RNA Abstractions for Structural Comparison and Classification

Michela Quadrini
joint work with Luca Tesei and Emanuela Merelli

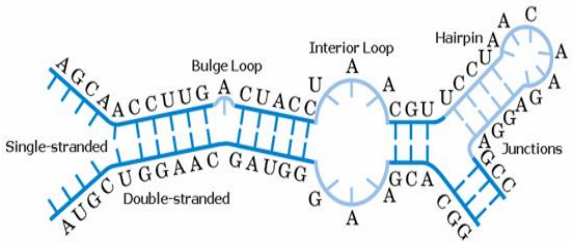
University of Camerino, Italy

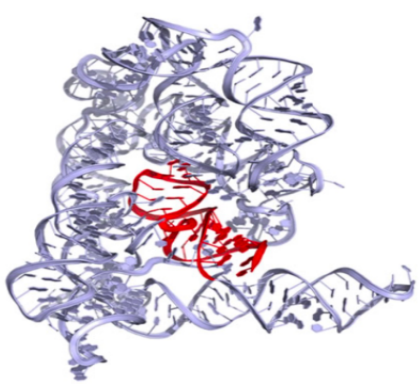
Workshop on Structure and topology of RNA in living systems

30 January 2023 to 2 February 2023

Trento, Italy

RNA Molecules: Hierarchical Structures





Primary Structure | Secondary Structure | Tertiary Structure

- *Primary Structure*: sequence of nucleotides (i.e., `A`, `C`, `G`, `U`)
- *Secondary Structure* (2D): set of hydrogen bonds (base pairs)
- *Tertiary Structure* (3D): spatial arrangement of atoms

Structural Comparison

Biological Hypothesis

The shape is the main predictor of the molecular functions and behaviours

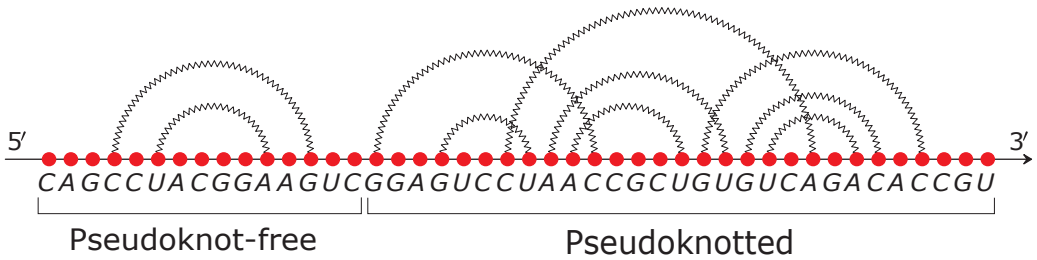
The **structural comparison** supports

- the measurement of the evolution stability
- the identification of the functions
- the classification of the molecules
- the study and prediction of the folding process

The performance (time, space and accuracy) of the methods of comparison depends on the RNA abstractions and representations

RNA Secondary Structures

Arc-Annotated Sequences Representation



Red dot: nucleotide

arc: hydrogen bond

- **Pseudoknot-free structures**
  - are motifs without any crossing among arcs
  - represent local patterns
- **Pseudoknotted structures**
  - are motifs with at least a crossing among arcs
  - are determined by interactions of local patterns

Pseudoknotted Structures Comparison and Classification

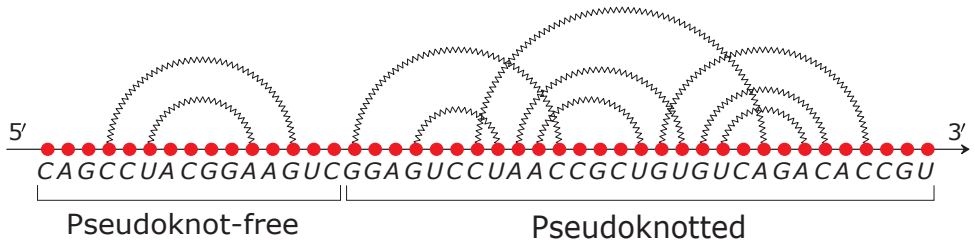
Some approaches introduced in the literature

- Topological-based classification
  - Genus
  - Shape
  - Crossing number
- Comparison
  - Progressive stem matching
  - Pseudoknots order
  - RAG based on Dual Graph
  - **ASPRA distance**

ASPRA Distance[1]

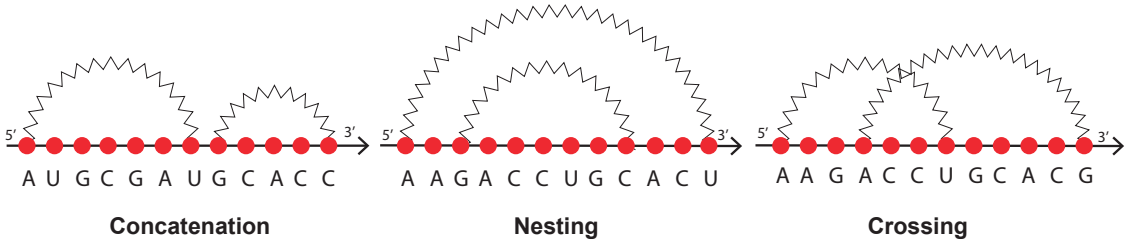
ASPRA Distance is a dissimilarity measure between RNA pairs by

- associating a real number
- neglecting the sequences of nucleotides



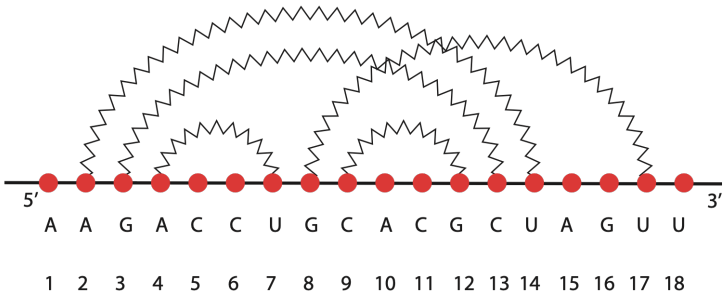
Starting points:

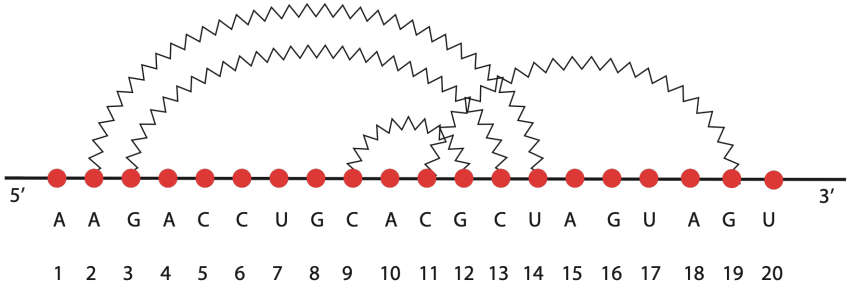
- The problem of pseudoknot-free structure comparison was solved as a **tree comparison**
- Only three relations between arcs are feasible

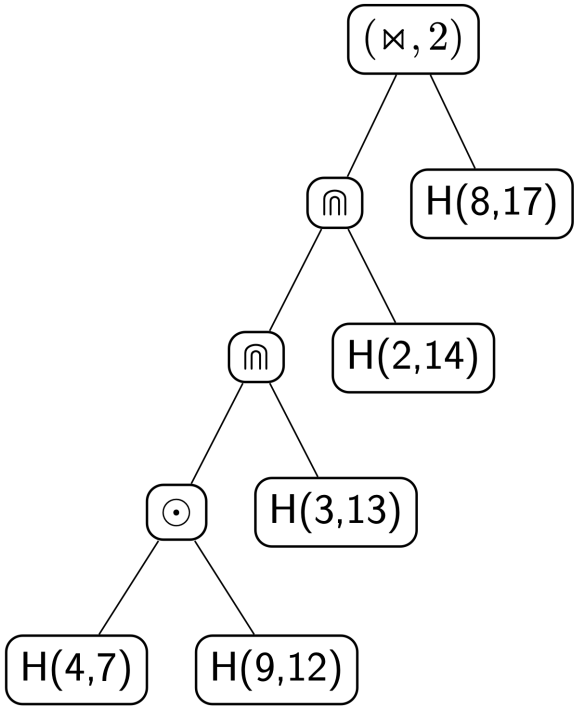


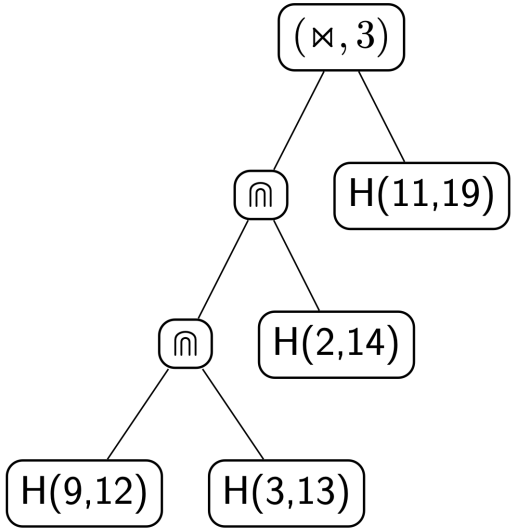
[1] Michela Quadrini, Luca Tesei, Emanuela Merelli, An algebraic language for RNA pseudoknots comparison, BMC Bioinformatics, 2019

Basic idea

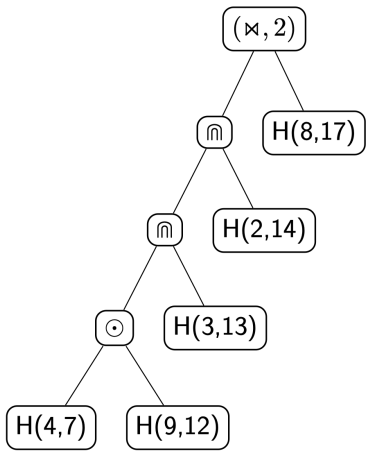

5′
3′
A A G A C C U G C A C G C U A G U U
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18


5′
3′
A A G A C C U G C A C G C U A G U A G U
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

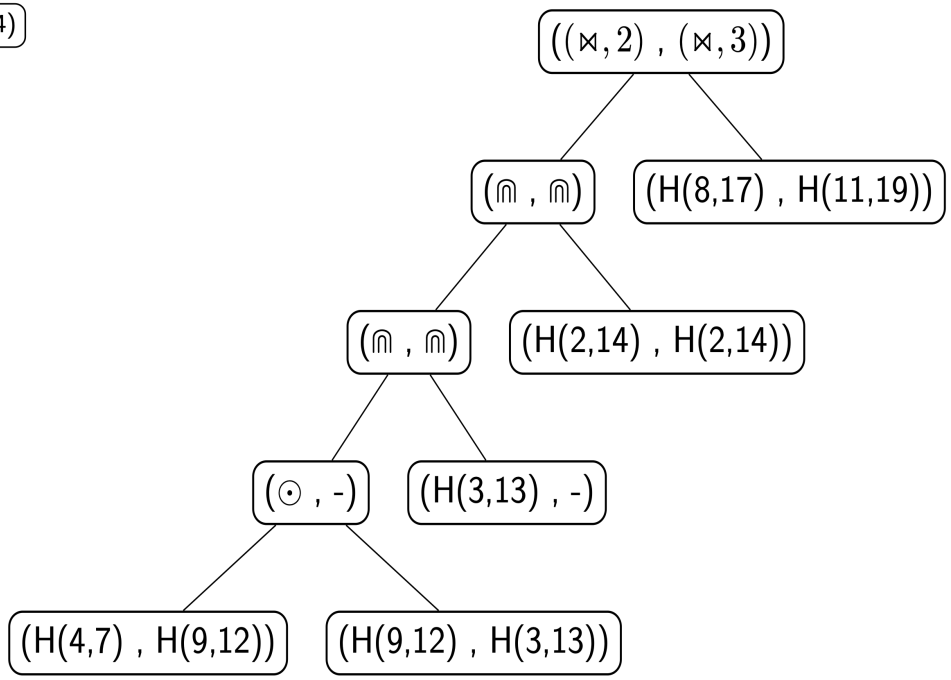

$(\bowtie, 2)$
$\Cap$
H(8,17)
$\Cap$
H(2,14)
$\odot$
H(3,13)
H(4,7)
H(9,12)


$(\bowtie, 3)$
$\Cap$
H(11,19)
$\Cap$
H(2,14)
H(9,12)
H(3,13)

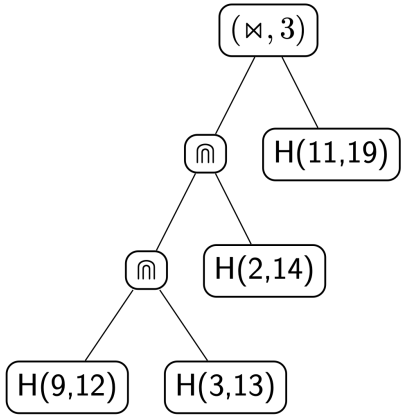
Alignment of Structural Trees



Structural Tree



Alignment of Structural Trees



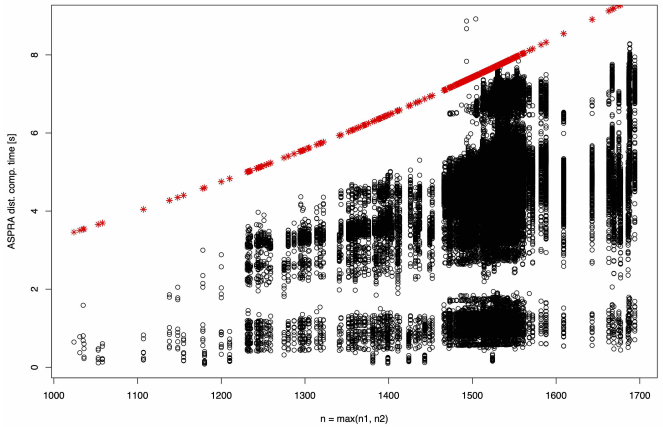
Structural Tree

$$d_{ASPRA}(t_1, t_2) = 201$$

ASPRAlign[2]

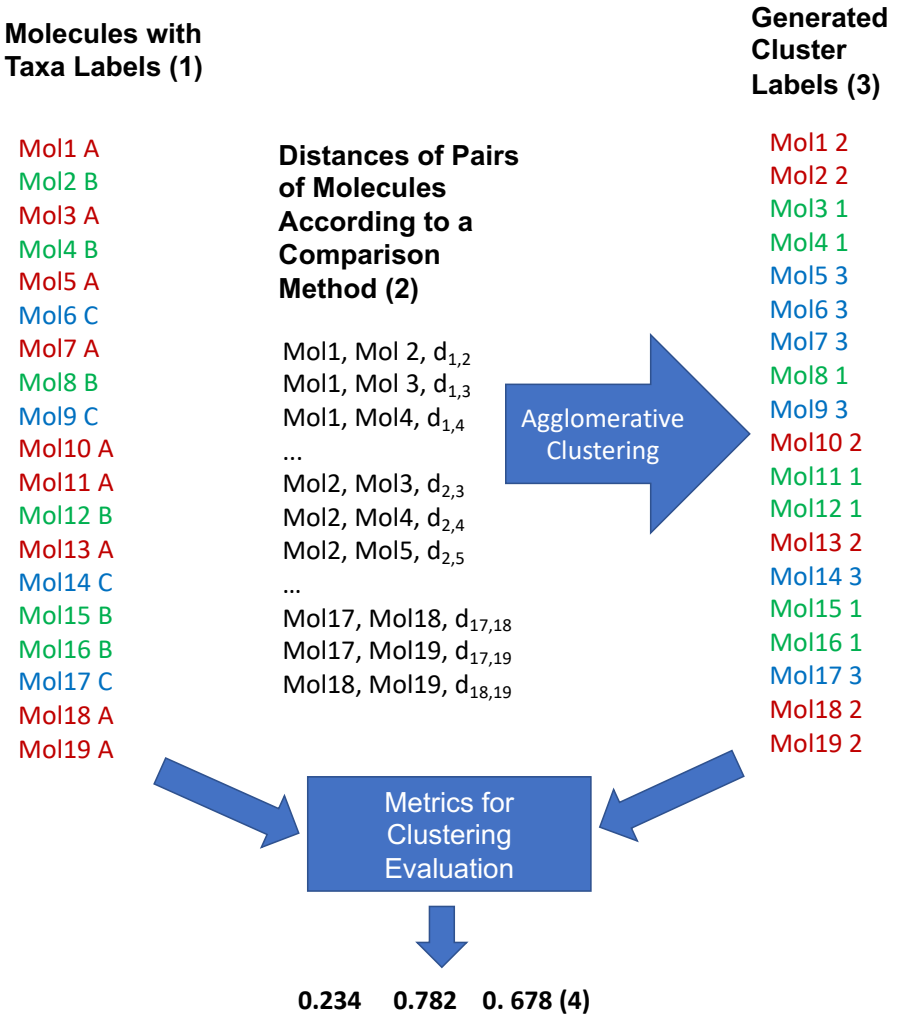
ASPRAlign is a Java tool for computing ASPRA distance that

- compares RNA 2D with arbitrary pseudoknots
- takes as input several formats
  - Extended Dot-Bracket Notation
  - Arc Annotated Sequence
  - BPSEQ
  - CT format
- is available at `https://github.com/bdslab/aspralign`
- works in $\mathcal{O}(n^2)$ time with respect to number of nucleotides



[2] Quadrini M, Tesei L, Merelli E. ASPRAlign: a tool for the alignment of RNA secondary structures with arbitrary pseudoknots. Bioinformatics. 2020

TaxonClassifier: a Framework to classify RNAs [3]



Available on GitHub at

`https://github.com/bdslab/TaxonClassifier`

[3] Michela Quadrini, Luca Tesei and Emanuela Merelli, Automatic Generation of Pseudoknotted RNAs Taxonomy, Under Review

TaxonClassifier: Methods and Clusters

Methods to compare RNA SS

- ASPRA distance
- Pseudoknots order
- Progressive stem matching
- RAG-2D
- genus

Clusters with machine learning approaches

- algorithm: Agglomerative Clustering
- methods: Single, Complete, Average

Reconstruction of Organism Taxonomy

TaxonClassifier has been applied to reconstruct the taxonomy

Defined benchmark

Life Domain	5S rRNA	16S rRNA	23S rRNA
Archaea	25	1486	177
Bacteria	219	1530	570
Eukaryota	351	1512	651

Experiments consider

- the rank phylum according to the European Nucleotide Archive (ENA) taxonomy
- three metrics (Rand Index, Homogeneity, Completeness) to evaluate the clusters

Reconstruction of TaxonClassifier

Some results of the TaxonClassifier

Archaea (16S rRNA)

	Genus	PSMAlign	ASPRAlign	PskOrder	RAG-2D
Rand Index	0.565	0.431	**0.746**	0.601	0.547
Homogeneity	0.134	0.180	**0.833**	0.118	0.128
Completeness	0.198	0.147	**0.559**	0.230	0.100

Bacteria (23S rRNA)

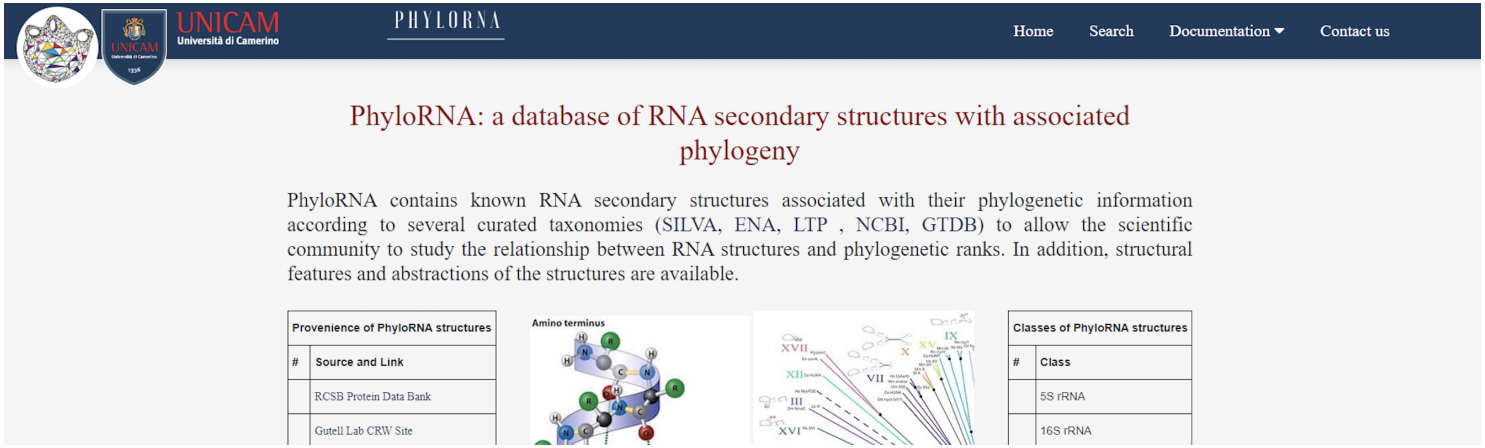
	Genus	PSMAlign	ASPRAlign	PskOrder	RAG-2D
Rand Index	0.458	0.403	0.403	0.470	**0.584**
Homogeneity	0.242	0.155	0.155	**0.248**	0.216
Completeness	0.500	0.320	0.320	**0.510**	0.230

Eukaryota (5S rRNA)

	Genus	PSMAlign	ASPRAlign	PskOrder	RAG-2D
Rand Index	0.486	**0.569**	0.486	0.486	0.486
Homogeneity	0.093	**0.269**	0.093	0.093	0.093
Completeness	0.323	**0.725**	0.323	0.323	0.323

Ongoing Works and Future Directions

- consider other taxonomies, like SILVA and LTP

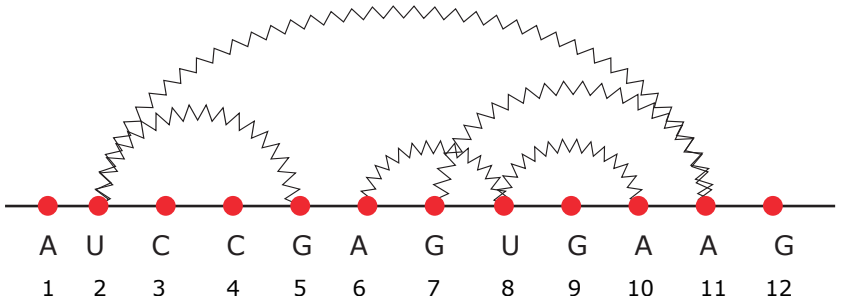


- applications of other methods to compare RNA 2D
- evaluation of the methods predicting the folding process
- classification of the molecules according to their functions (relevant for non-coding RNAs)

Comparison of Tertiary Structures[4]

RNA 2Ds may be not expressive enough to capture dissimilarities: different tertiary structure may create different interactions

We introduce an abstraction of 3D structure based on spatial proximity among bases



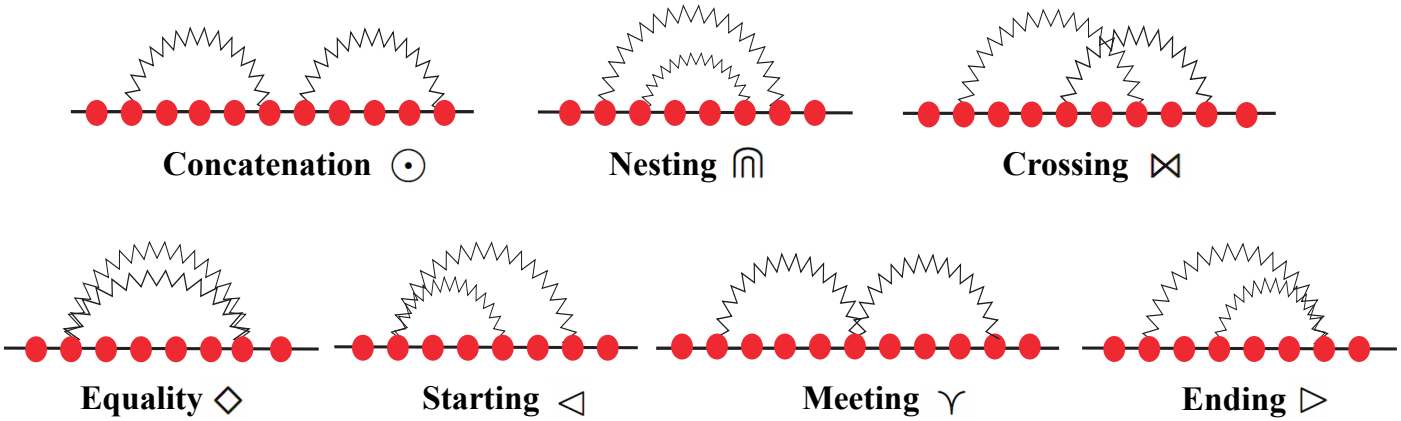
Red dot: nucleotide

arc: proximity (distance less than a threshold)

[4] Michela Quadrini, Luca Tesei and Emanuela Merelli, Proximity-based Comparison of RNA and Protein 3D Structures, Submitted

ASA Distance[5]

From three to seven relations between arcs in 3D structures

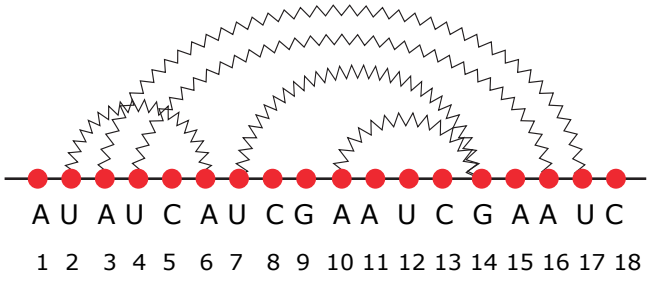


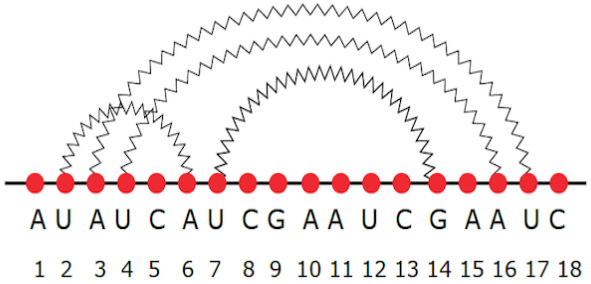
ASA Distance is a measure of dissimilarity between RNA pairs by

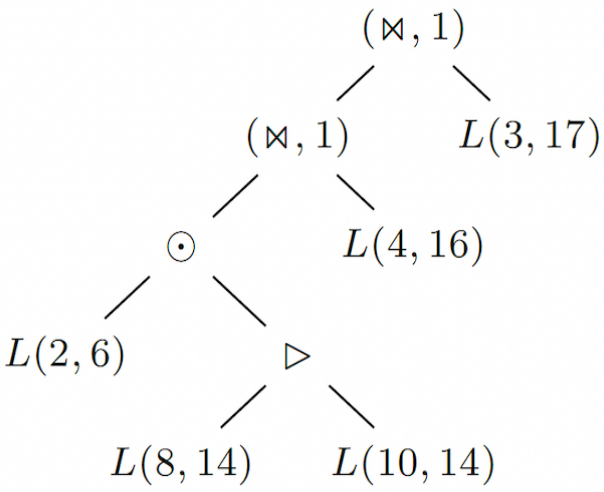
- extending ASPRA distance
- associating a real number
- considering the proximity among bases
- neglecting the sequences of nucleotides and hydrogen bonds

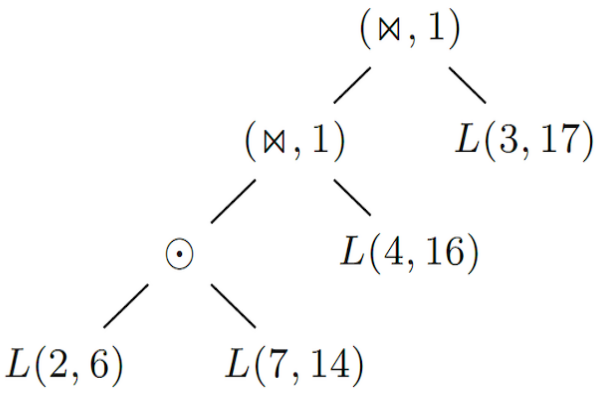
[5] Michela Quadrini, Luca Tesei and Emanuela Merelli, Proximity-based Comparison of RNA and Protein 3D Structures, Submitted

Structural Trees for 3D

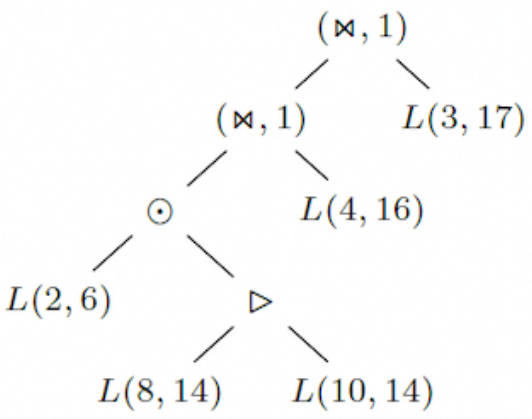




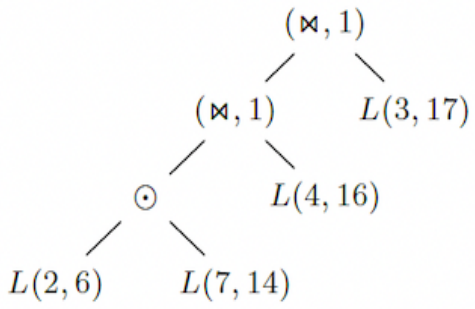




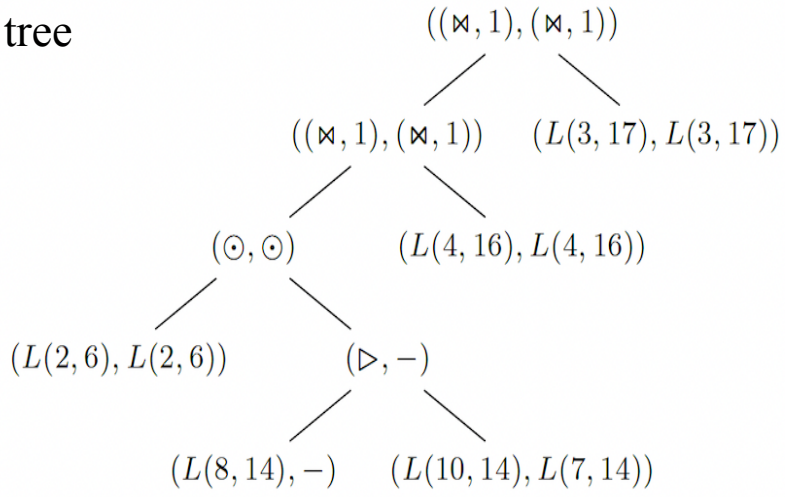
Alignment of Structural Trees



Structural tree



Structural tree



Alignment of the two Structural trees

STAling[6]

STAling is a Java tool for computing ASA distance that

- compares tertiary structures of RNAs, proteins, and complexes
- takes as input
  - PDB ID
  - PDB files
  - Arc Annotated Sequences
- is available at `https://github.com/bdslab/STAlign`
- works in $\mathcal{O}(n^4)$ time with respect to number of nucleotides

[6] Filippo Lampa, Michela Quadrini, Marco Serenelli, and Luca Tesei, STAlign Structural Tree Alignment, In Preparation

Future Directions

We intend to

- compare STAlign respect to other RNA 3D comparison tools
- define a framework to evaluate folding algorithms
- define other dissimilarities on trees
- design pipelines for understanding the relation sequence - secondary structure - tertiary structure - functions

"We shall need for this the creation of a new breed of mathematical professionals able to mediate between pure mathematics and applied science. The cross fertilization of ideas is crucial for the health of the science and mathematics." (Gromov, 1998, p. 847)

Thank you for the attention!

michela.quadrini@unicam.it