION

EvryRNA

BiokoP
Home
Prediction

IpiRId
piRPred
miRBoost
ncRNAclassifier
miRNAFold
Tfold
SSCA
P-DCFold

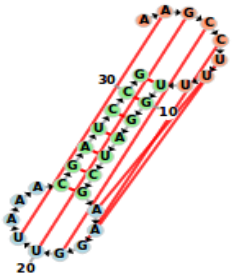
RNA sequence :

- **Name** : No name
- **Sequence** : AAGCCUUUUGGAUCGAAGGUUAAACGAUCCG

Download Result File

Optimal solutions:

.((((((.[[[[[[)))))))...]]]]]] Energy objective: -20.90 Probability objective: -0.00

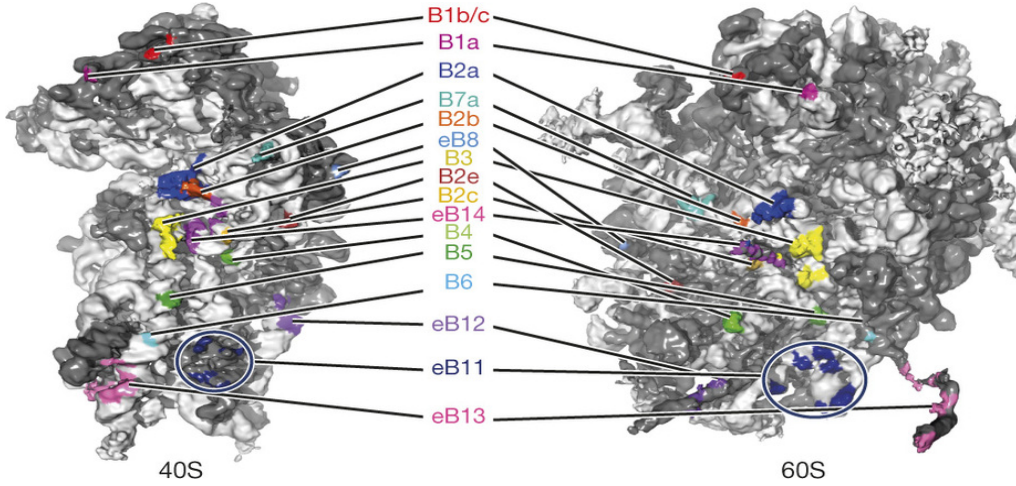


.((((((..[[[[[[)))))))...]]]]]]. Energy objective: -18.50 Probability objective: -5.58

Projets en cours

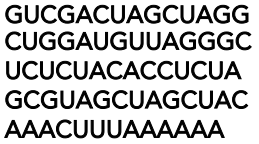
Outil interactif pour la Prédiction de structures “quaternaires” d’ARN

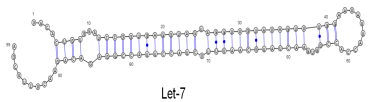
- Thèse (Audrey Legendre, 2016)
- Collaborations :
  - E. Angel (IBISC)
  - O. Namy (I2BC)

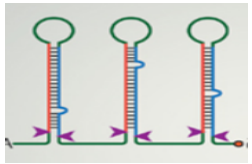


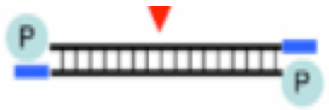
Prédiction et Classification des ARNncs par des approches holistiques

- Thèse (Ludovic Platon, octobre 2015)
- Collaborations :
  - F. Zehraoui (IBISC)
  - A. Bendahmane (IPS2)









Remerciements

Doctorants

Audrey Legendre
Anouar Boucheham
Ludovic Platon

Stéfan Engelen

Post-doctorants

Sébastien Tempel
Van Du Tran

Stagiaires

Christophe Tav
Anis Demigha
Vivien Sommard
Jocelyn Brayet
Mederich Besnard
Mikael Trellet
Gabriel Chandesris
…

IBISC
Eric Angel
Laurent Poligny
Farida Zehraoui

ISSB
Nicolas Pollet
Mohamed Elati

Institut Curie
François Radvenyi

Généthon
David Israeli
Laurence Jaenson-Leh

/IPS2
Abdelhafid Bendahmane
Adnane Boualem
Martin Crespi

I2BC
Olivier Namy

INRA
Eric Barrey

Université de Constantine 2
Mohamed Chawki Batouche

THANK YOU

fariza.tahi@univ-evry.fr

http://EvryRNA.ibisc.univ-evry.fr

Mesures

- TP, True Positive : nombre d'appariements correctement prédits.
- FP, False Positive : nombre d'appariements présents dans la prédiction mais pas dans la structure de référence.
- FN, False Negative : nombre d'appariements non prédits.
- TN, True Negative : $\frac{n(n-1)}{2} - TP - FN - FP$.

Sensitivité

$$\text{Sensitivité} = \frac{TP}{TP+FN}$$

Spécificité

$$\text{Spécificité} = \frac{TP}{TP+FP}$$

Matthews Correlation Coefficient MCC

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

PPV

$$\text{PPV} = \frac{TP}{TP+FP}$$

F_1-score

$$F_1\text{-score} = 2 \times \frac{Sensitivity \times PPV}{Sensitivity + PPV}$$