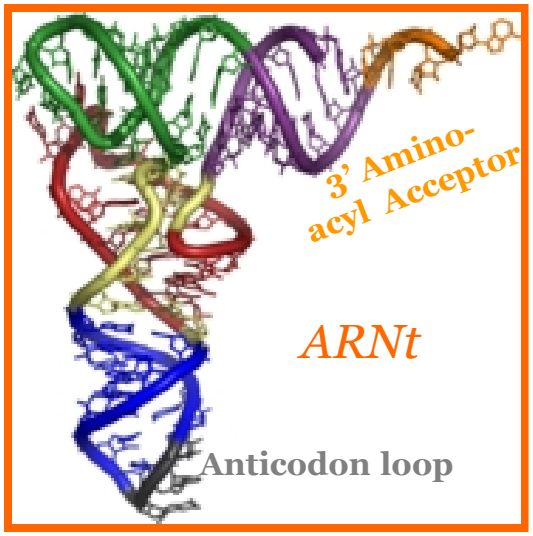
Extraction et classification automatiques de motifs dans les structures tertiaires

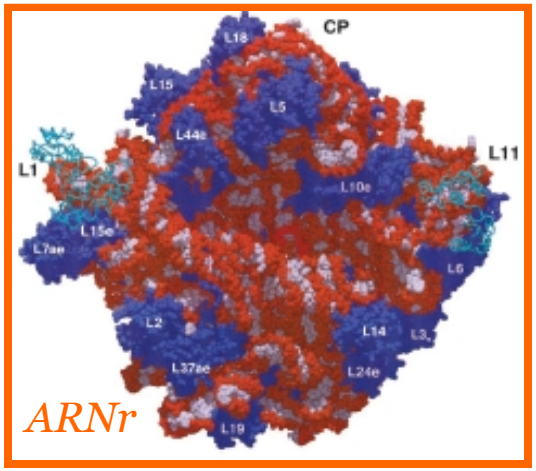
Mahassine Djelloul & Alain Denise

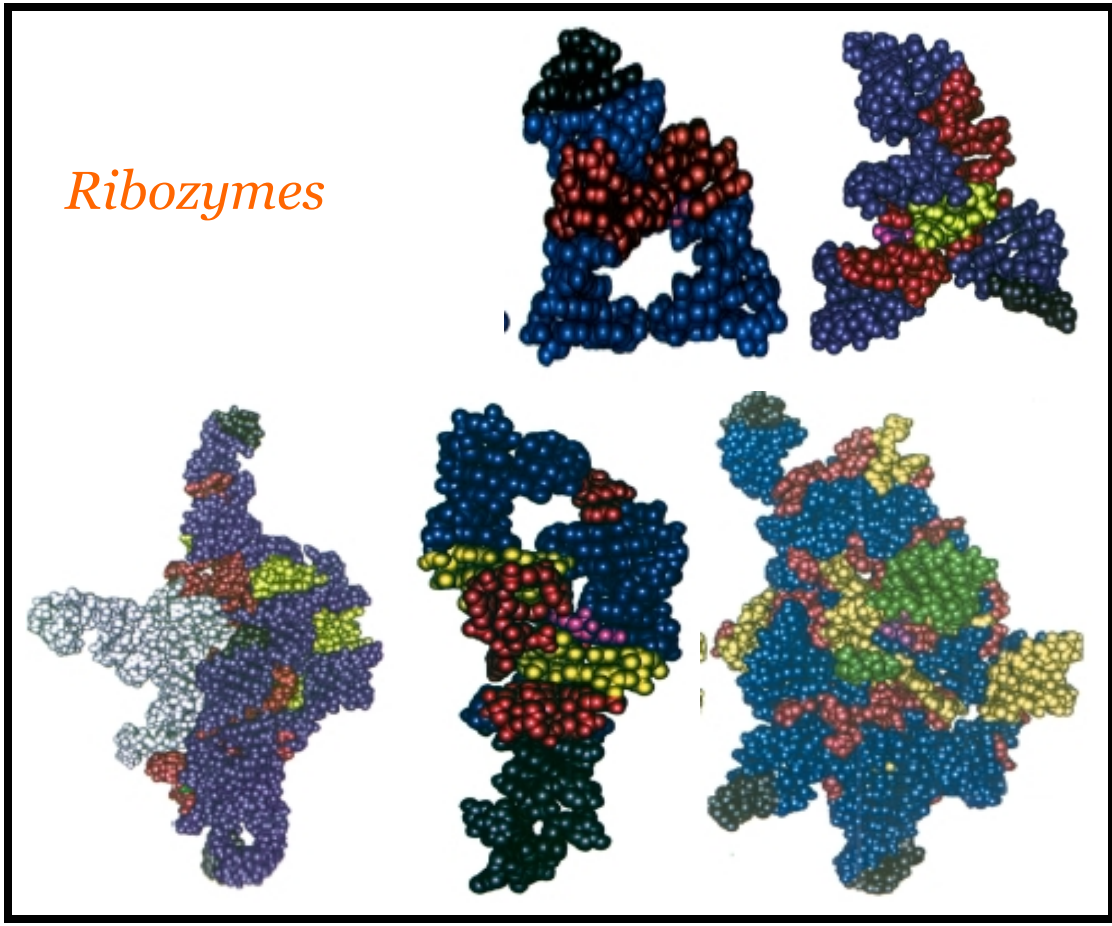
LRI et IGM Orsay, Université Paris-Sud 11

PRES UniverSud Paris

Structure de l'ARN





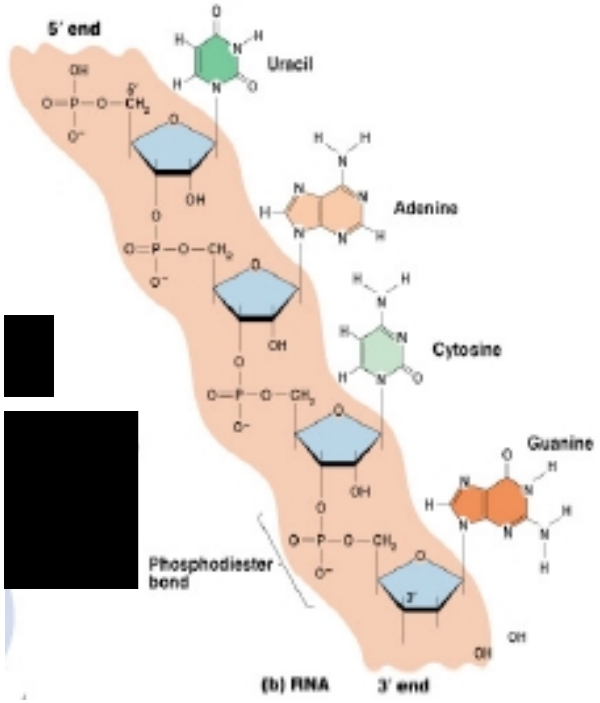


[Catalytic RNA, F. Walter & E. Westhof, els, 2002]

Structure de l'ARN

ARN = une chaîne de nucléotides qui

5' end



3' end

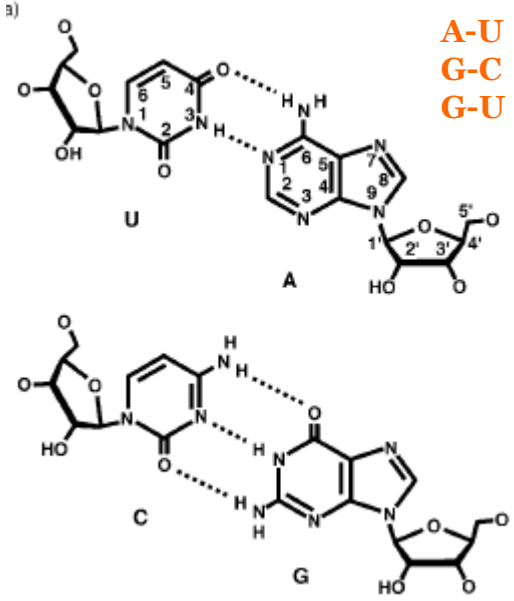
[© 2003 Pearson Education, Inc., Publishing as Benjamin Cummings]

5'....UACGUCGAU...3'

Structure de l'ARN

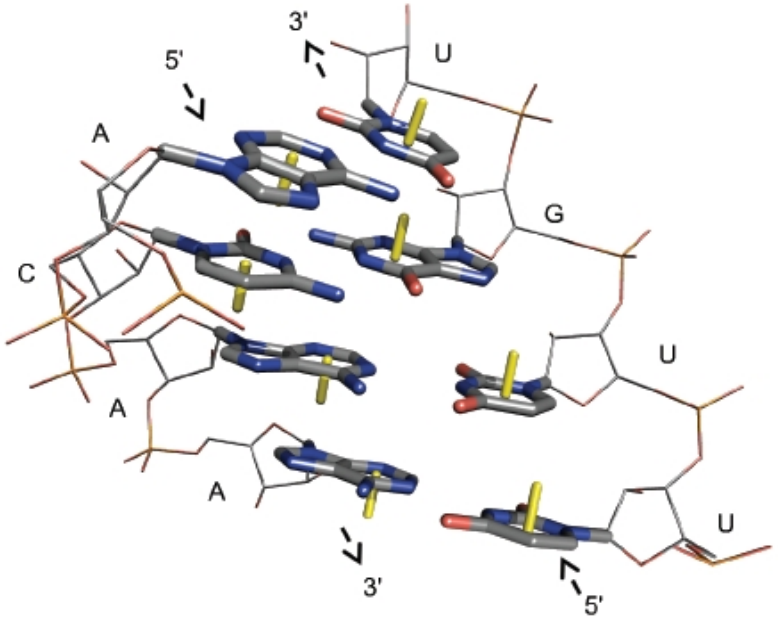
qui se replie en hélices de type A

Appariements WC
canoniques



ET

Stacking



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

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

Structure de l'ARN

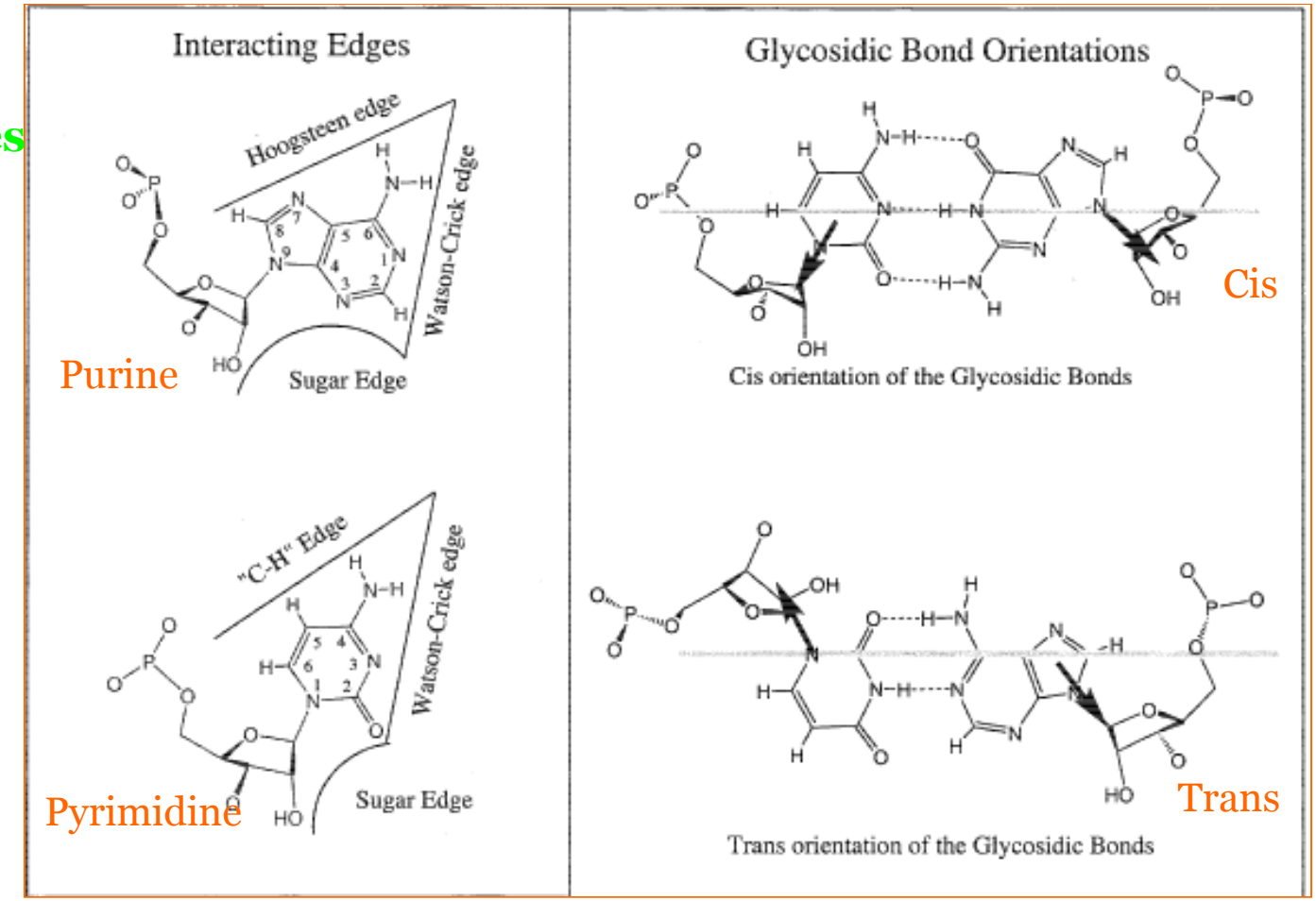
Nomenclature Leontis-Westhof (LW)

3 Interacting Edges

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

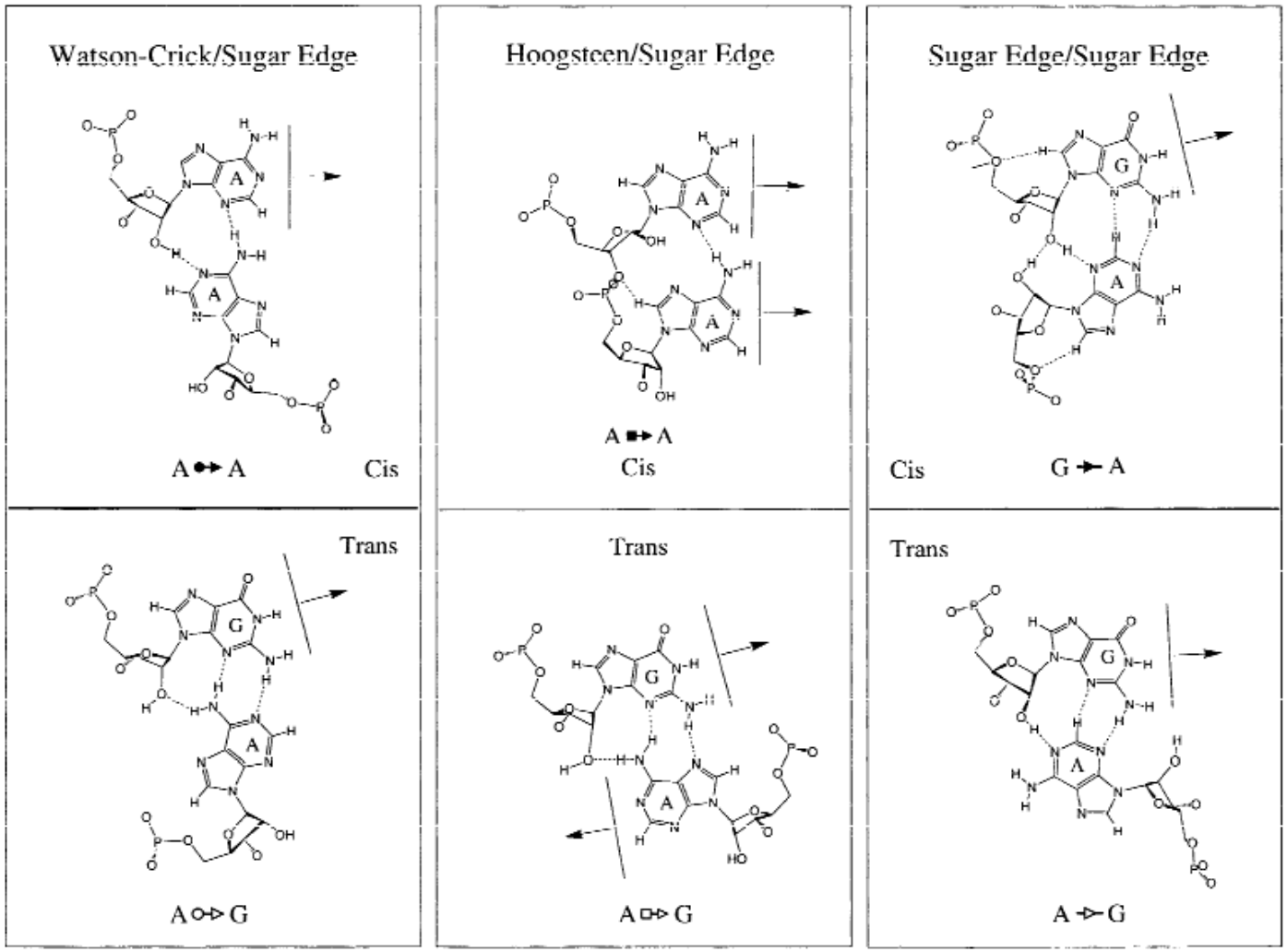
2 Orientations

- Cis
- Trans



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

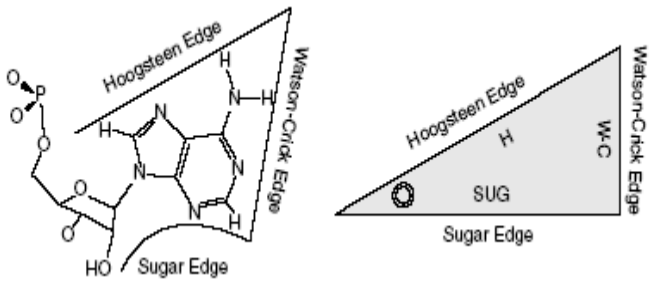
Liaisons non canoniques

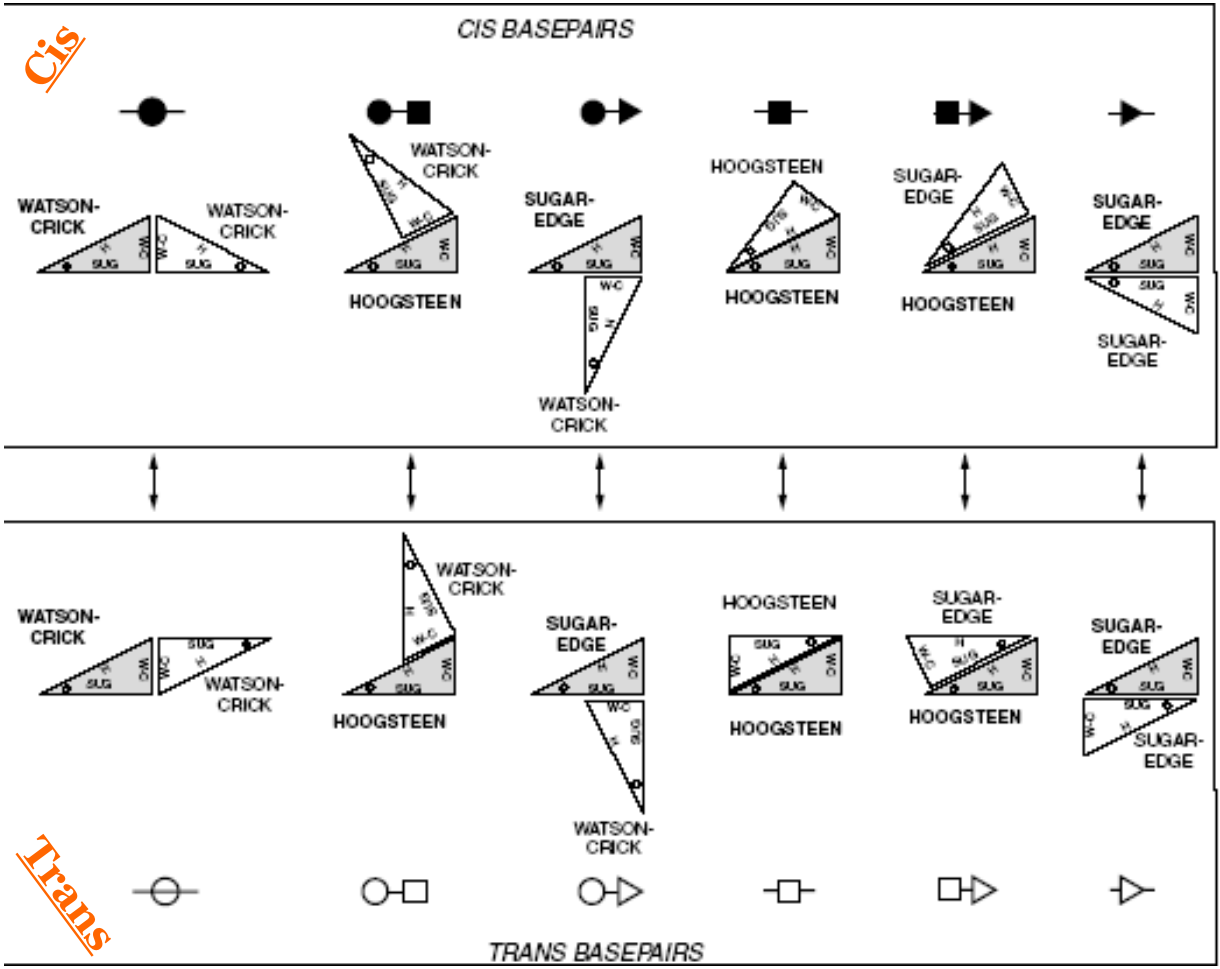


[Leontis & Westhof, RNA 2001]

Nomenclature Leontis Westhof (LW)

12 Familles





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

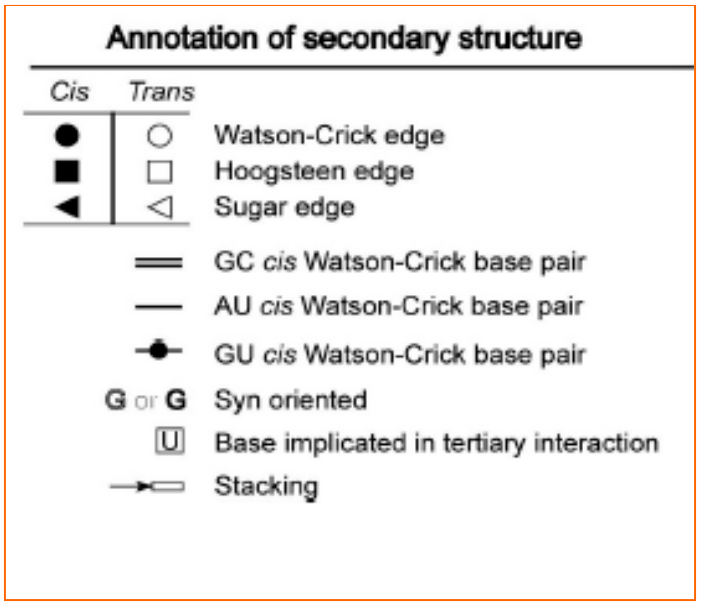
Nomenclature Leontis Westhof (LW)

Annotation

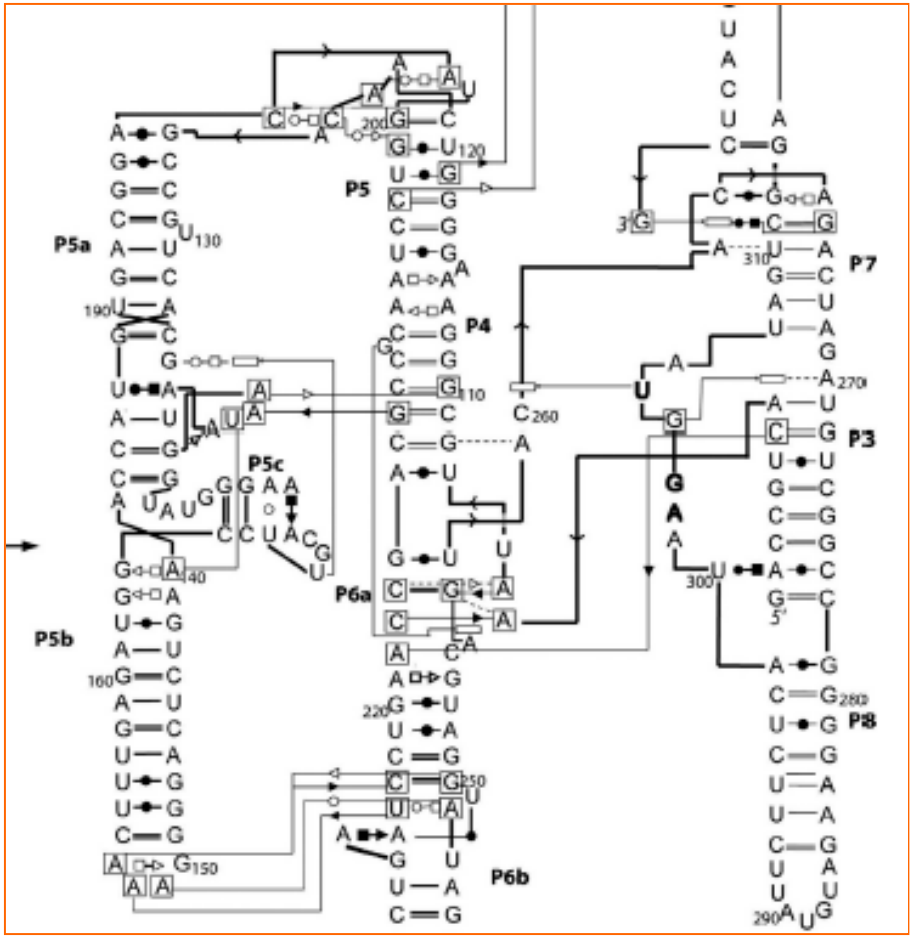
No.	Glycosidic bond orientation	Interacting edges	Symbol	Default local strand orientation
1	*cis*	Watson–Crick/Watson–Crick		Anti-parallel
2	*trans*	Watson–Crick/Watson–Crick		Parallel
3	*cis*	Watson–Crick/Hoogsteen		Parallel
4	*trans*	Watson–Crick/Hoogsteen		Anti-parallel
5	*cis*	Watson–Crick/Sugar edge		Anti-parallel
6	*trans*	Watson–Crick/Sugar edge		Parallel
7	*cis*	Hoogsteen/Hoogsteen		Anti-parallel
8	*trans*	Hoogsteen/Hoogsteen		Parallel
9	*cis*	Hoogsteen/Sugar edge		Parallel
10	*trans*	Hoogsteen/Sugar edge		Anti-parallel
11	*cis*	Sugar edge/Sugar edge		Anti-parallel
12	*trans*	Sugar edge/Sugar edge		Parallel

[The annotation of RNA Motifs, NB. Leontis & E. Westhof, Comparative and functional genomics, 2002]

Nomenclature Leontis Westhof (LW)

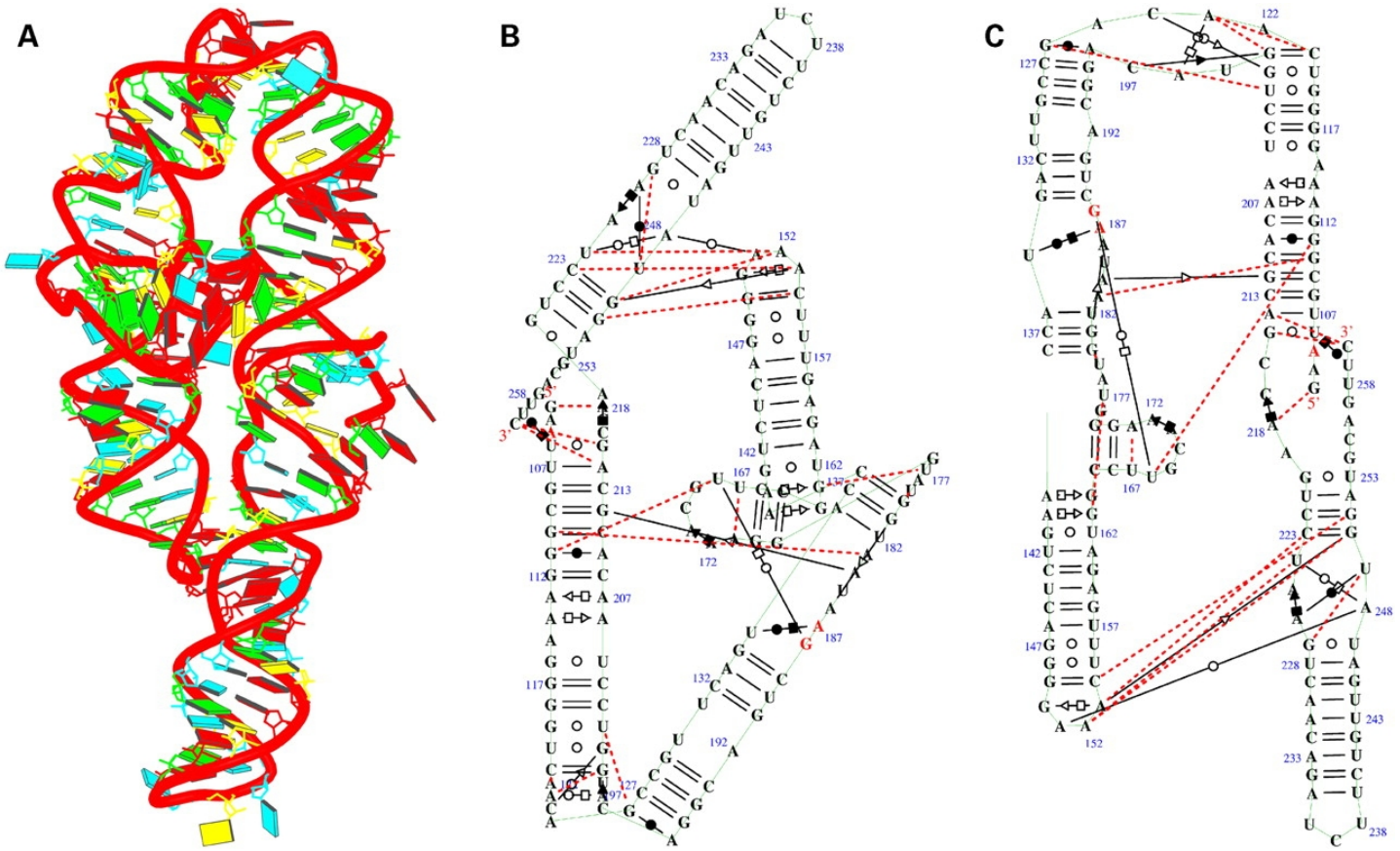


Graphe d'ARN = graphe de degré borné, étiqueté sur les sommets et sur les arêtes, contenant un chemin hamiltonien (connu).



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

De la 3D au graphe d'ARN



RNAView [H. Yang et al., NAR 2003]

(aussi: MC-annotate [P. Gendron et al., J Mol Biol 2001])

Comment l'ARN se replie-t-il ?

- Vers une conformation d'énergie libre minimale
- De façon hiérarchique :

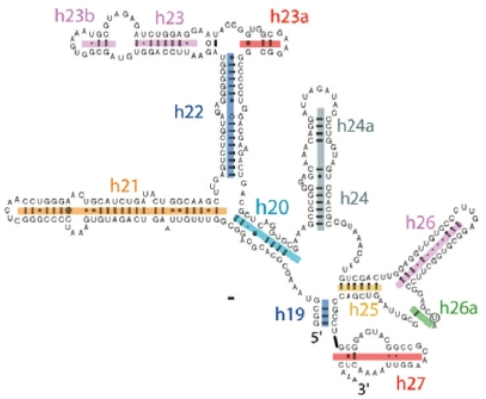
1. les interactions fortes et « locales »

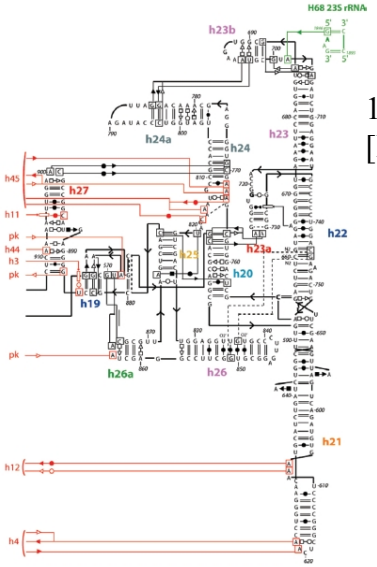
   = structure secondaire
   (sans pseudo-noeud)

2. les interactions faibles
   les interactions à longue portée

   = structure 3D

[Tinoco et al. J. Mol. Biol. 1999]

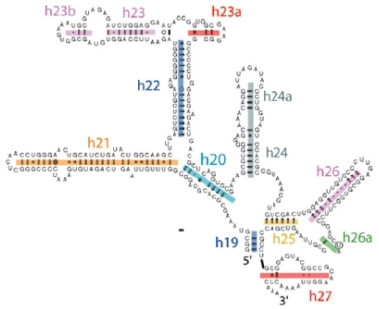




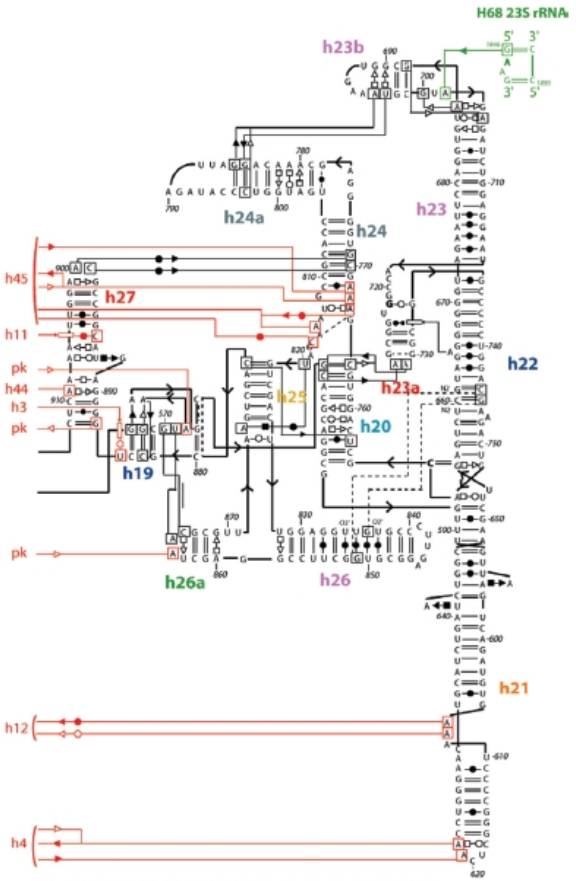
16S central domain
[Lescoute & Westhof 2006]

Interactions canoniques et non canoniques

• Les interactions canoniques forment les hélices et déterminent la structure secondaire.



• Les interactions non canoniques
  • forment les **motifs structuraux,**
  • sont responsables de la plupart des interactions entre les éléments de la structure secondaire.

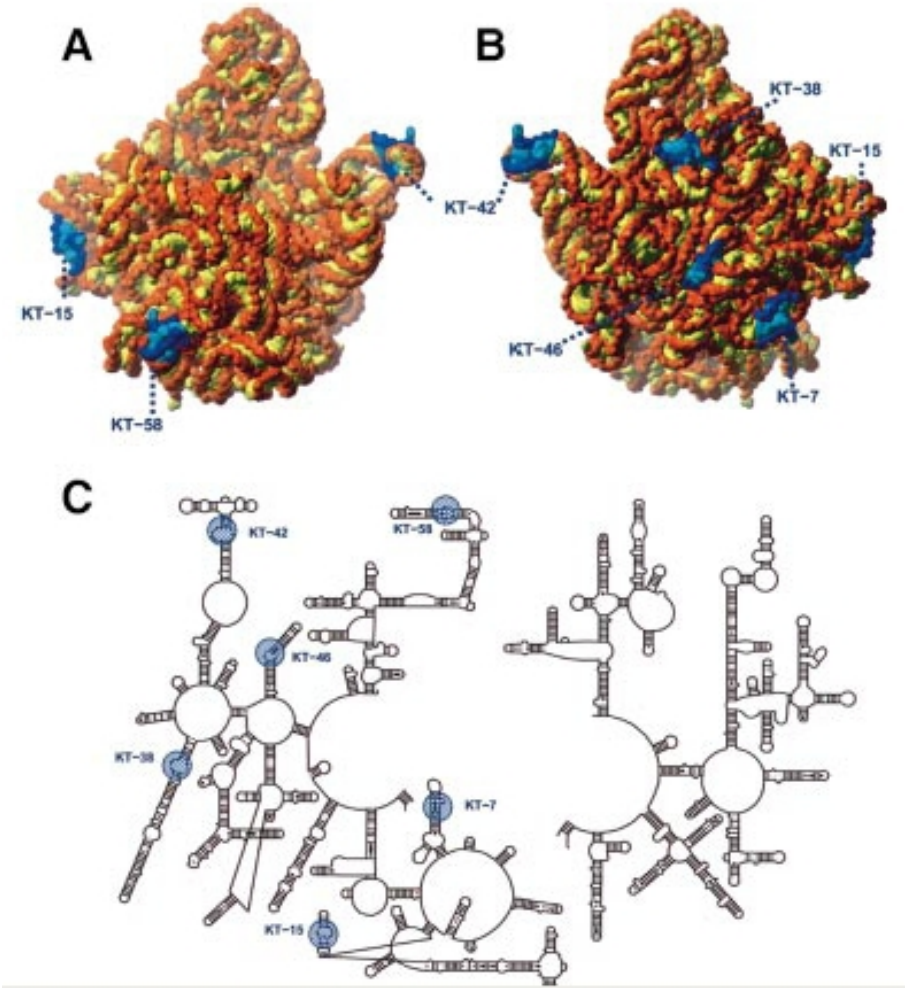


16S central domain
[Lescoute & Westhof 2006]

Motifs structuraux

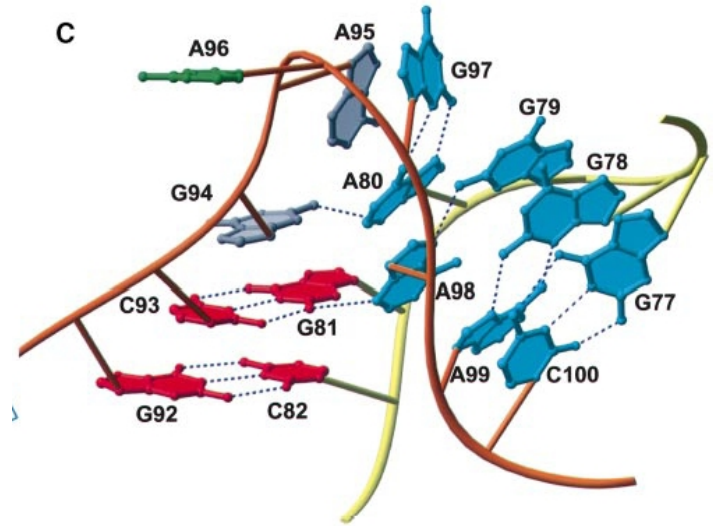
« Small, recurrent, stacked arrays of isosteric basepairs that intersperse the 2D structural elements (internal, junction, terminal loops) and fold into essentially identical 3D structures »

[Leontis & Westhof 2001]

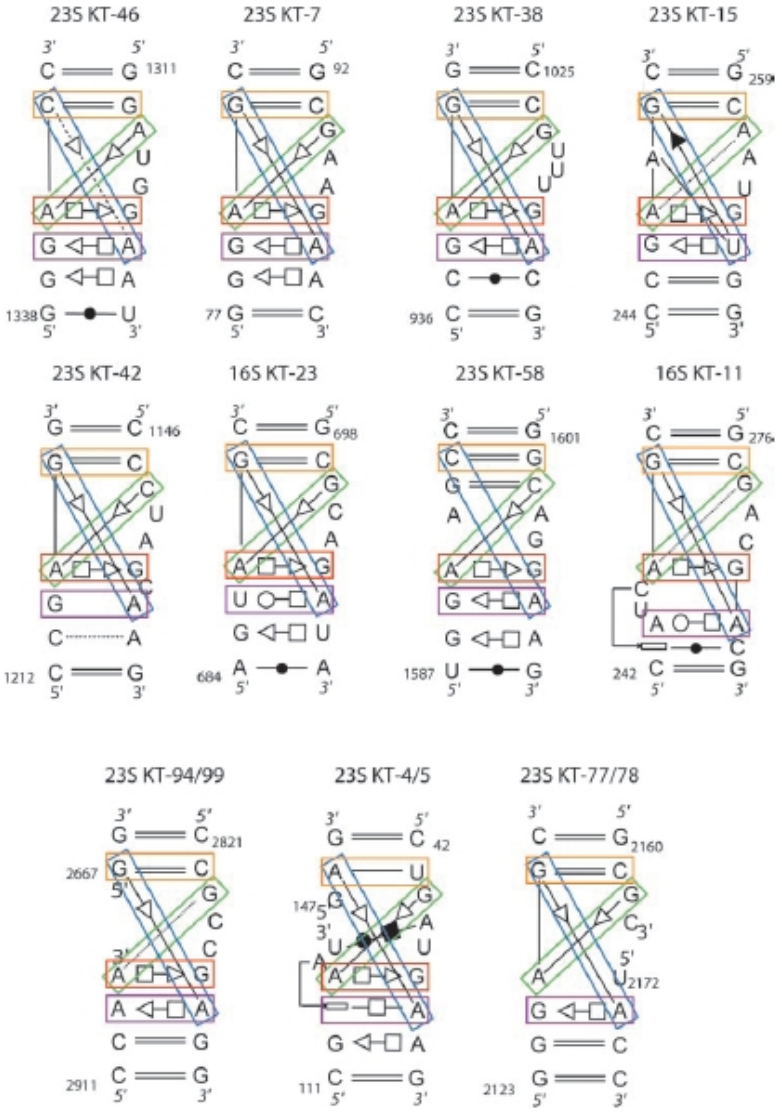


[The kink-turn: a new RNA secondary structure motif, D.J. Klein et al., The EMBO journal, 2001]

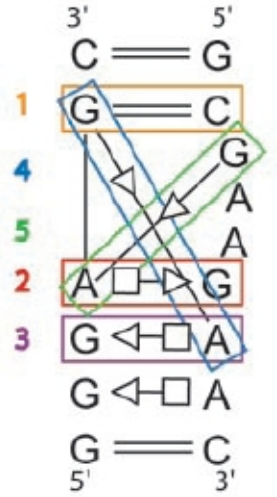
Exemple : le Kink-turn



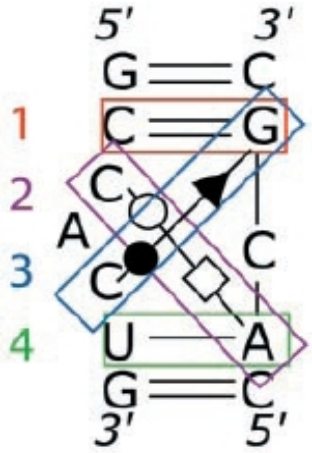
rmsd=1.7 Å



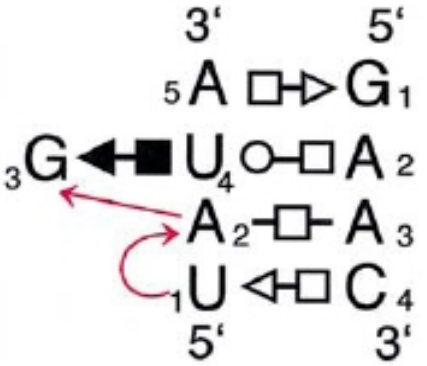
Quelques motifs structuraux



Kink-turn



C-loop



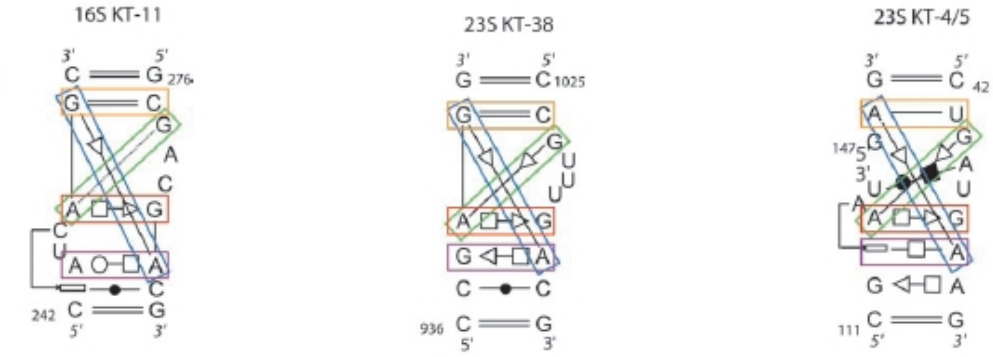
Sarcin-ricin

Découvrir de nouveaux motifs ?

Les motifs connus ont été découverts dans les structures PDB

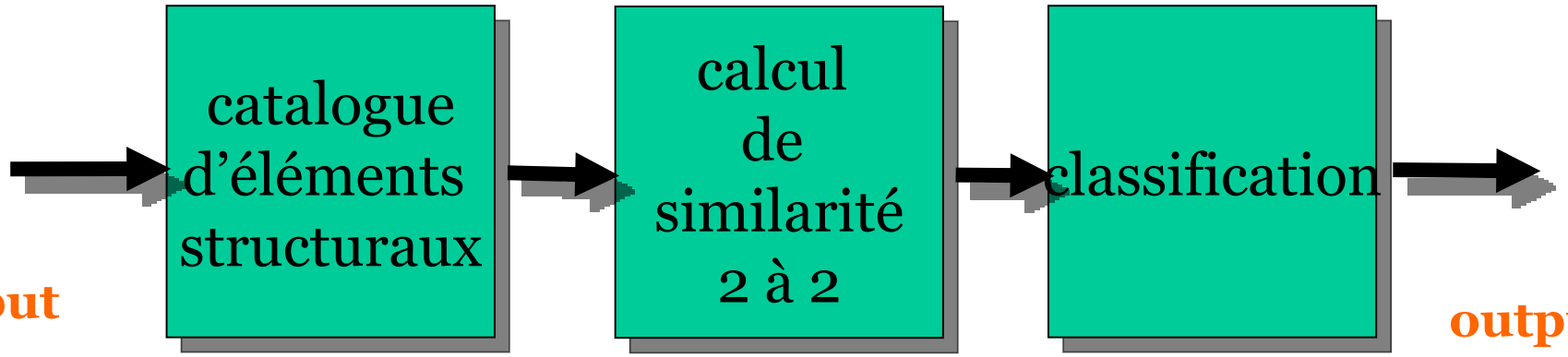
- soit à l'oeil,

- soit automatiquement par comparaison des formes géométriques. Approche très sensible aux différences de structure primaire/secondaire (insertions/suppressions)



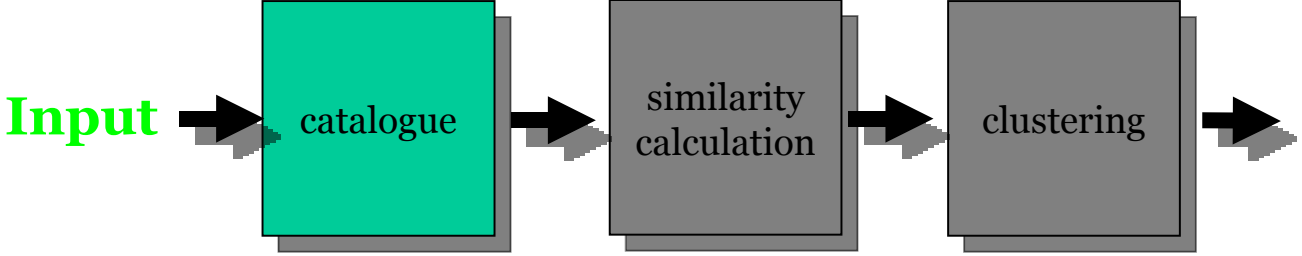
⇨ Une nouvelle approche basée sur le graphe de la structure.

Principe général

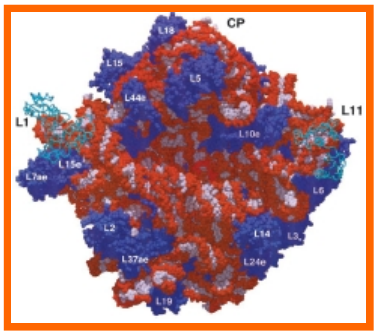


RNA graph
(from PDB file)

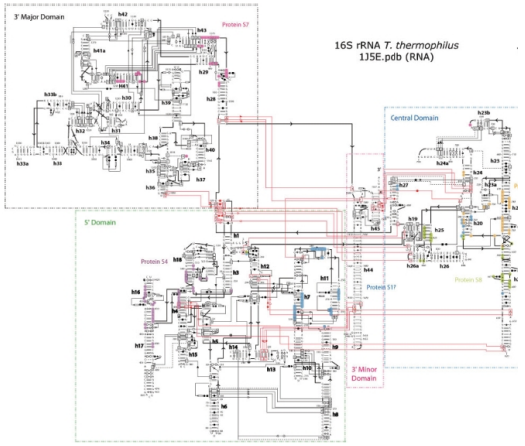
Potential RNA motifs



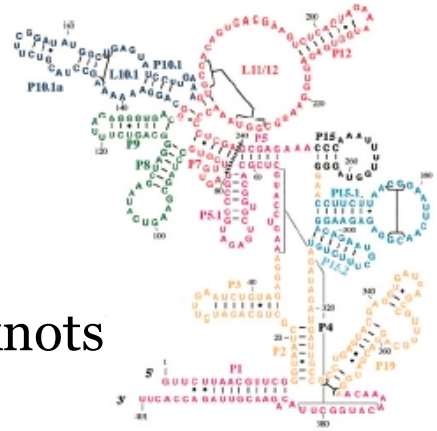
3D structure (PDB file)



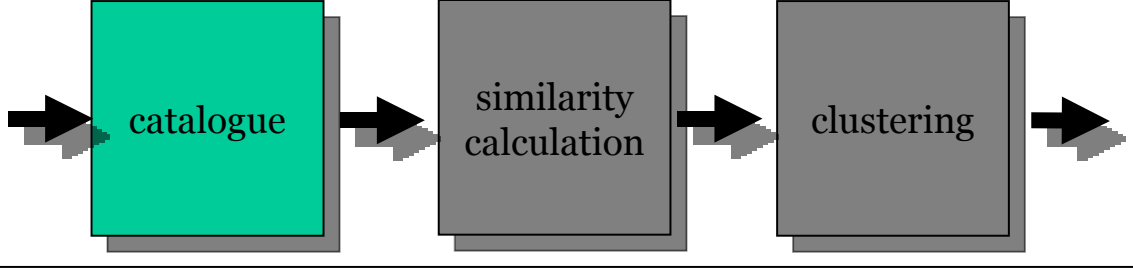
RNA graph



1. Remove pseudoknots + non-WC basepairs



2D structure without pseudoknots

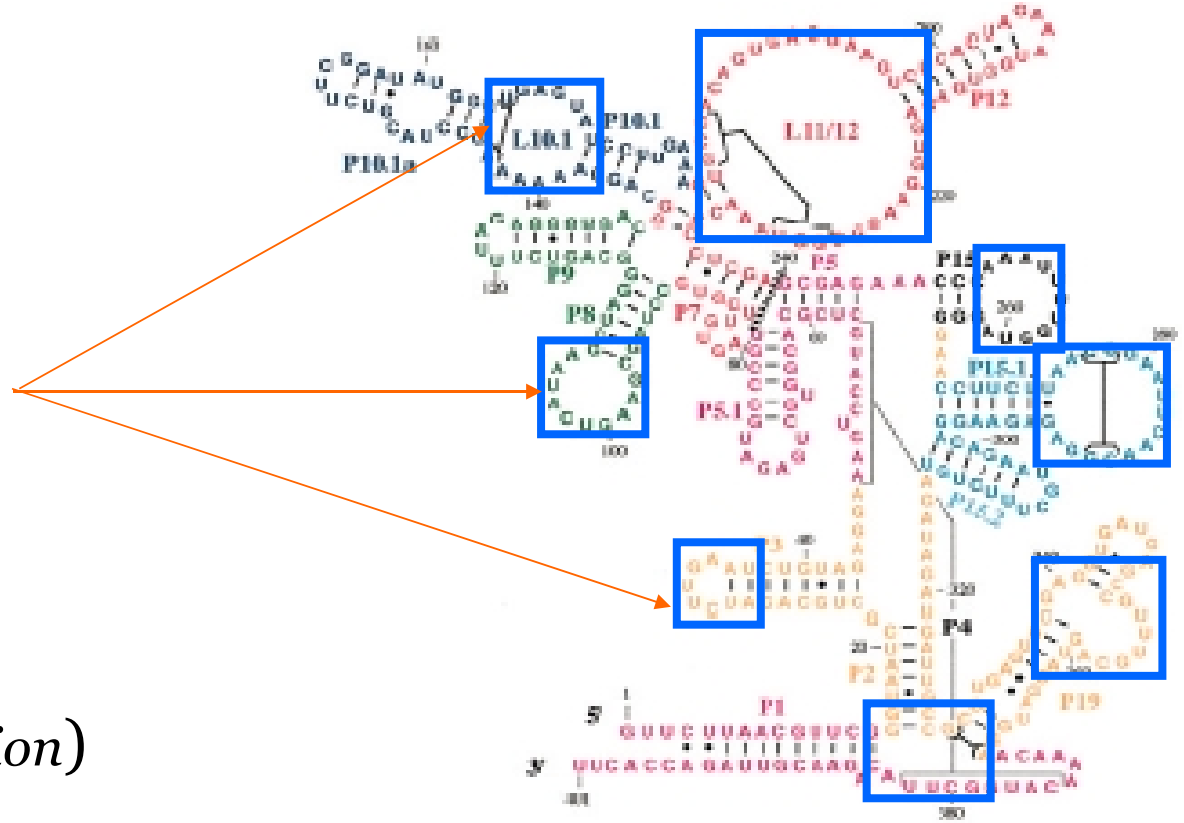


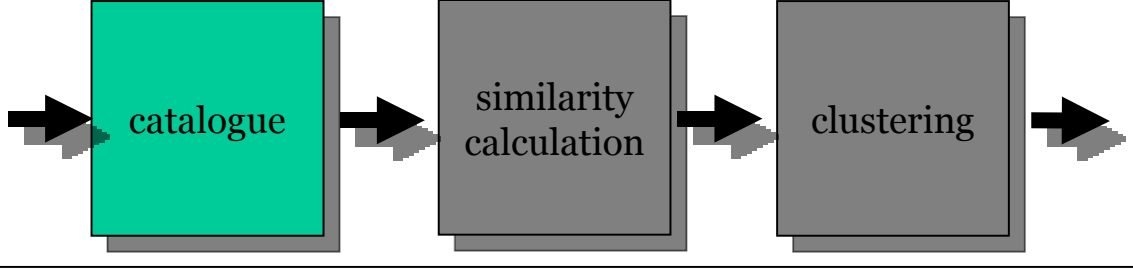
2. Identify 2D structural elements

- Bulges
- Internal
- Junction
- Terminal

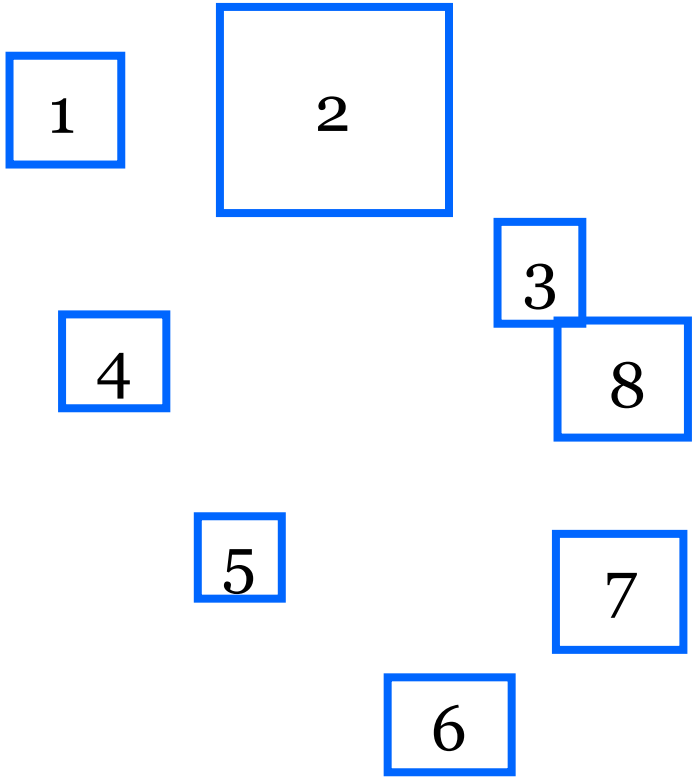
} loops

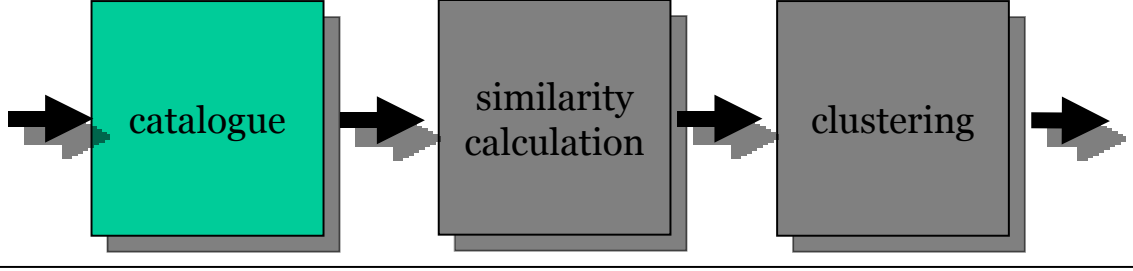
(*use tree representation*)





3. For each 2D structural element, restore local non-WC basepairs

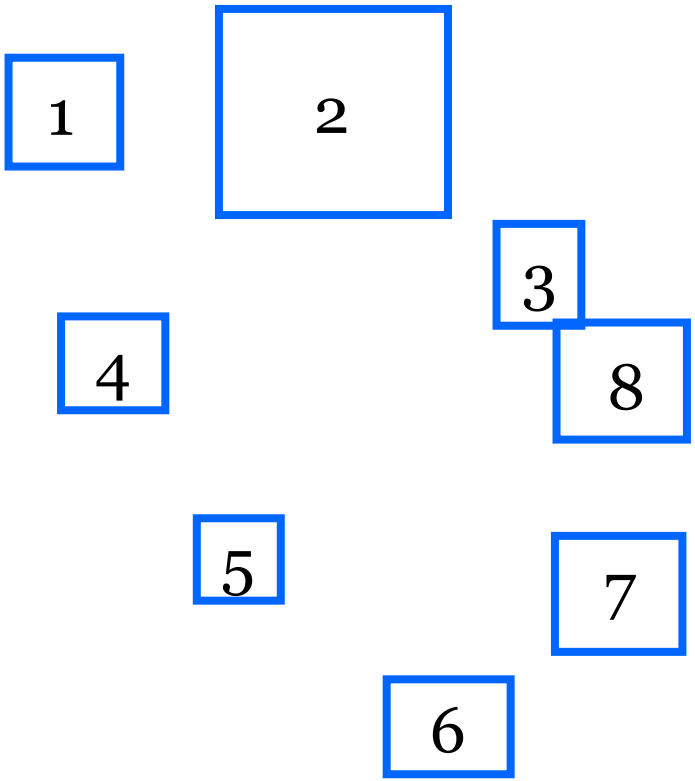


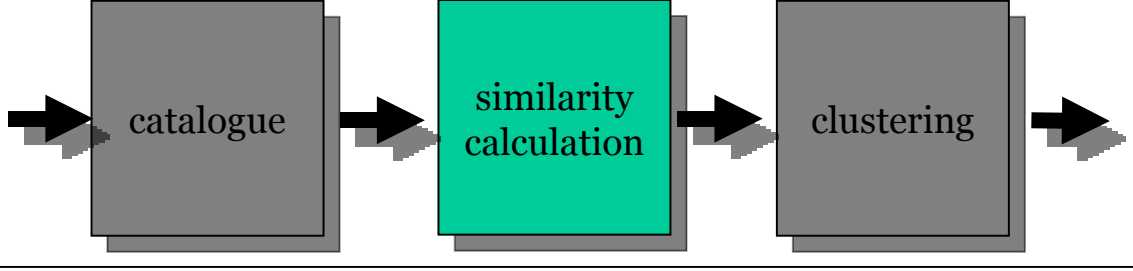


Output

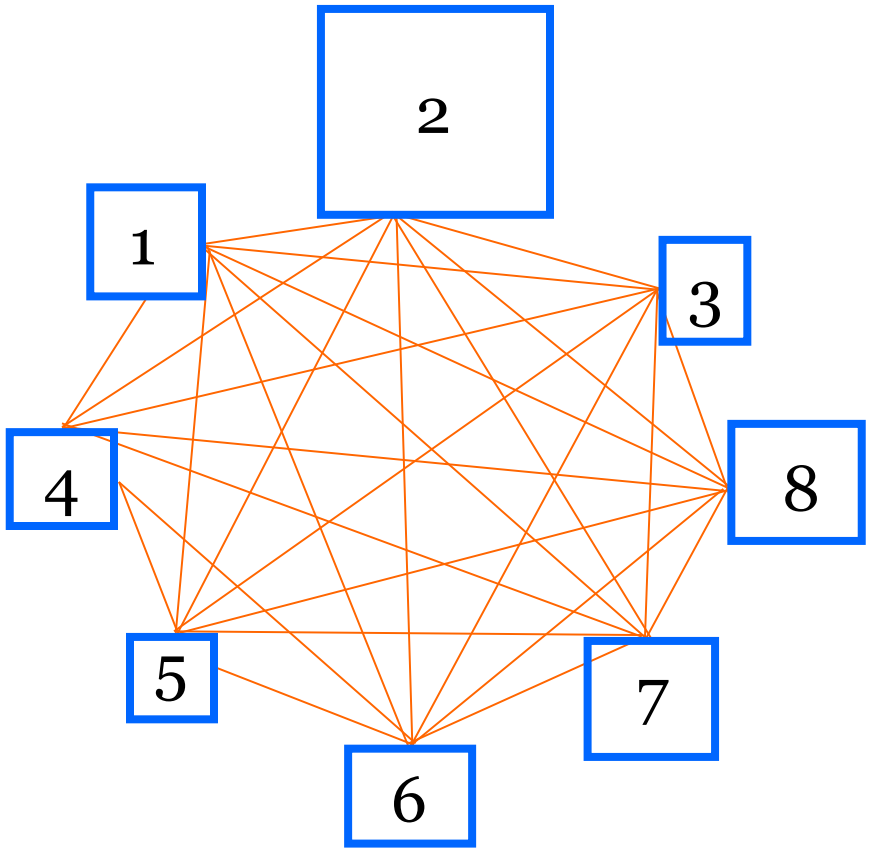
catalogue
List of RNA subgraphs

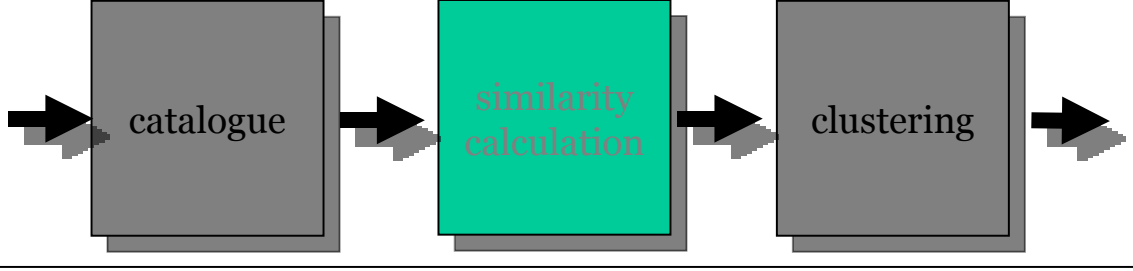
3. For each 2D structural element, restore local non-WC basepairs





4. Calculate a pairwise similarity measure between subgraphs i and j

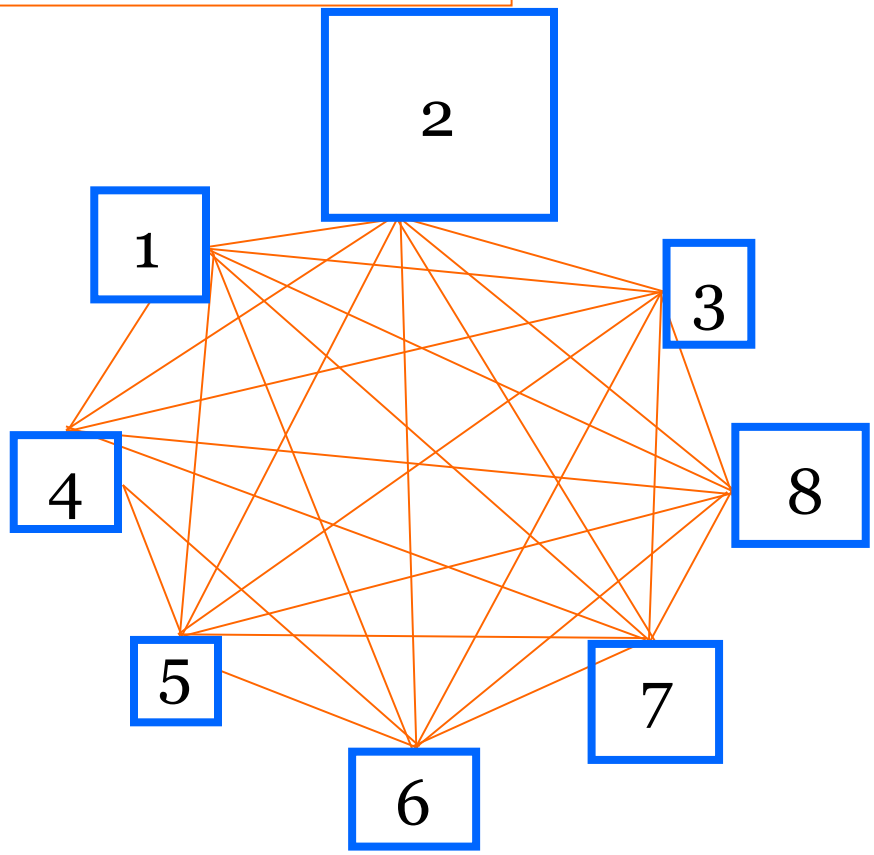


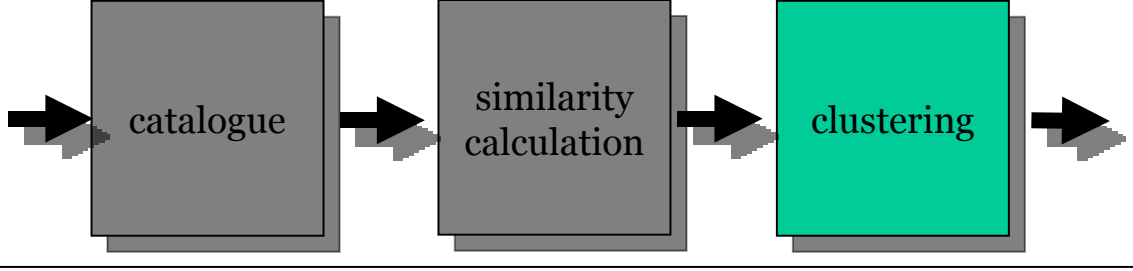


Output

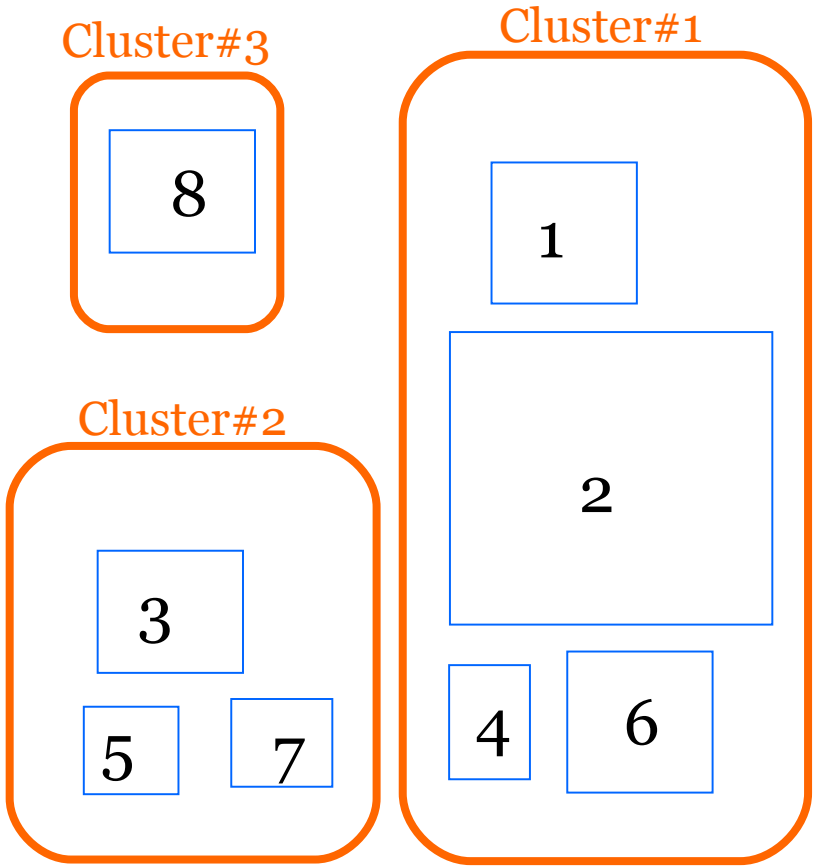
Similarity matrix

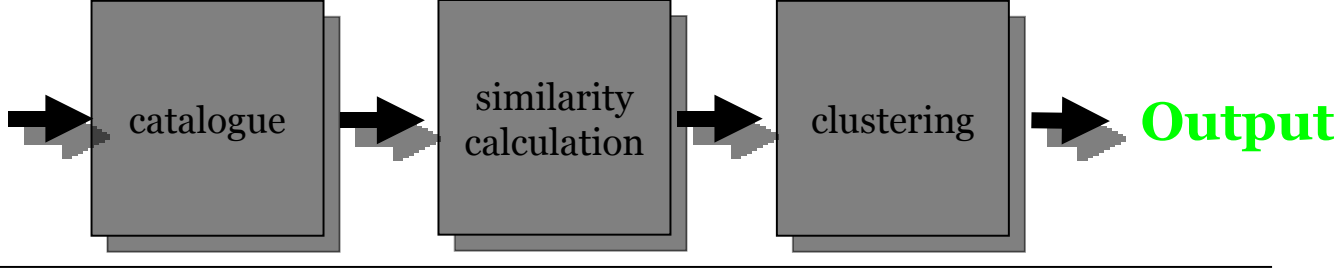
4. Calculate a pairwise similarity measure between subgraphs i and j





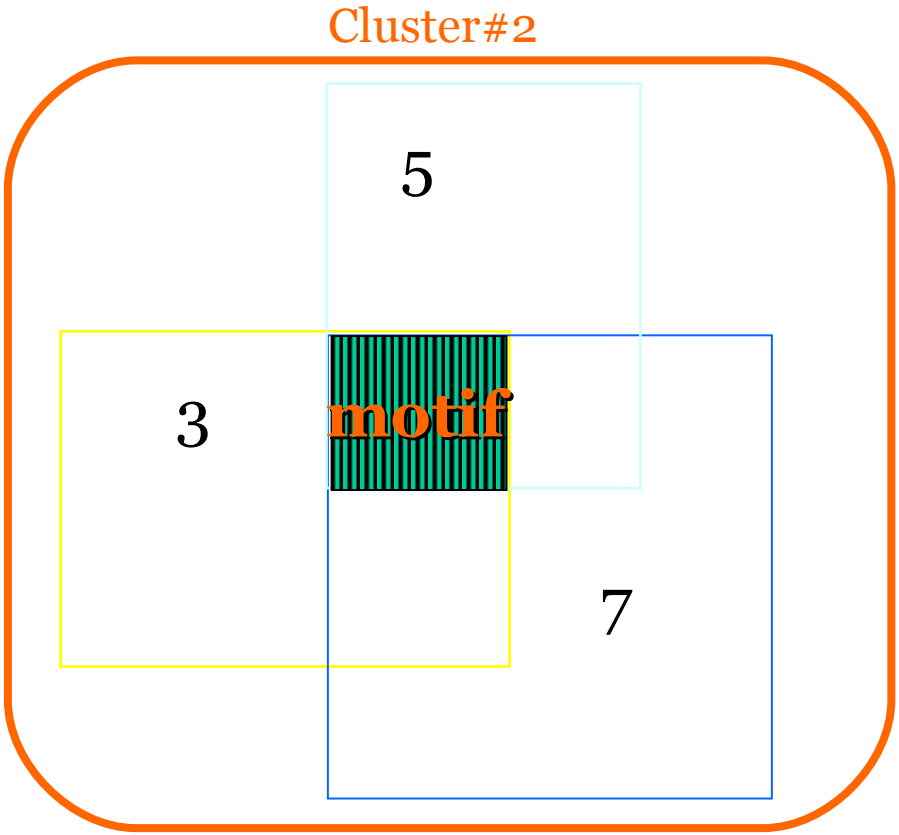
5. Cluster subgraphs with **high** similarity value



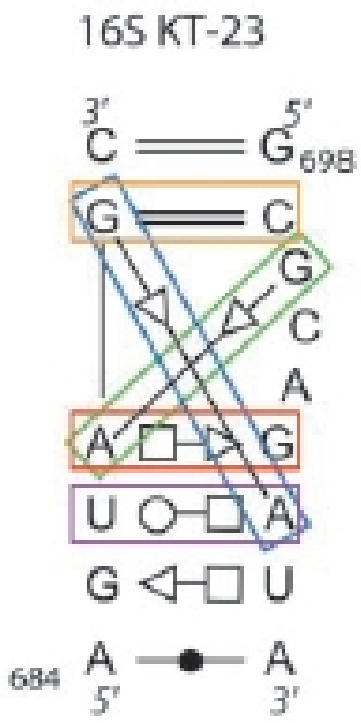


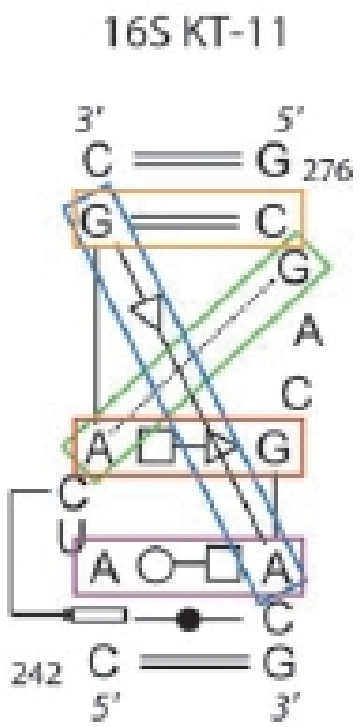
motif :

subgraph common to members of the cluster



Example

Similarity (

, 

) **= ?**

➔ *some definitions …*

Definitions

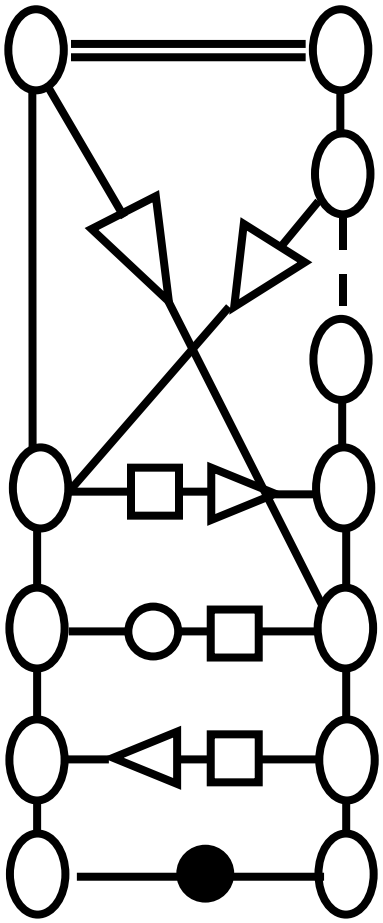
Non-canonical size of G :

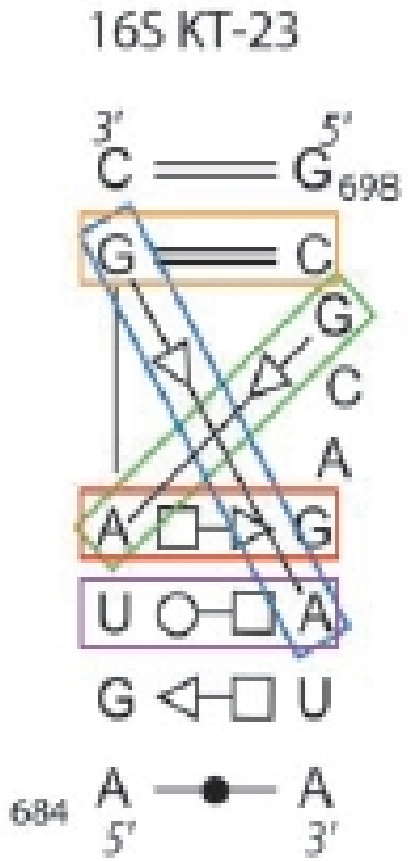
|| G || = Number of its non-canonical edges

Example:

|| 16S KT-23 || = 6

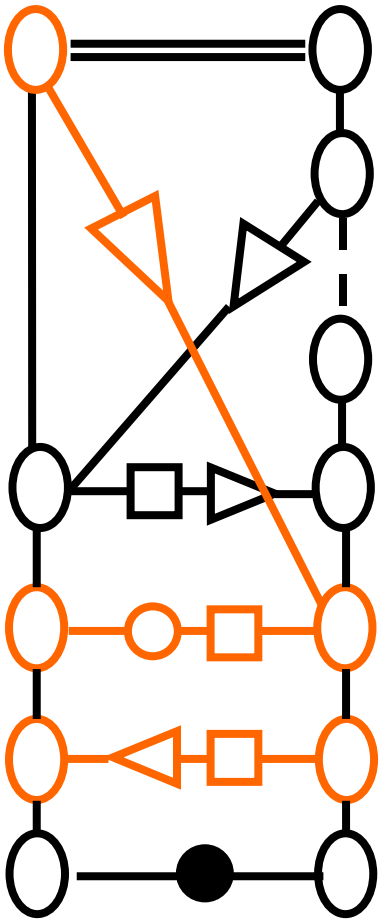
16S KT-23





Definitions

16S KT-23

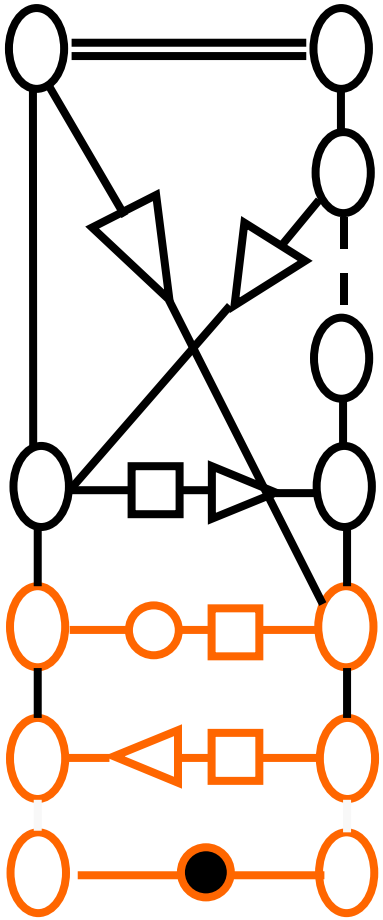


A **Non-canonical subgraph**
=
Subgraph whose all edges are non-canonical

Example 1

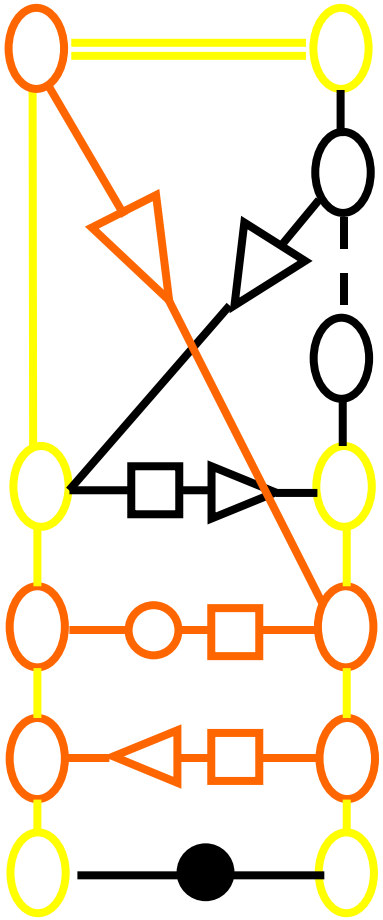
Example 2

16S KT-23



Definitions

16S KT-23



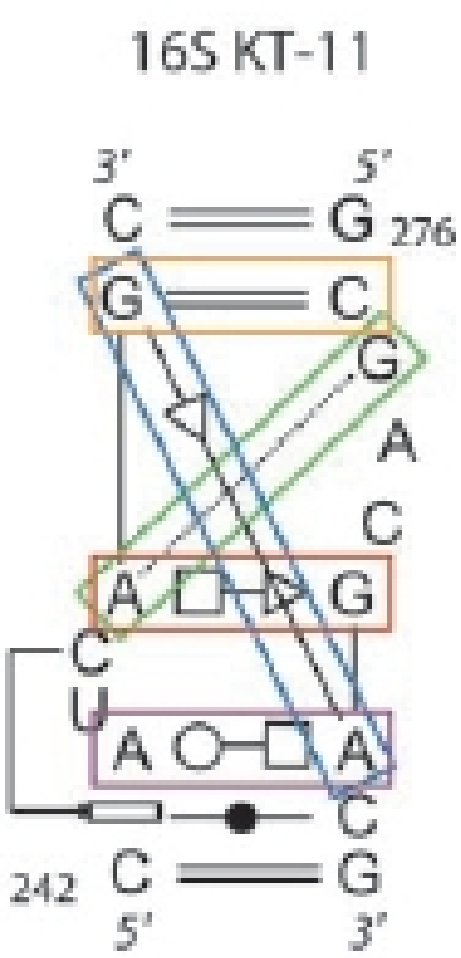
A **completion** of a non-canonical subgraph H of G

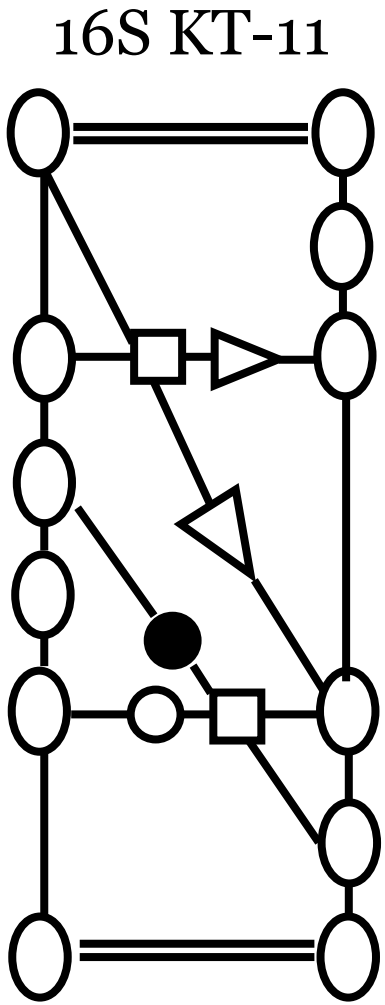
=

H + {all canonical + backbone edges of G having at least one end in H}

Definitions

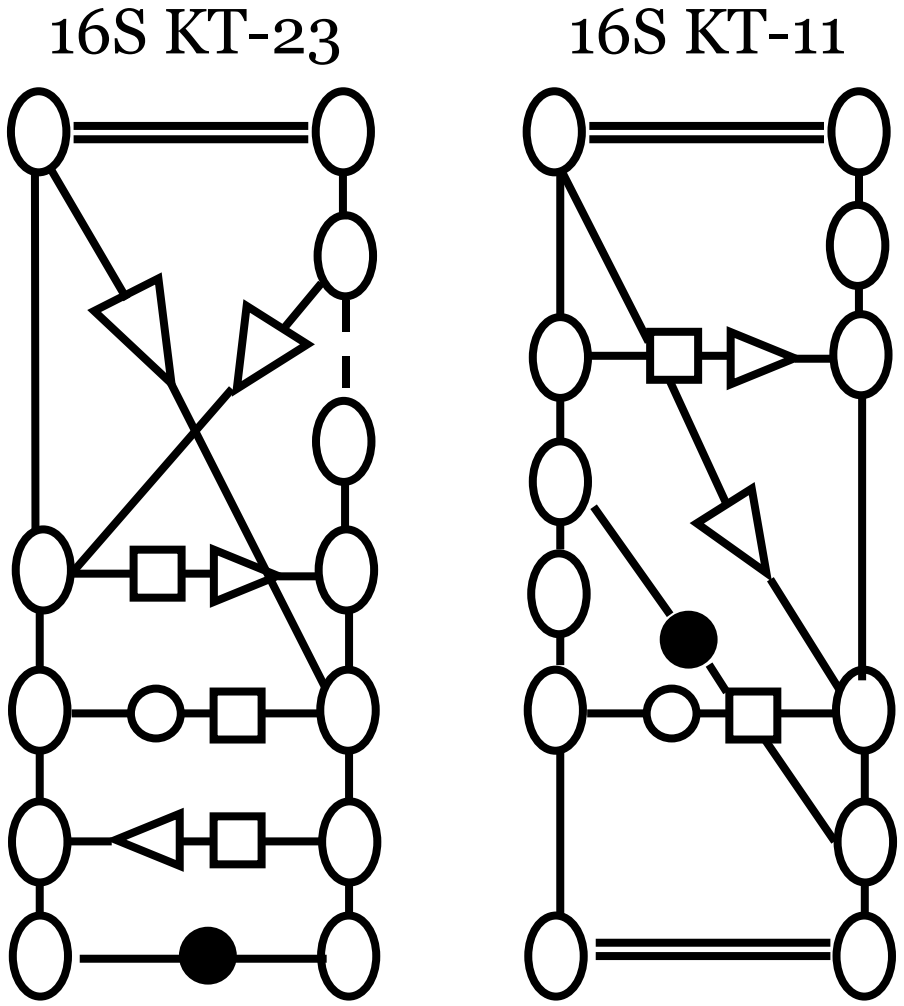
Another graph…





Definitions

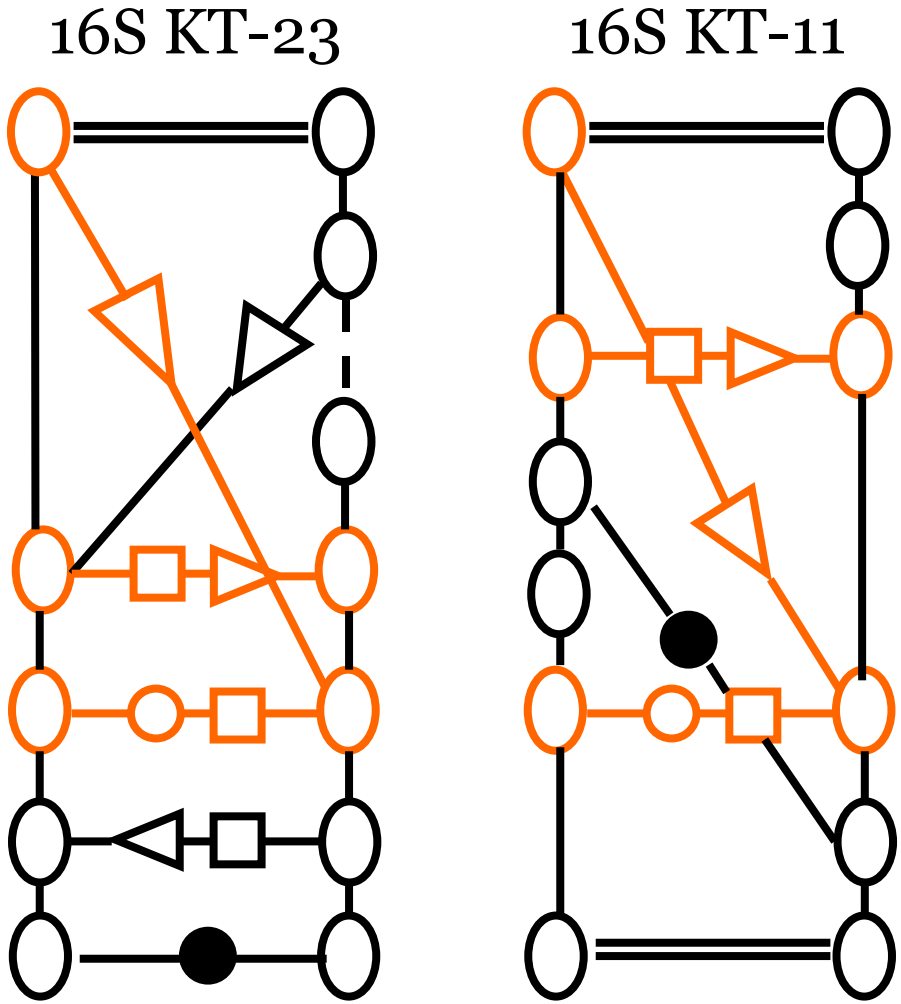
A **common non-canonical** subgraph H to G1 and G2 ....



Definitions

A **common non-canonical** subgraph H to G1 and G2

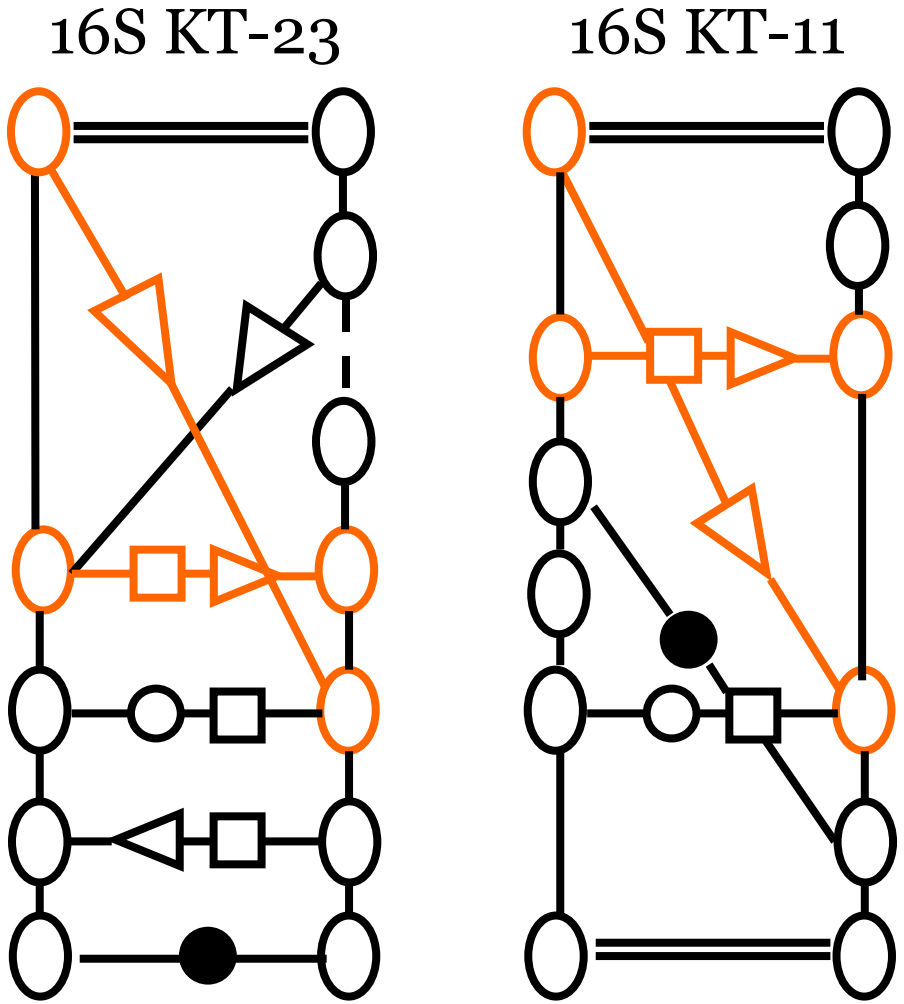
Example 1



Definitions

A **common non-canonical** subgraph H to G1 and G2

Example 2



Definitions

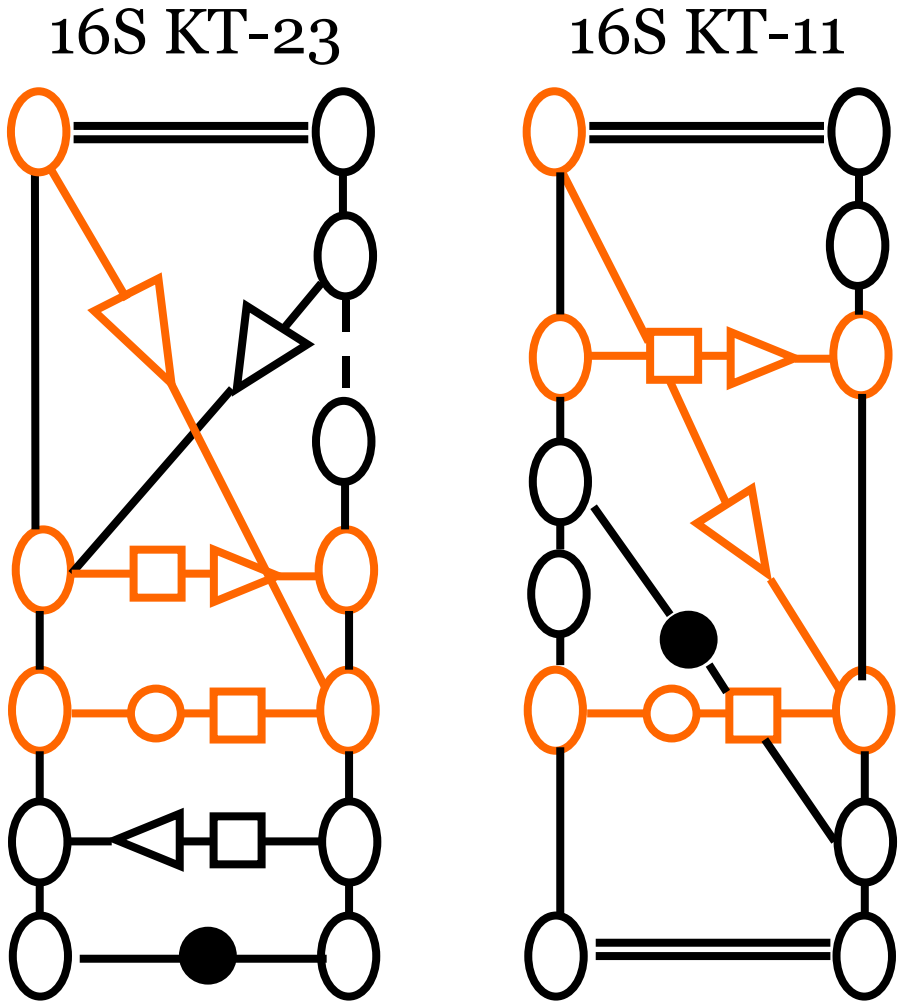
An **extensible common non-canonical** subgraph to G1 and G2

=

A **non-canonical** subgraph **common** to G1 and G2 whose **completions** in G1 and G2 are **isomorphic**

Example 1:

- common non-canonical subgraph

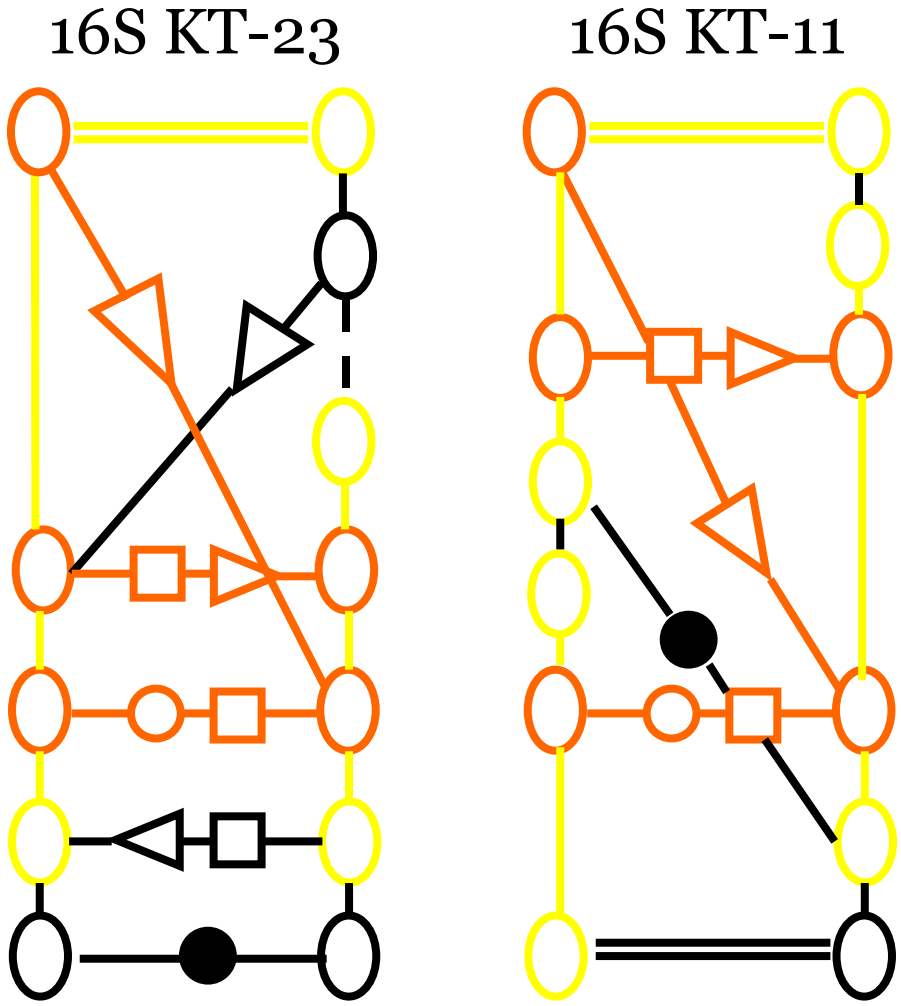


Definitions

An **extensible common non-canonical** subgraph to G1 and G2

=

A **non-canonical** subgraph **common** to G1 and G2 whose **completions** in G1 and G2 are **isomorphic**



Example 1:

- common non-canonical subgraph
- completions

Definitions

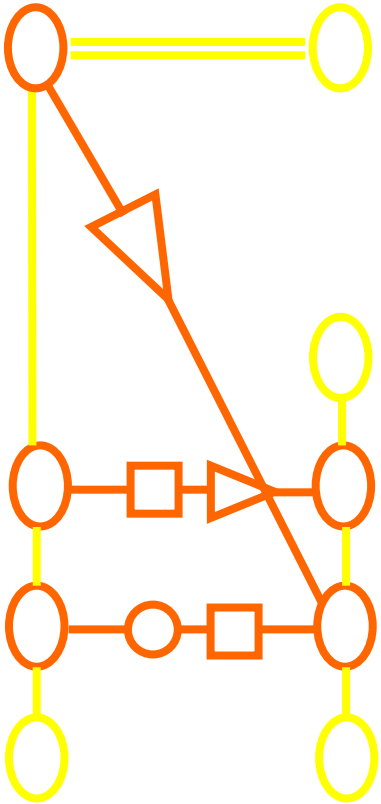
An **extensible common non-canonical** subgraph to G1 and G2

=

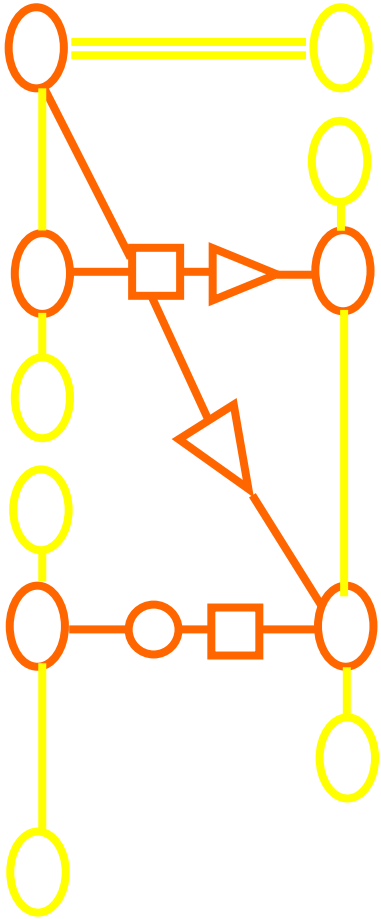
A **non-canonical** subgraph **common** to G1 and G2 whose **completions** in G1 and G2 are **isomorphic**

Example 1:

16S KT-23



16S KT-11

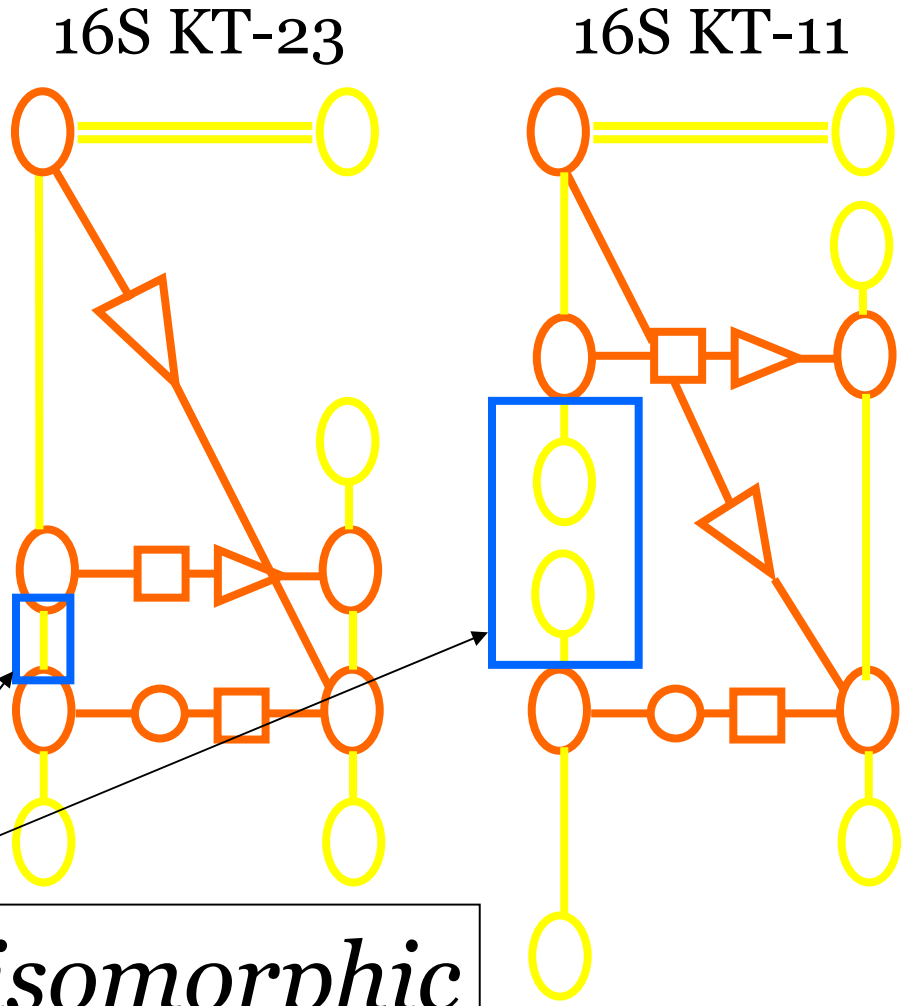


Definitions

An **extensible common non-canonical** subgraph to G1 and G2

=

A **non-canonical** subgraph **common** to G1 and G2 whose **completions** in G1 and G2 are **isomorphic**

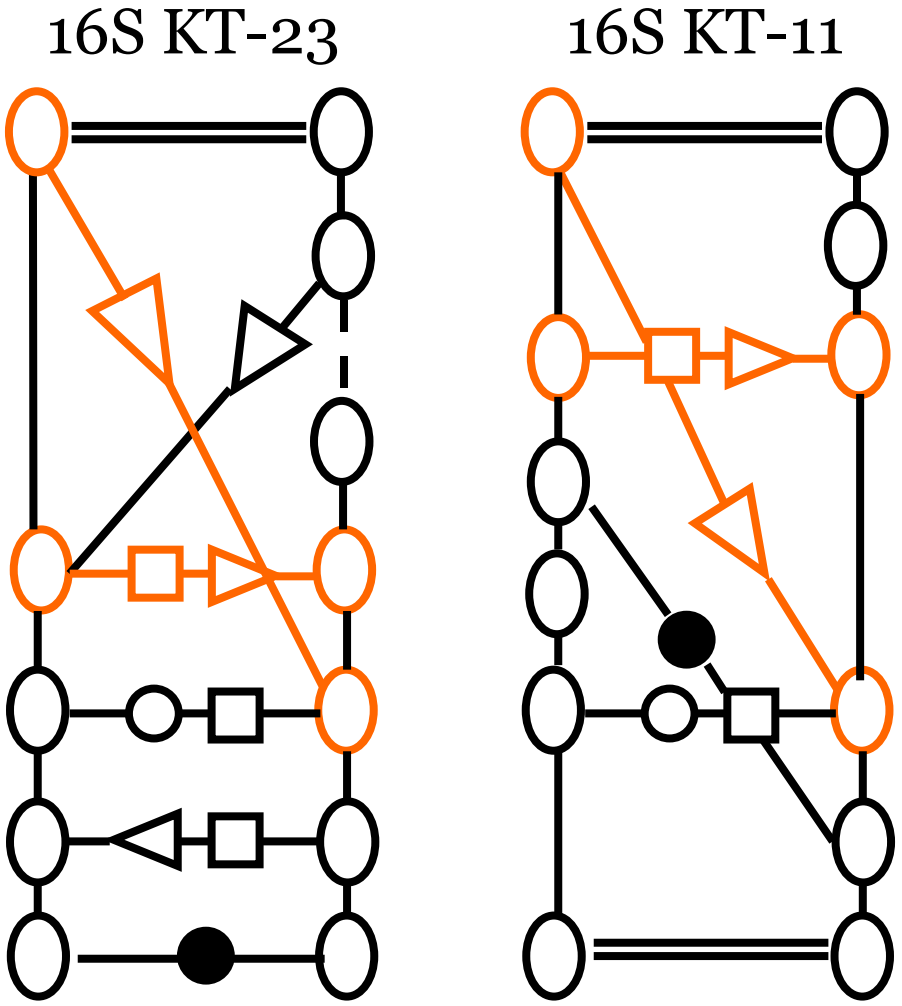


Example 1:

Completions NOT isomorphic

Definitions

An **extensible common non-canonical** subgraph to G1 and G2

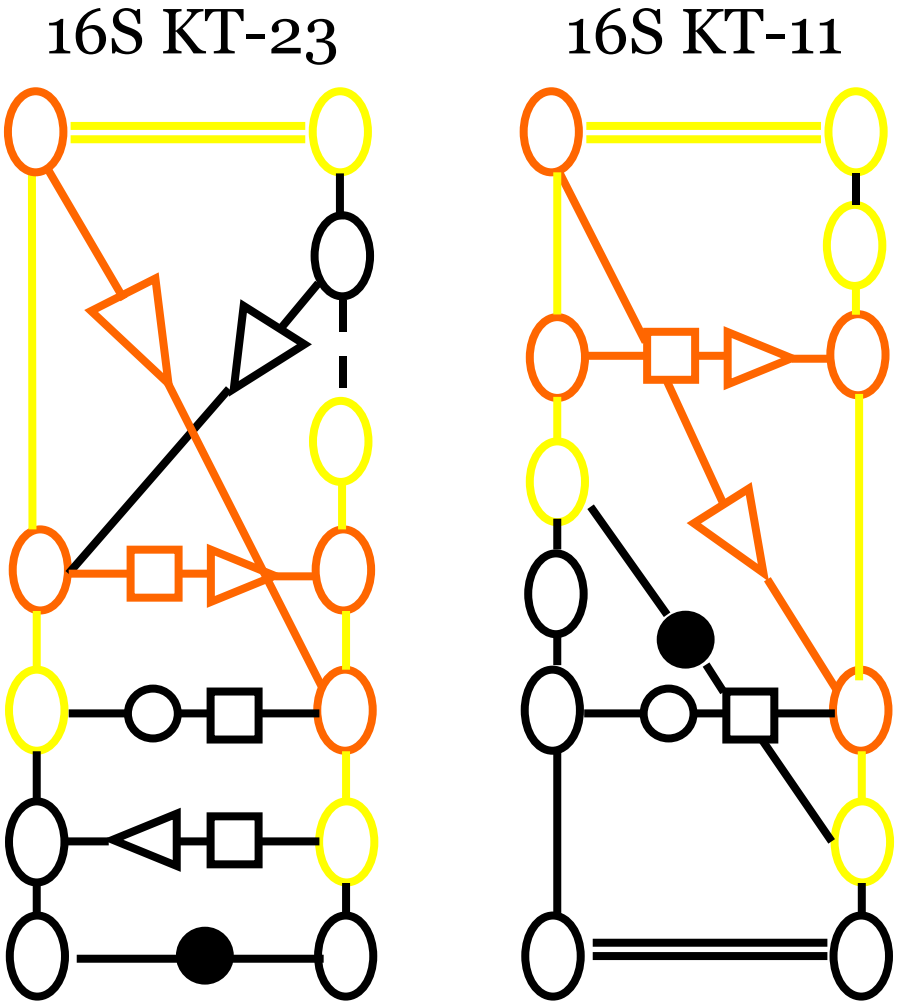


Example 2:

- common non-canonical subgraph

Definitions

An **extensible common non-canonical** subgraph to G1 and G2

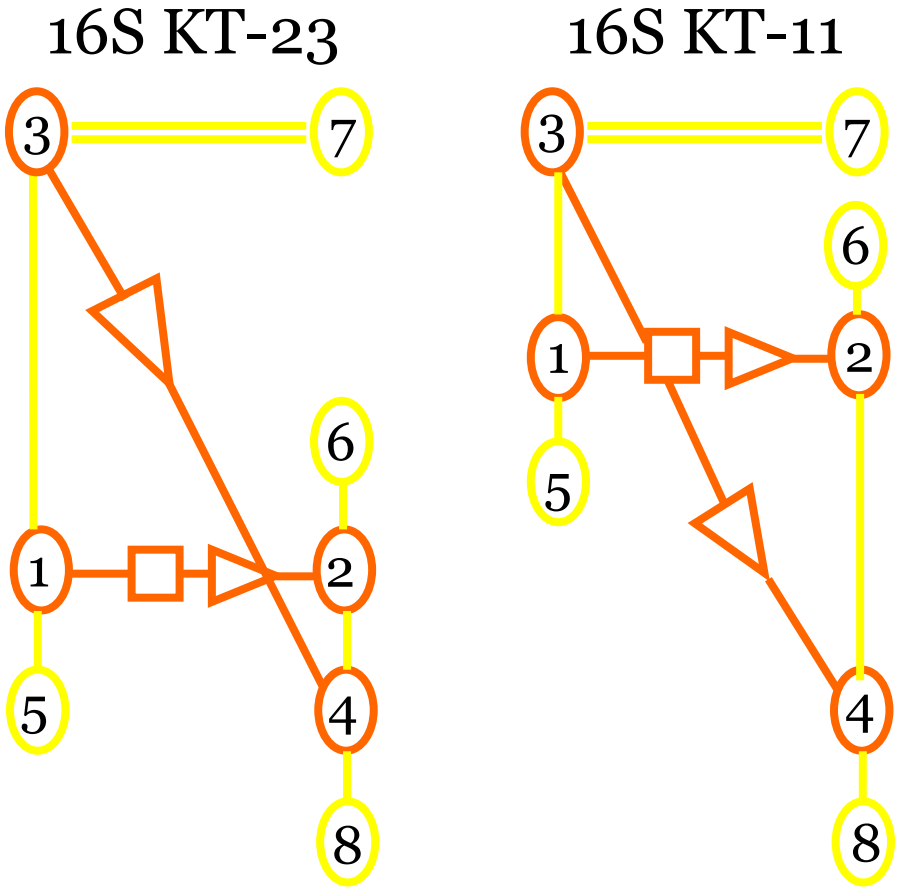


Example 2:

- common non-canonical subgraph
- completions

Definitions

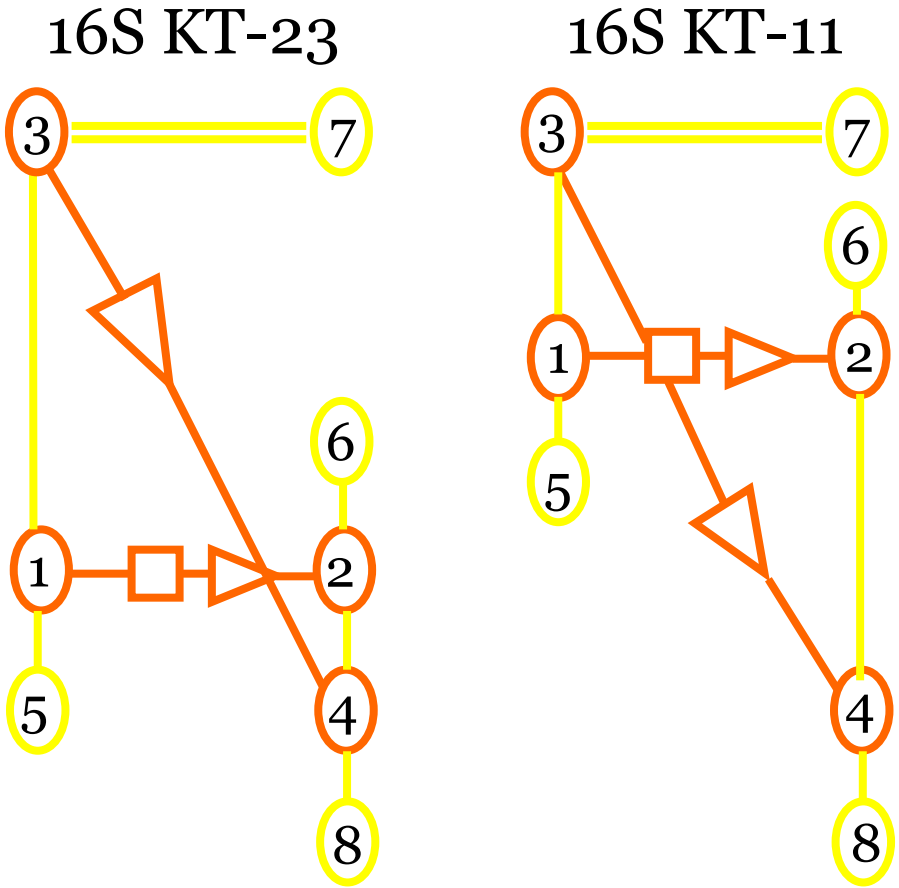
An **extensible common non-canonical** subgraph to G1 and G2



Example 2:

Definitions

An **extensible common non-canonical** subgraph to G1 and G2



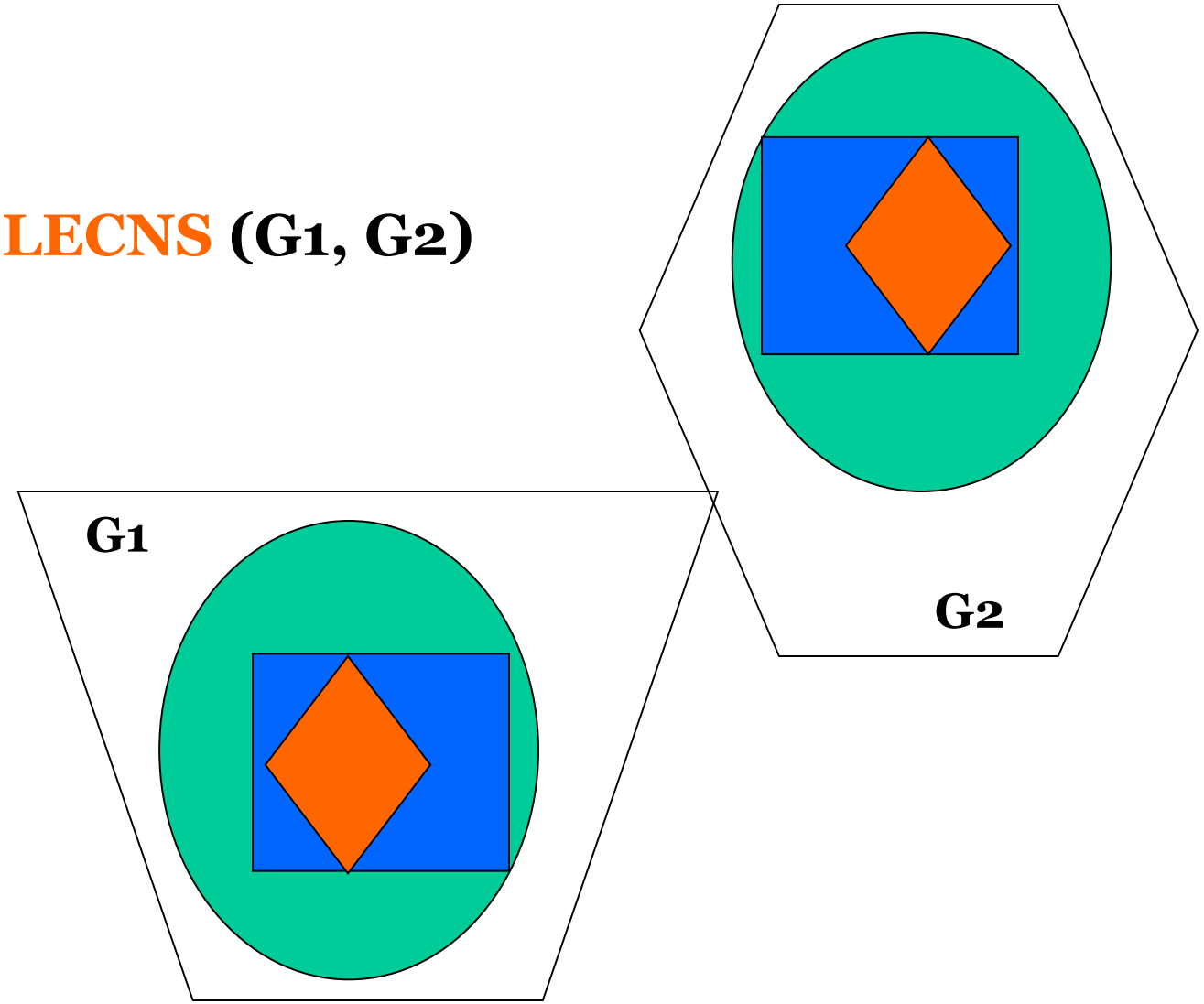
Example 2:

Completions ARE isomorphic

Definitions

Largest
Extensible
Common
Non-canonical
Subgraph

LECNS (G1, G2)

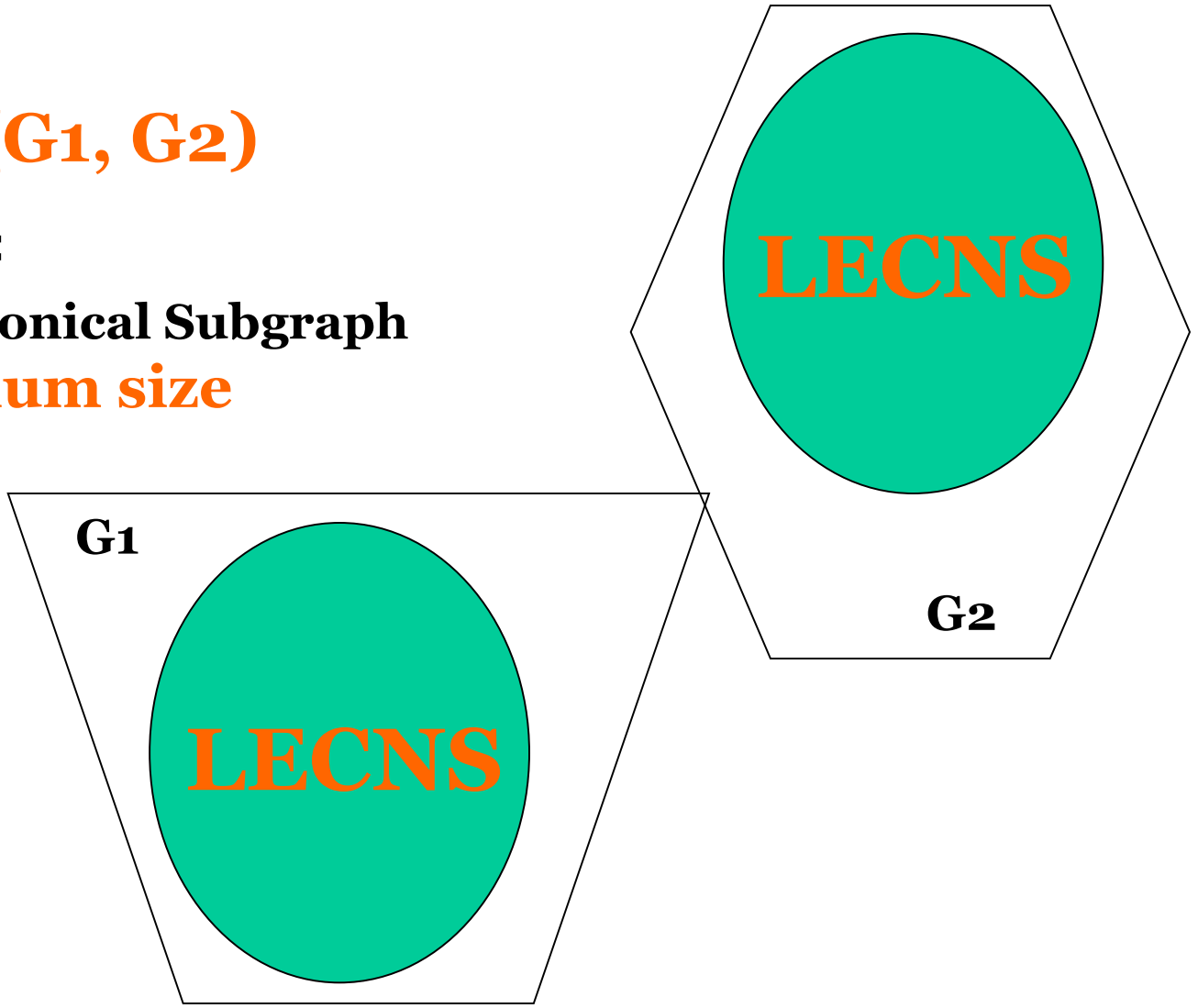


Definitions

LECNS (G1, G2)

=

Common Non-canonical Subgraph of maximum size



Measure

$$\text{Sim}(G1, G2) = \frac{|| \text{LECNS}(G1, G2) ||}{\max(|| G1 ||, || G2 ||)}$$

Measure

$$Sim(G1, G2) = \frac{|| LECNS(G1, G2) ||}{\max(|| G1 ||, || G2 ||)}$$

Properties:

- $0 \leq sim \leq 1$
- $sim(G1, G2) = sim(G2, G1)$

Learning set

catalogue of *H.marismortui.23S (reference structure)*

209 elements

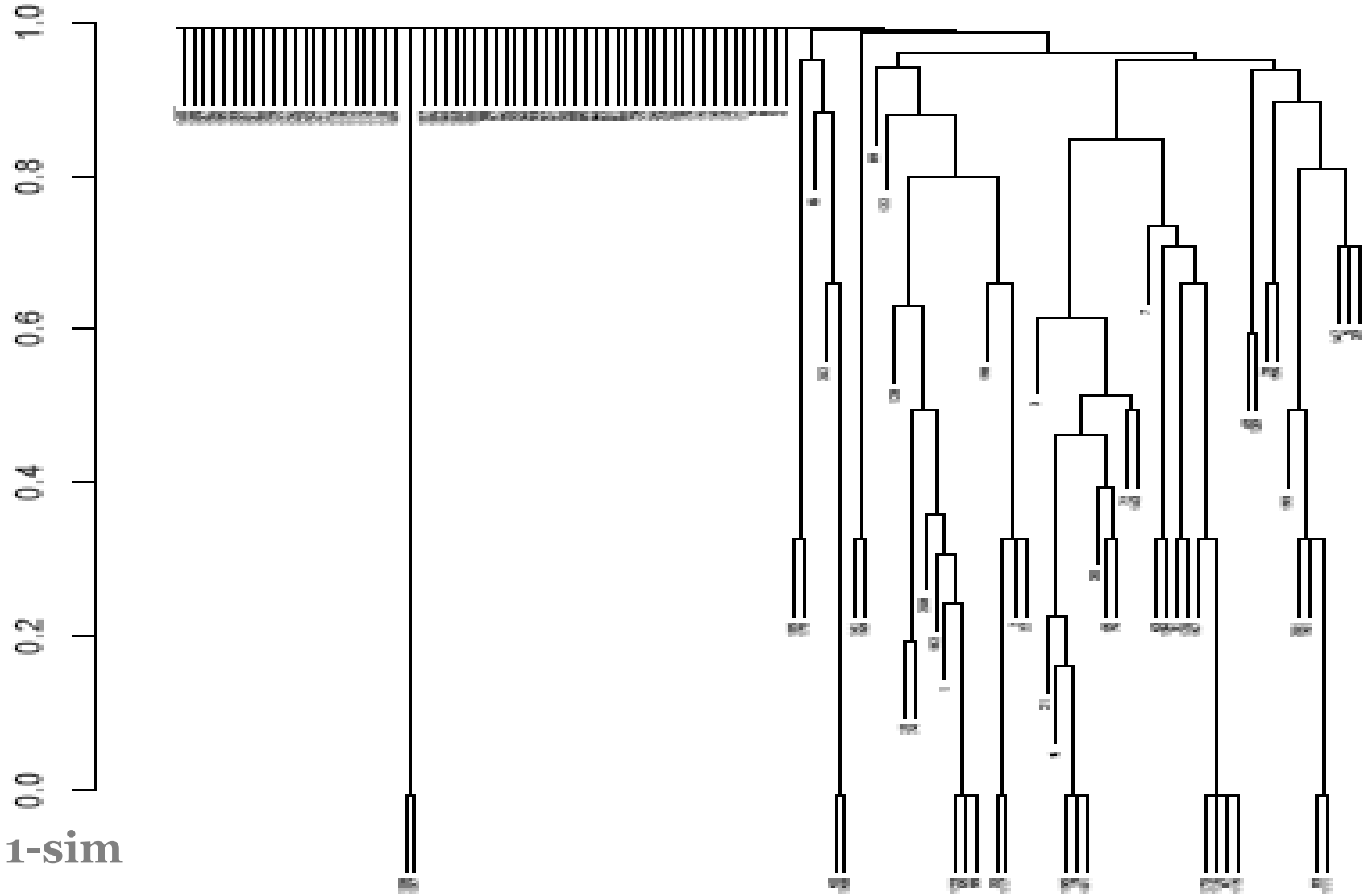
Method

Hierarchical clustering with average linkage (*UPGMA algorithm*)

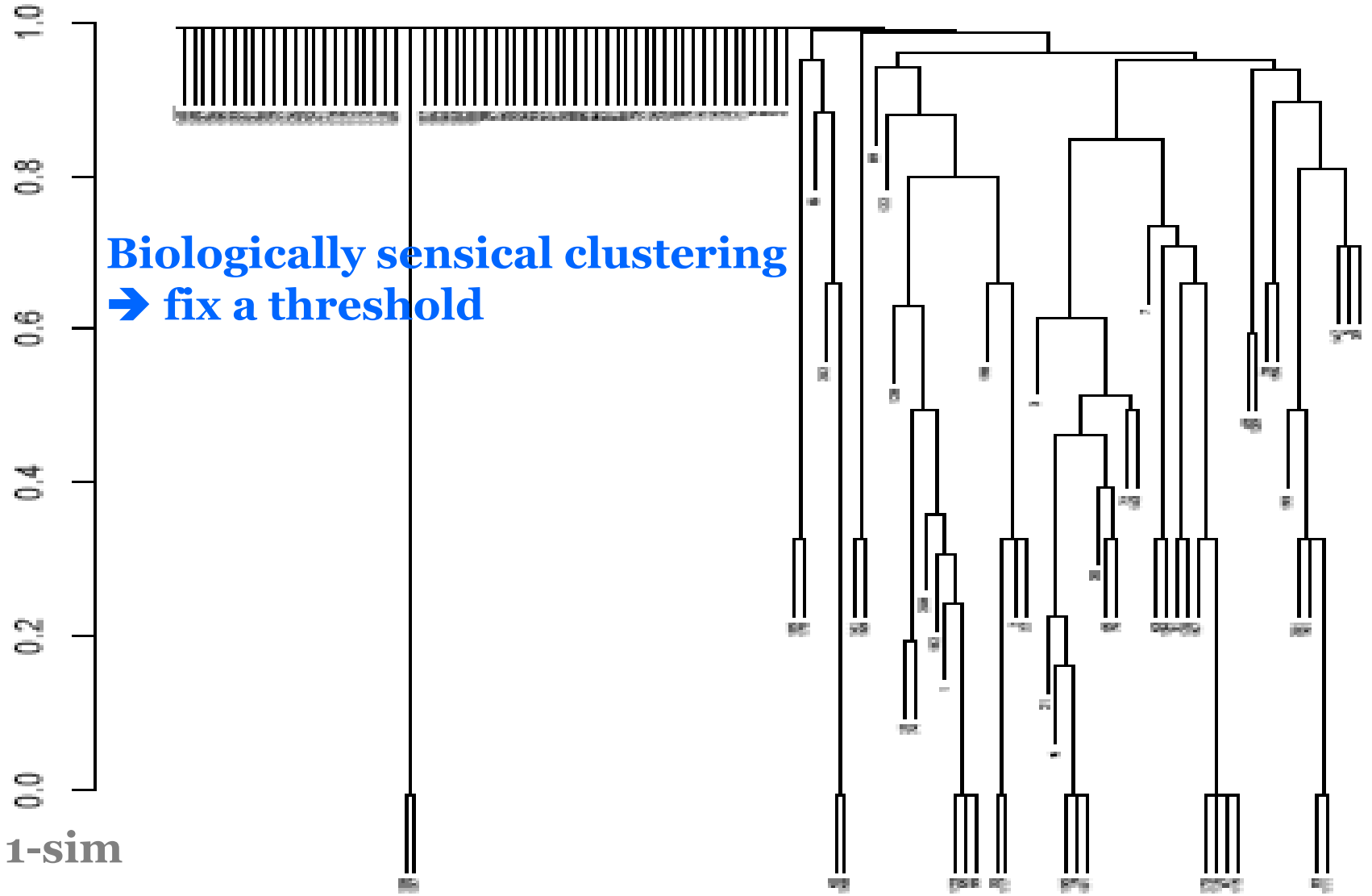
Output

Dendrogram of 2D structural elements

Hierarchical dendrogram *H.m 23S*

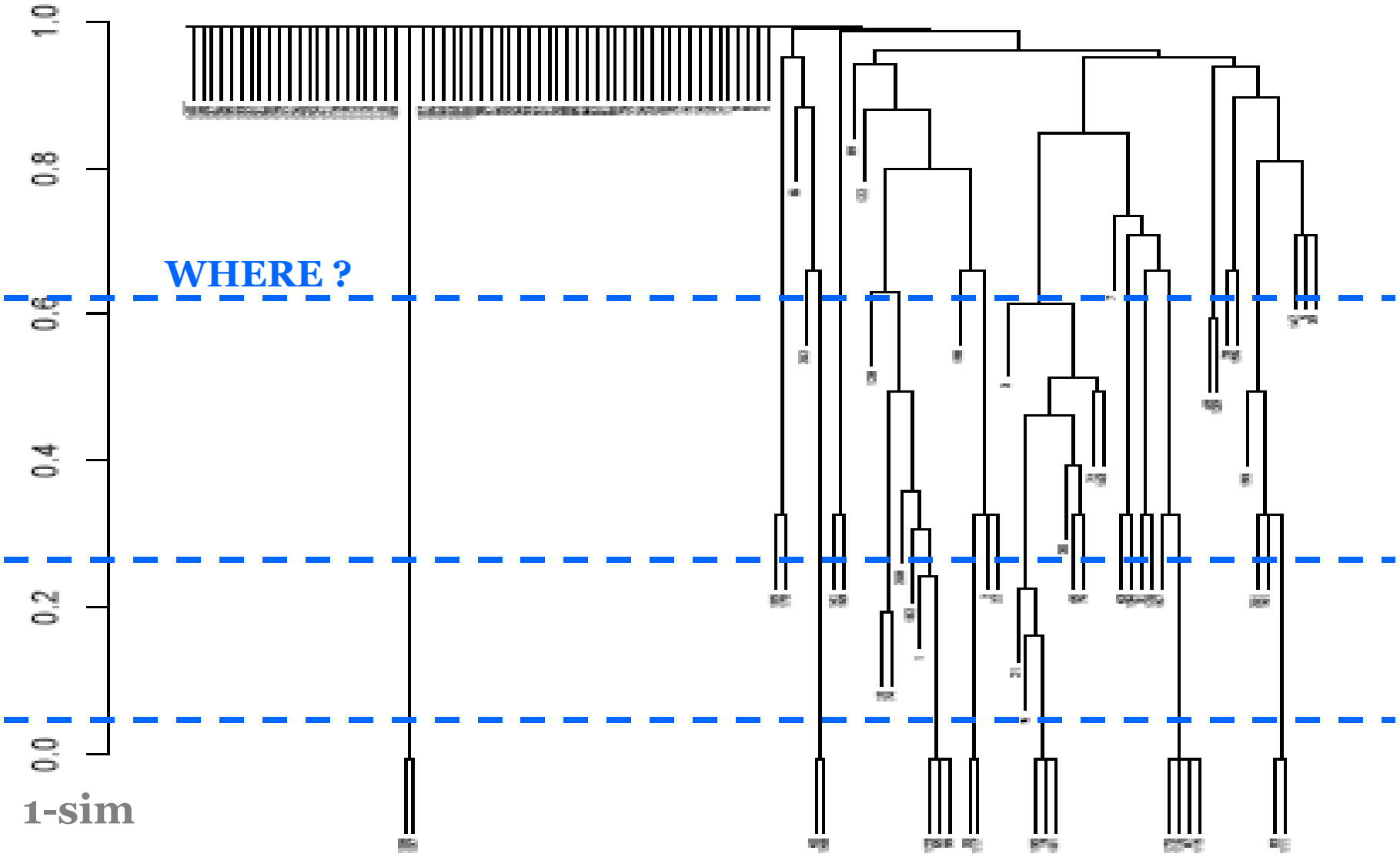


Hierarchical dendrogram *H.m 23S*

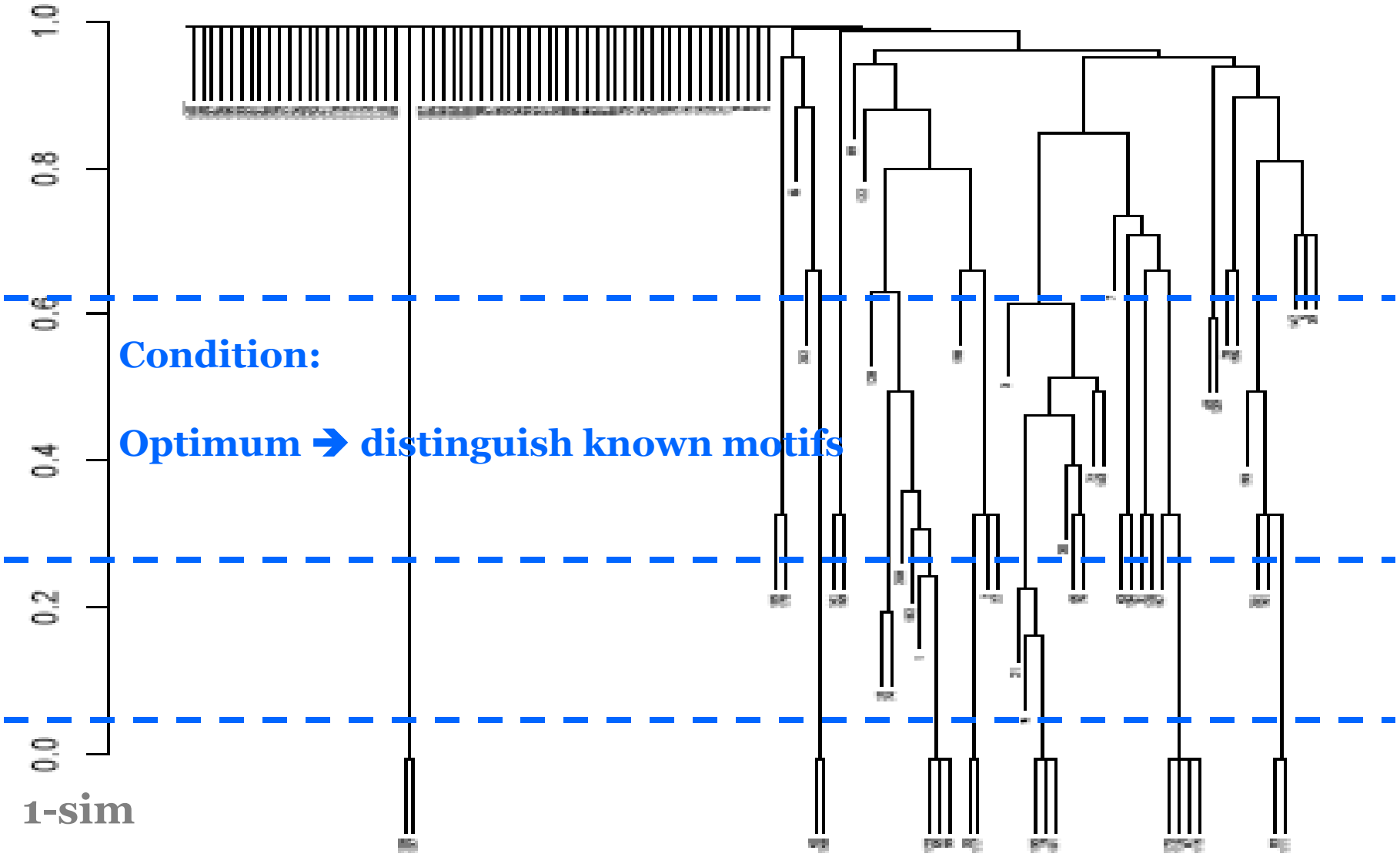


Biologically sensical clustering
➔ fix a threshold

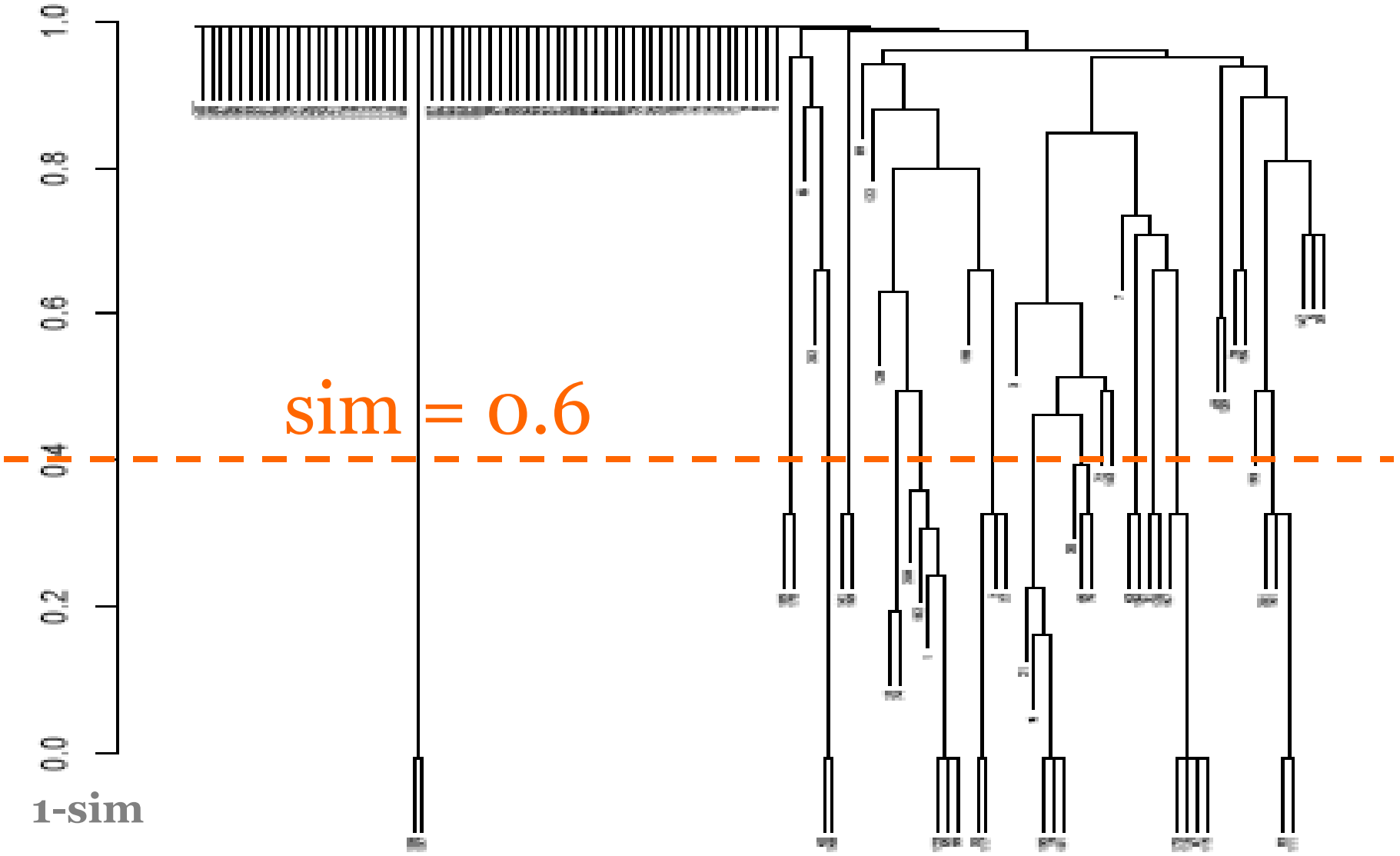
Hierarchical dendrogram H.m 23S



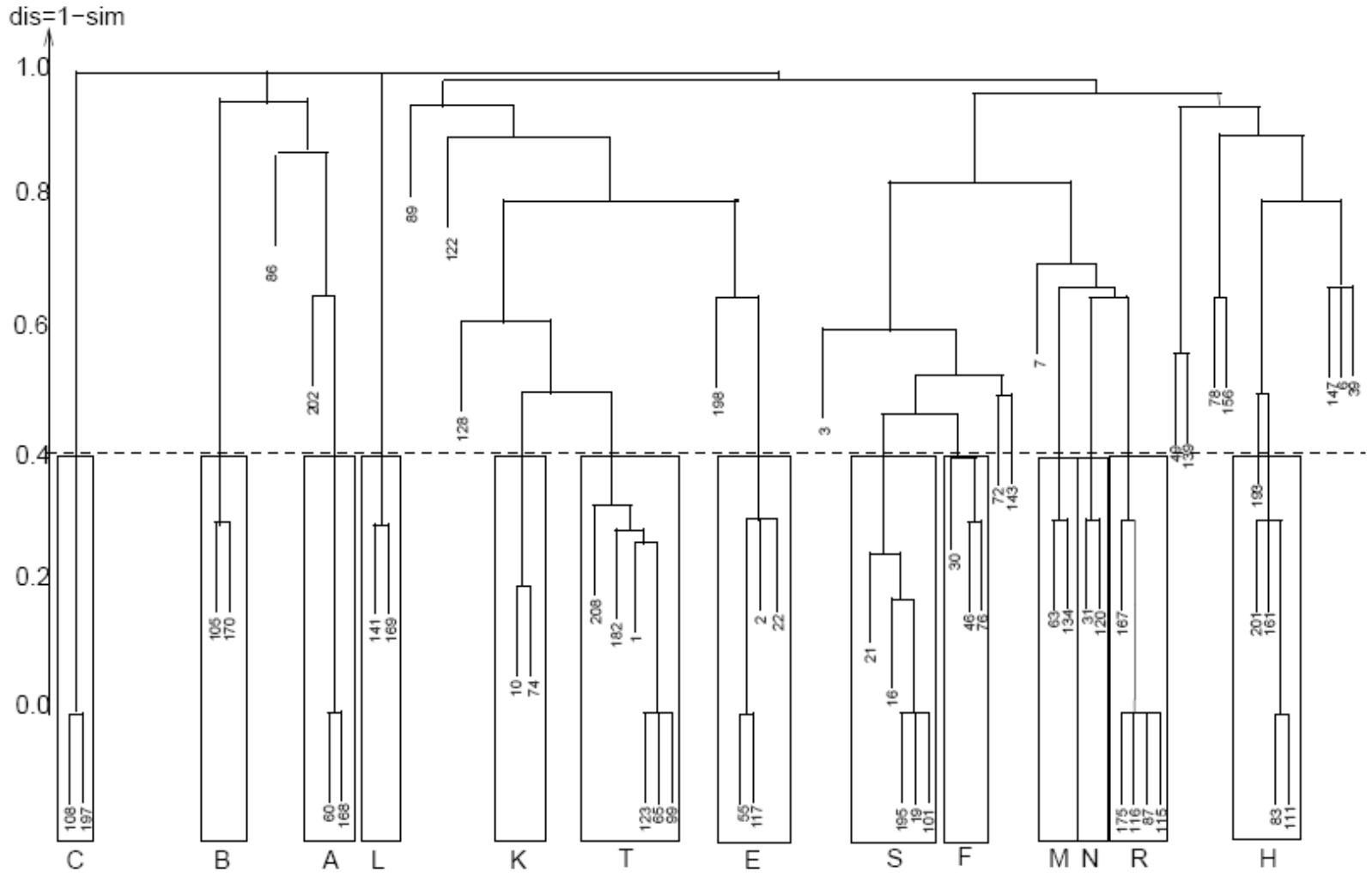
Hierarchical dendrogram *H.m 23S*



Hierarchical dendrogram *H.m 23S*



Hierarchical dendrogram *H.m 23S*

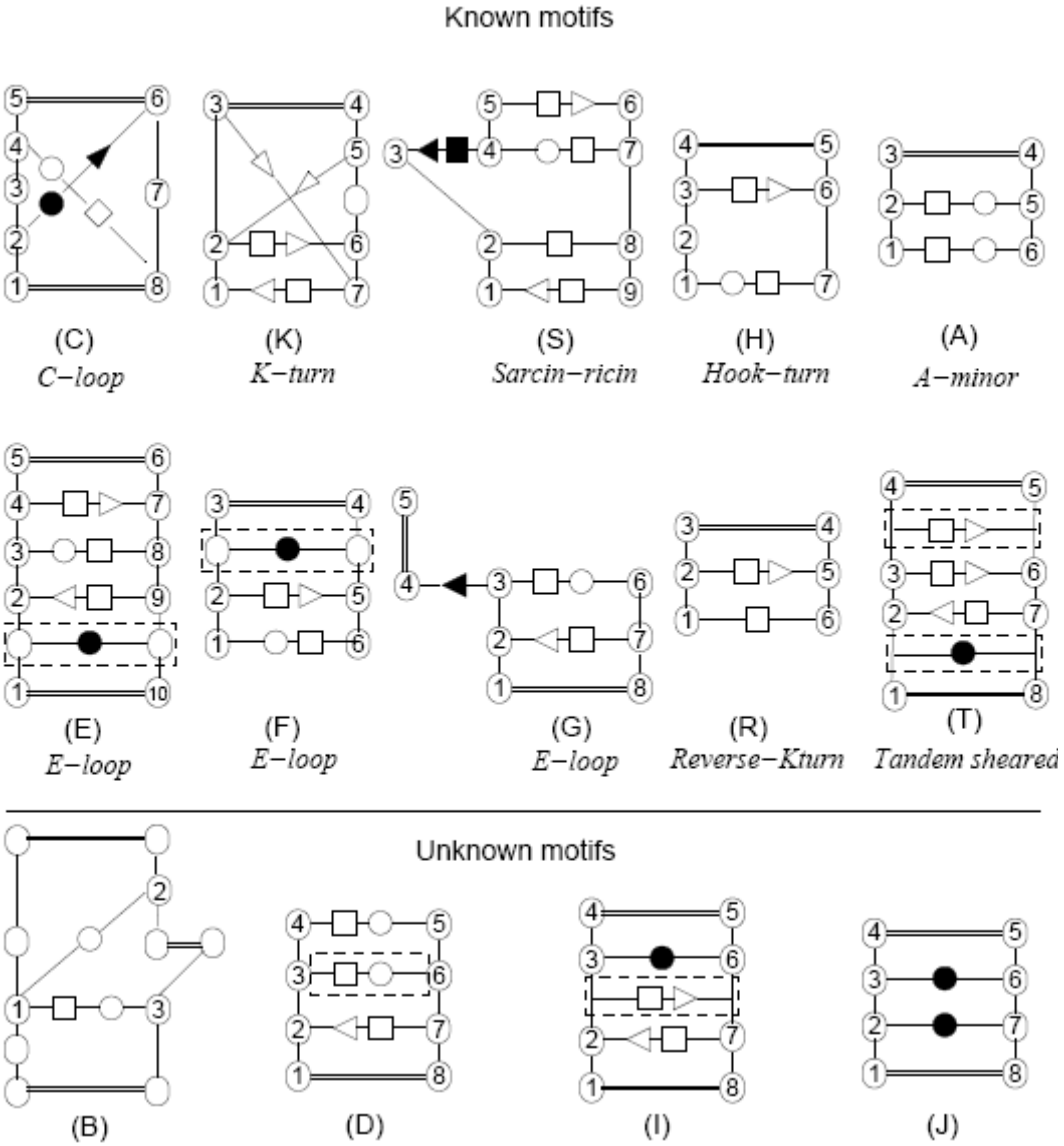


Results

3 organisms:

- 50S H.marismortui
- 50S E.coli
- 16S T.thermophilus

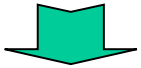
- 10 known motifs
- 4 putative new motifs



Results

3 organisms:

- 50S H.marismortui
- 50S E.coli
- 16S T.thermophilus



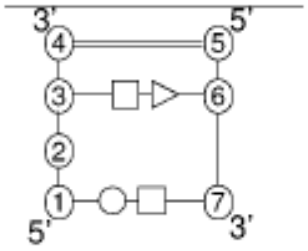
- 10 known motifs
- 4 putative new motifs

Motifs	Molecule	PDB file	Occur.	Known/Unknown
(C)	*H.m*23S	1s72	2	C-loop [22]
	*E.coli*23S	2aw4	2	C-loop [22]
(K)	*H.m*23S	1s72	2	Kturns KT-7, KT-38 [22]
(S)	*H.m*23S	1s72	6	Sarcin-ricin [18]
	*E.coli*23S	2aw4	5	Sarcin-ricin [18]
	*T.th*16S	1j5e	2	Sarcin-ricin [18]
(H)	*H.m*23S	1s72	5	Hook-turn [33]
	*E.coli*23S	2aw4	6	Hook-turn [33]
(A)	*H.m*23S	1s72	3	A-minor [23]
(E)	*H.m*23S	1s72	3	23S E-loop [18]
	*T.th*16S	1j5e	4	23S E-loop [18]
(F)	*E.coli*23S	2aw4	5	23S E-loop comprising sarcin G2664 [18]
	*H.m*23S	1s72	5	23S E-loop comprising composite sarcin G911 [18]
(G)	*E.coli*23S	2aw4	2	23S E-loop [18]
(R)	*H.m*23S	1s72	7	Reverse-Kturn [17]
	*E.coli*23S	2aw4	6	Reverse-Kturn [17]
(T)	*E.coli*23S	2aw4	8	Tandem sheared
	*H.m*23S	1s72	6	Tandem sheared comprising KT-46, KT-58 [22]
	*T.th*16S	1j5e	2	Tandem sheared
(B)	*H.m*23S	1s72	2	Unknown
(D)	*E.coli*23S	2aw4	2	Unknown
(I)	*T.th*16S	1j5e	2	Unknown
(J)	*T.th*16S	1j5e	2	Unknown

Table 1: List of the clusters formed in *H.m 23S*, *E.coli 23S* and *T.th 16S*.

An RNA motif

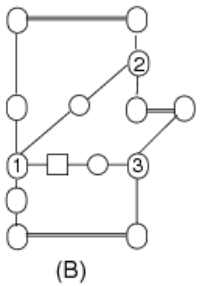
Hook turn

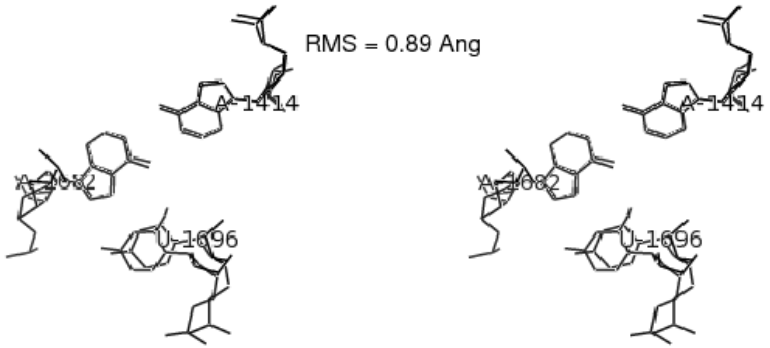


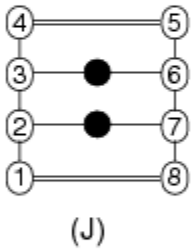
(H)

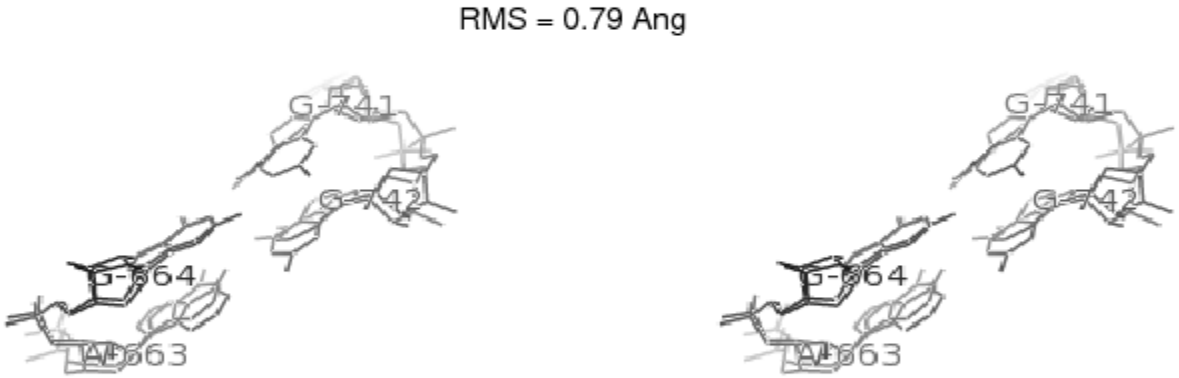
PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	Catalogue	RMSD (ref.H83)
1s72	H83	1096_U	1097_A	1098_A	1099_G	1257_C	1258_G	1259_A	Internal L.(83)	0.00
	H111	1457_U	1458_A	1459_A	1460_G	1483_C	1484_G	1485_A	Internal L.(111)	0.76
	H201	2774_U	2775_A	2776_A	2777_G	2797_C	2798_G	2799_A	Internal L.(201)	0.33
	H161	2242_U	2243_C	2244_A	2245_C	2256_G	2257_G	2258_A	Junction L.(161)	1.61
	H193	2673_U	2674_G	2675_A	2676_C	2809_G	2810_G	2811_A	Internal L.(193)	1.61
2aw4	H73	999_U	1000_A	1001_A	1002_G	1153_C	1154_G	1155_A	Internal L.(73)	0.42
	H101	1352_U	1353_A	1354_A	1355_G	1376_C	1377_G	1378_A	Internal L.(101)	0.69
	H106	1578_U	1579_A	1580_A	1581_G	1417_C	1418_G	1419_A	Internal L.(106)	0.31
	H205	2739_U	2740_A	2741_A	2742_G	2762_C	2763_G	2764_A	Internal L.(205)	0.52
	H161bis	2197_U	2198_A	2199_A	2200_C	2223_G	2224_G	2225_A	Junction L. (161)	1.60
	H196	2637_U	2638_G	2639_A	2640_G	2774_C	2775_G	2776_A	Internal L.(196)	1.66

Deux nouveaux motifs potentiels









Conclusions

- Une mesure de similarité sur les sous-graphes d'ARN qui capte bien la notion de motif structural : elle permet de retrouver les motifs connus, sans aucun a priori sur leur structure, forme, position.

- La suite : **motifs d'interaction**.

Merci à...

Dominique Barth

Daniel Gautheret

Alexis Lamiable

Eric Westhof

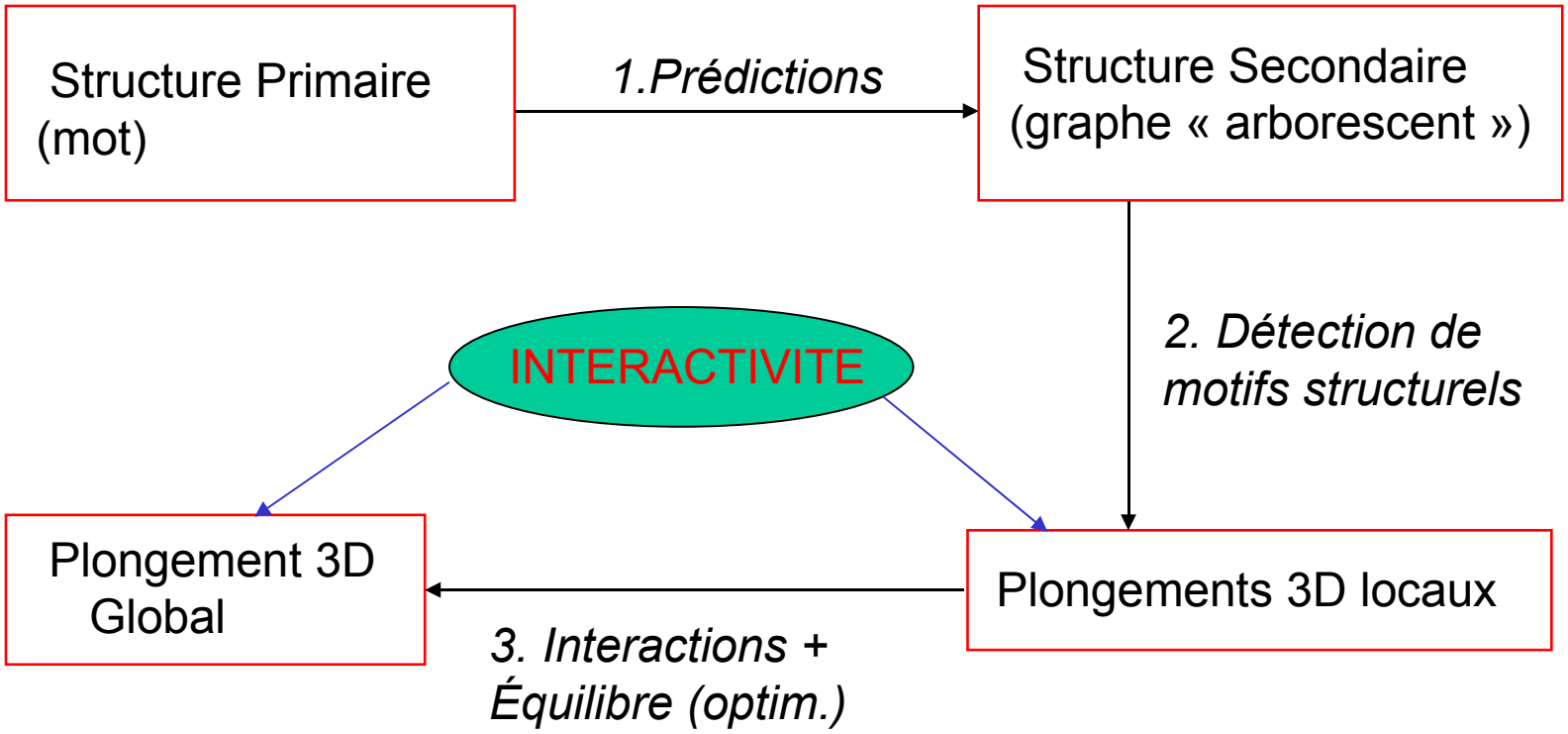
Franck Quessette

François Major

Fabrice Jossinet

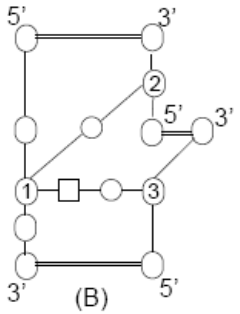
Sandrine Vial

Une approche pour la prédiction de l'architecture 3D de l'ARN

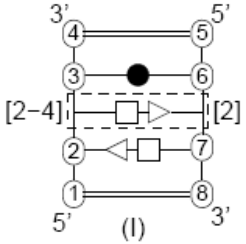


Projet PASAPAS (PRISM, LRI, IGM + IBMC)

Deux nouveaux motifs potentiels



PDB	Inst.	(1)	(2)	(3)	Catalogue	RMSD (ref.B105)
1s72	B170	2369_A	2356_A	2330_U	Junction L.(170)	0.89
	B105	1682_A	1414_A	1696_U	Junction L.(105)	0.00



PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Catalogue	RMSD (ref. I47)
1j5e	I99	1303_C	1304_G	1307_U	1308_U	1329_A	1330_U	1333_A	1334_G	Internal L.(99)	1.85
	I47	605_U	606_G	611_A	612_C	628_G	629_G	632_A	633_G	Internal L.(47)	0.00