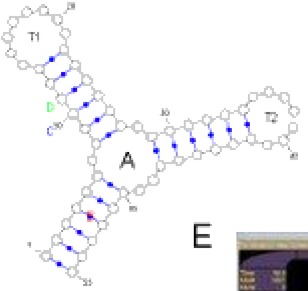
Algorithms for analysing and predicting RNA 3D structures

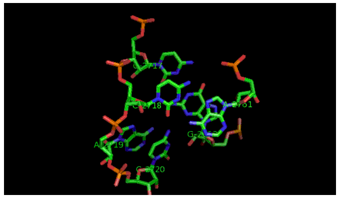
Alain Denise
LRI and I2BC
Université Paris-Sud – CNRS – Université Paris-Saclay

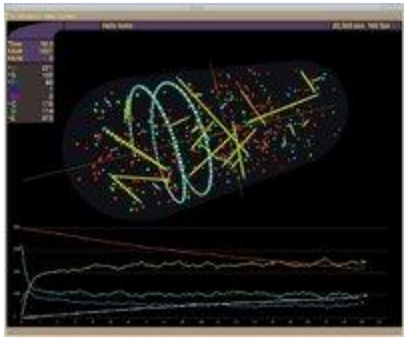
Bioinfo team at LRI / Paris-Sud

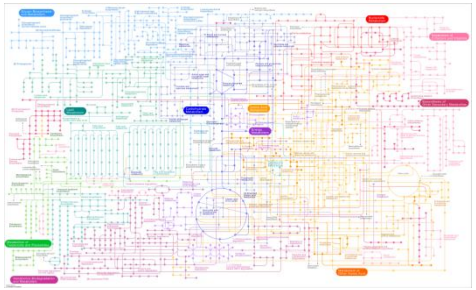
Main themes :

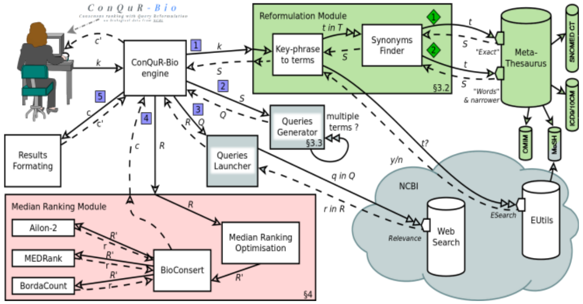
- RNA structural bioinformatics
- Computational systems biology
- Biological data integration
- Evolution

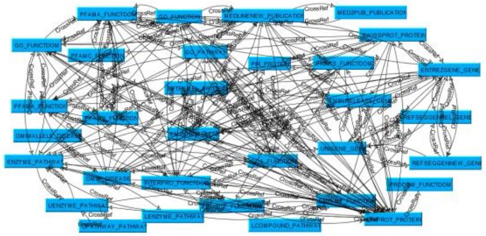






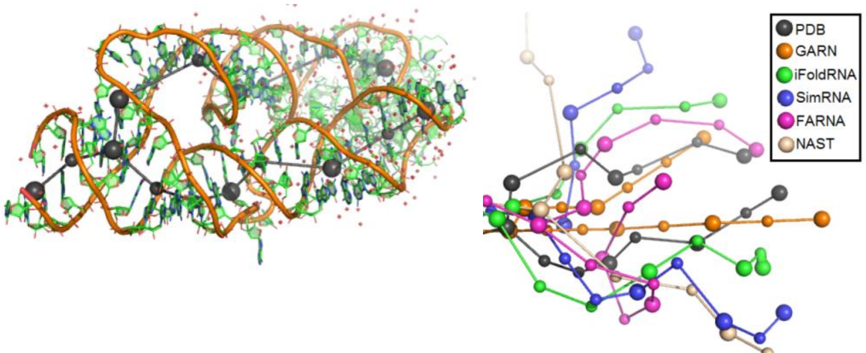




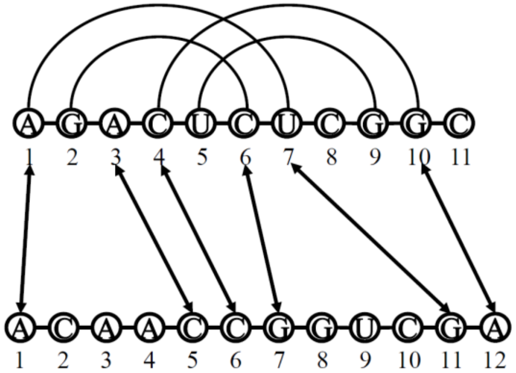


Computer science issues: Algorithmics and combinatorics, Database integration, Machine learning, Simulation

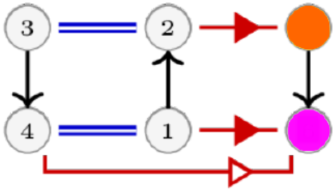
RNA structural bioinformatics

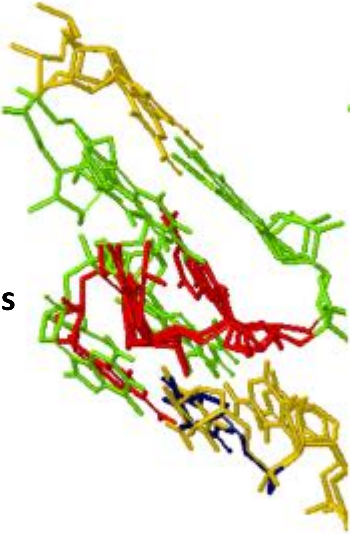


Game theory for coarse-grained 3D structure prediction
[Boudard et al. PloS one 2015, Bioinformatics 2017]
http://garn.lri.fr



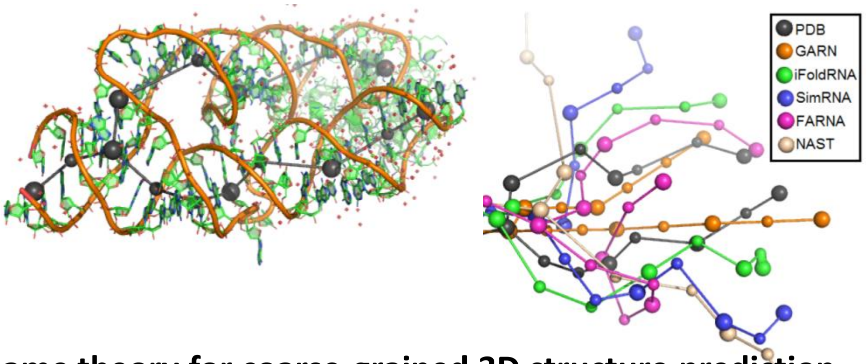
Structure- sequence alignment including pseudoknots
[Rinaudo et al. WABI 2012]
[Wei WANG's thesis Dec. 2017, with Y. Ponty]
https://licorna.lri.fr/



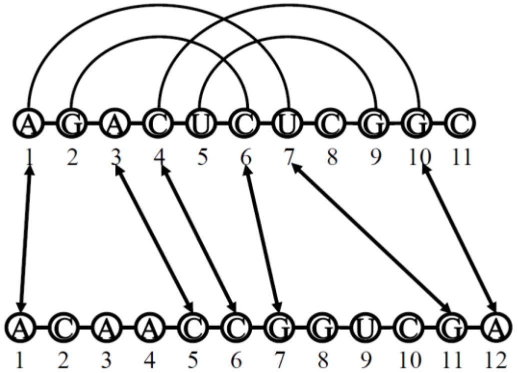


Mining for recurrent motifs in RNA structures
[Djelloul et al. RNA 2008]
[Reinharz et al., submitted]

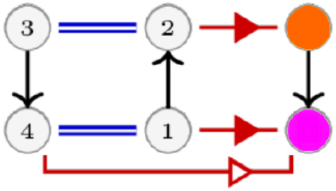
RNA structural bioinformatics

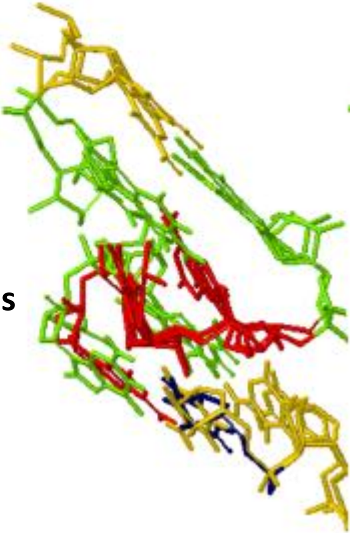


Game theory for coarse-grained 3D structure prediction
[Boudard et al. PloS one 2015, Bioinformatics 2017]
http://garn.lri.fr



Structure- sequence alignment including pseudoknots
[Rinaudo et al. WABI 2012]
[Wei WANG's thesis Dec. 2017]
https://licorna.lri.fr/





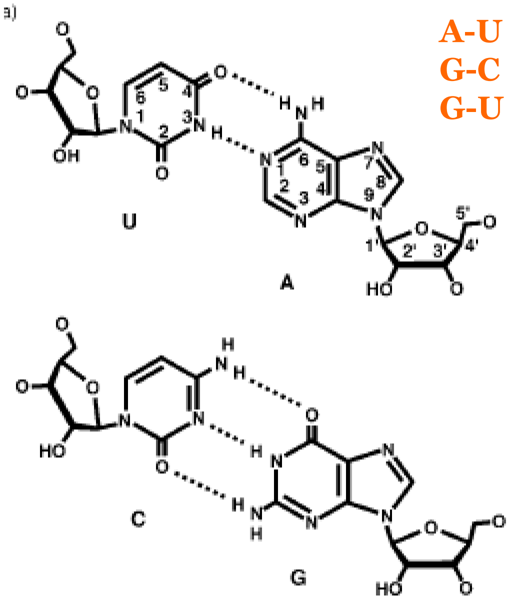
Mining for recurrent motifs in RNA structures
[Djelloul et al. RNA 2008]
[Reinharz et al., submitted]

RNA structure: canonical interactions

which folds in type A helices

WC canonical basepairs

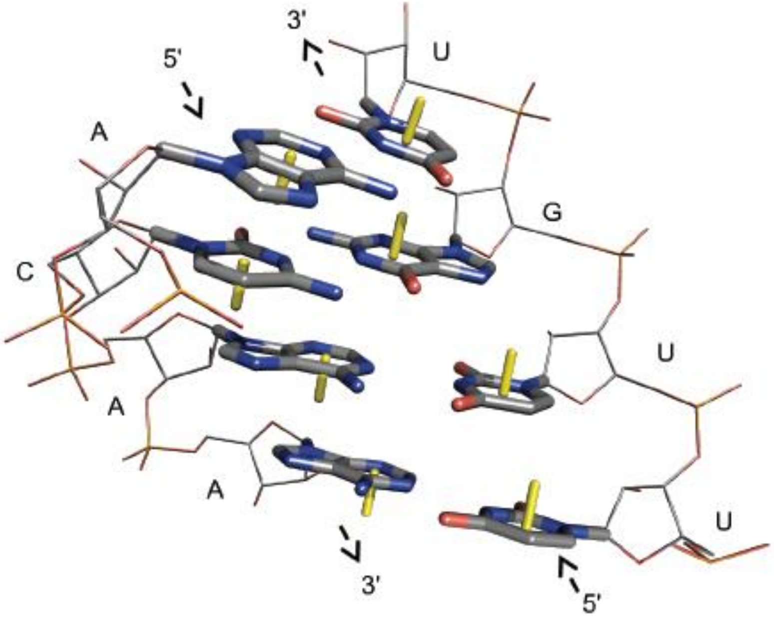
A-U
G-C
G-U



[Tertiary motifs in RNA structure and folding, J. Doudna et al., Angew Chem Int Ed Engl. 1999]

ET

Stacking



[Base stacking annotation, F. Major & P. Thibault, presented at the RNA ontology consortium workshop, RNA society meeting, Seattle WA, June 19-20 2006]

RNA structure: non canonical interactions

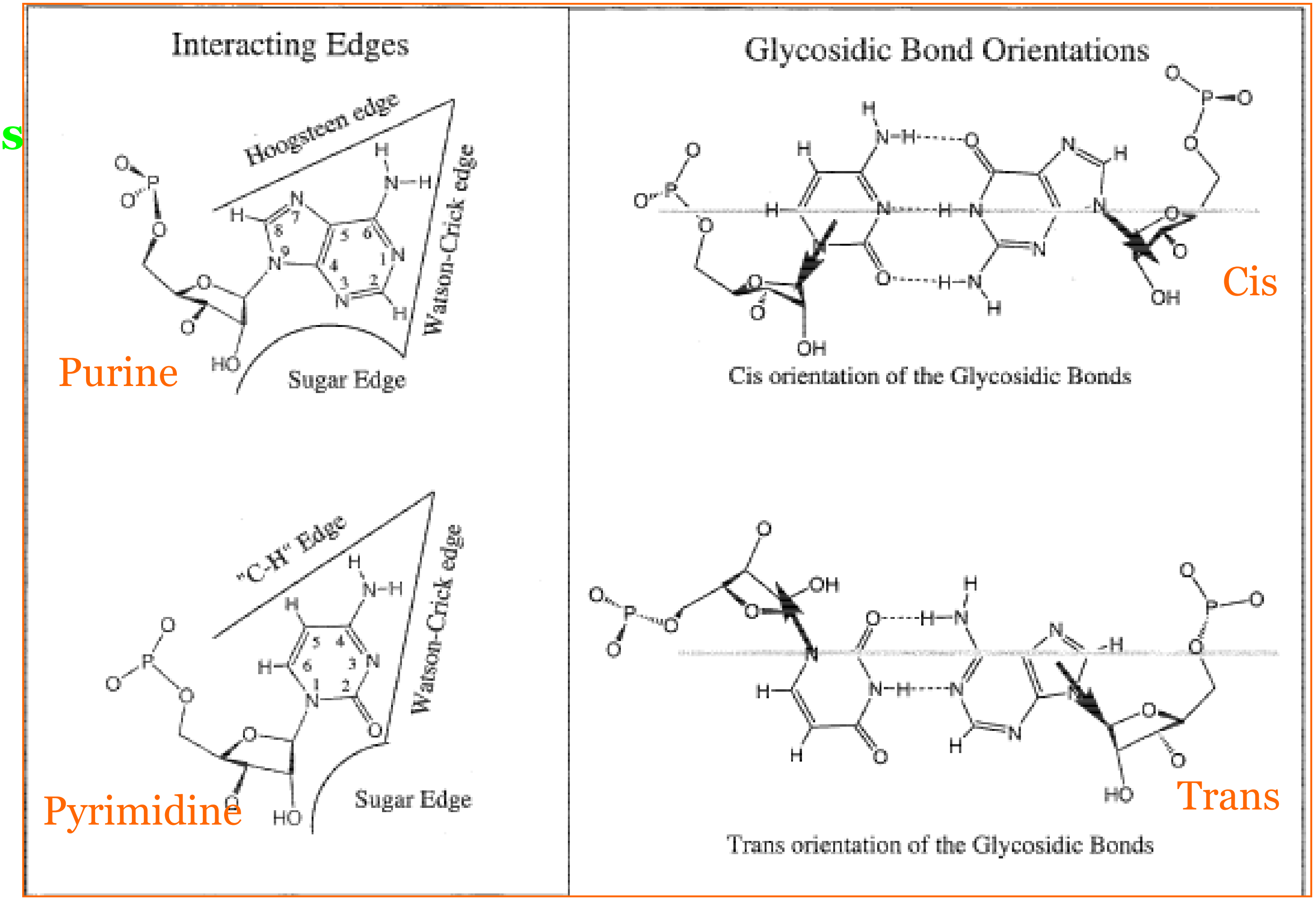
Leontis-Westhof (LW) nomenclature

3 Interacting Edges

- Hoogsteen (H)
- Watson-Crick (W)
- Sugar (S)

2 Orientations

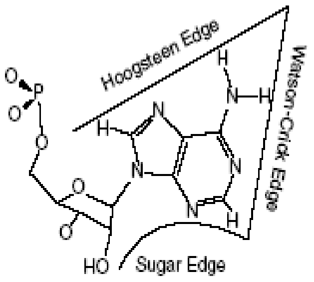
- Cis
- Trans

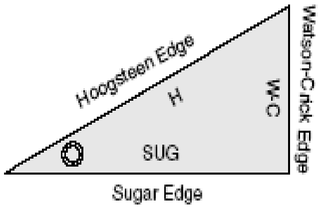


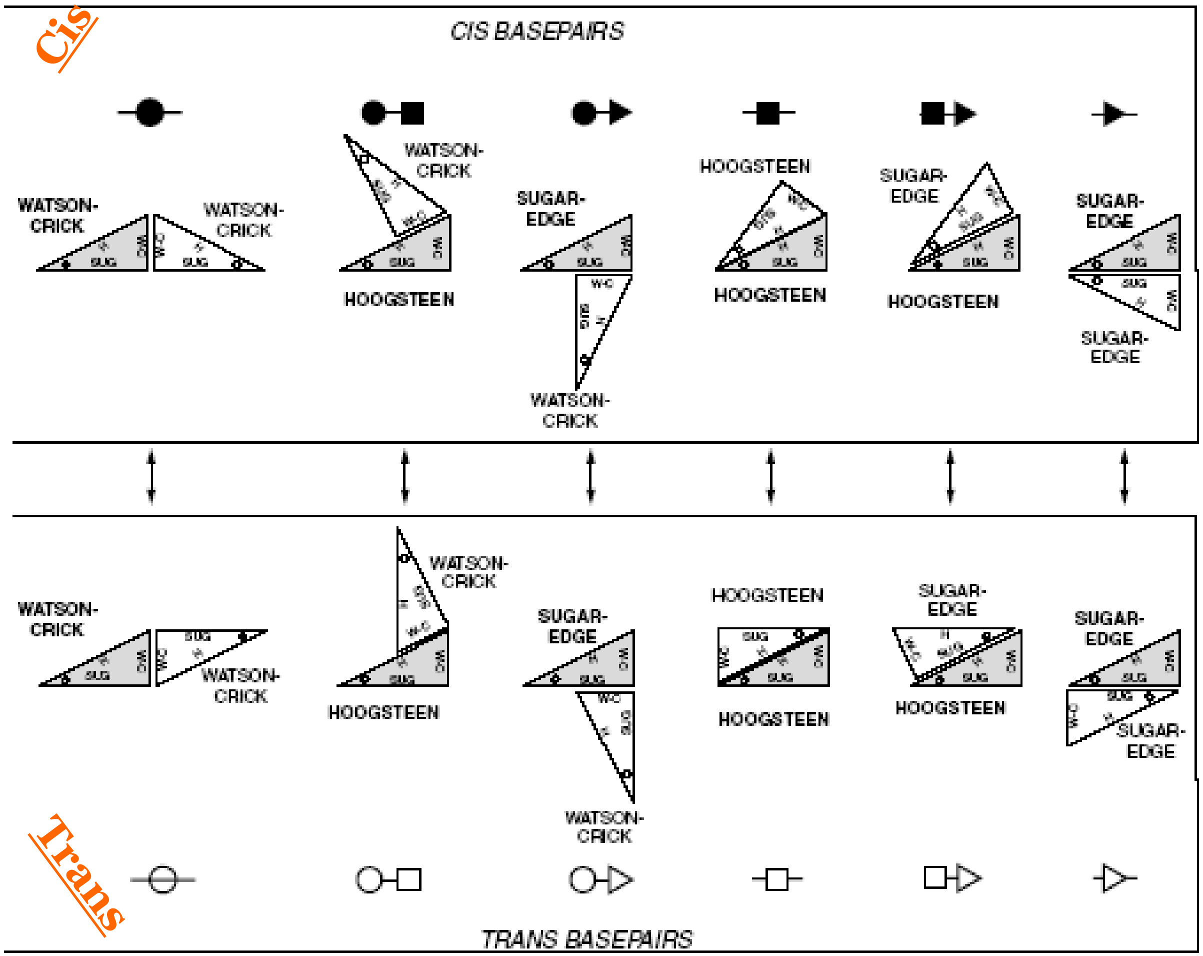
[The Non-WC base pairs and their isostericity matrices, Leontis et al., NAR 2002]

Leontis Westhof (LW) nomenclature

12
Families

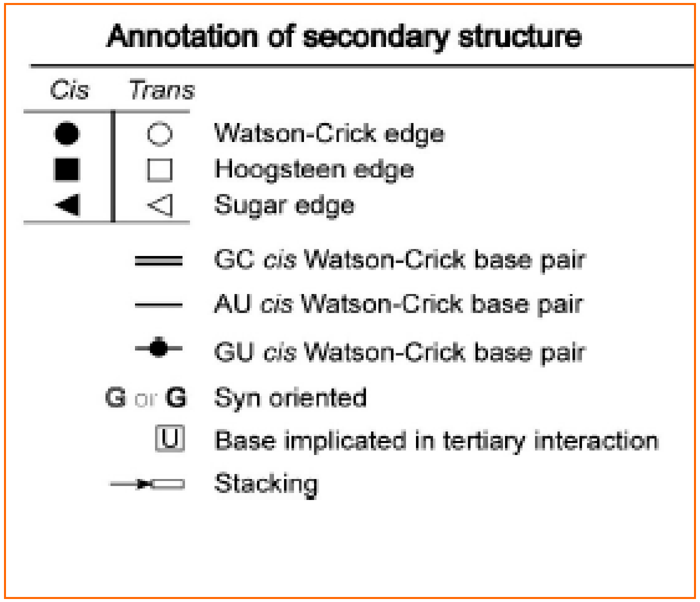






[The annotation of RNA Motifs, N.B. Leontis & E. Westhof, Conference Review 2000]

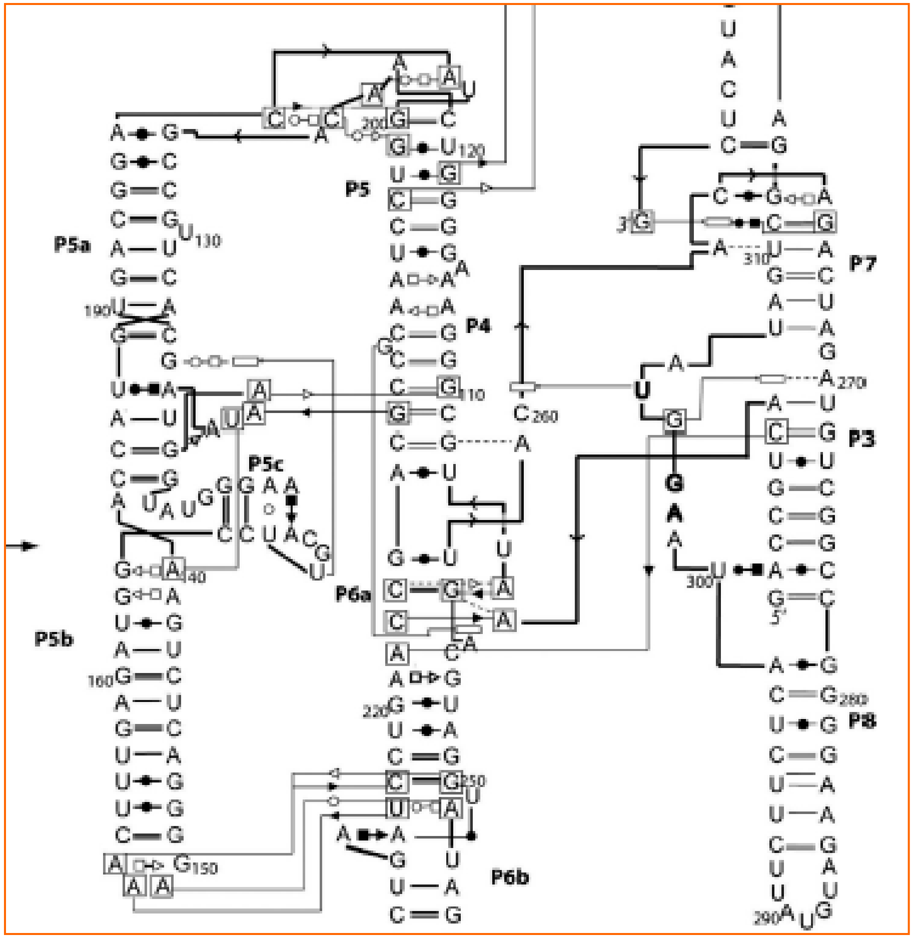
Leontis Westhof (LW) nomenclature



Annotation of secondary structure

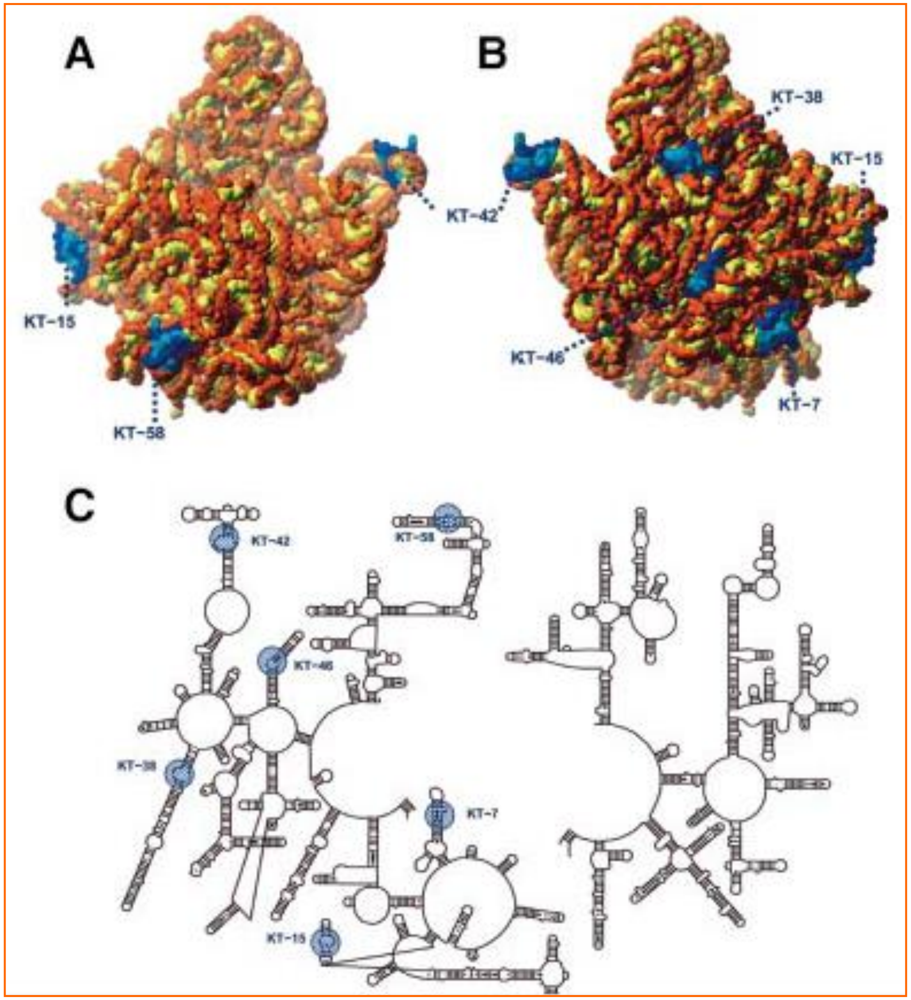
Cis	Trans	
●	○	Watson-Crick edge
■	□	Hoogsteen edge
◀	◁	Sugar edge
	═	GC *cis* Watson-Crick base pair
	—	AU *cis* Watson-Crick base pair
	—●—	GU *cis* Watson-Crick base pair
	G or **G**	Syn oriented
	[U]	Base implicated in tertiary interaction
	→▭	Stacking

RNA graph = graph with bounded degree, whose vertices and edges are labeled.



Group I intron (detail). [The interaction Networks of structured RNAs, A. Lescoute & E. Westhof, NAR 2006]

RNA tertiary motifs



Ex: Kink-turn

They are mostly composed of noncanonical interactions

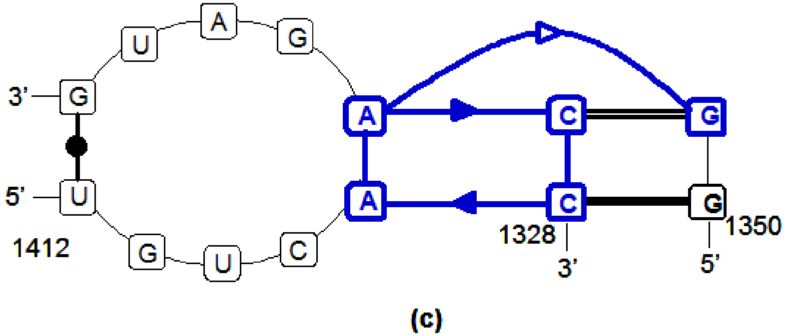
They can mediate the 3D folding of the molecule, they can also be sites for chemical synthesis.

RNA tertiary motifs

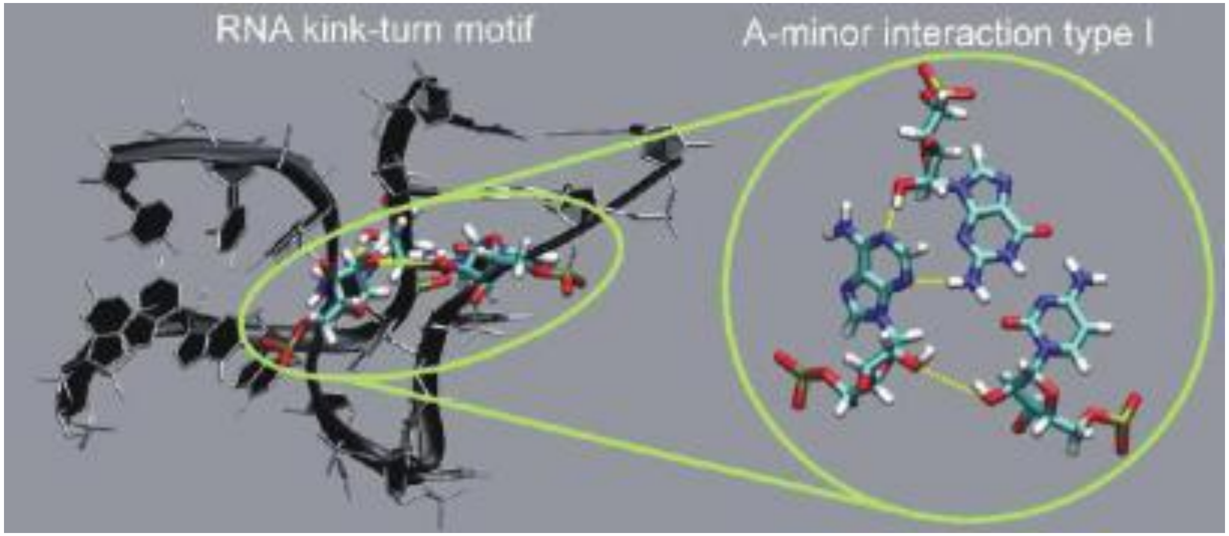
- Knowing the RNA tertiary motifs is essential to understand how the molecule folds into its 3ary structure.
- Problem : how to detect these motifs (including **unknown** motifs) **automatically** ?
  - Local motifs : [Djelloul, Denise RNA 2008]
  - **Interaction networks : this presentation**

RNA interaction networks

An interaction network connects two distinct secondary structure elements (SSEs)



2D diagram of an *A-minor motif typeI/II* which connects a terminal loop and an helix.

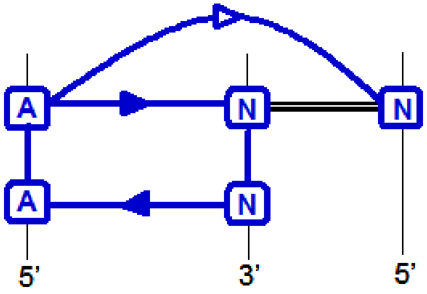


[Reblova et al. 2011]

How to define an interaction network?

(from a ‘graph theoretical’ approach)

- **Hints:**
  - An interaction network **connects two secondary structure elements (SSEs)**
  - An interaction network is **recurrent** : at least two occurrences in a non redondant set of RNAs.
  - The **context** is important (flanking interactions and nucleotides)
  - An interaction network can be **modular**, *i.e.* it can contain smaller interaction networks.



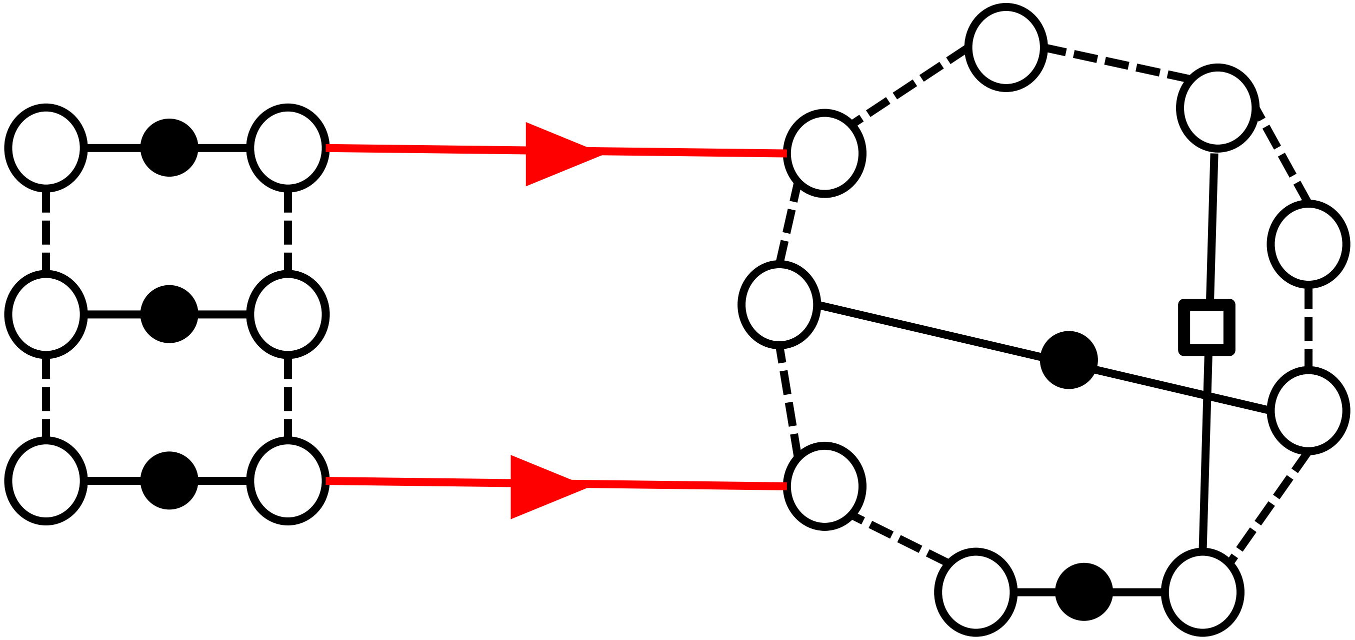
- **Validation:**
  - **Distinct occurrences** of a same interaction network**must have similar 3D shapes**.

How to define an interaction network?

(from a ‘graph theoretical’ approach)

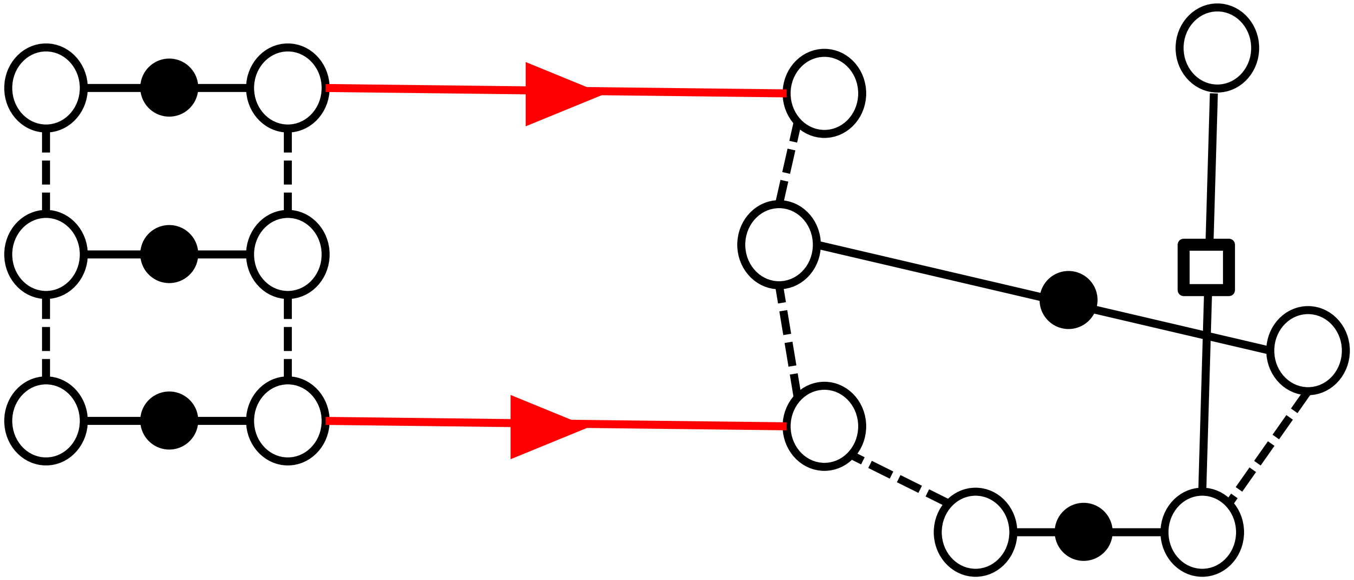
- **Definition**, in two steps :
  - Interaction graphs
  - Recurrent interaction networks (RINs)

Interaction graphs



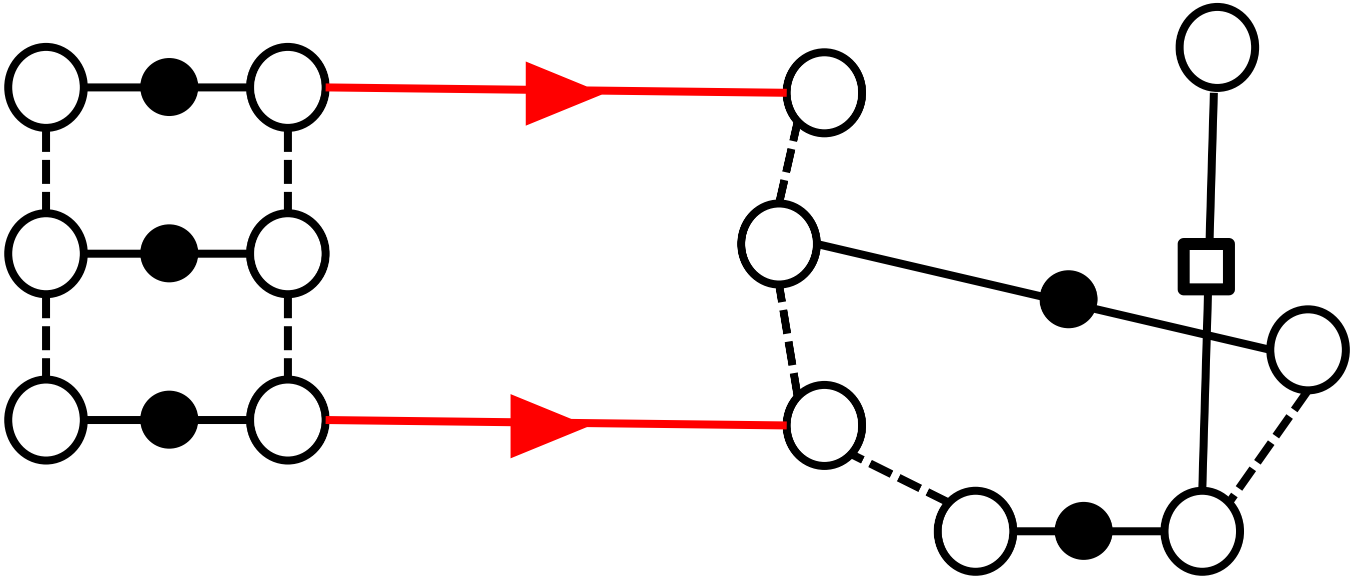
- Let G be the graph of two SSEs with all their inner interactions and mutual interactions.

Interaction graphs



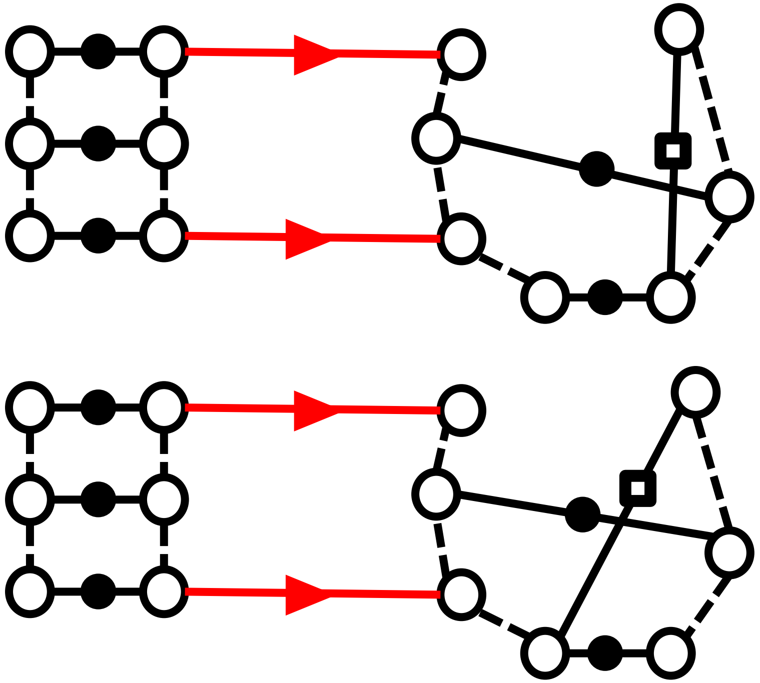
- Let G’ be the subgraph of G obtained by removing the vertices which have only backbone interactions.

Interaction graphs



- The set of **interaction graphs** is the set of the largest connected subgraphs of G’.

Comparing two interaction graphs

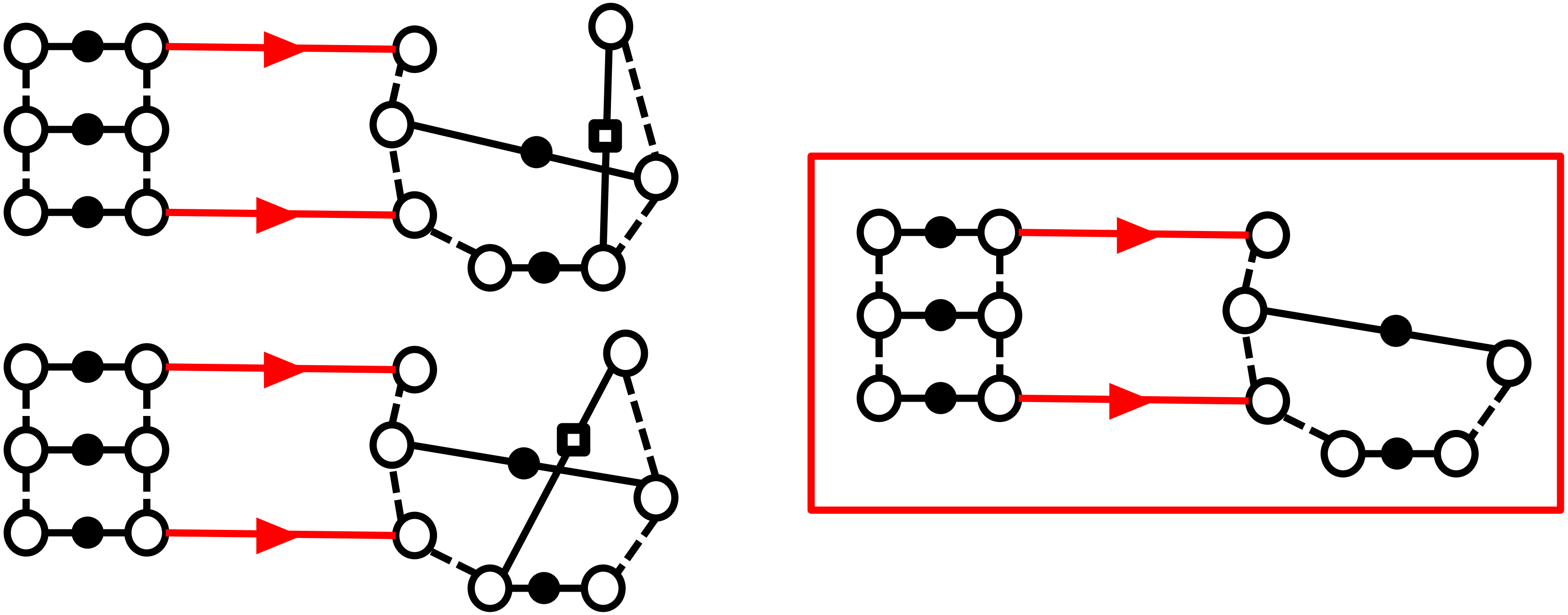


Compute the largest common connected subgraphs

- wich contain at least two red edges.
- and where each vertex belongs to a cycle.

(There may be several such subgraphs)

Comparing two interaction graphs



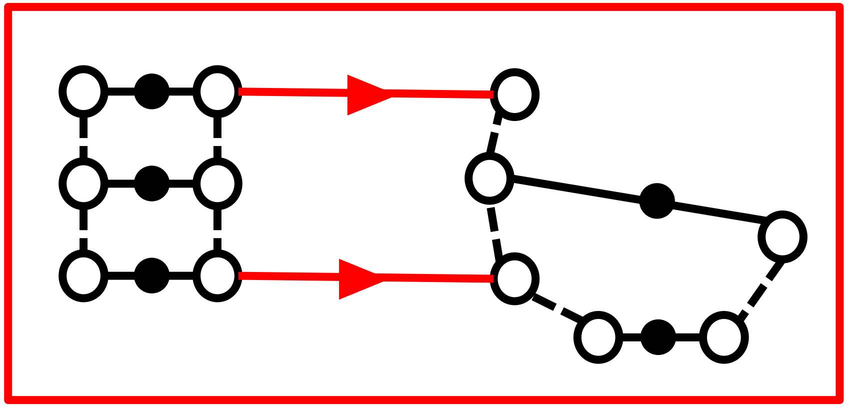
Compute the largest common connected subgraphs

- wich contain at least two red edges.
- and where each vertex belongs to a cycle.

(There may be several such subgraphs)

Recurrent interaction networks

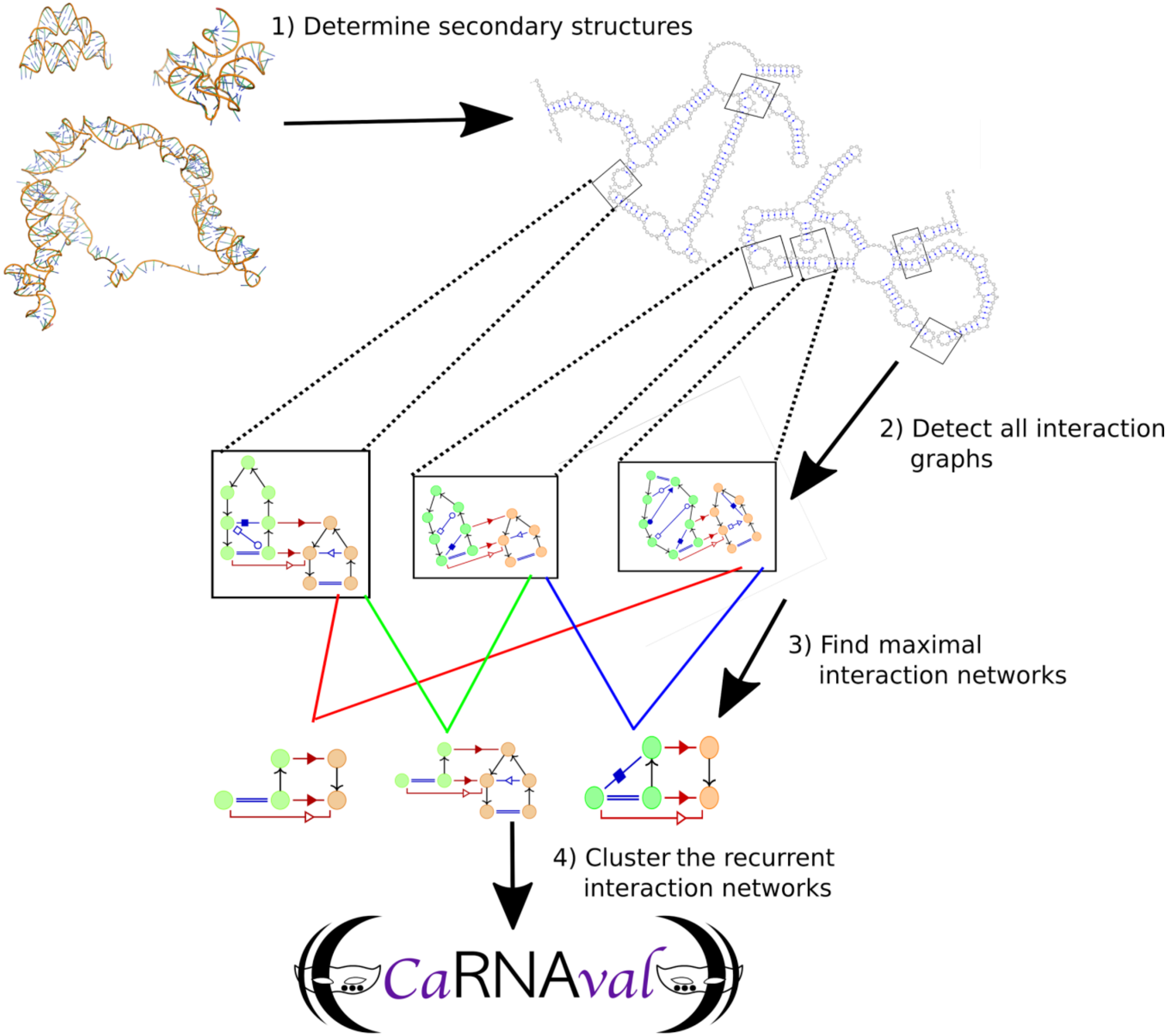
This is what we call a **recurrent interaction network (RIN)**.



Compute the largest common connected subgraphs
- wich contain at least two red edges.
- and where each vertex belongs to a cycle.

(There may be several such subgraphs)

Overview



1) Determine secondary structures

2) Detect all interaction graphs

3) Find maximal interaction networks

4) Cluster the recurrent interaction networks

Computational issues

- The problem of finding a largest common connected subgraph of two graphs is *NP-hard*.

- We developed an *ad hoc* algorithm for this purpose.

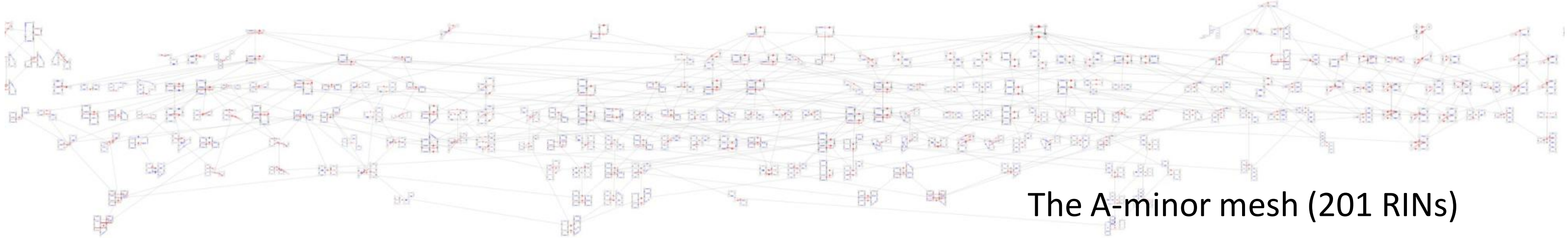
- It takes time!

AN OVERVIEW OF THE RESULTS

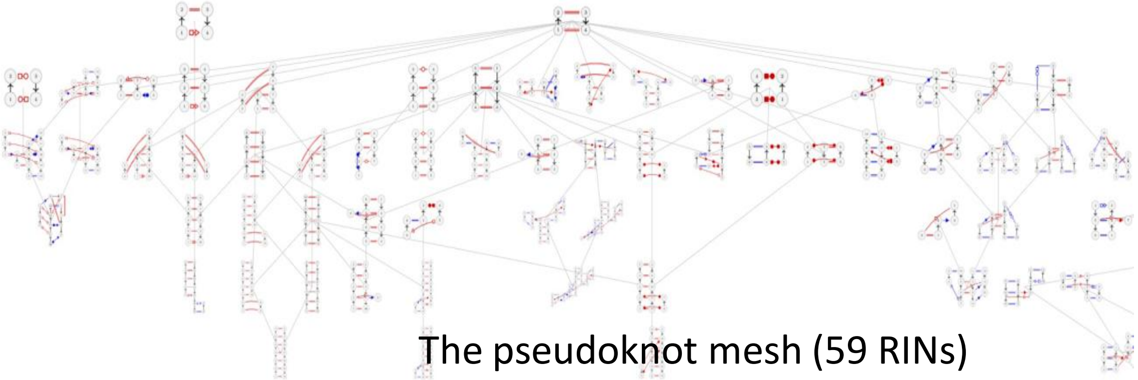
Data and statistics

- All non-redondant structures in RNA3DHub (http://rna.bgsu.edu/rna3dhub) version 2.92, September 2016, at 3.0 Å resolution.
- Some statistics:
  - 845 structures extracted from the PDB, containing 912 RNA chains identified as non-redundant
  - 1426 pairs of SSEs connected by long range interactions
  - 337 recurrent interaction networks (RINs) fund; from 2 to 257 occurrences of each.

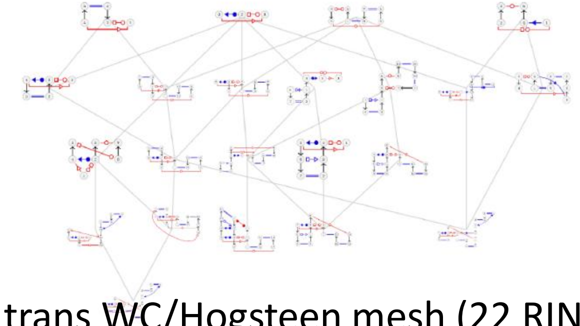
The 337 RINs with their inclusion relations



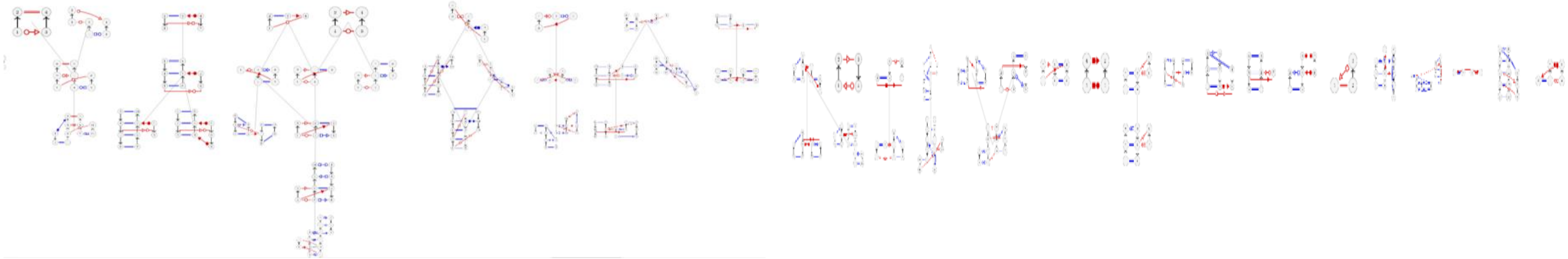
The A-minor mesh (201 RINs)



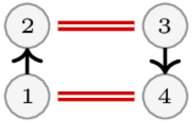
The pseudoknot mesh (59 RINs)



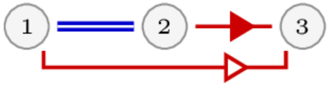
The trans WC/Hogsteen mesh (22 RINs)



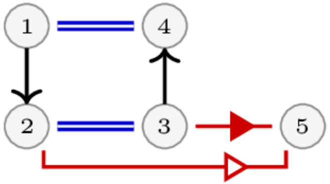
The first 12 RINs (with #occurrences)



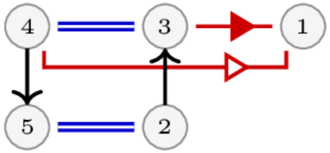
257



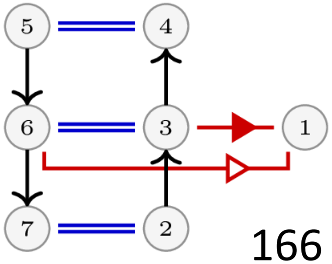
194



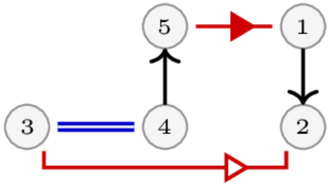
177



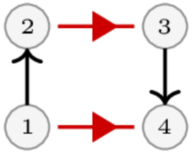
176



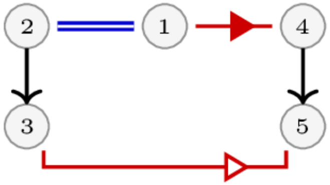
166



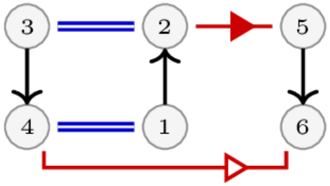
154



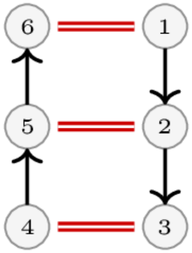
142



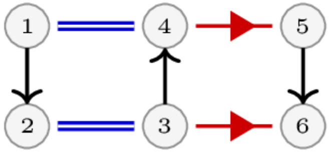
139



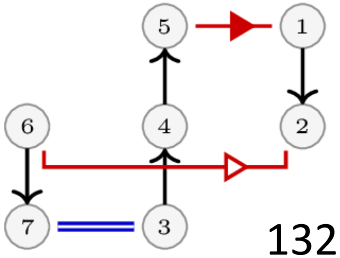
139



135



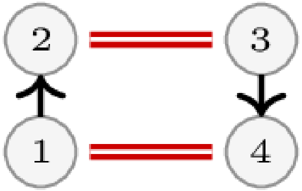
133

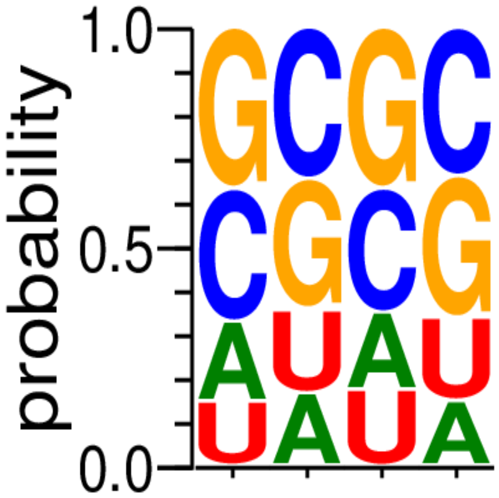


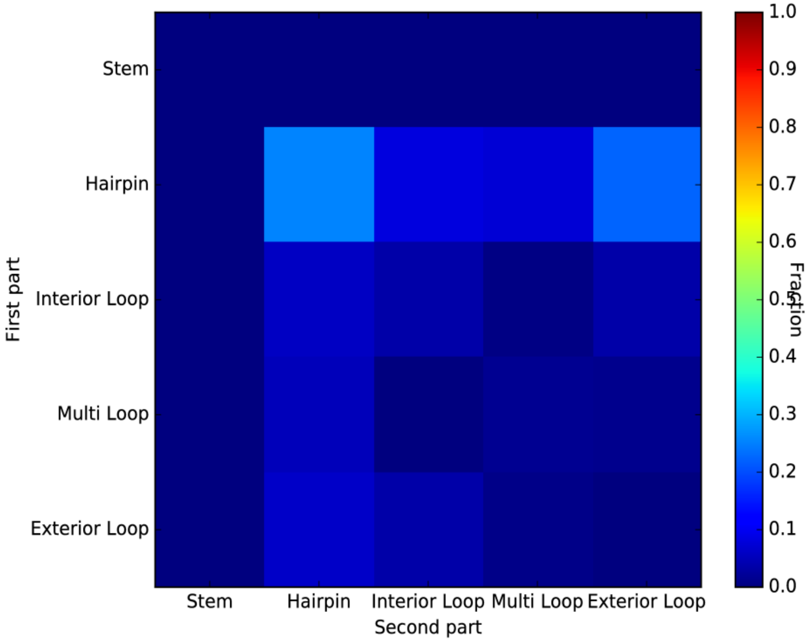
132

First RIN : 257 occurrences

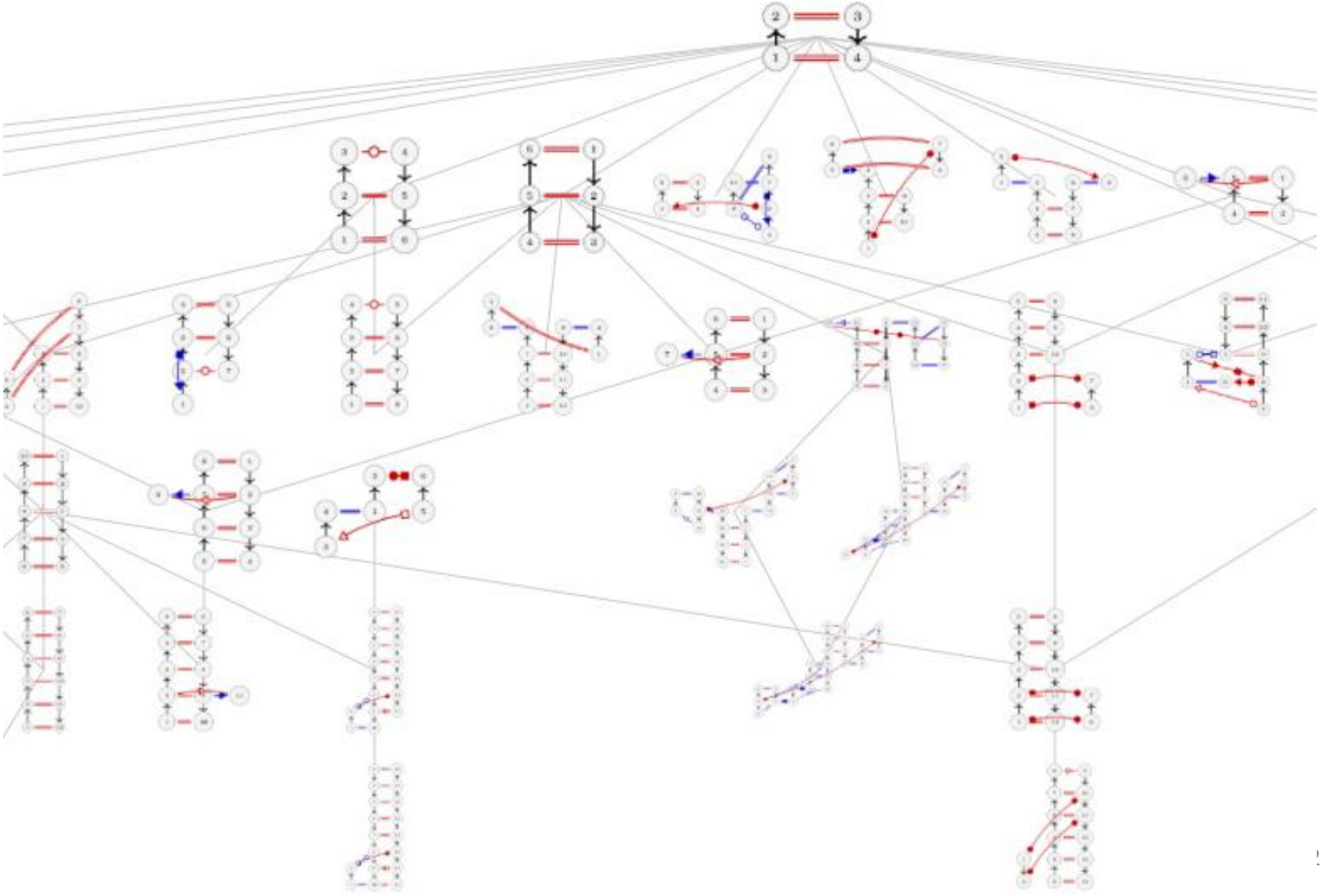
The smallest ‘standard’ pseudoknot motif





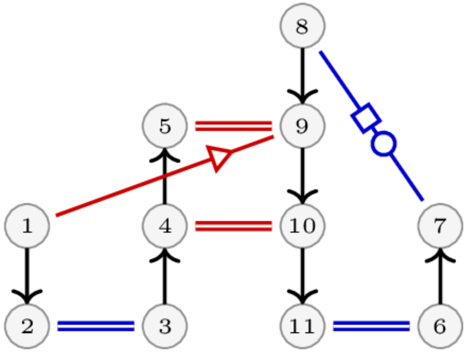


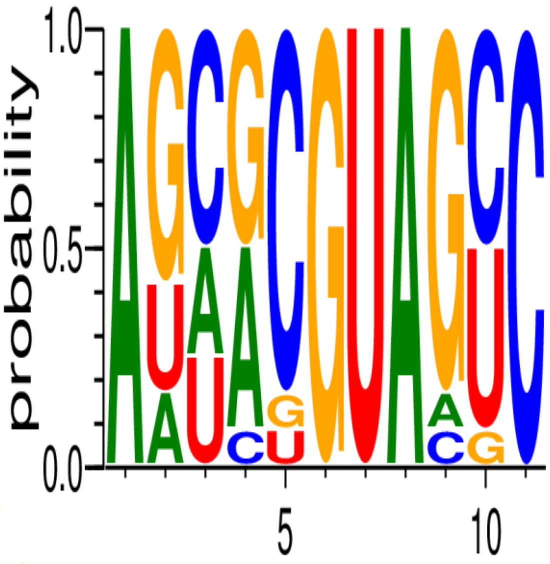
(Part of) the pseudoknot mesh

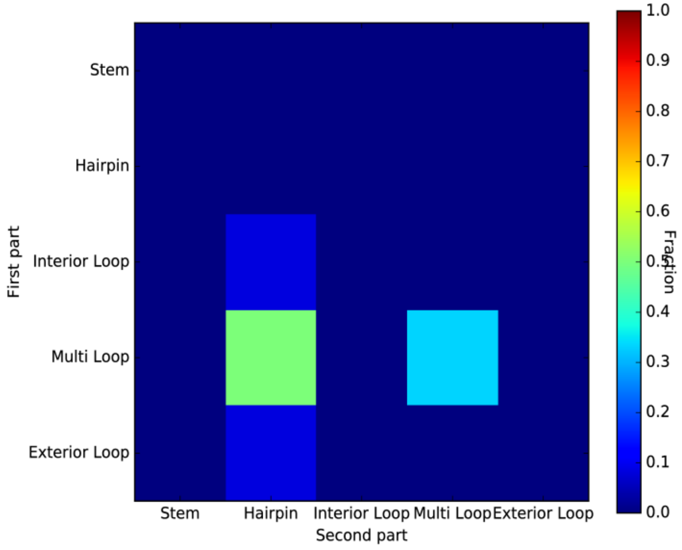


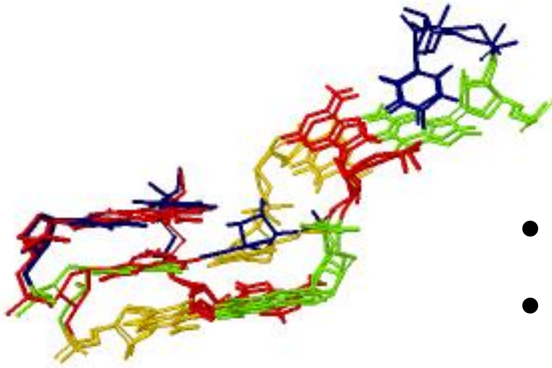
RIN 78 : 12 occurrences

From a structural point of view, the motifs whose occurrences can be found in non homologous molecules are particularily interesting.



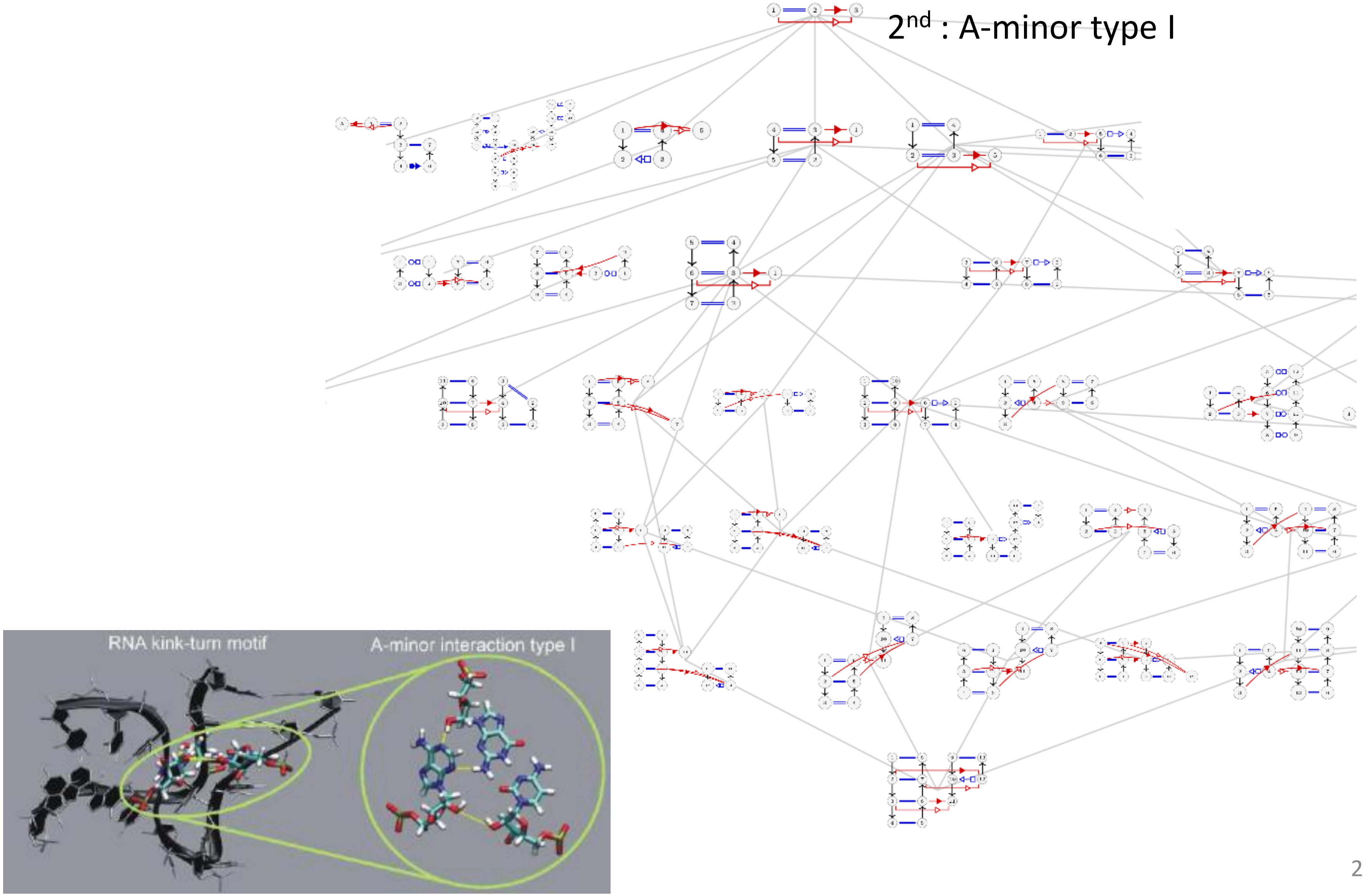






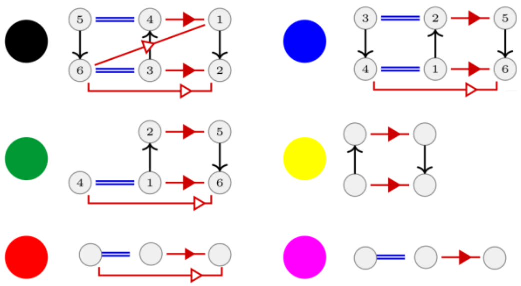
- 8 in ribosomes,
- 4 in riboswitches (colabamin, twister, fmn).

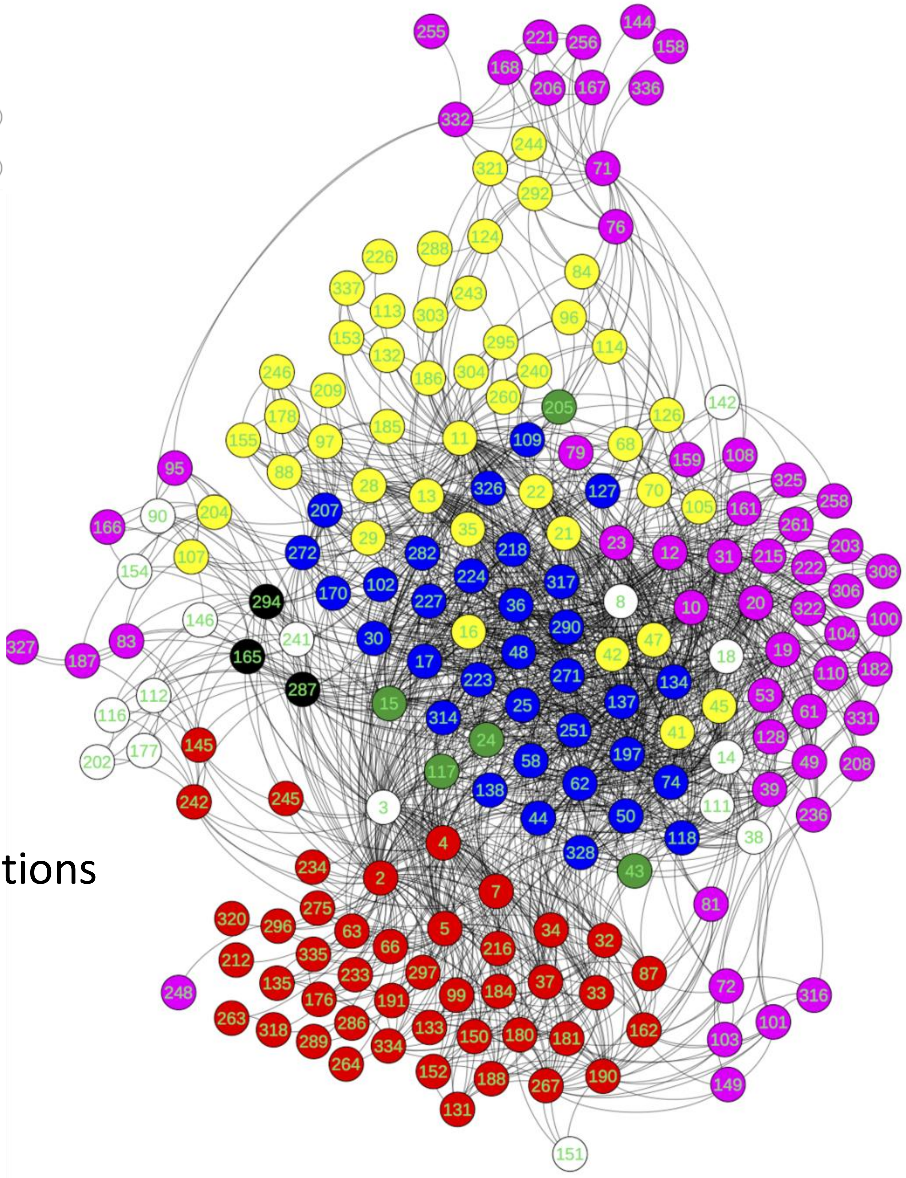
(Part of) the A-minor mesh



2nd : A-minor type I

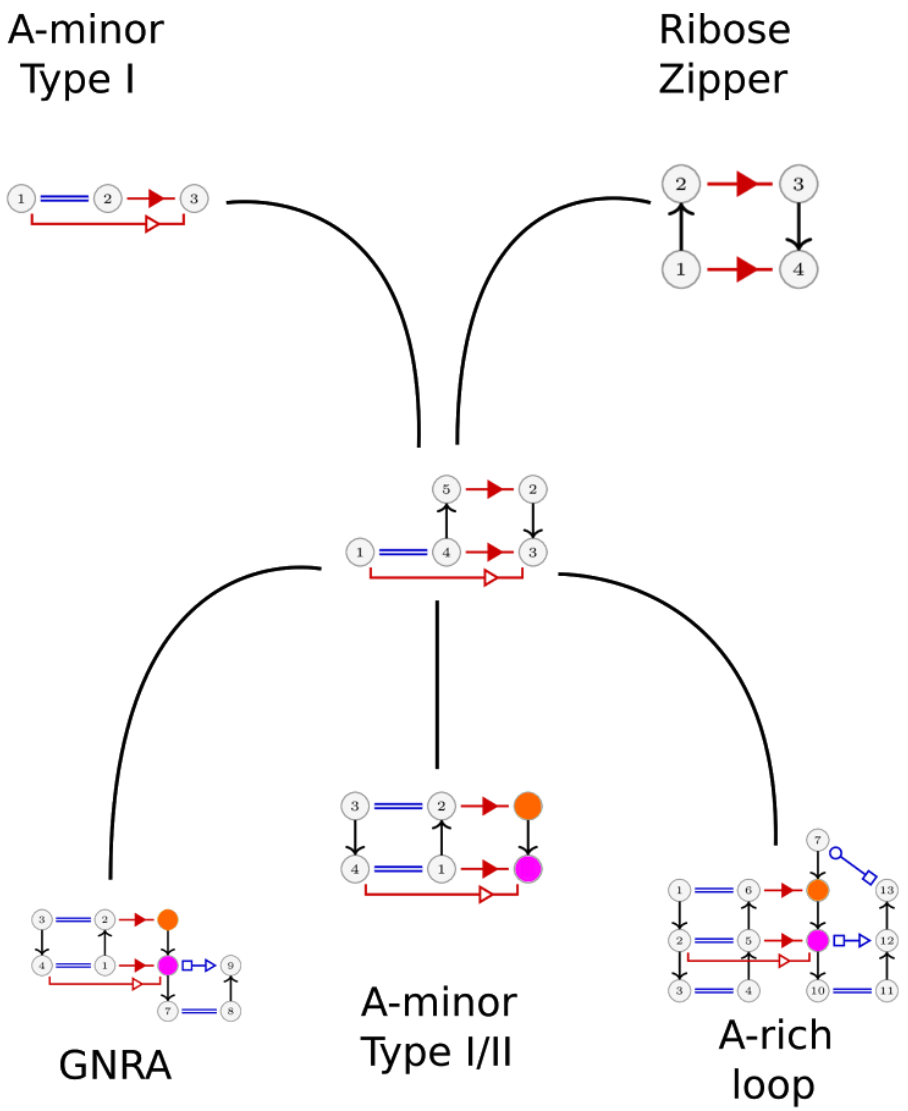
The A-minor mesh





Graph of inclusion relations

Combination of networks



A-minor
Type I

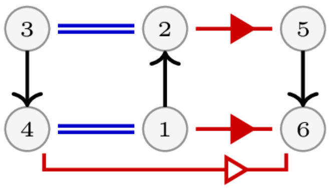
Ribose
Zipper

GNRA

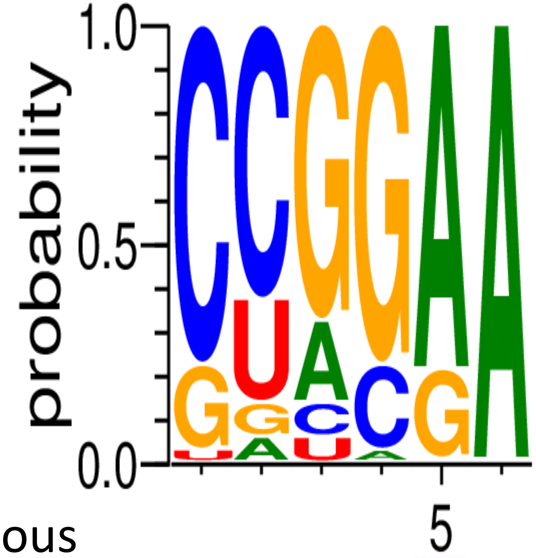
A-minor
Type I/II

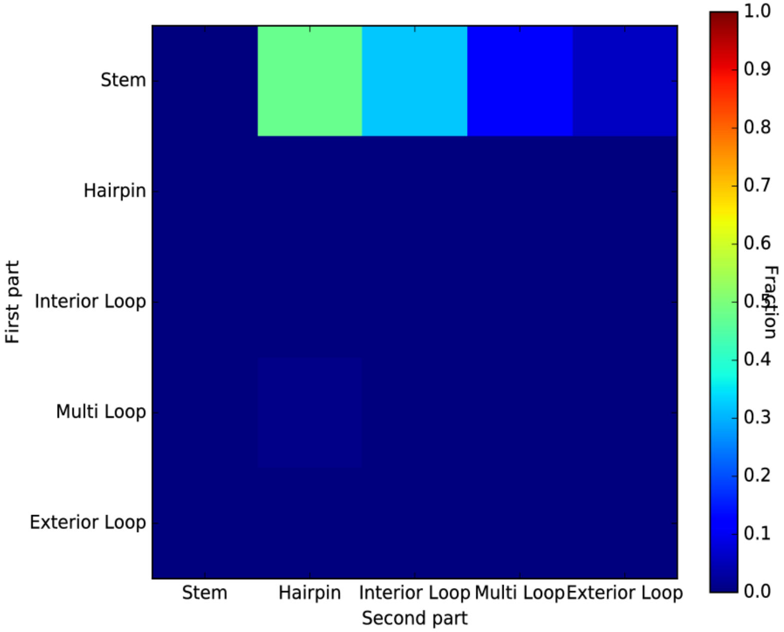
A-rich
loop

RIN 17 : 102 occurrences

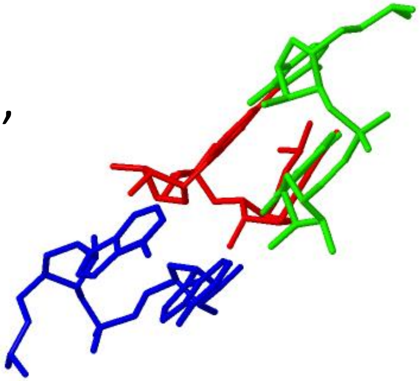


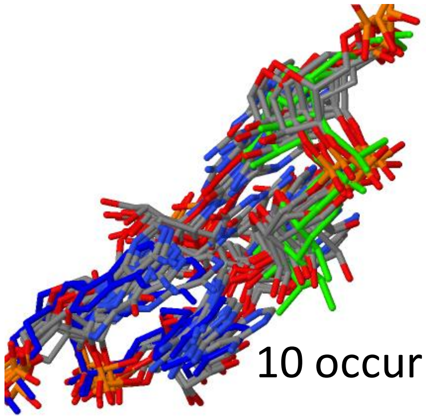
A-minor type I/II





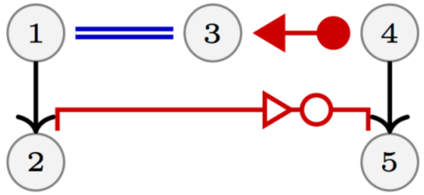
In many non homologous molecules : ribosomes, ribozymes, riboswitches, group II introns, ribonuclease P

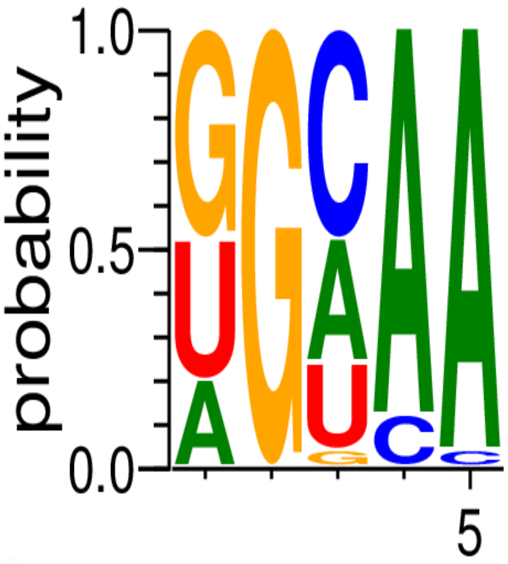


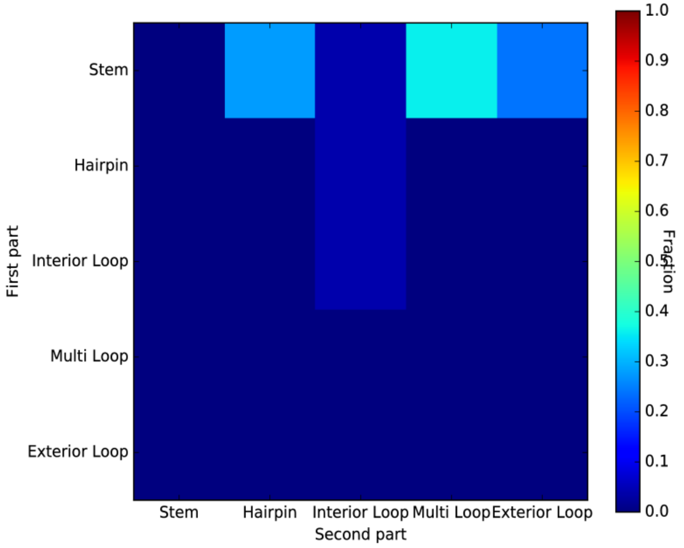


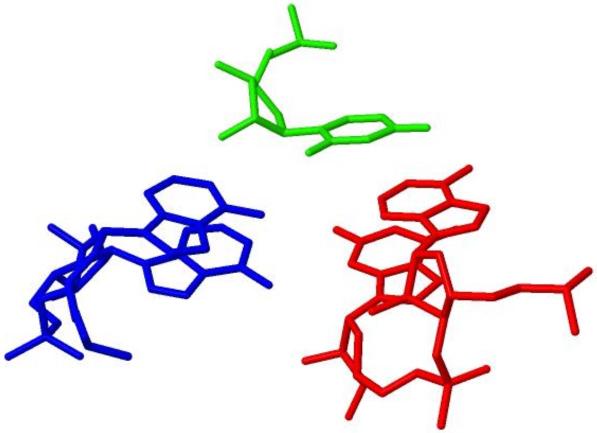
10 occurrences

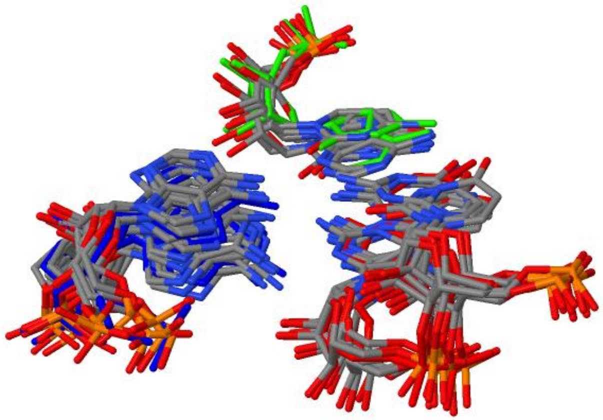
A new RIN: RIN 56, 25 occurrences











10 occurrences

Found in Group I introns, riboswitches, ribosomes

Conclusion

- The first fully automated method for de novo retrieving and clustering RNA recurrent interaction networks.
- New RINs found, and a full map of the modular network of RINs : inclusion relations, combination of RINs for forming new RINs.
- Online database which will be periodically updated.
- Perspective: using RINs for predicting tertiary interactions from secondary structures.

Thanks!

- Collaborations :
  - Interaction motifs :
    - McGill University : Vladimir Reinharz, Jérôme Waldispühl
    - Université de Strasbourg / CNRS : Eric Westhof
    - Ecole Polytechnique (+ McGill) : Antoine Soulé
    - Université Paris-Sud : Mahassine Djelloul
  - Game theory for structure prediction :
    - Université de Versailles – St Quentin : Alexis Lamiable (+ Paris-Sud), Dominique Barth, Franck Quessette, Sandrine Vial
    - Ecole Polytechnique /INRIA : Julie Bernauer
    - Université Paris-Sud : Mélanie Boudard (+ Versailles), Johanne Cohen
  - Structure-sequence alignment :
    - Ecole Polytechnique / CNRS: Yann Ponty
    - Université de Versailles – St Quentin : Dominique Barth
    - Université Paris-Sud : Philippe Rinaudo, Wei Wang, Matthieu Barba