Comparaison de structures secondaires d'ARN

J. Allali

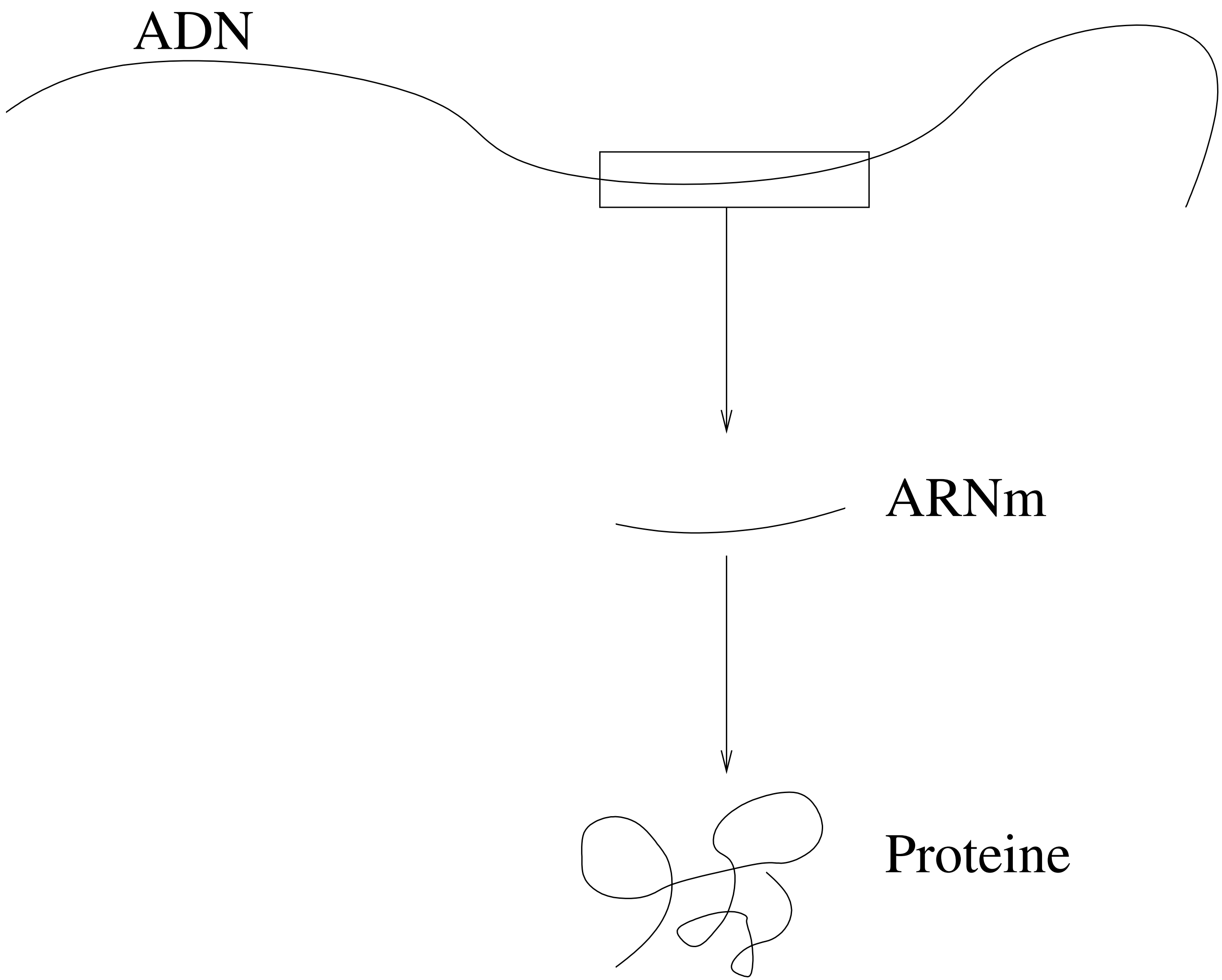
IGM, Marne la Vallée, INRIA Rhône-Alpes

Plan

- ARN, structures secondaires
- Modélisations : arbres et séquences annotées
- L'édition d'arbre en pratique : résultats et problèmes
- Deux nouvelles opérations d'édition
- Nouveaux résultats pratiques
- MiGaL
- Conclusion

ARN, structures secondaires

Introduction (1)

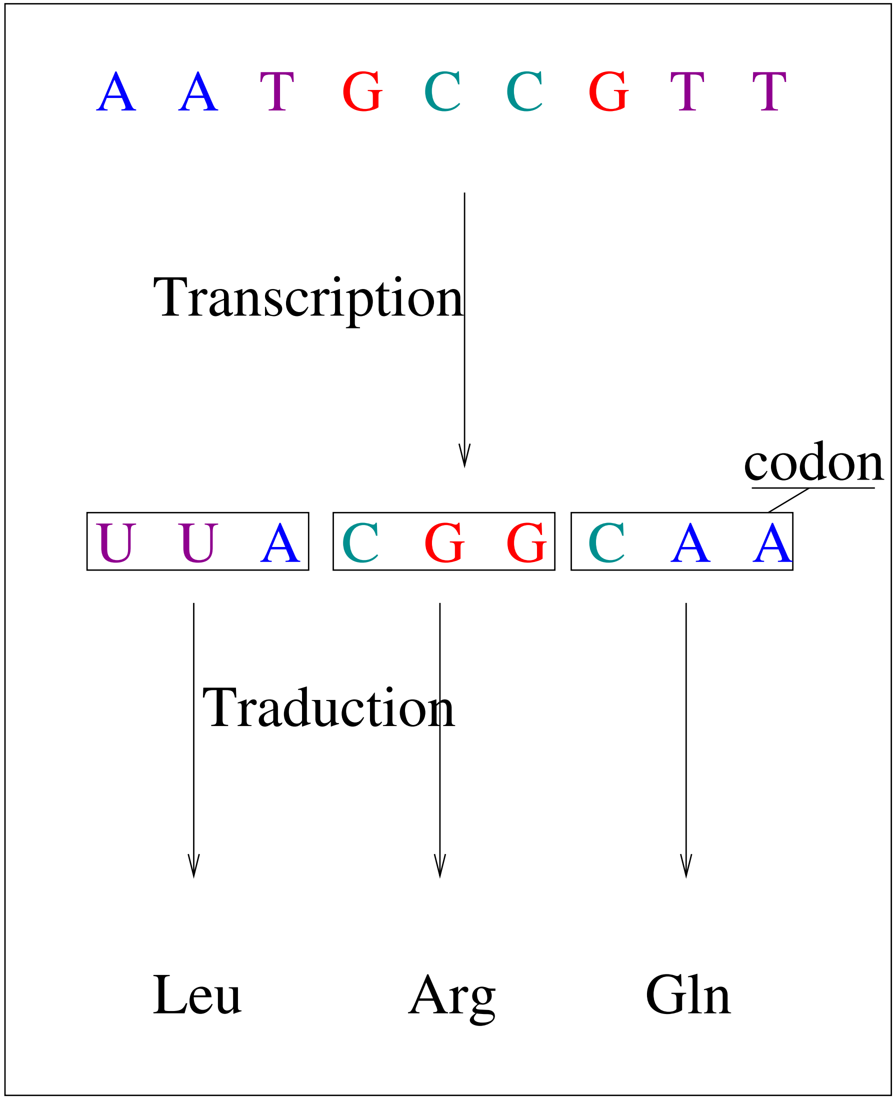


Introduction (1)

ADN (nucléotides)

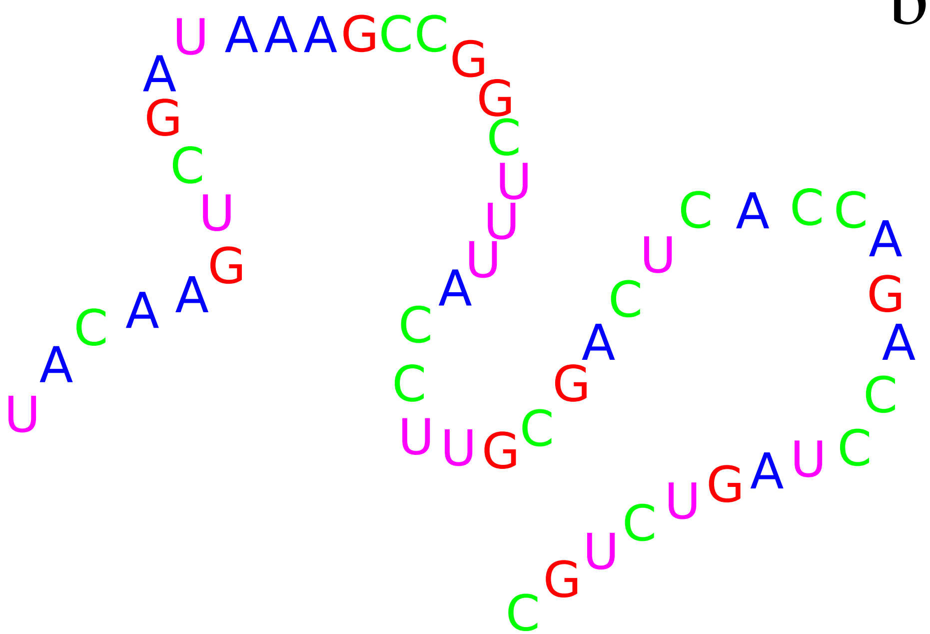
ARN (nucléotides)

Proteine (acides aminés)



Introduction (2)

Repliement :

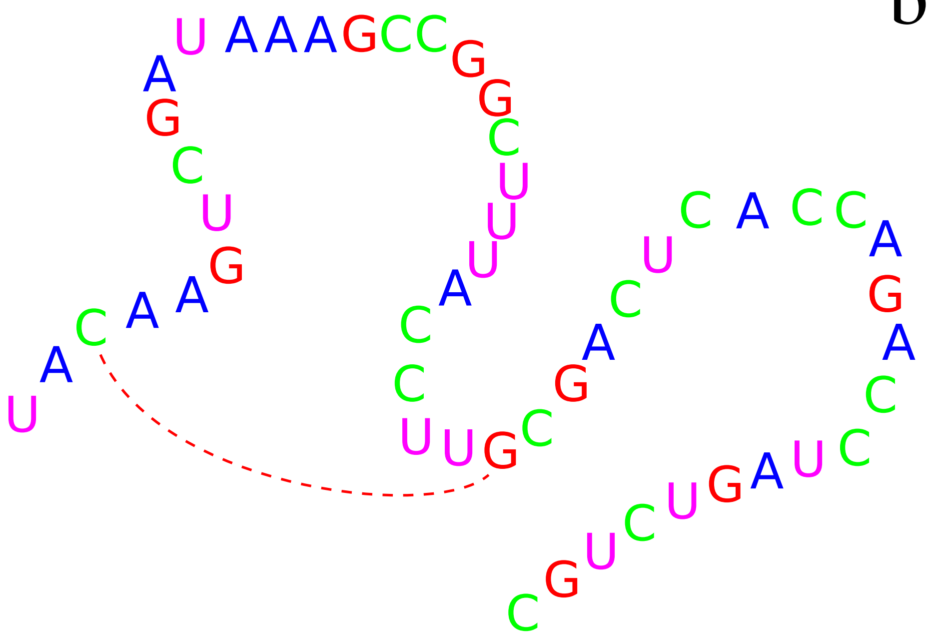


Appariement, paire de bases :

- Watson Crick : (A-U) (C-G)
- Wobble : (G-U)
- Non-canonique : Holbrook, 91 (C-U) Noller, 84 (G-A)

Introduction (2)

Repliement :

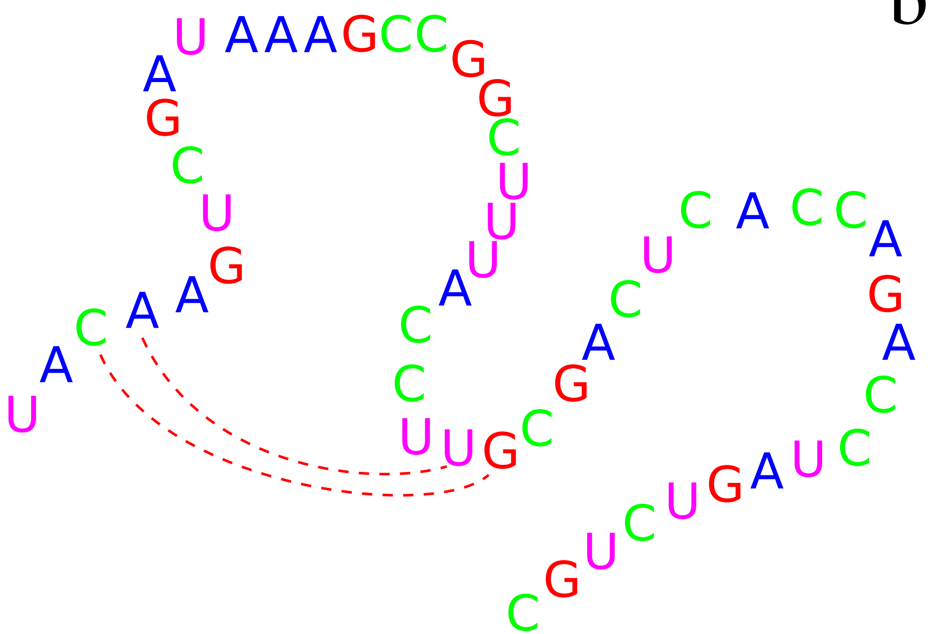


Appariement, paire de bases :

- Watson Crick : (A-U) (C-G)
- Wobble : (G-U)
- Non-canonique : Holbrook, 91 (C-U) Noller, 84 (G-A)

Introduction (2)

Repliement :

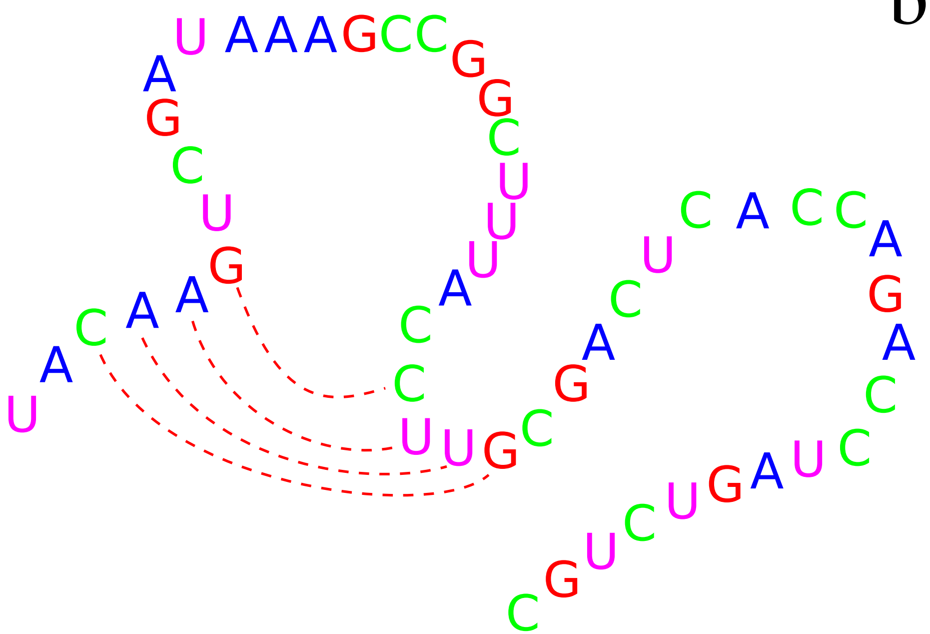


Appariement, paire de bases :

- Watson Crick : (A-U) (C-G)
- Wobble : (G-U)
- Non-canonique : Holbrook, 91 (C-U) Noller, 84 (G-A)

Introduction (2)

Repliement :

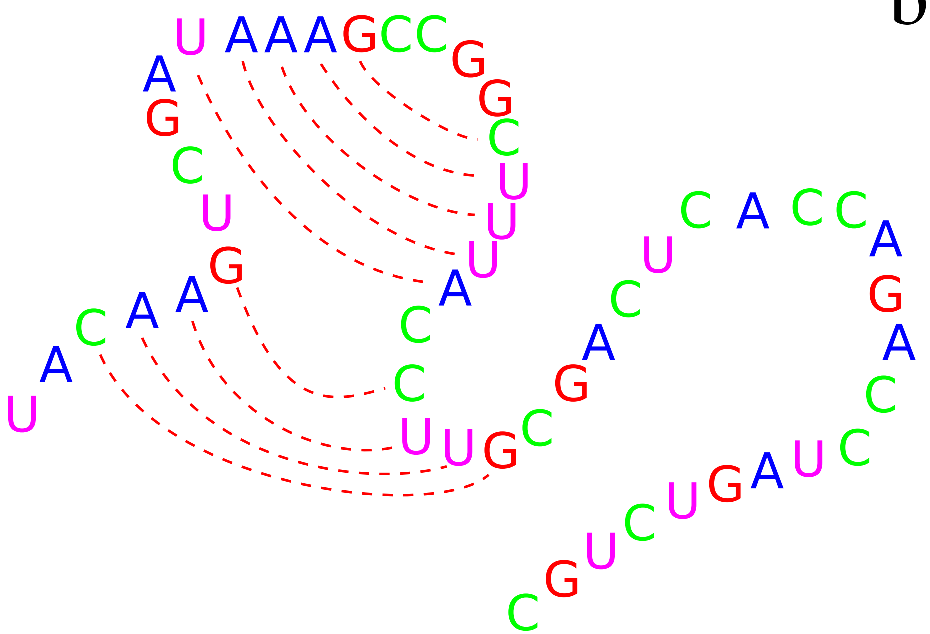


Appariement, paire de bases :

- Watson Crick : (A-U) (C-G)
- Wobble : (G-U)
- Non-canonique : Holbrook, 91 (C-U) Noller, 84 (G-A)

Introduction (2)

Repliement :

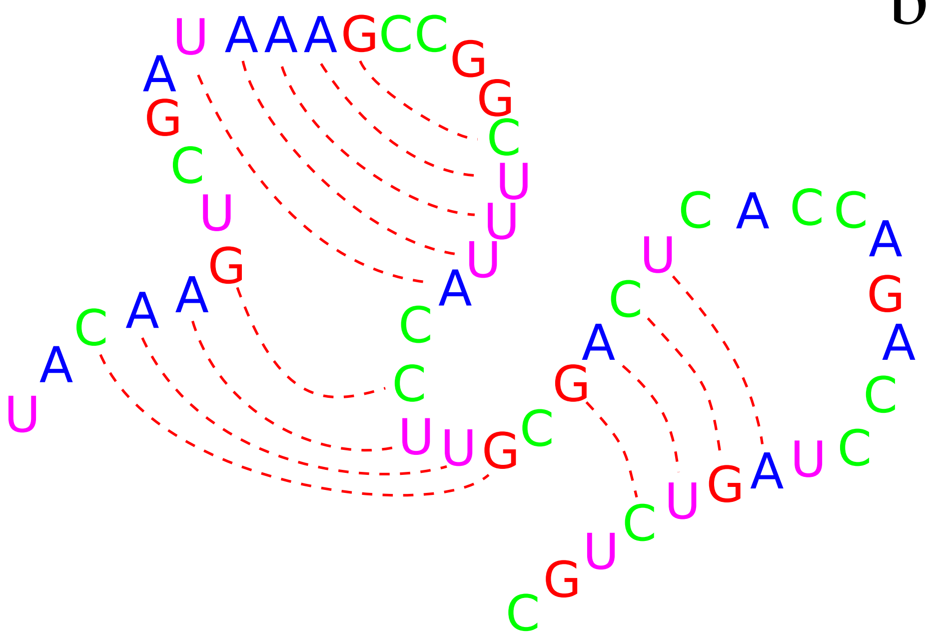


Appariement, paire de bases :

- Watson Crick : (A-U) (C-G)
- Wobble : (G-U)
- Non-canonique : Holbrook, 91 (C-U) Noller, 84 (G-A)

Introduction (2)

Repliement :

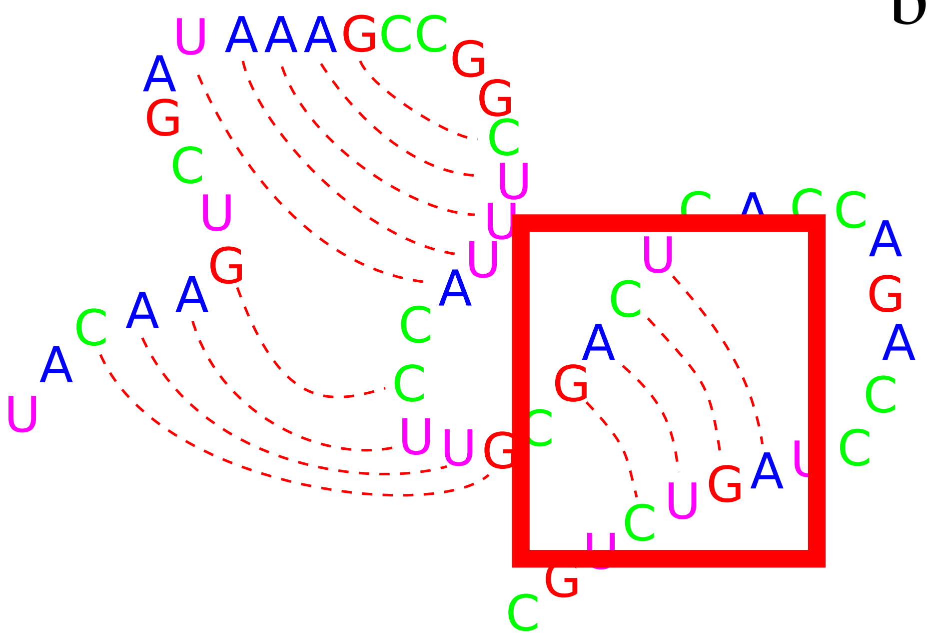


Appariement, paire de bases :

- Watson Crick : (A-U) (C-G)
- Wobble : (G-U)
- Non-canonique : Holbrook, 91 (C-U) Noller, 84 (G-A)

Introduction (2)

Repliement :

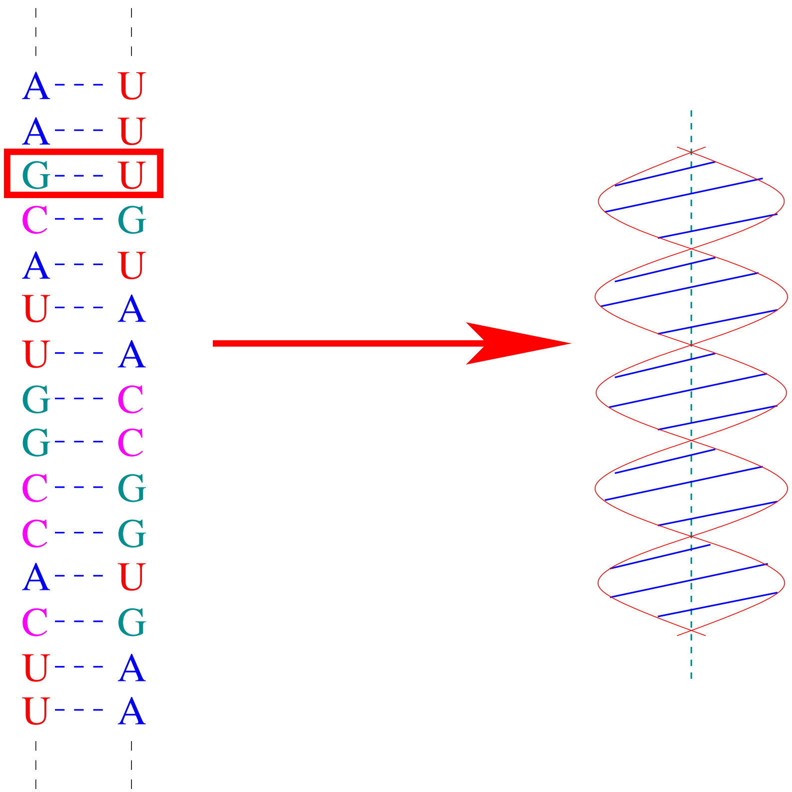


Appariement, paire de bases :

- Watson Crick : (A-U) (C-G)
- Wobble : (G-U)
- Non-canonique : Holbrook, 91 (C-U) Noller, 84 (G-A)

Introduction (2)

Repliement :

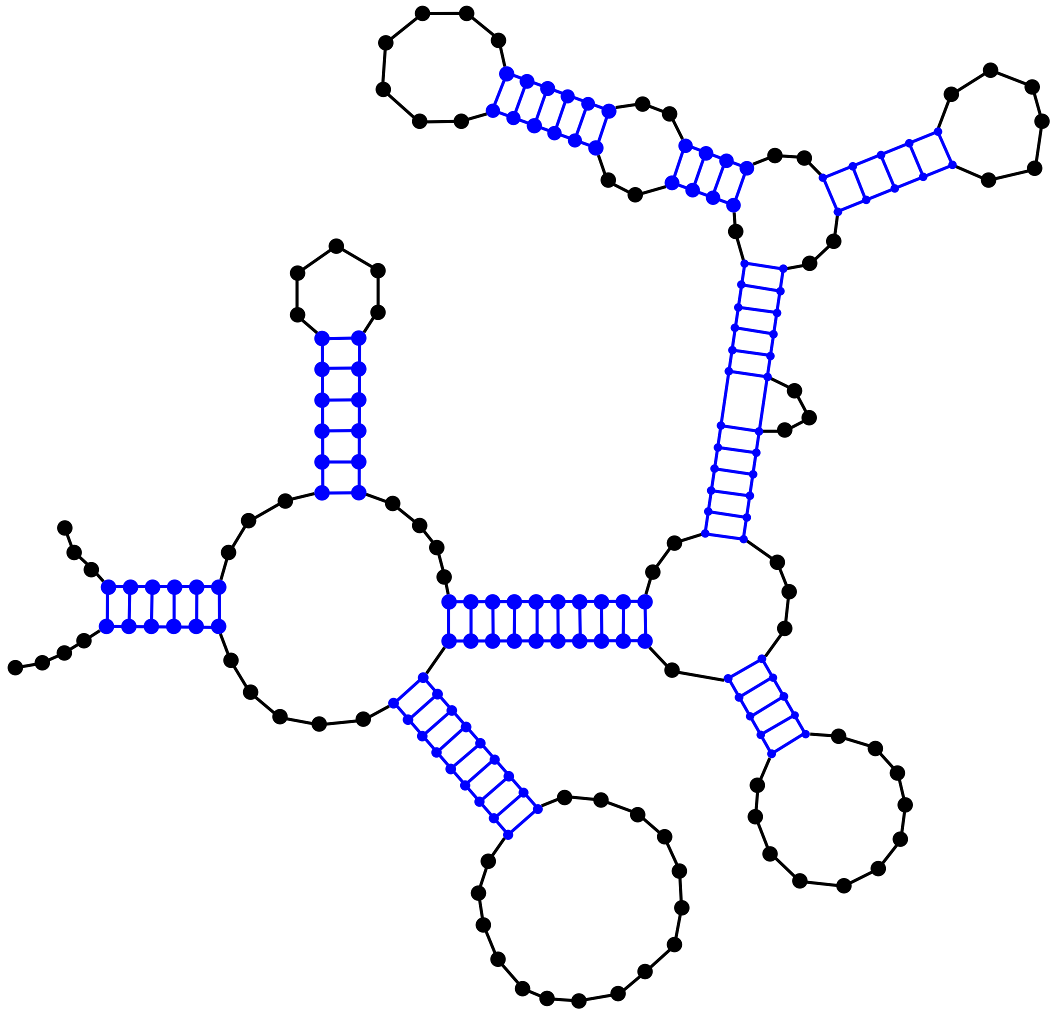


Appariement, paire de bases :

- Watson Crick : (A-U) (C-G)
- Wobble : (G-U)
- Non-canonique : Holbrook, 91 (C-U) Noller, 84 (G-A)

Introduction (3)

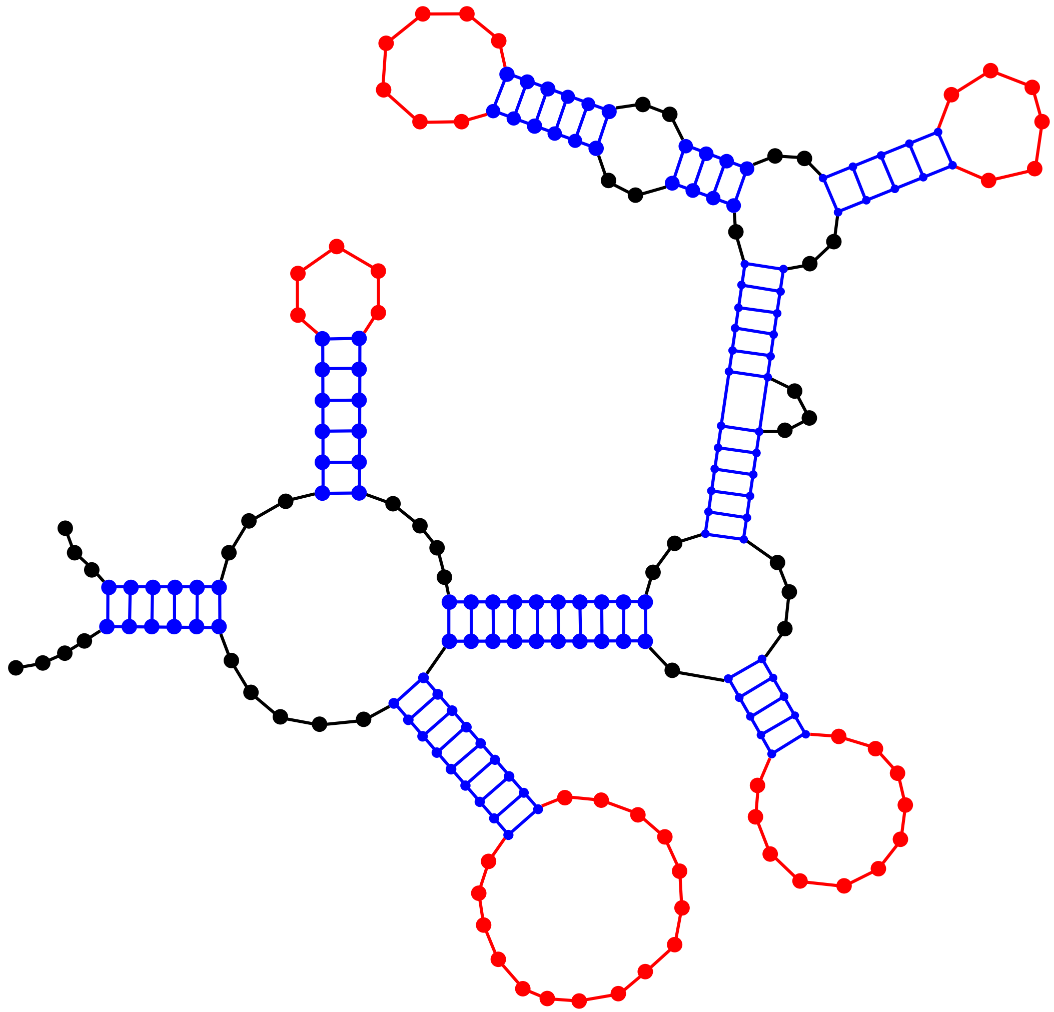
Structure secondaire :



- Hélice

Introduction (3)

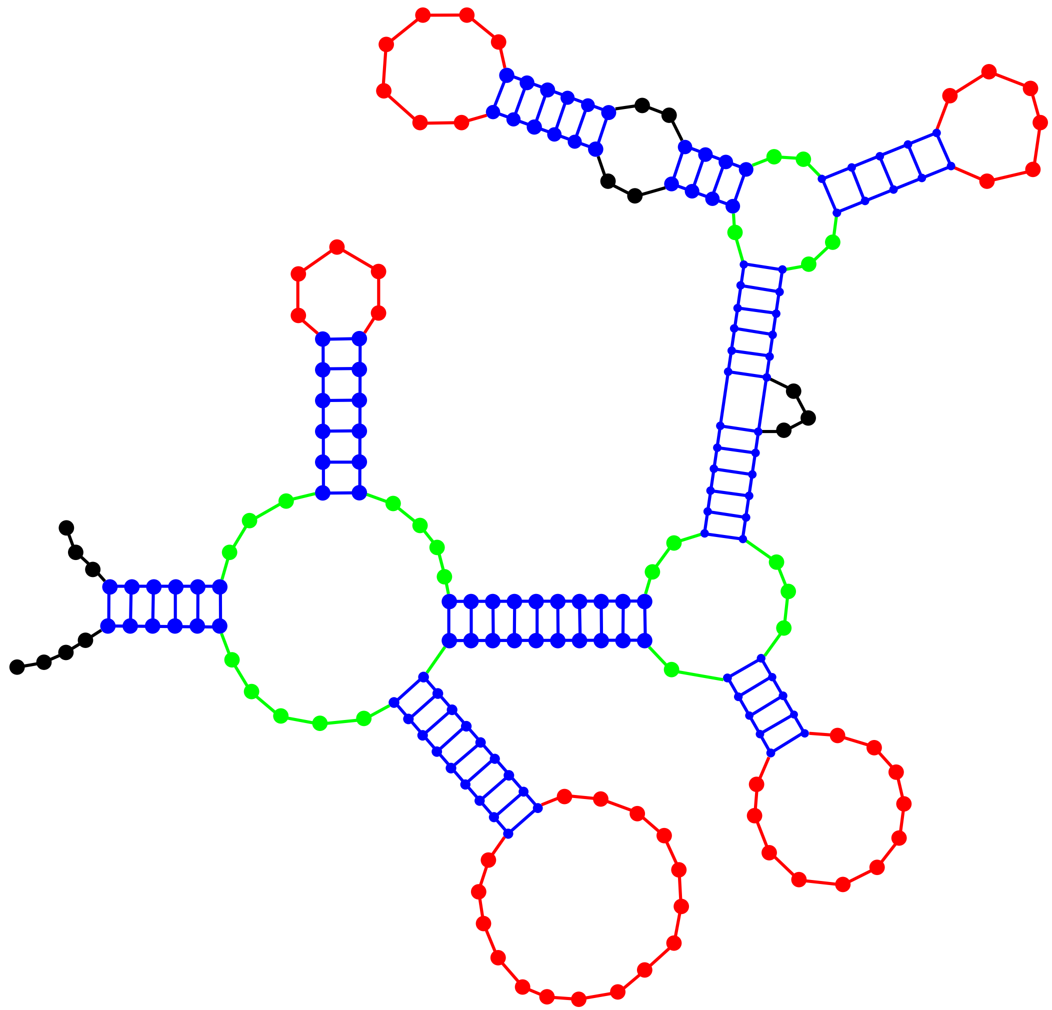
Structure secondaire :



- Hélice
- Boucle terminale

Introduction (3)

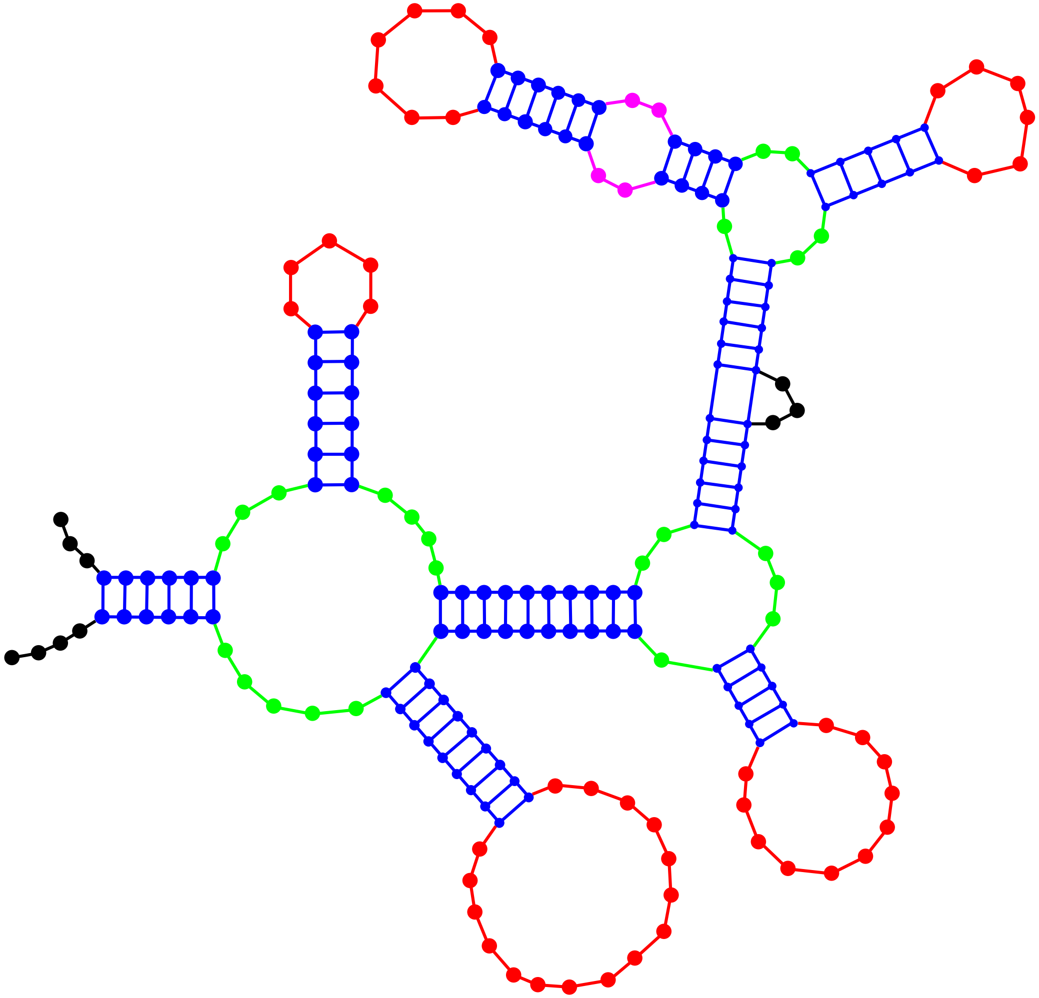
Structure secondaire :



- Hélice
- Boucle terminale
- Boucle multiple

Introduction (3)

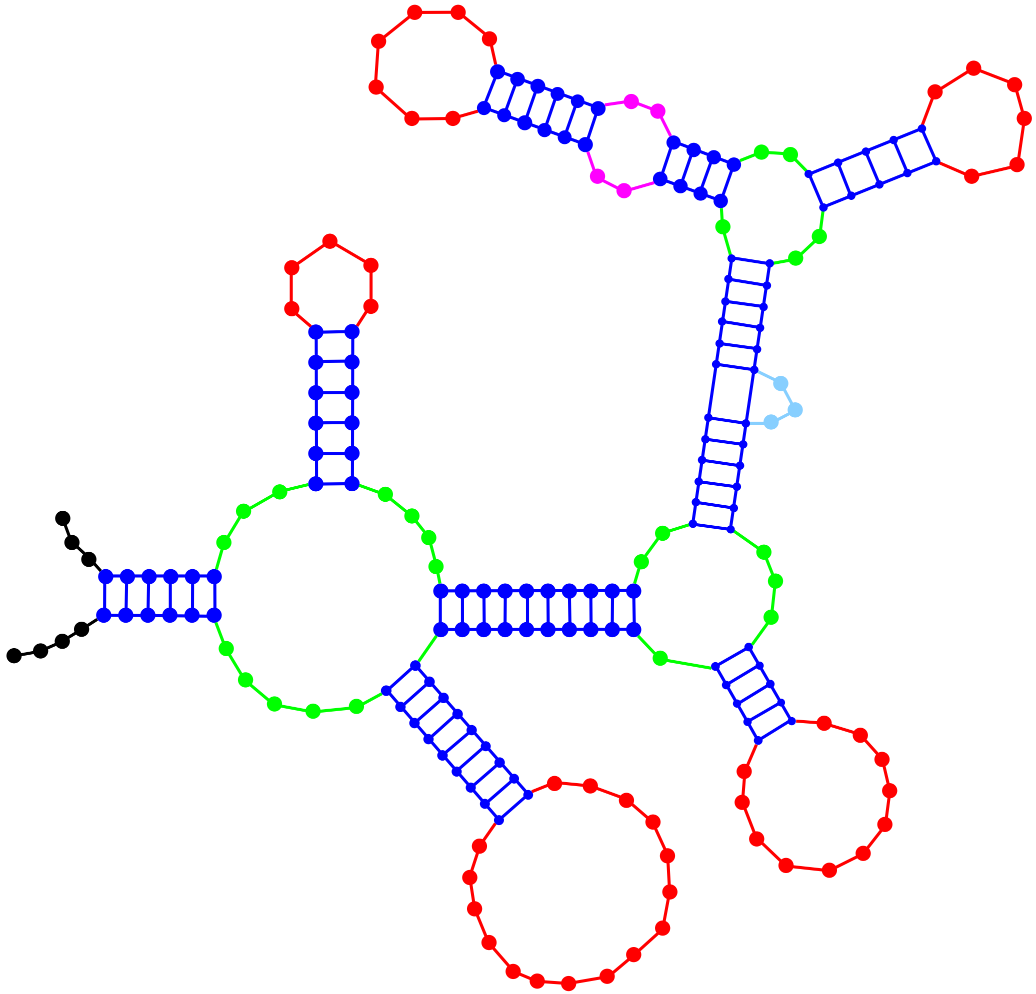
Structure secondaire :



- Hélice
- Boucle terminale
- Boucle multiple
- Boucle interne

Introduction (3)

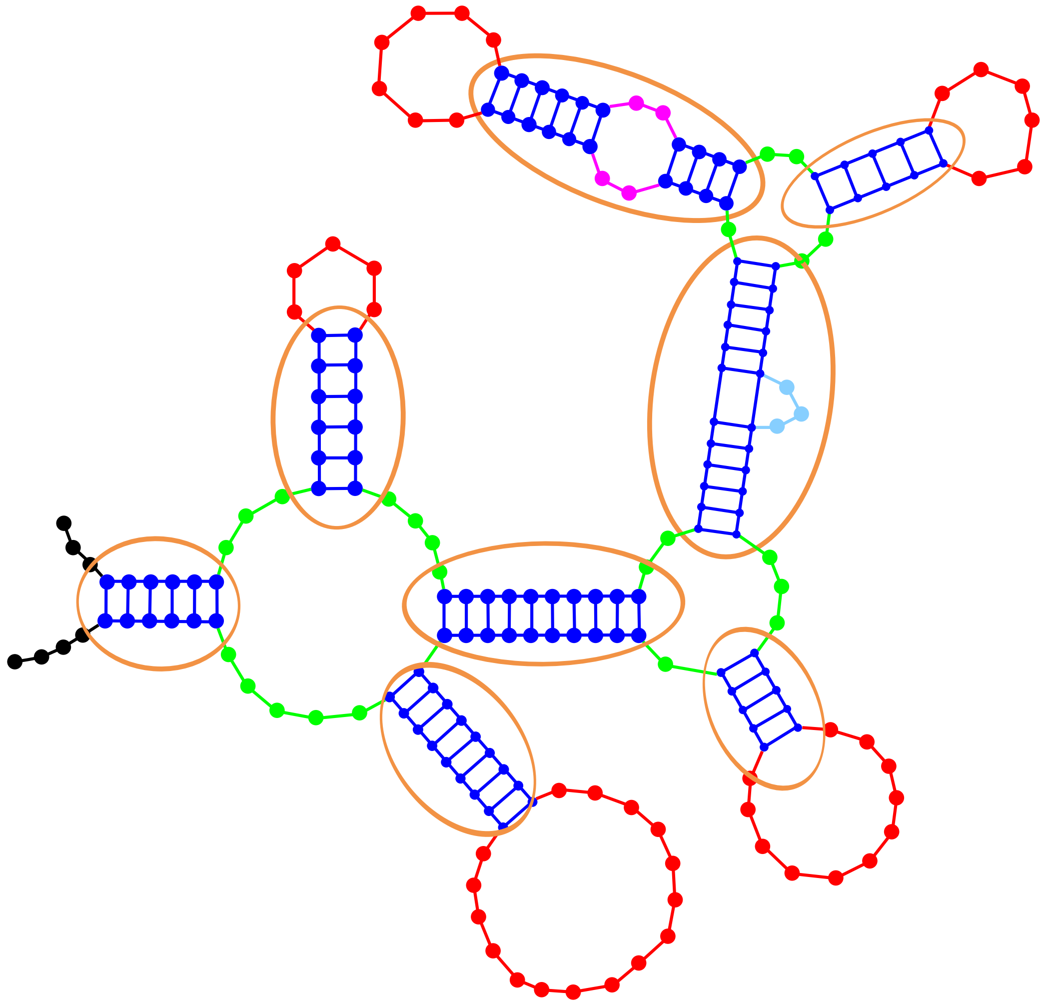
Structure secondaire :



- Hélice
- Boucle terminale
- Boucle multiple
- Boucle interne
- Renflement

Introduction (3)

Structure secondaire :



- Hélice
- Boucle terminale
- Boucle multiple
- Boucle interne
- Renflement
- Tige

Introduction (4)

Deux ARNs ont la même fonction :

- $\Rightarrow$ structures secondaires proches

Introduction (4)

Deux ARNs ont la même fonction :

- $\Rightarrow$ structures secondaires proches
- $\not\Rightarrow$ séquences (nucléotides) proches

Introduction (4)

Deux ARNs ont la même fonction :

- $\Rightarrow$ structures secondaires proches
- $\not\Rightarrow$ séquences (nucléotides) proches

Comparer des ARNs pour :

- calcul de repliement

Introduction (4)

Deux ARNs ont la même fonction :

- $\Rightarrow$ structures secondaires proches
- $\not\Rightarrow$ séquences (nucléotides) proches

Comparer des ARNs pour :

- calcul de repliement
- classification

Introduction (4)

Deux ARNs ont la même fonction :

- $\Rightarrow$ structures secondaires proches
- $\not\Rightarrow$ séquences (nucléotides) proches

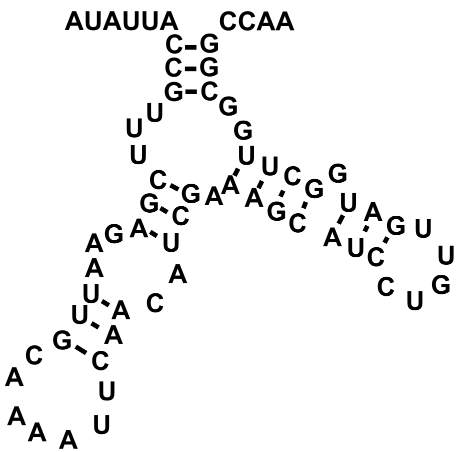
Comparer des ARNs pour :

- calcul de repliement
- classification
- phylogénie
- …

Modélisations : arbres et séquences annotées

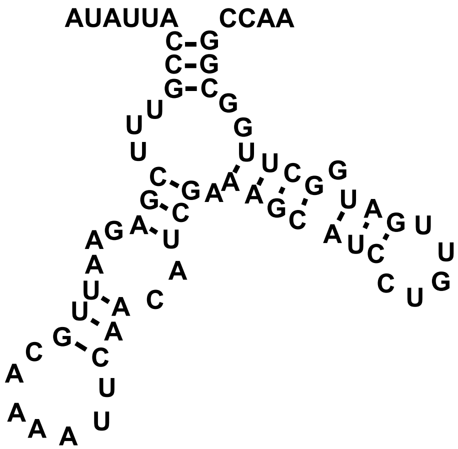
Arbres

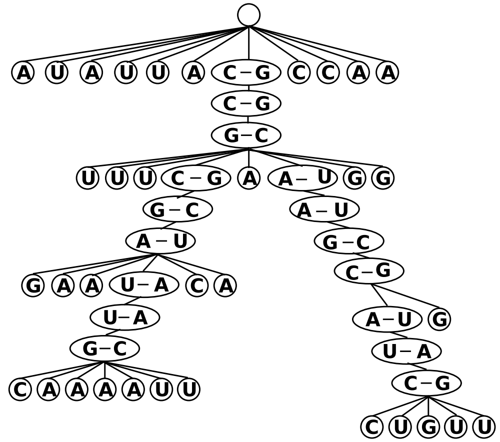
[Zuker & Sankoff, 84] [Shapiro, 88]



Arbres

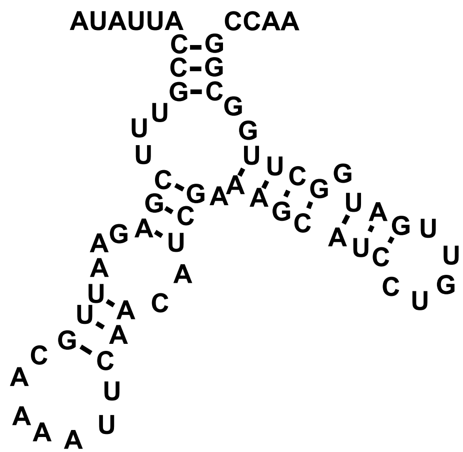
[Zuker & Sankoff, 84] [Shapiro, 88]

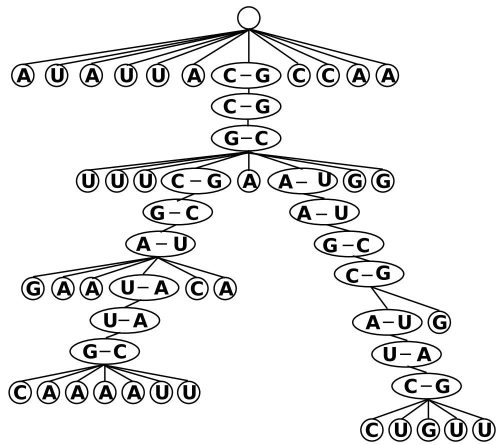


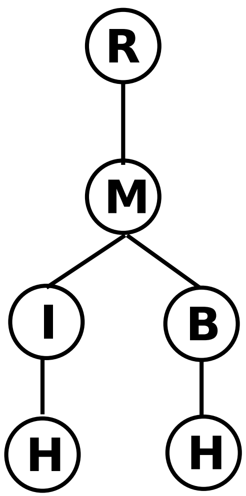


Arbres

[Zuker & Sankoff, 84] [Shapiro, 88]

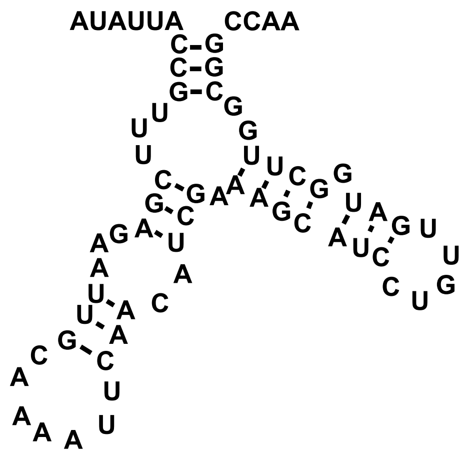


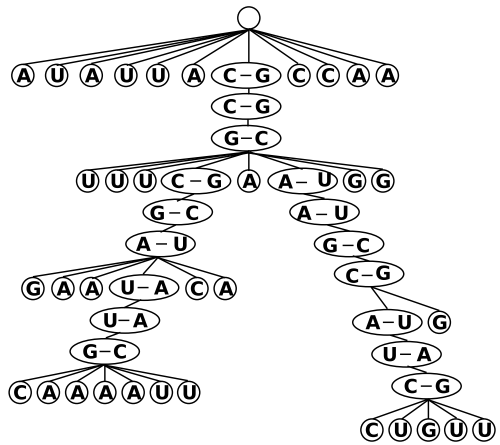


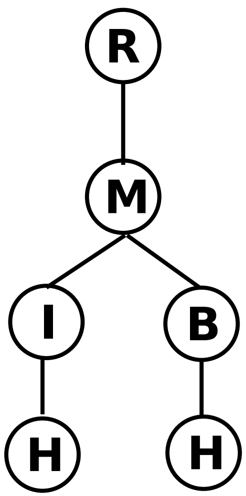


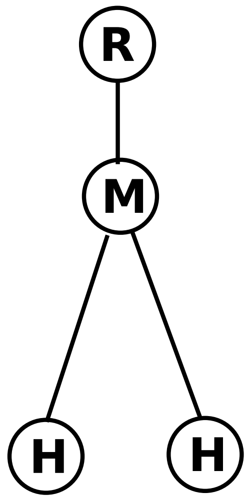
Arbres

[Zuker & Sankoff, 84] [Shapiro, 88]

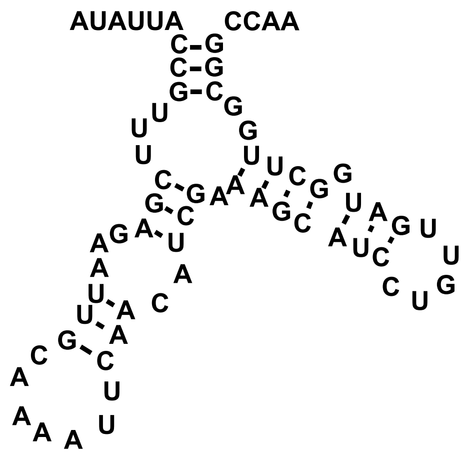


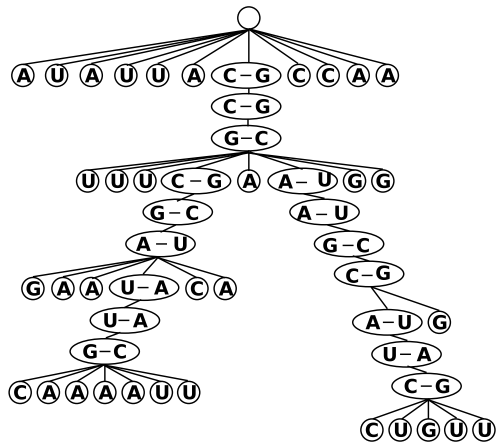




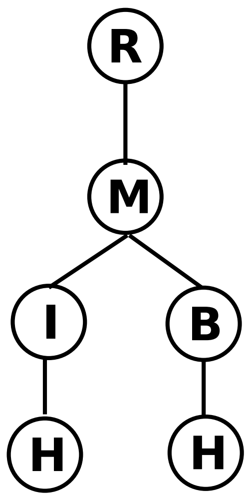


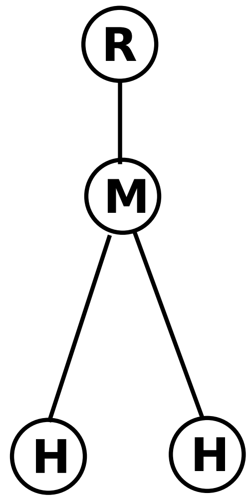
Arbres [Zuker & Sankoff, 84] [Shapiro, 88]





tous sont des arbres enracinés, **ordonnés**





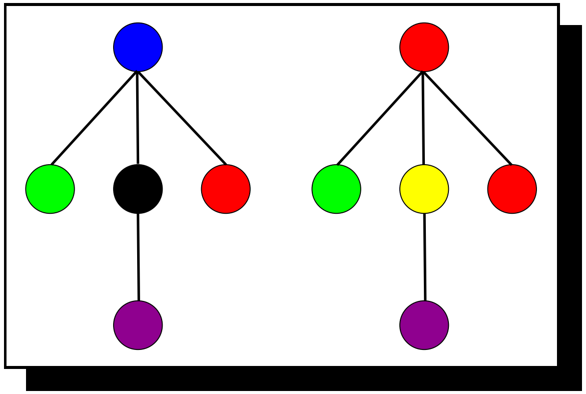
Arbres : édition [Selkow, 77] [Tai, 79]

Comparaison de deux arbres ⇒ distance d'édition :

Arbres : édition [Selkow, 77] [Tai, 79]

Comparaison de deux arbres ⇒ distance d'édition :

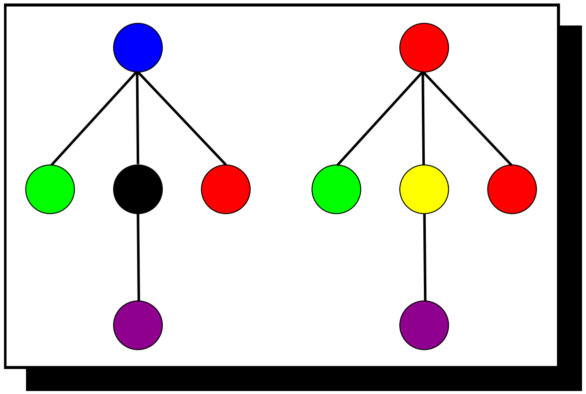
Substitution



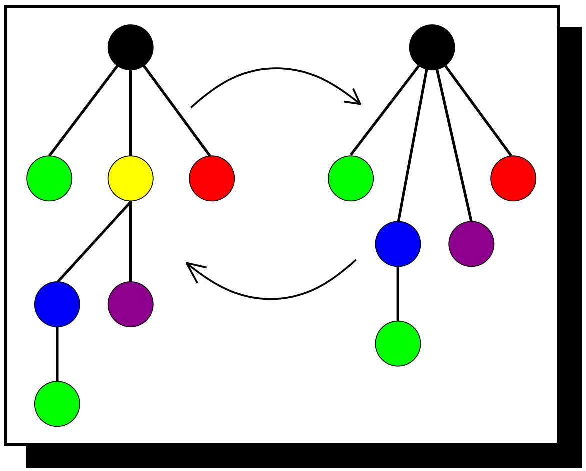
Arbres : édition [Selkow, 77] [Tai, 79]

Comparaison de deux arbres ⇒ distance d'édition :

Substitution



Délétion / Insertion

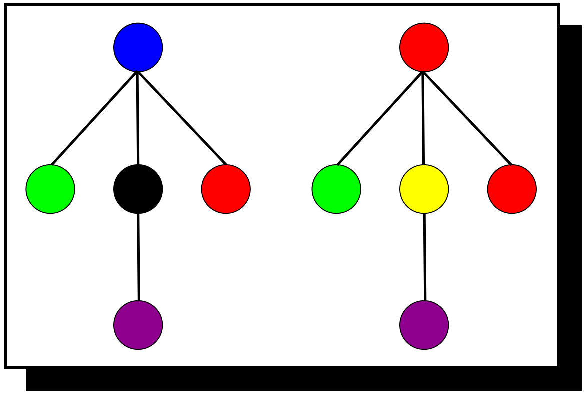


Arbres : édition

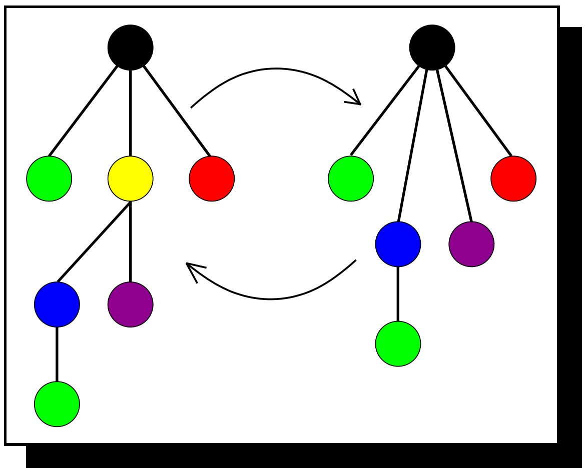
[Selkow, 77] [Tai, 79]

Comparaison de deux arbres $\Rightarrow$ distance d'édition :

Substitution



Délétion / Insertion



- T_1,T_2 : arbres ordonnés
- $S = \{s_1; \ldots s_k\}$: suite d'opérations
- $T_1 \xrightarrow{s_1} {T^1}_1 \cdots \xrightarrow{s_k} {T^k}_1 = T_2$
- $T_1 \overset{S}{\rightsquigarrow} T_2$

Arbres : distance [Tai, 79]

Fonction de coût : $cost : s \longrightarrow c \in \mathbb{R}$

$$E = \{S | T_1 \overset{S}{\rightsquigarrow} T_2\}$$

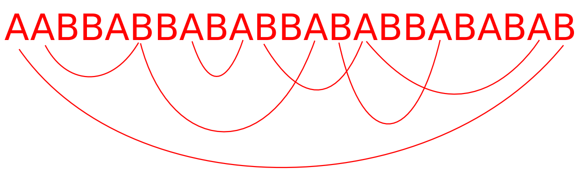
Distance d'édition :

$$D(T_1, T_2) = min\{cost(S) | S \in E\}$$

$\Rightarrow$ Association $A = \{(i, j)\}$

Séquences annotées [Evans, 99]

- Texte t
- Ensemble A de couples de positions de t

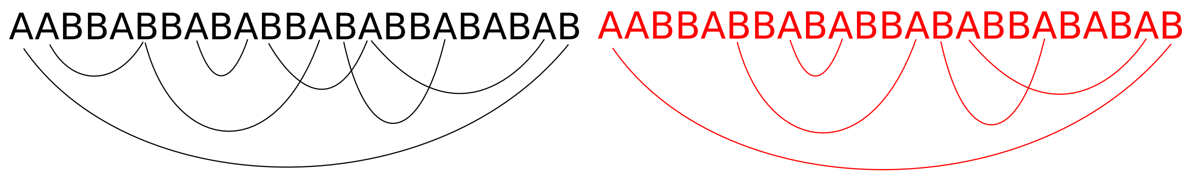


5 types :

- général

Séquences annotées [Evans, 99]

- Texte t
- Ensemble A de couples de positions de t

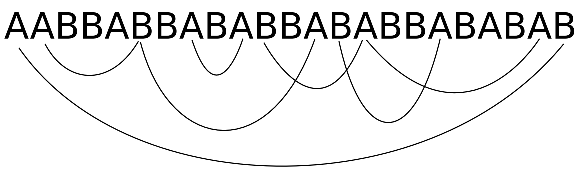


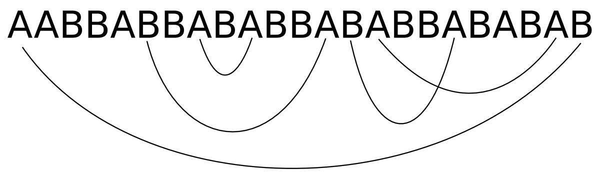
5 types :

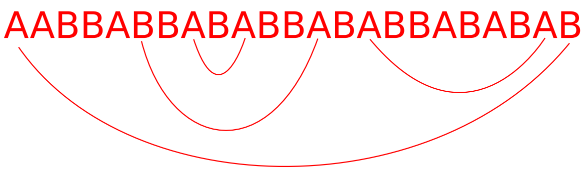
- général
- croisé

Séquences annotées [Evans, 99]

- Texte t
- Ensemble A de couples de positions de t





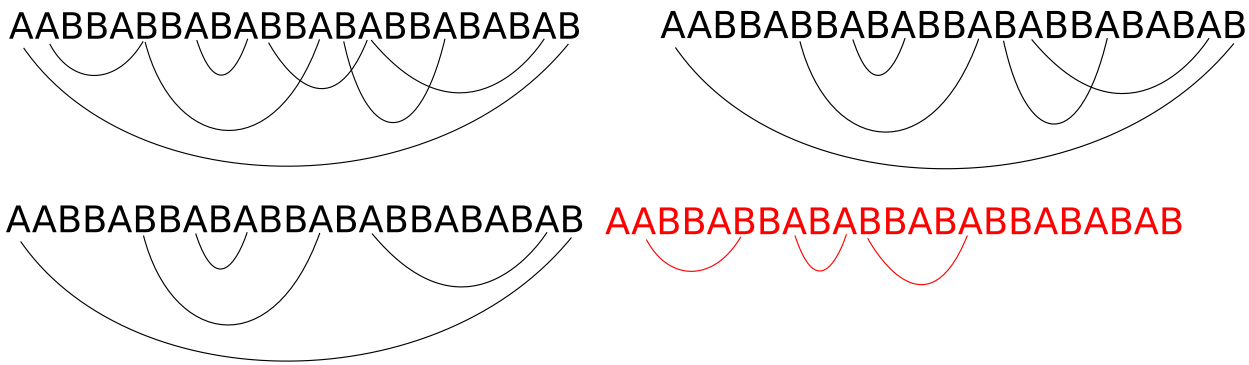


5 types :

- général
- croisé
- imbriqué (arbres)

Séquences annotées [Evans, 99]

- Texte t
- Ensemble A de couples de positions de t

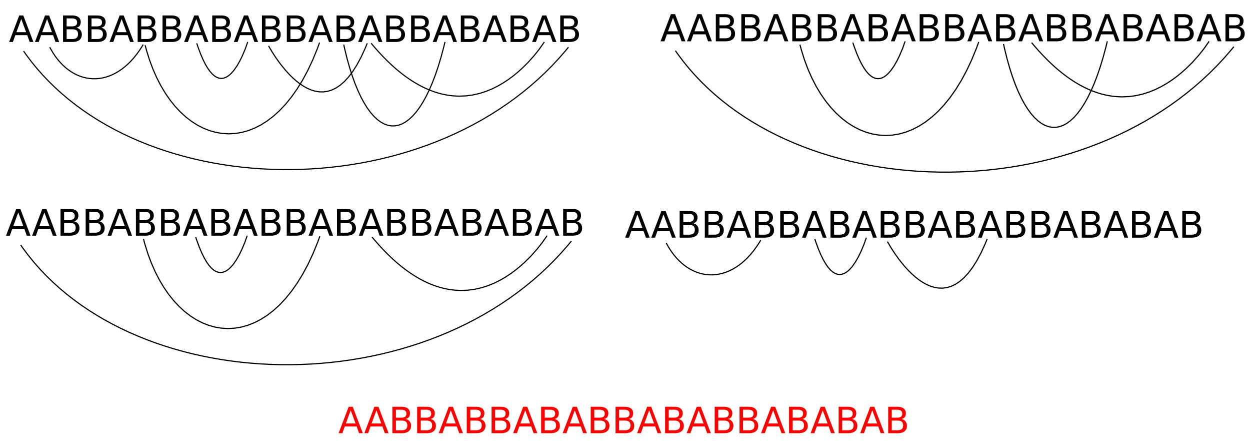


5 types :

- général
- croisé
- imbriqué (arbres)
- successif

Séquences annotées [Evans, 99]

- Texte t
- Ensemble A de couples de positions de t

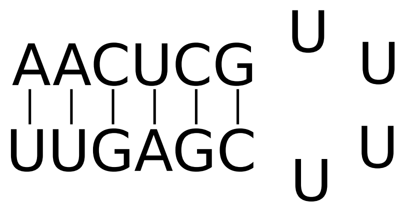


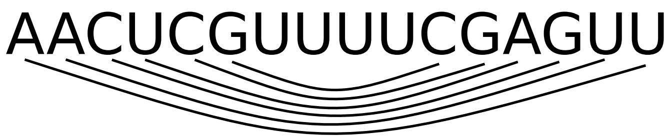
5 types :

- général
- croisé
- imbriqué (arbres)
- successif
- sans arc

Séquences annotées [Evans, 99]

- Texte t
- Ensemble A de couples de positions de t





5 types :

- général
- croisé
- imbriqué (arbres)
- successif
- sans arc

Séquences annotées : Comparaison

3 problèmes étudiés à ce jour pour la comparaison des séquences annotées :

- LAPCS[Evans, 99][Lin, 02] : NP-C pour (général,*), (croisé,*) et (imbriqué, imbriqué)

Séquences annotées : Comparaison

3 problèmes étudiés à ce jour pour la comparaison des séquences annotées :

- LAPCS[Evans, 99][Lin, 02] : NP-C pour (général,*), (croisé,*) et (imbriqué, imbriqué)
- APS[Blin, 04][Guo, 02] : NP-C pour (général,*), (croisé,*)

Séquences annotées : Comparaison

3 problèmes étudiés à ce jour pour la comparaison des séquences annotées :

- LAPCS[Evans, 99][Lin, 02] : NP-C pour (général,*), (croisé,*) et (imbriqué, imbriqué)
- APS[Blin, 04][Guo, 02] : NP-C pour (général,*), (croisé,*)
- distance d'édition $\Rightarrow$ arbres

L'édition d'arbre en pratique :

Résultats et problèmes

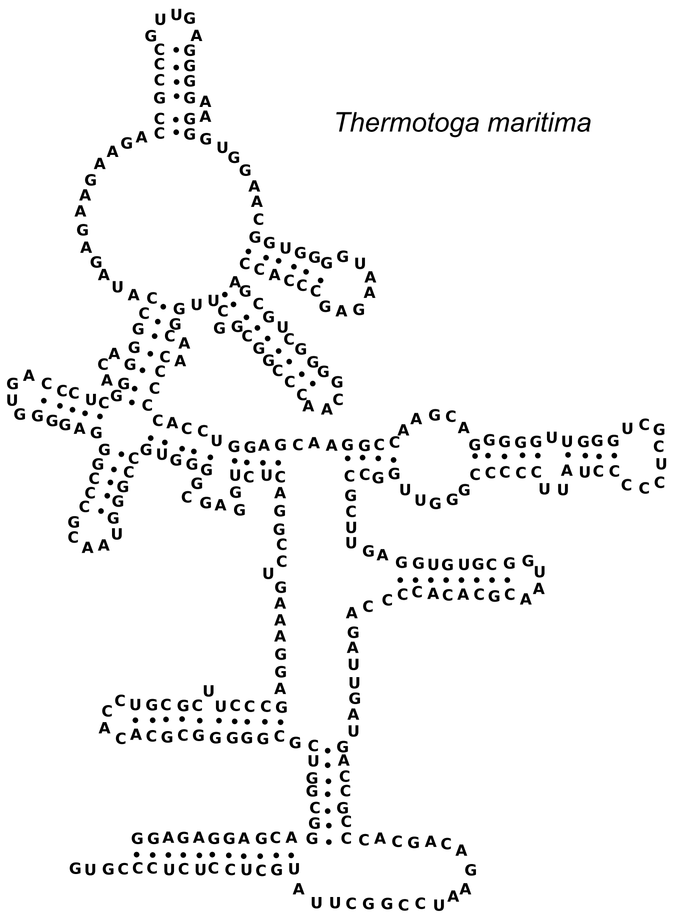
Problèmes

Problèmes de l'édition d'arbre pour la comparaison de structure secondaire d'ARN :

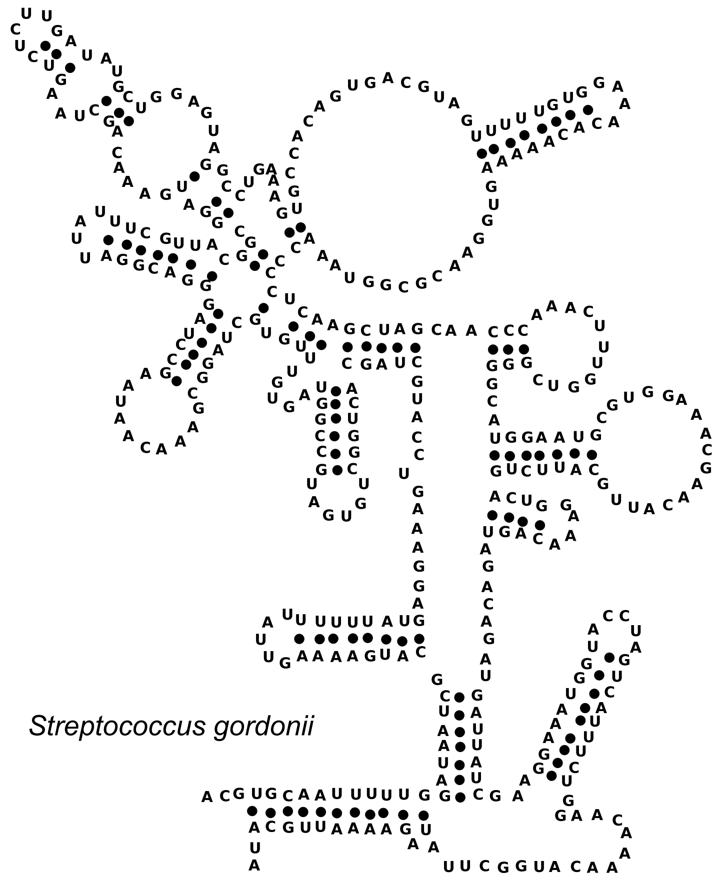
1. dispersion
2. association entre les arcs
3. association entre les noeuds

Résultats expérimentaux (1-a)

Thermotoga maritima :
RNase P

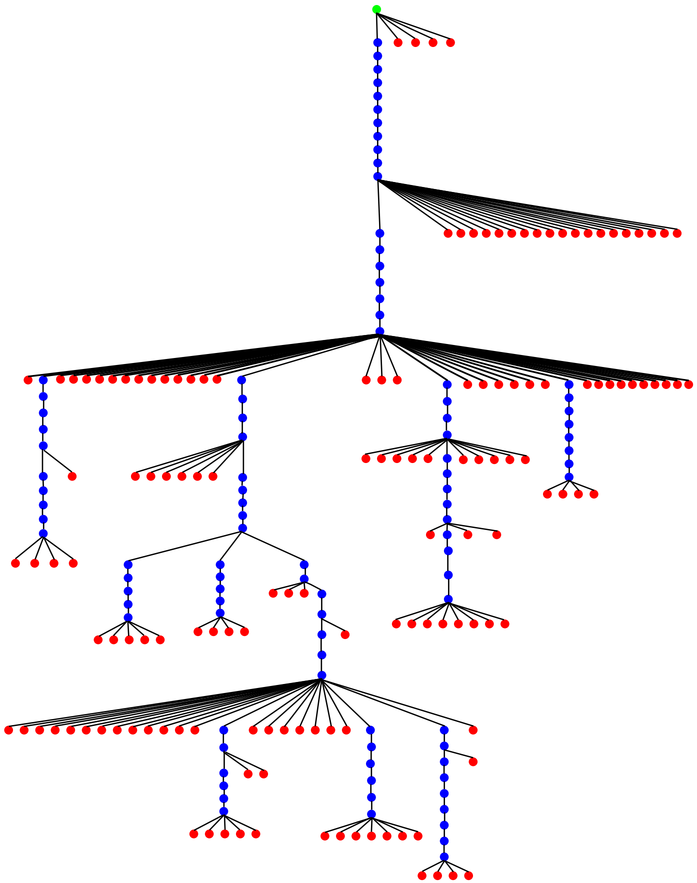


Streptococcus gordonii :
RNase P

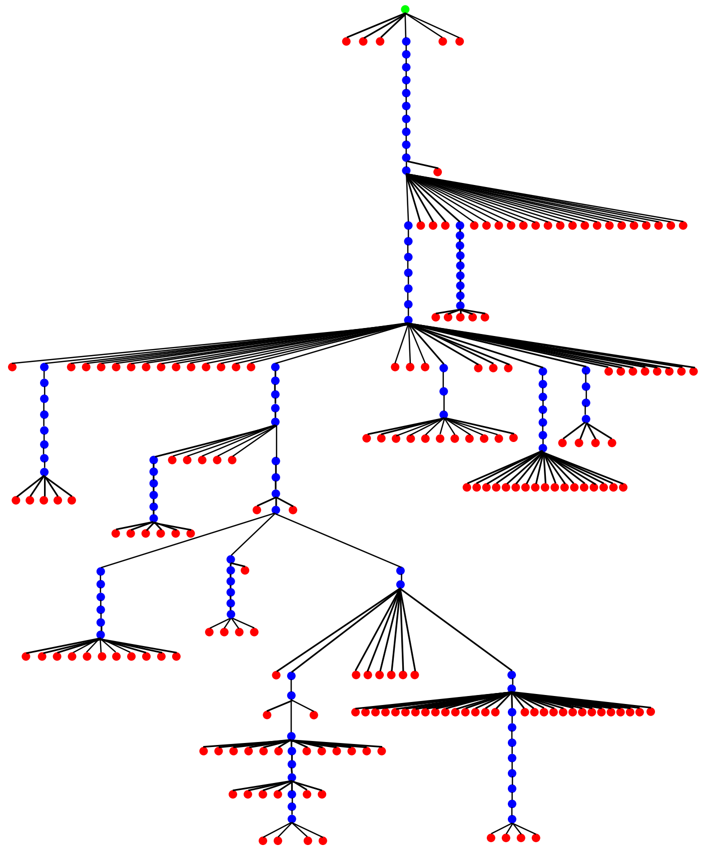


Résultats expérimentaux (1-b)

Thermotoga maritima :
RNase P

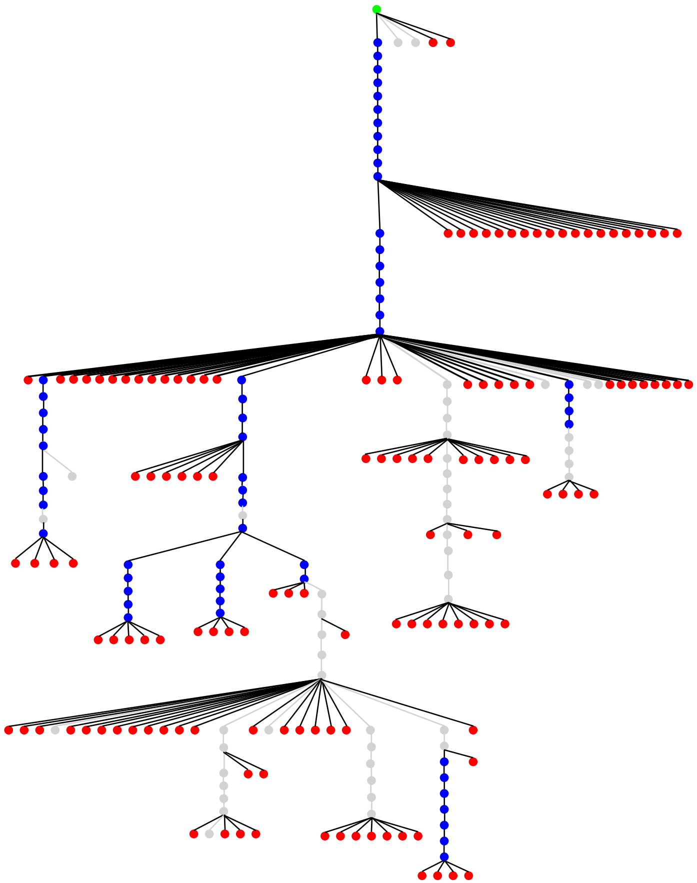


Streptococcus gordonii :
RNase P

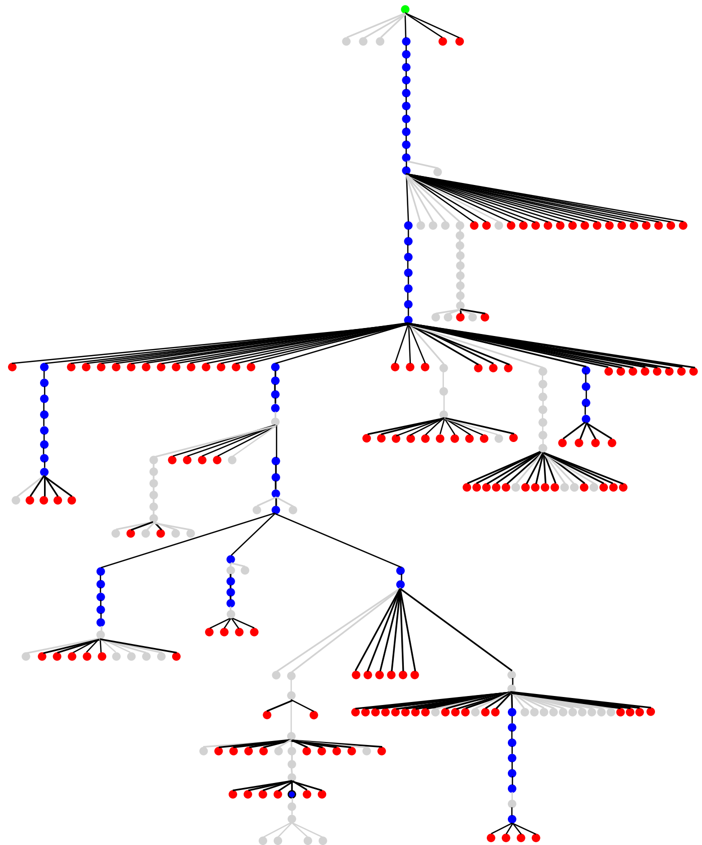


Résultats expérimentaux (1-c)

Thermotoga maritima :
RNase P

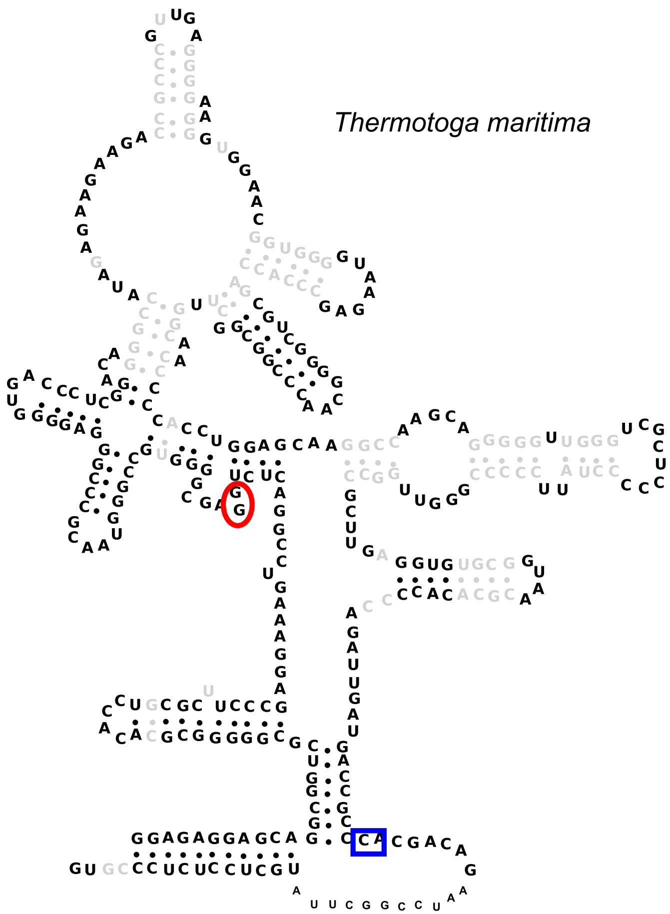


Streptococcus gordonii :
RNase P

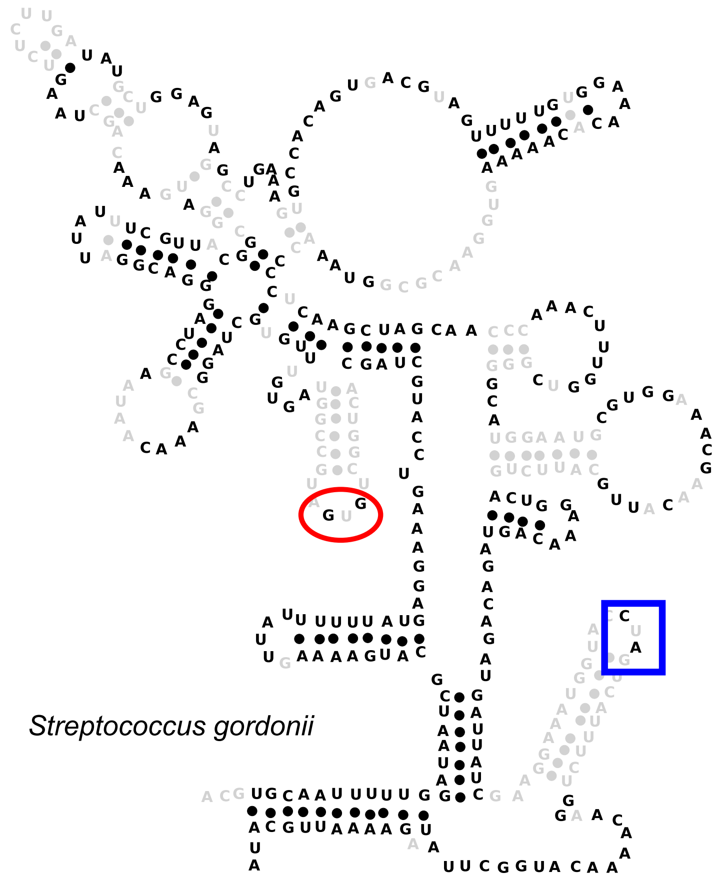


Résultats expérimentaux (1-d)

Thermotoga maritima :
RNase P



Streptococcus gordonii :
RNase P



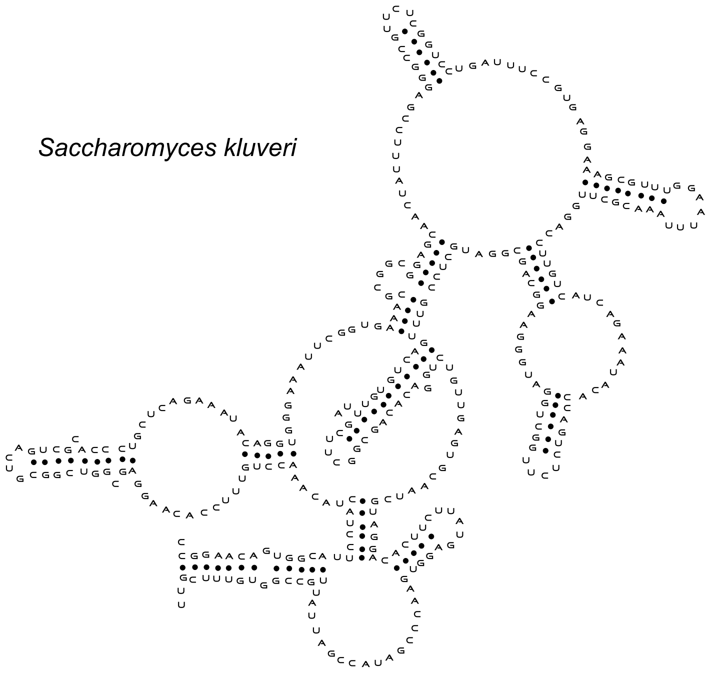
Problèmes

Problèmes de l'édition d'arbre pour la comparaison de structure secondaire d'ARN :

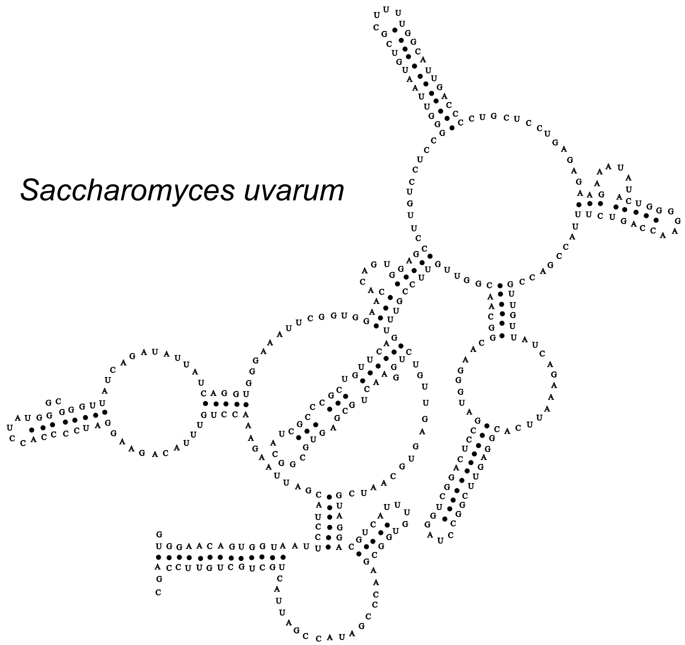
1. dispersion
2. association entre les arcs
3. association entre les noeuds

Résultats expérimentaux (2-a)

Saccharomyces kluveri : RNase P



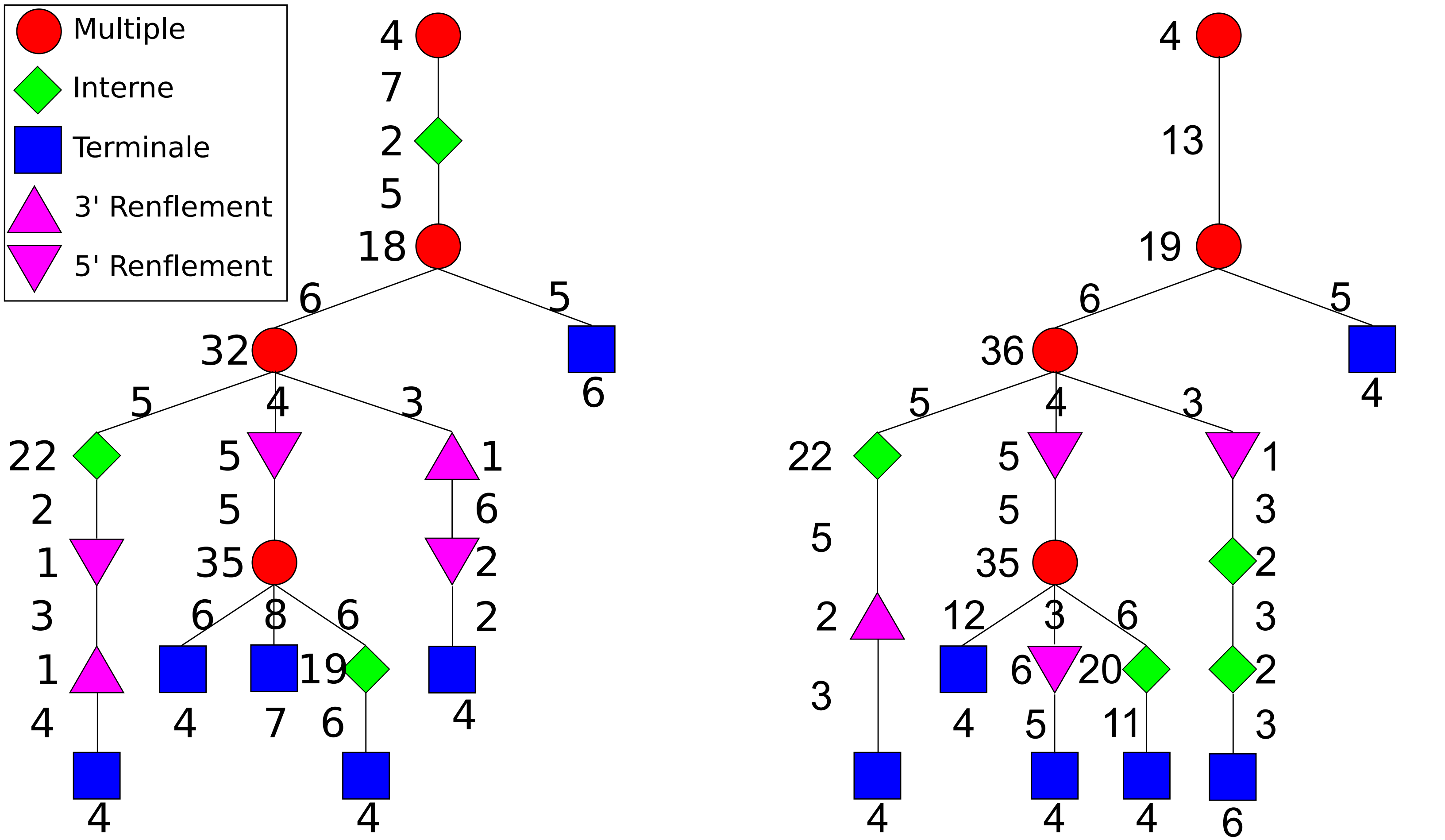
Saccharomyces uvarum : RNase P



Résultats expérimentaux (2-b)

Saccharomyces kluveri : RNase P

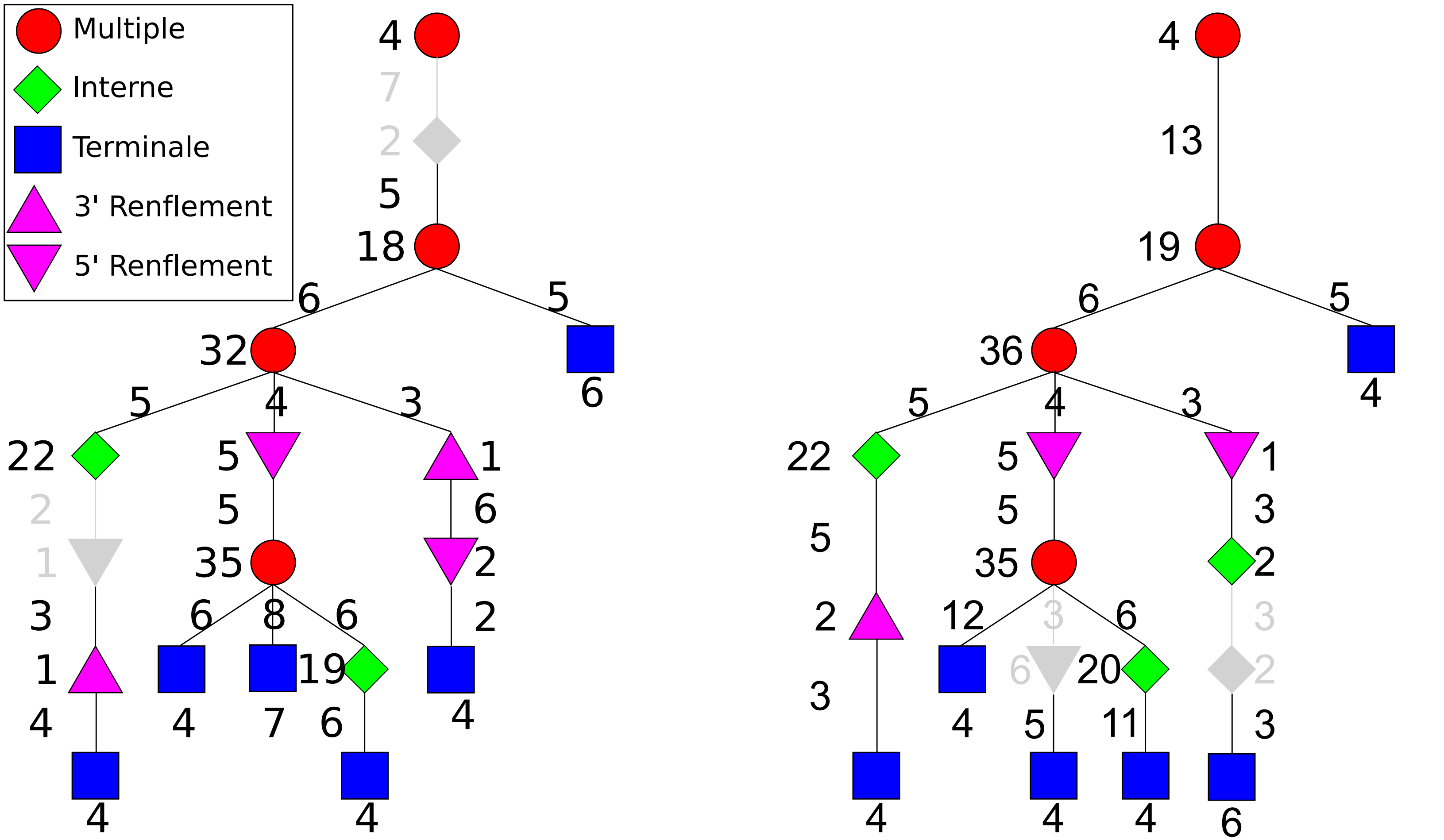
Saccharomyces uvarum : RNase P



Résultats expérimentaux (2-c)

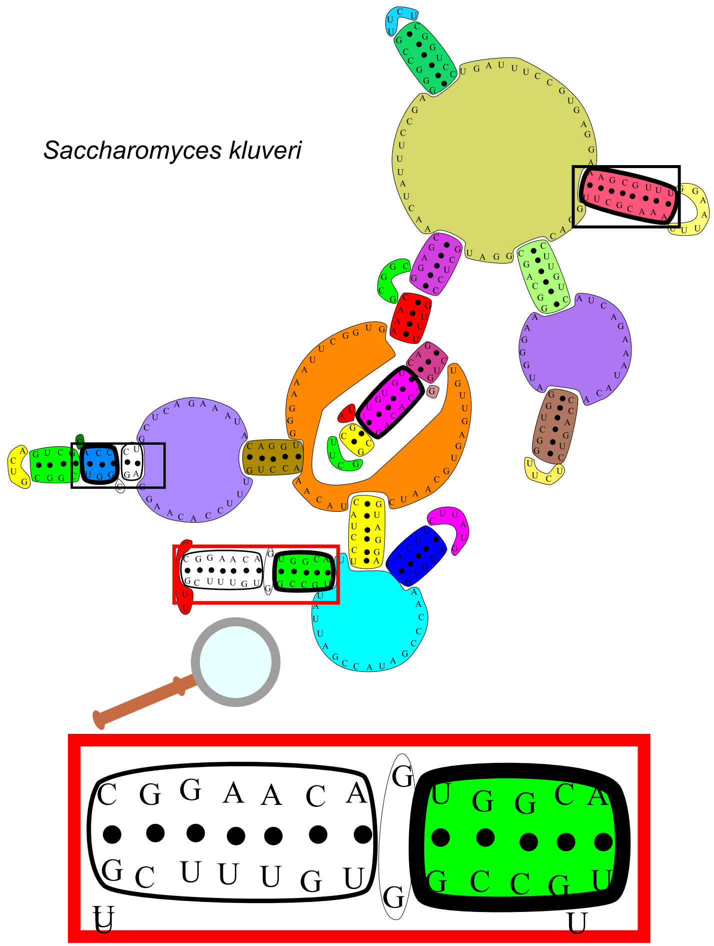
Saccharomyces kluveri : RNase P

Saccharomyces uvarum : RNase P

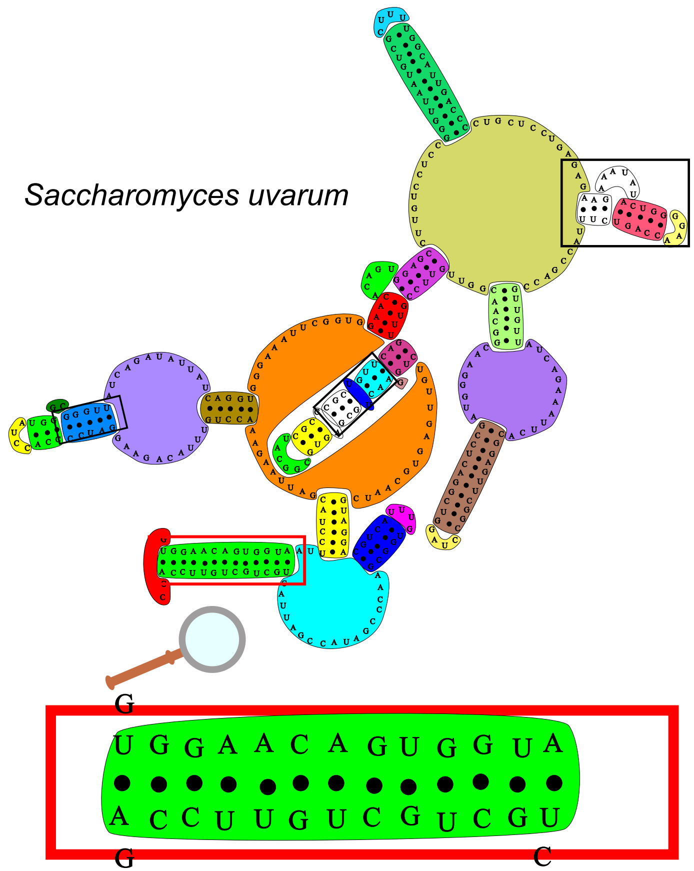


Résultats expérimentaux (2-d)

Saccharomyces kluveri :
RNase P



Saccharomyces uvarum :
RNase P



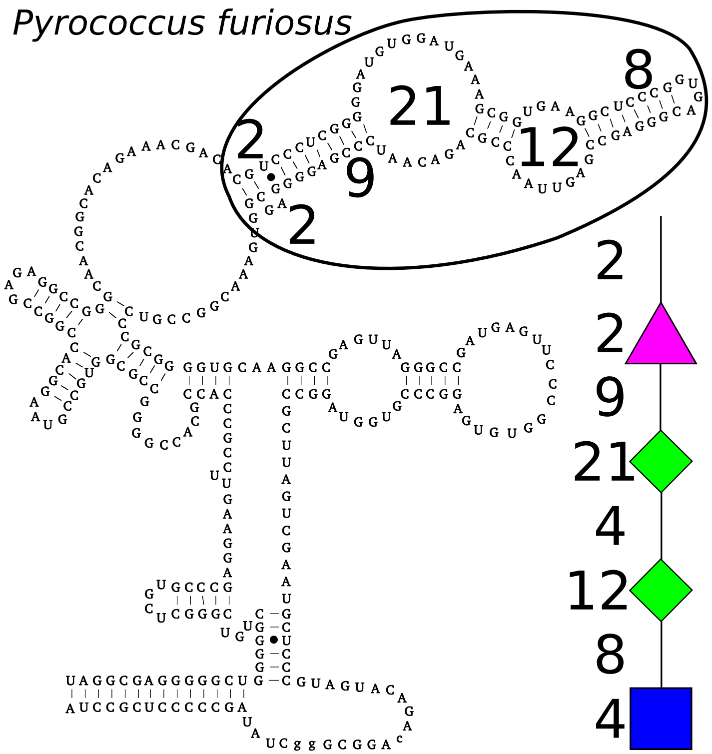
Problèmes

Problèmes de l'édition d'arbre pour la comparaison de structure secondaire d'ARN :

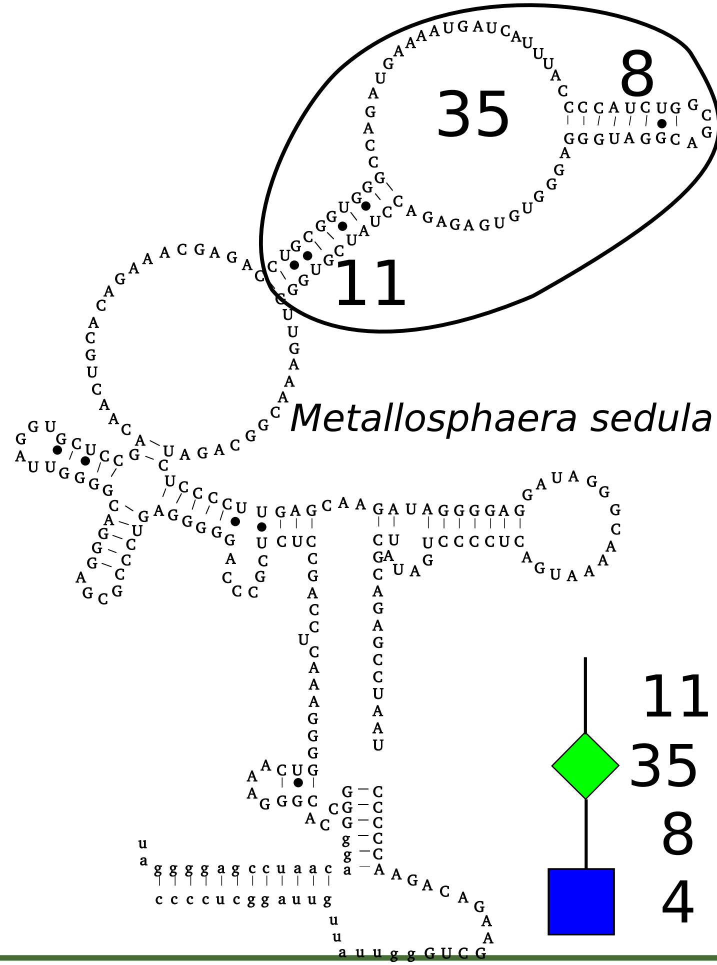
1. dispersion
2. association entre les arcs
3. association entre les noeuds

Résultats expérimentaux (3)

Pyrococcus furiosus : RNase P



Metallosphaera sedula : RNase P



Problèmes

Problèmes de l'édition d'arbre pour la comparaison de structure secondaire d'ARN :

1. dispersion
2. association entre les arcs
3. association entre les noeuds

Problèmes

Problèmes de l'édition d'arbre pour la comparaison de structure secondaire d'ARN :

1. dispersion
2. association entre les arcs
3. association entre les noeuds

$\Rightarrow$ nouvelle distance

Deux nouvelles opérations d’édition

[Allali, 04]

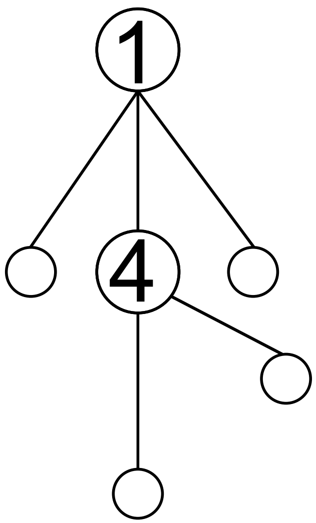
La fusion et la scission

Deux nouvelles opérations :

La fusion et la scission

Deux nouvelles opérations :

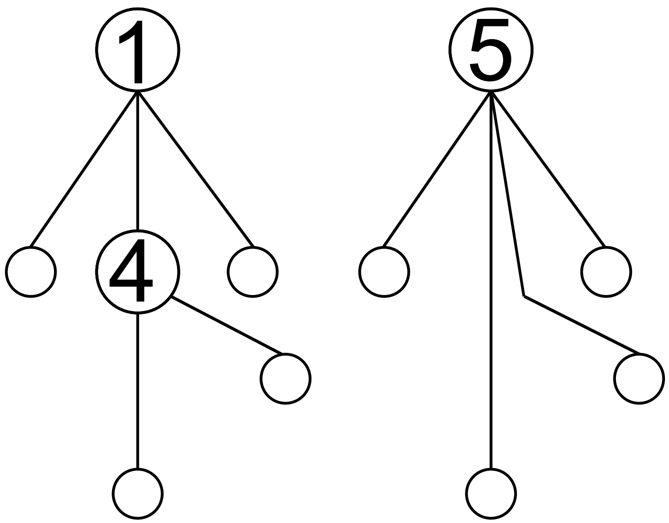
fusion/scission de noeuds :



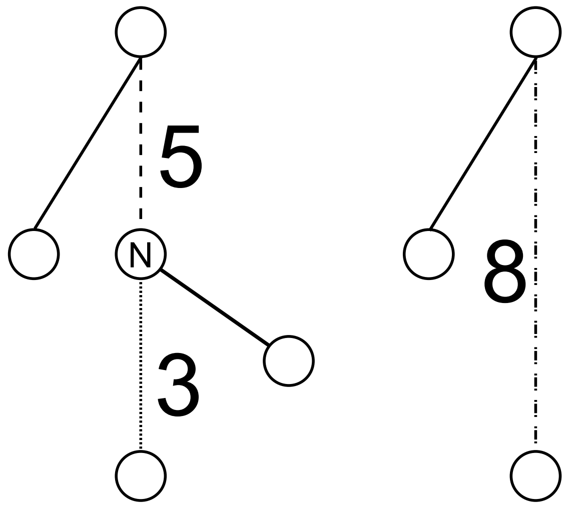
La fusion et la scission

Deux nouvelles opérations :

fusion/scission de noeuds :

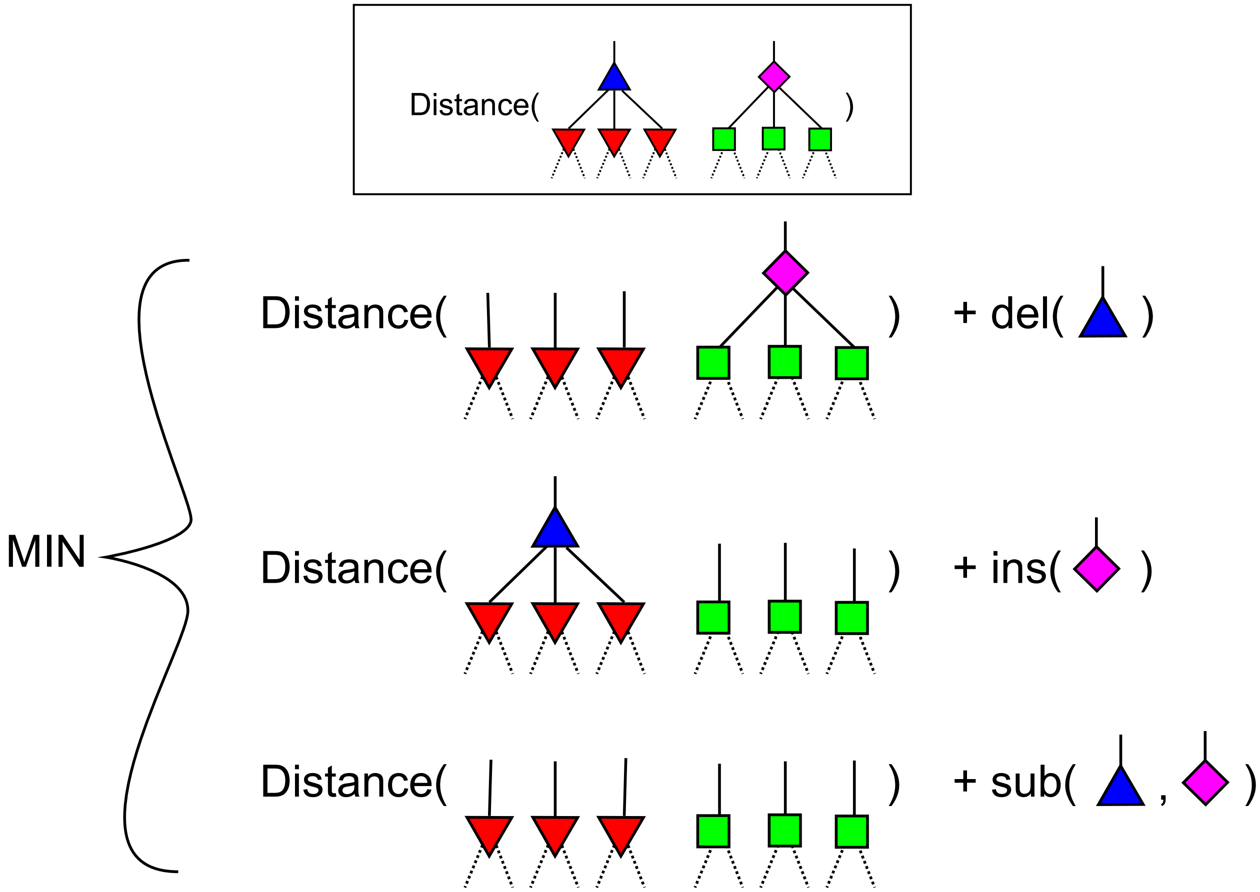


fusion/scission d’arcs :



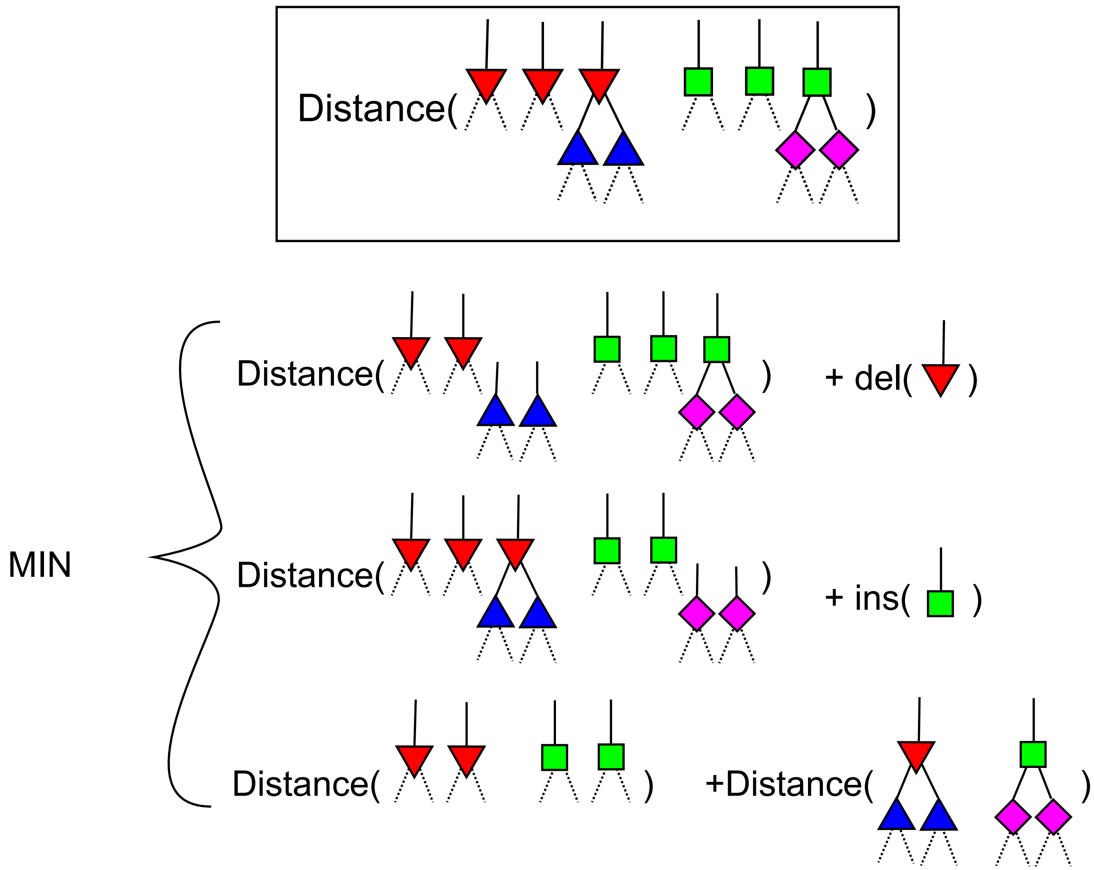
Algorithme (1)

Algorithme de Zhang and Shasha (89) :



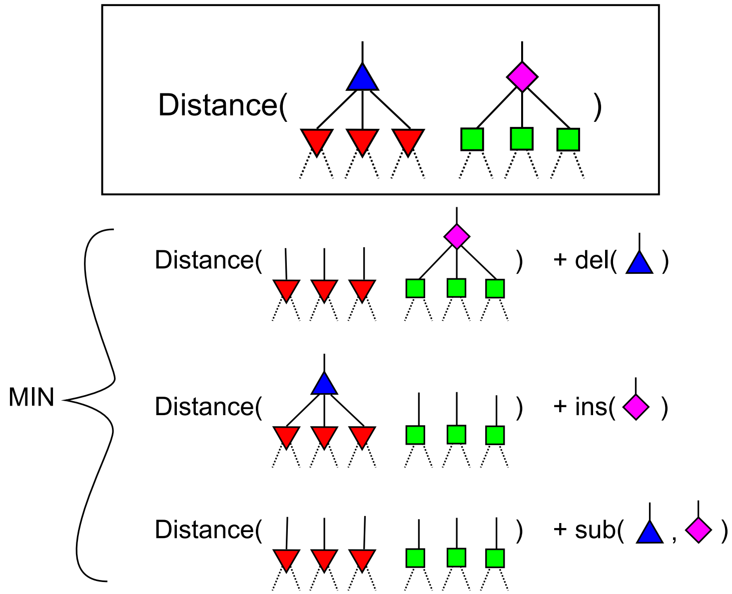
Algorithme (2)

Algorithme de Zhang and Shasha (89) :



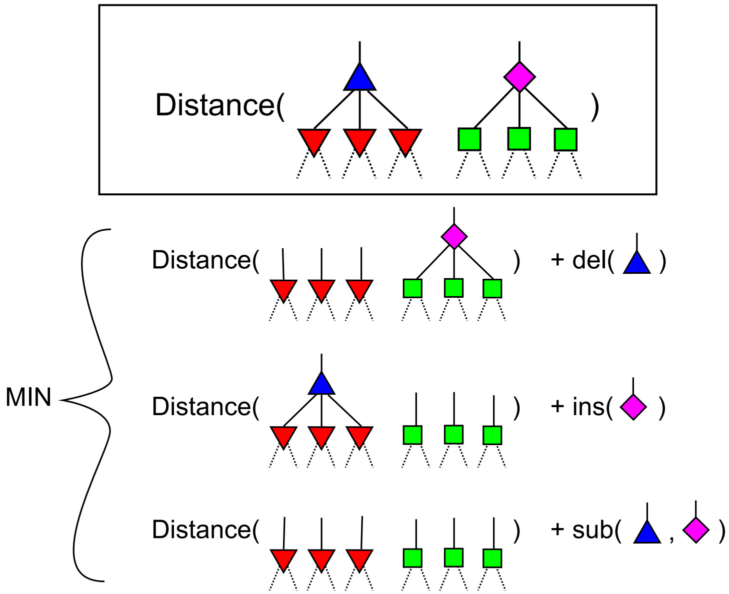
Algorithme (3)

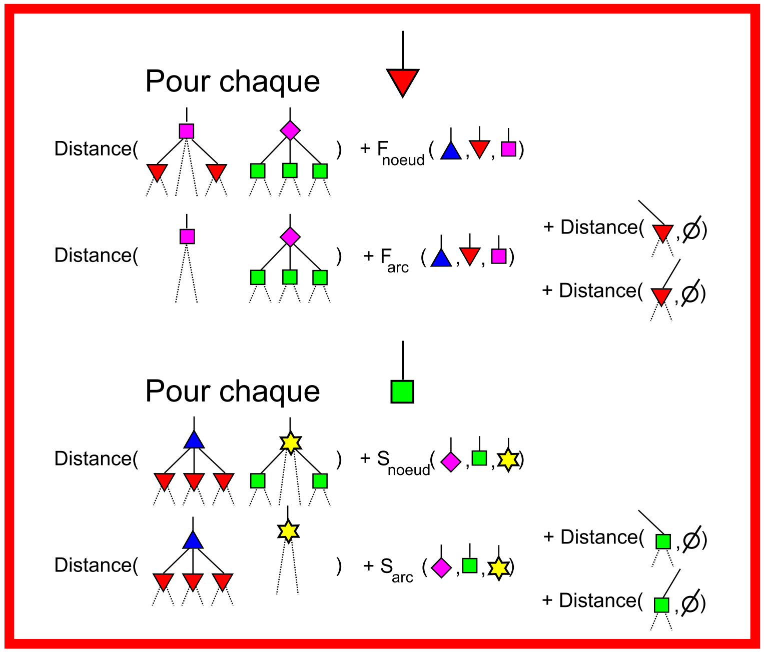
Algorithme de calcul de distance avec fusions :



Algorithme (3)

Algorithme de calcul de distance avec fusions :





Complexité

Zhang and Shasha :

$$\underline{\text{Temps :}}\; O(|T_1|^2 * |T_2|^2)$$

$$\underline{\text{Espace :}}\; O(|T_1| * |T_2|)$$

Complexité

Zhang and Shasha :

$$\underline{\text{Temps :}}\ O(|T_1|^2 * |T_2|^2)$$

$$\underline{\text{Espace :}}\ O(|T_1| * |T_2|)$$

Avec les opérations de fusions :

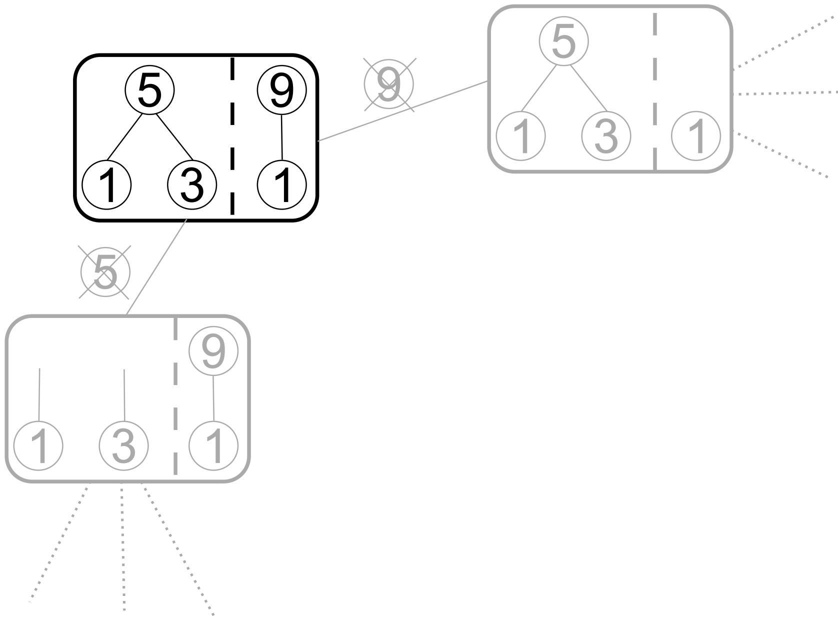
$$\underline{\text{Temps :}}\ O((2d_1)^l * (2d_2)^l * |T_1|^2 * |T_2|^2)$$

$$\underline{\text{Espace :}}\ O((2d_1)^l * (2d_2)^l * |T_1| * |T_2|)$$

d_i : degré maximal des noeuds de T_i

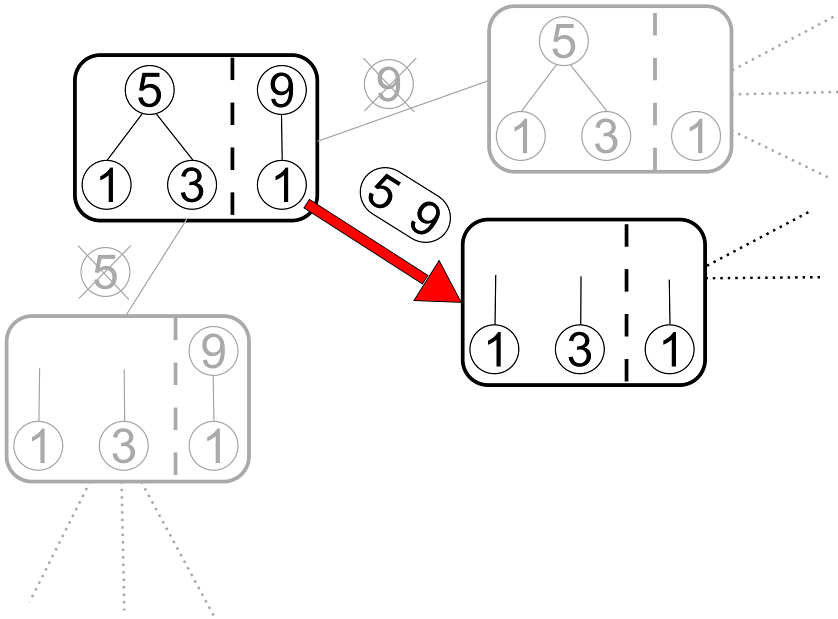
l : nombre de fusions **successives** par noeud

Fonctions de score



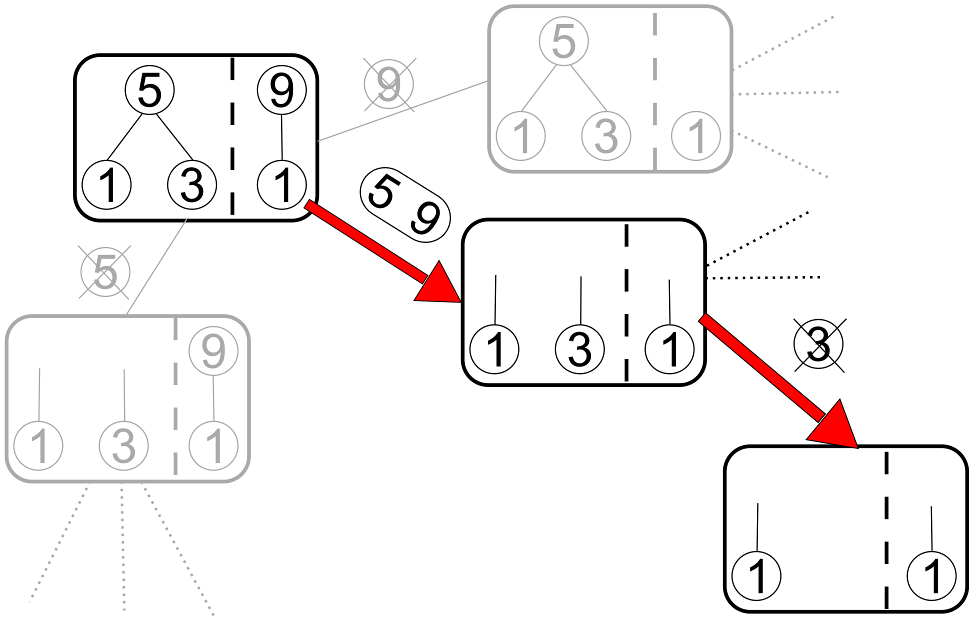
$$noeud_{fusion}(n, m) = del(m) + t$$

Fonctions de score



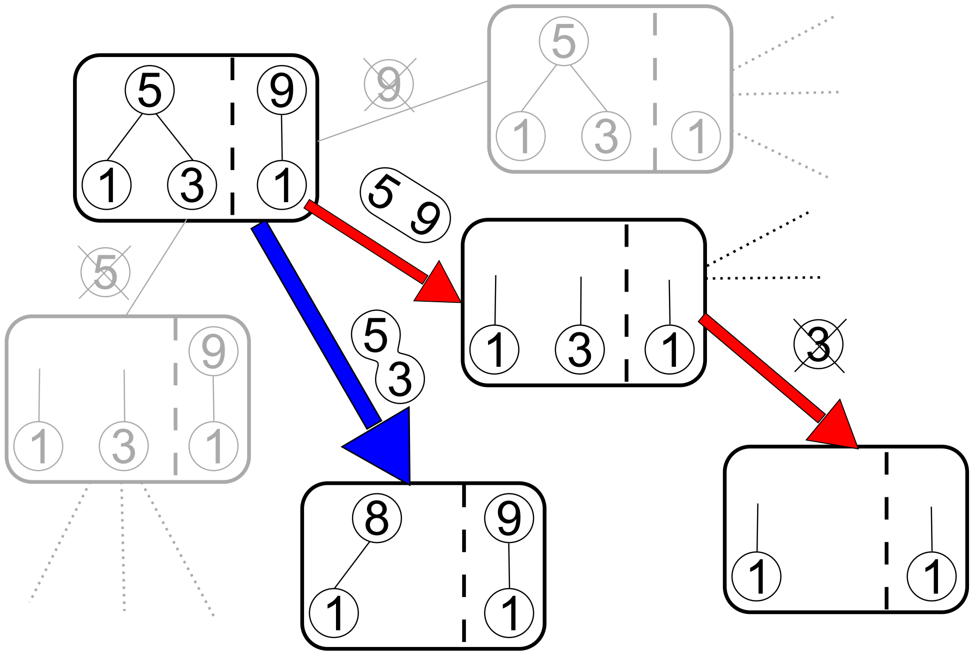
$$noeud_{fusion}(n, m) = del(m) + t$$

Fonctions de score



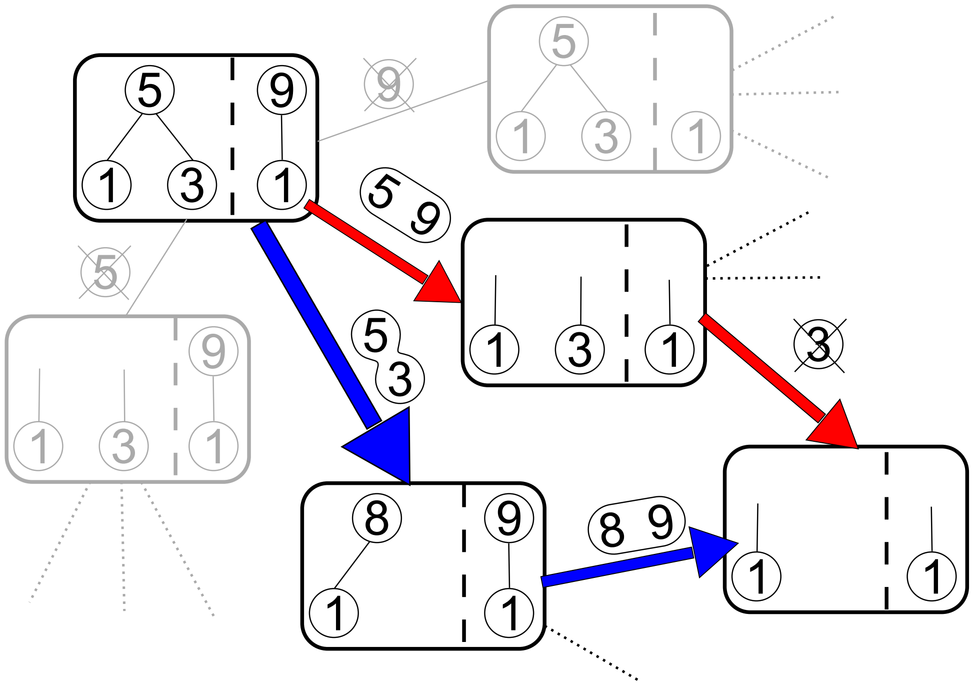
$$noeud_{fusion}(n, m) = del(m) + t$$

Fonctions de score



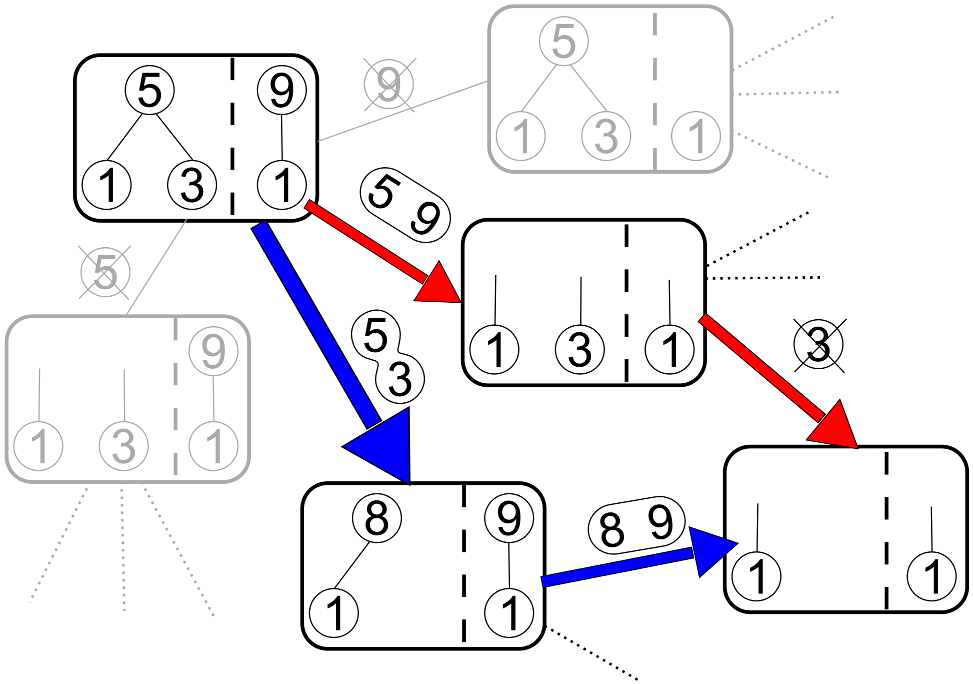
$$noeud_{fusion}(n, m) = del(m) + t$$

Fonctions de score



$$noeud_{fusion}(n, m) = del(m) + t$$

Fonctions de score

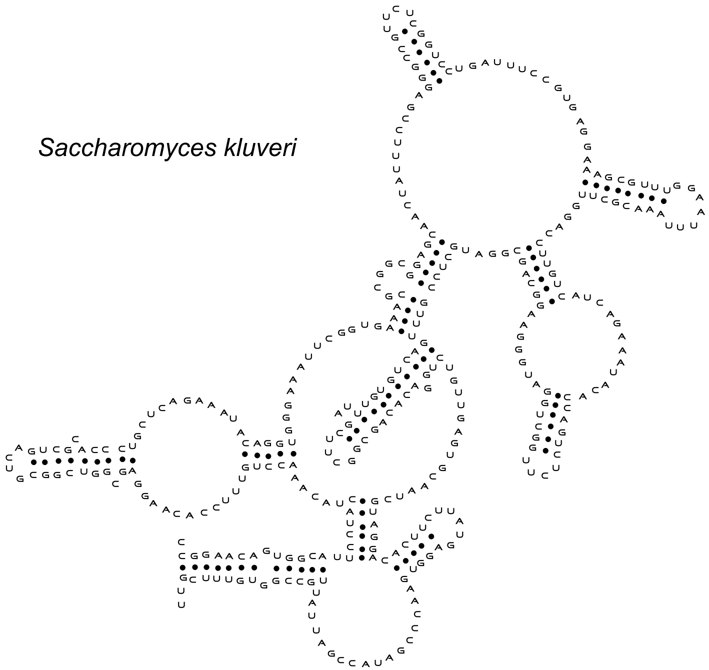


$$noeud_{fusion}(n, m) = del(m) + t$$
$$arc_{fusion}(n, m) = del(n) + t + \sum del_{arbre}(m_b)$$

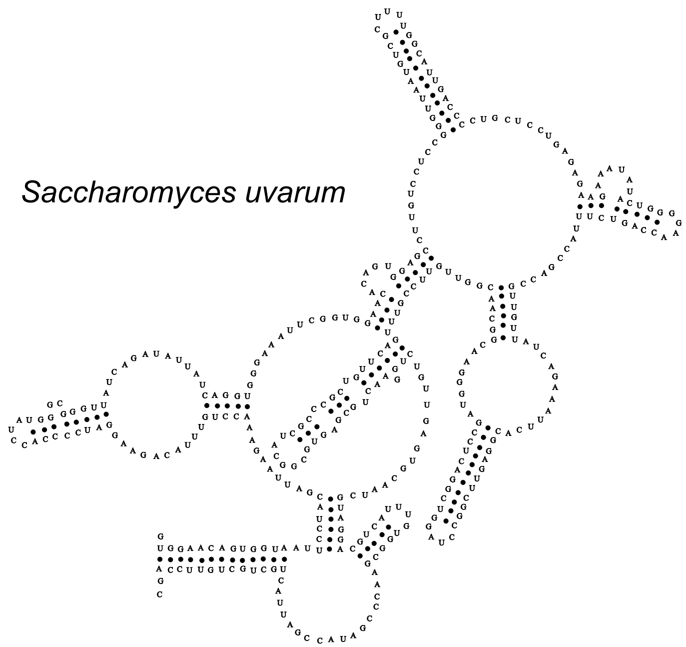
Nouveaux résultats pratiques

Résultats (1-a)

Saccharomyces kluveri :
RNase P

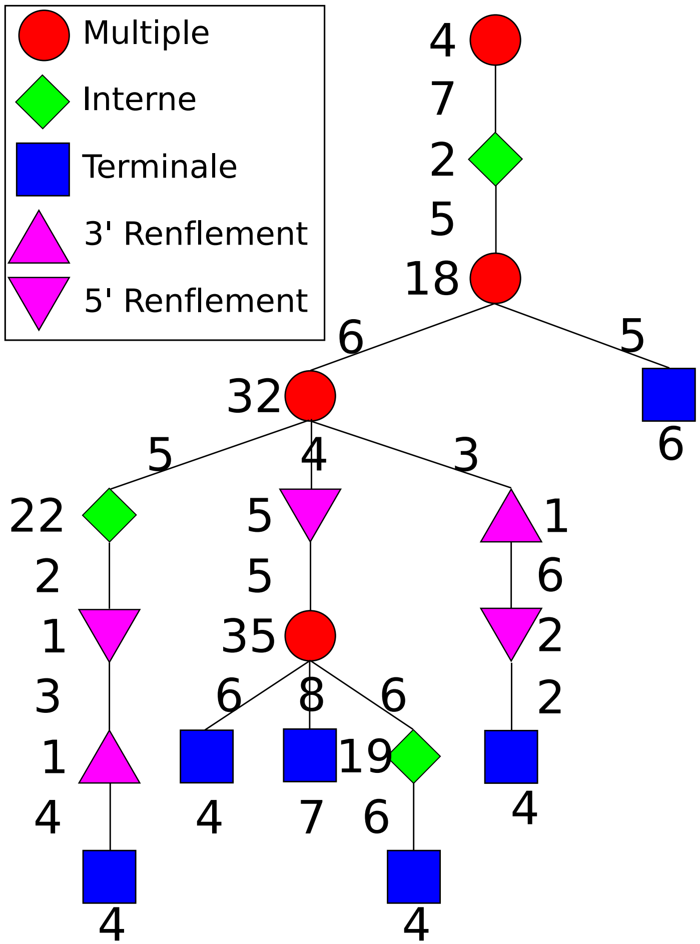


Saccharomyces uvarum :
RNase P

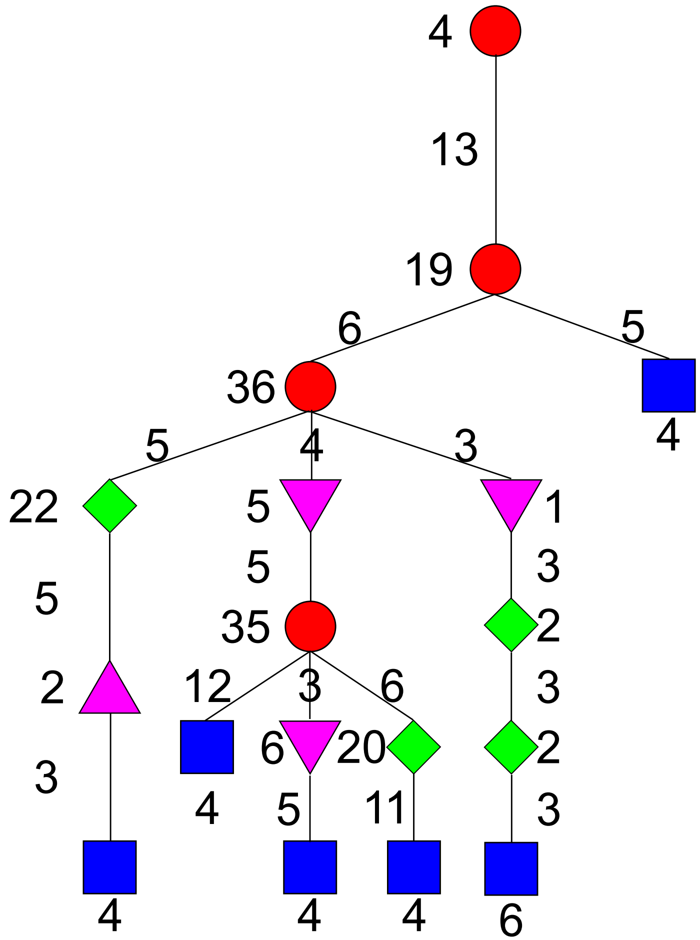


Résultats (1-b)

Saccharomyces kluveri :
RNase P

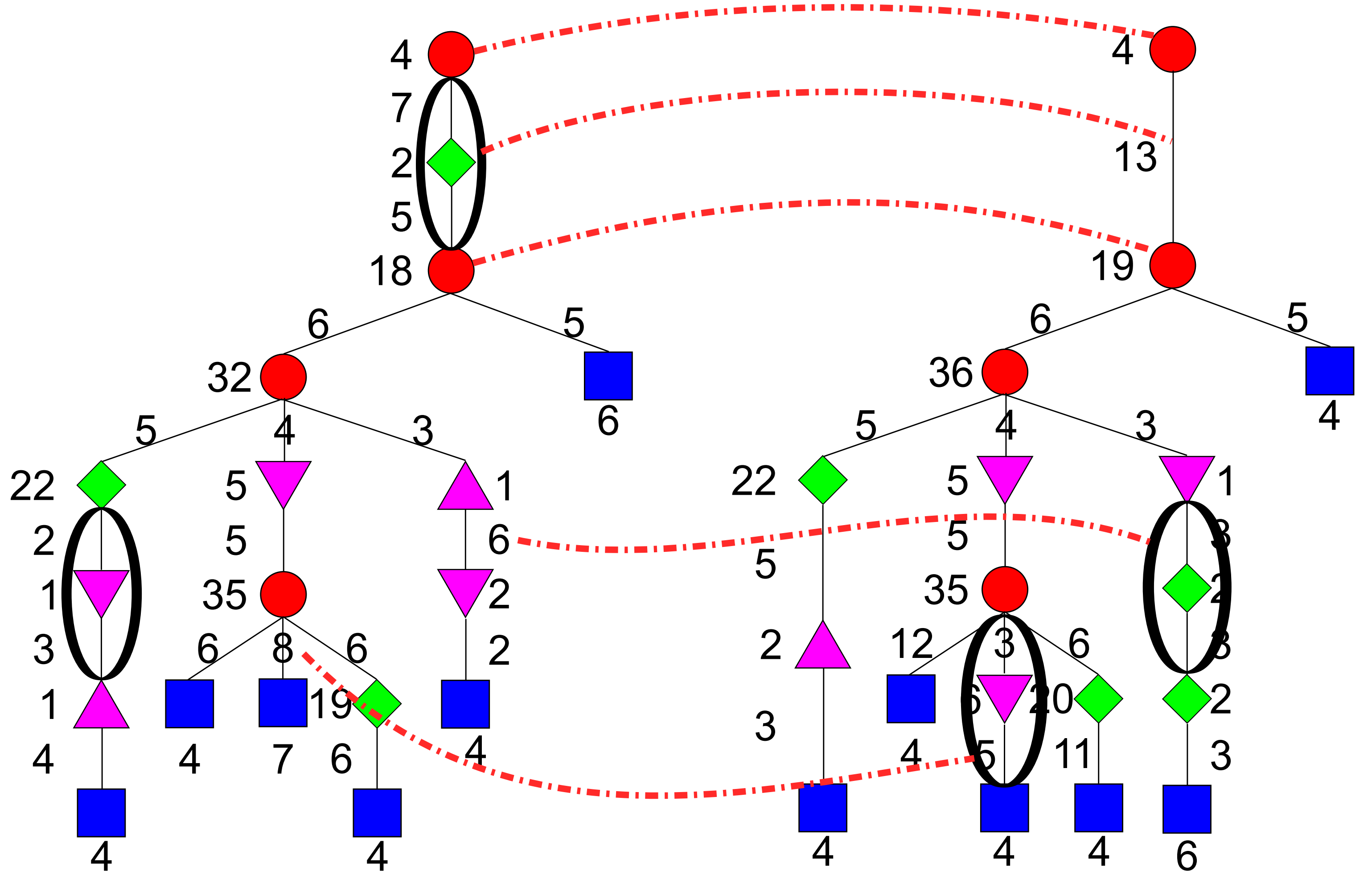


Saccharomyces uvarum :
RNase P



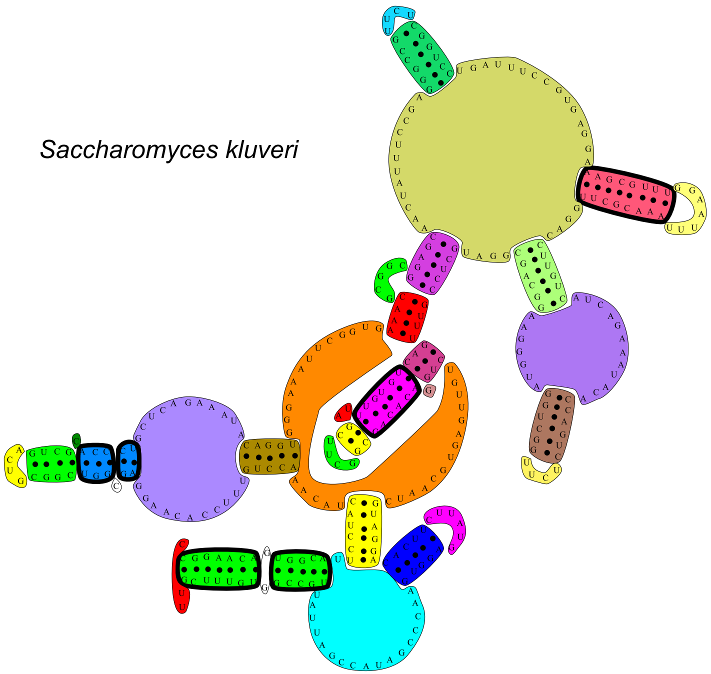
Résultats (1-c)

Saccharomyces kluveri *Saccharomyces uvarum*

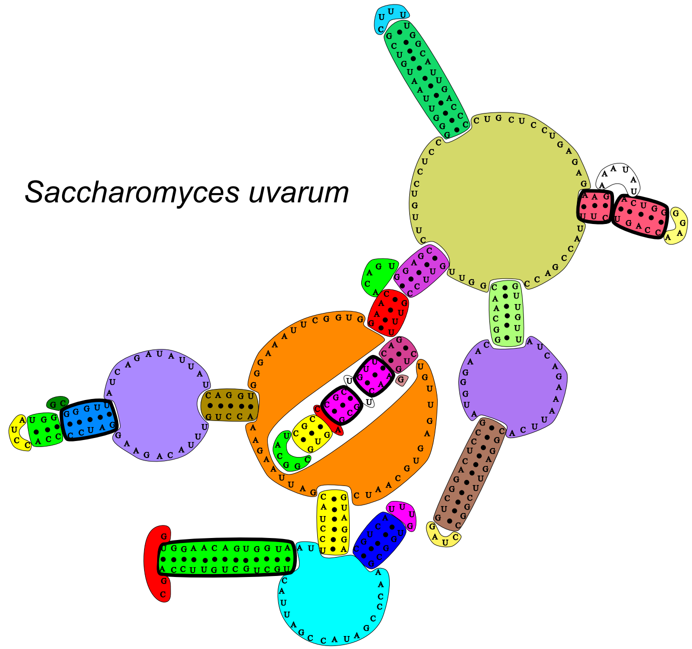


Résultats (1-d)

Saccharomyces kluveri :
RNase P



Saccharomyces uvarum :
RNase P



MiGaL

[Allali, 04]

Faits biologiques

Faits biologiques

- La structure secondaire est plus conservée que la séquence

Faits biologiques

- La structure secondaire est plus conservée que la séquence
- Le réseau de boucles multiples constitue le squelette de la structure

Faits biologiques

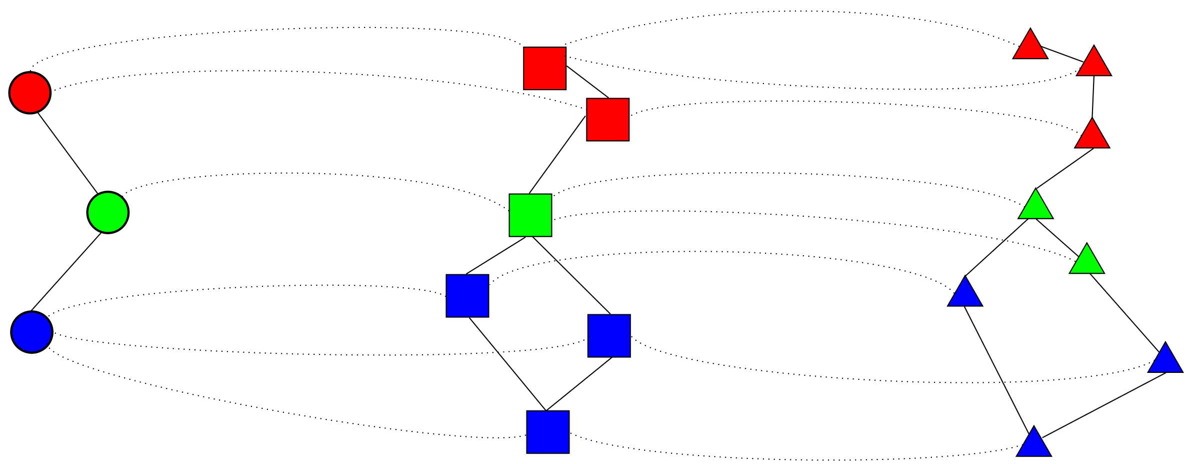
- La structure secondaire est plus conservée que la séquence
- Le réseau de boucles multiples constitue le squelette de la structure

Nouvelle modélisation basée sur différentes vues de la structure secondaire

MiGaL

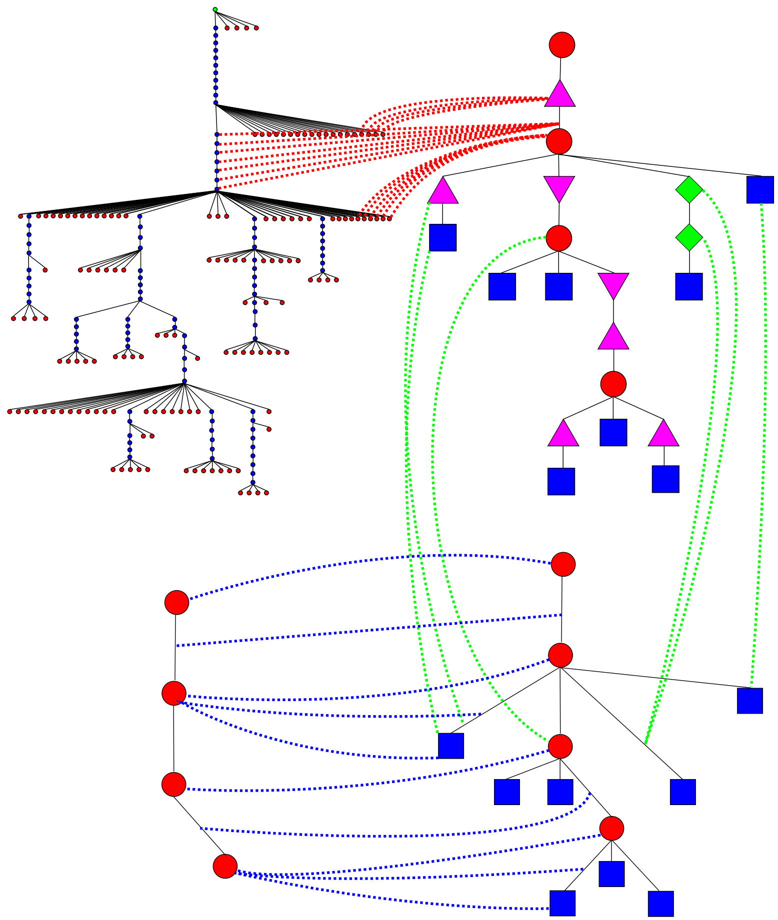
Structure MiGaL (*Multiple Graph Layer*) $M(G, R)$:

- G ensemble de N graphes $G_1, \ldots, G_N$
- R ensemble de $N-1$ raffinements $\alpha_1, \ldots, \alpha_{N-1}$:
  $\alpha_i : S_i \longrightarrow \mathcal{P}(S_{i+1})$



RNA-MiGaL

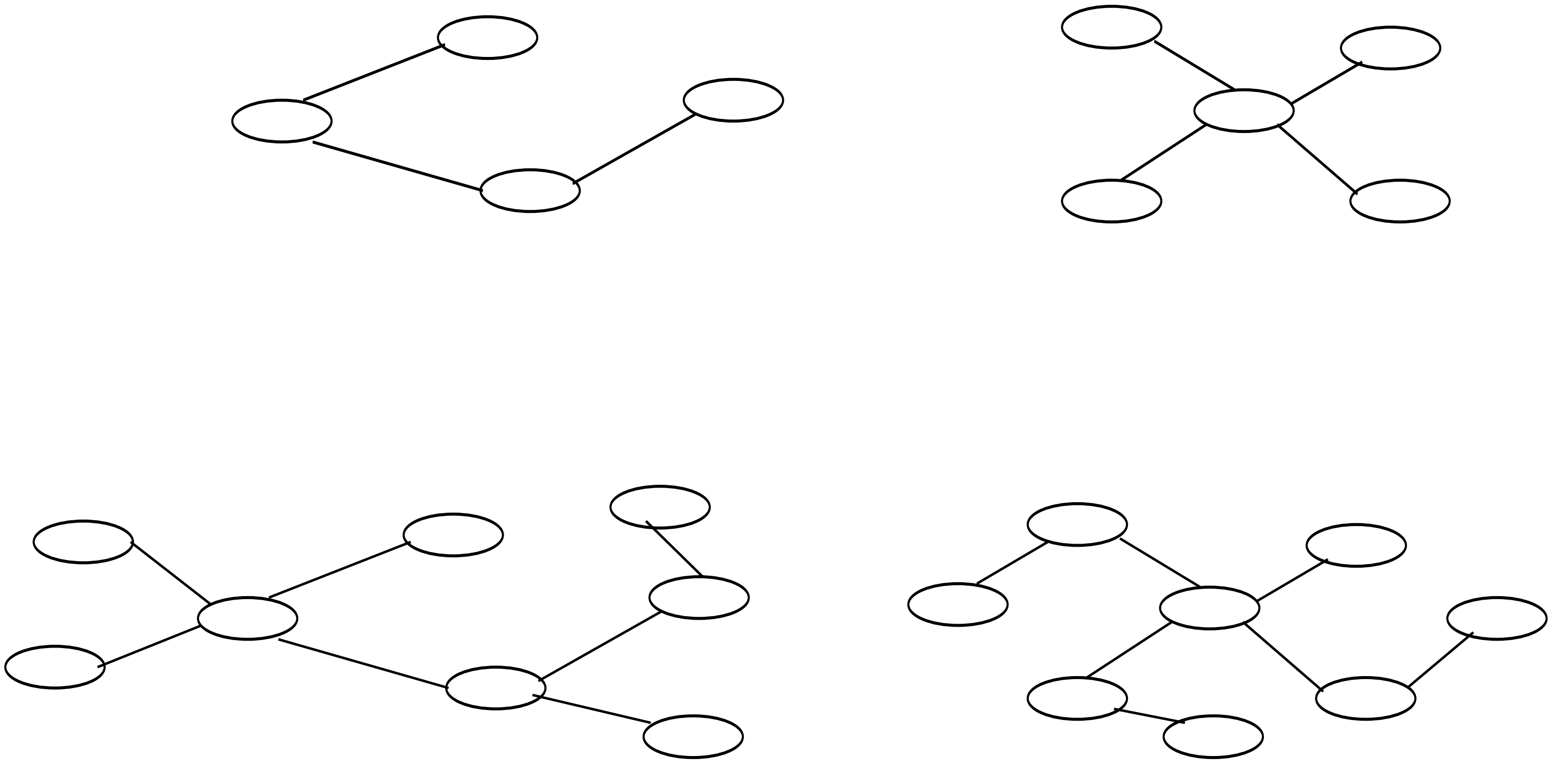
4 niveaux d'arbres enracinés, ordonnés :



- T_0 : réseau de boucles multiples
- T_1 : tiges
- T_2 : hélices
- T_3 : paires de bases

Algorithme de comparaison

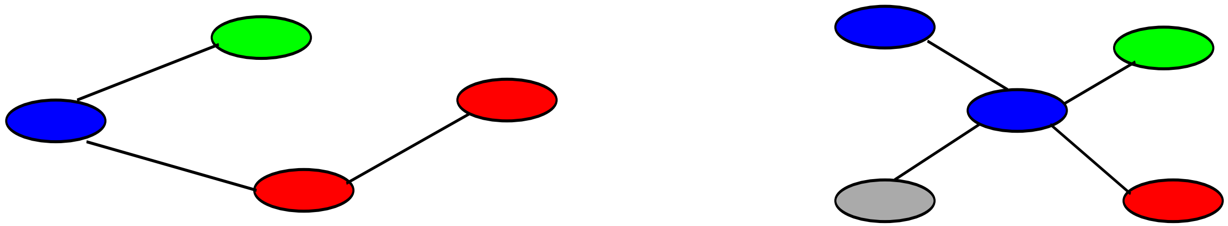
niveau par niveau de G_1 à G_N :

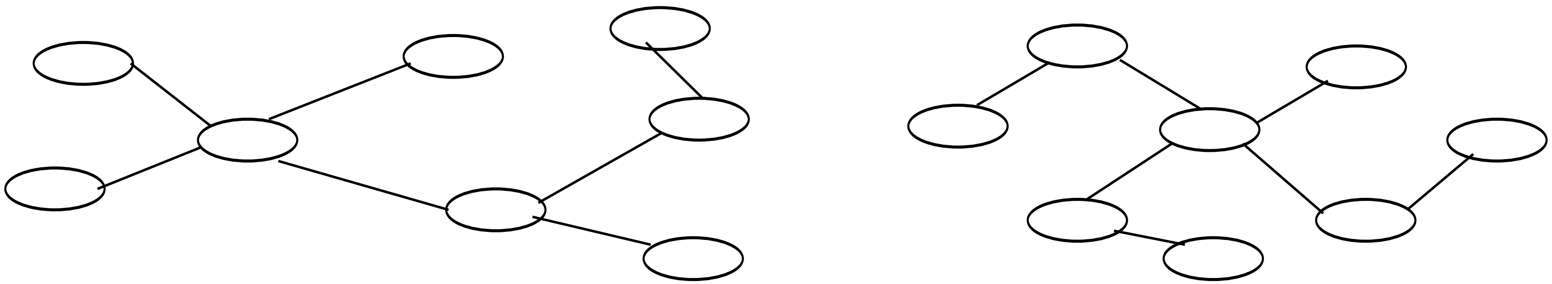


Les noeuds des graphes G_0 sont tous de la même couleur

Algorithme de comparaison

niveau par niveau de G_1 à G_N :

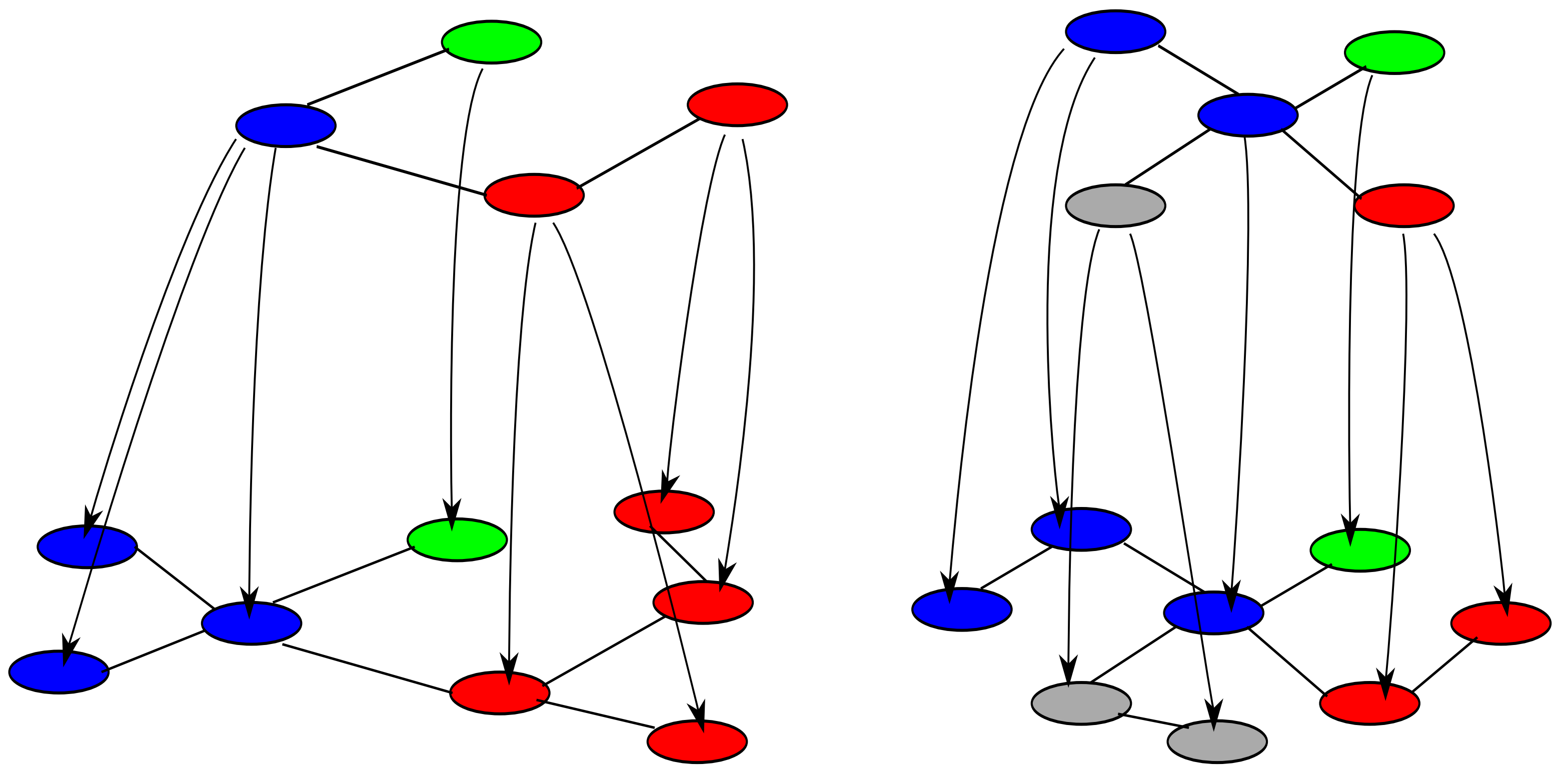




association entre les sommets des G_i

Algorithme de comparaison

niveau par niveau de G_1 à G_N :



Coloriage et report des couleurs

Comparaison de RNA-MiGaL(1)

Utilisation de la distance d’édition avec fusion pour la comparaison de chaque niveau :

$\Rightarrow$ La substitution et la fusion sont conditionnées par la couleur des noeuds

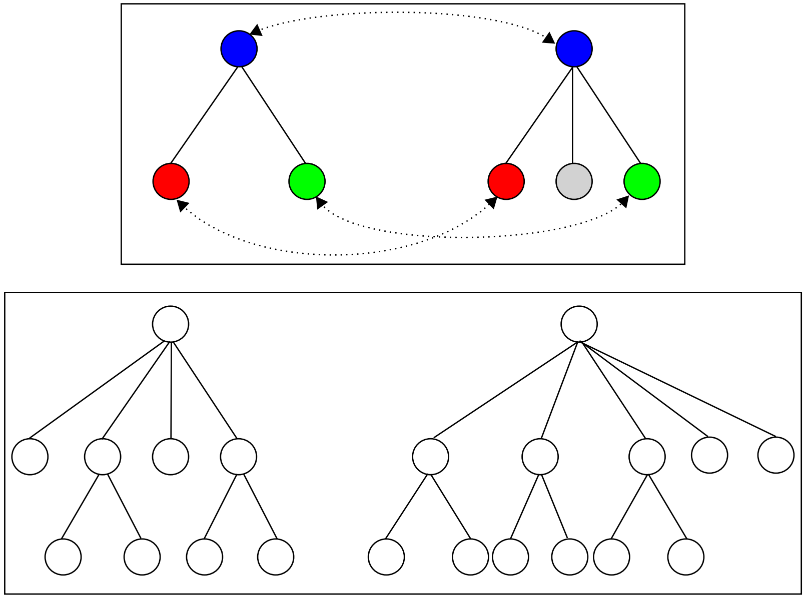
Comparaison de RNA-MiGaL(1)

Utilisation de la distance d'édition avec fusion pour la comparaison de chaque niveau :

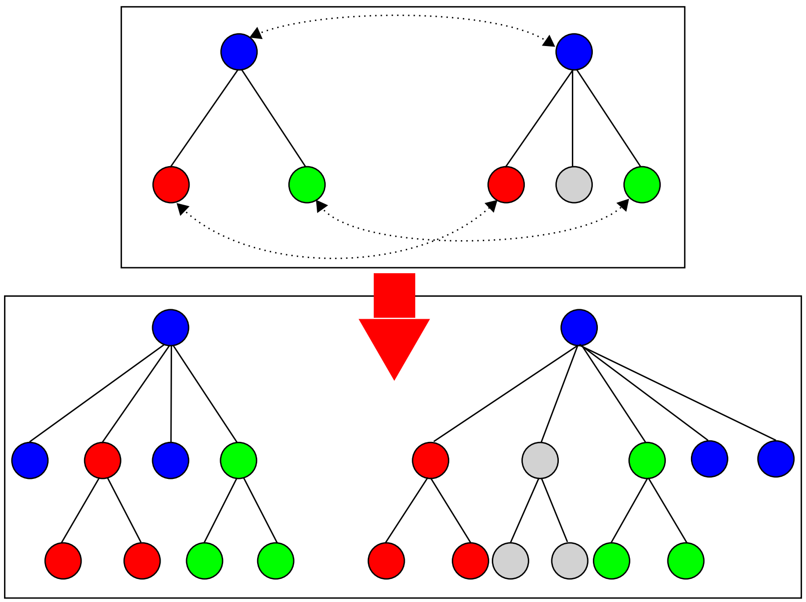
$\Rightarrow$ La substitution et la fusion sont conditionnées par la couleur des noeuds

> Amélioration du temps d'exécution par calcul de la distance entre chaque sous-arbre de même couleur de façon séparée?

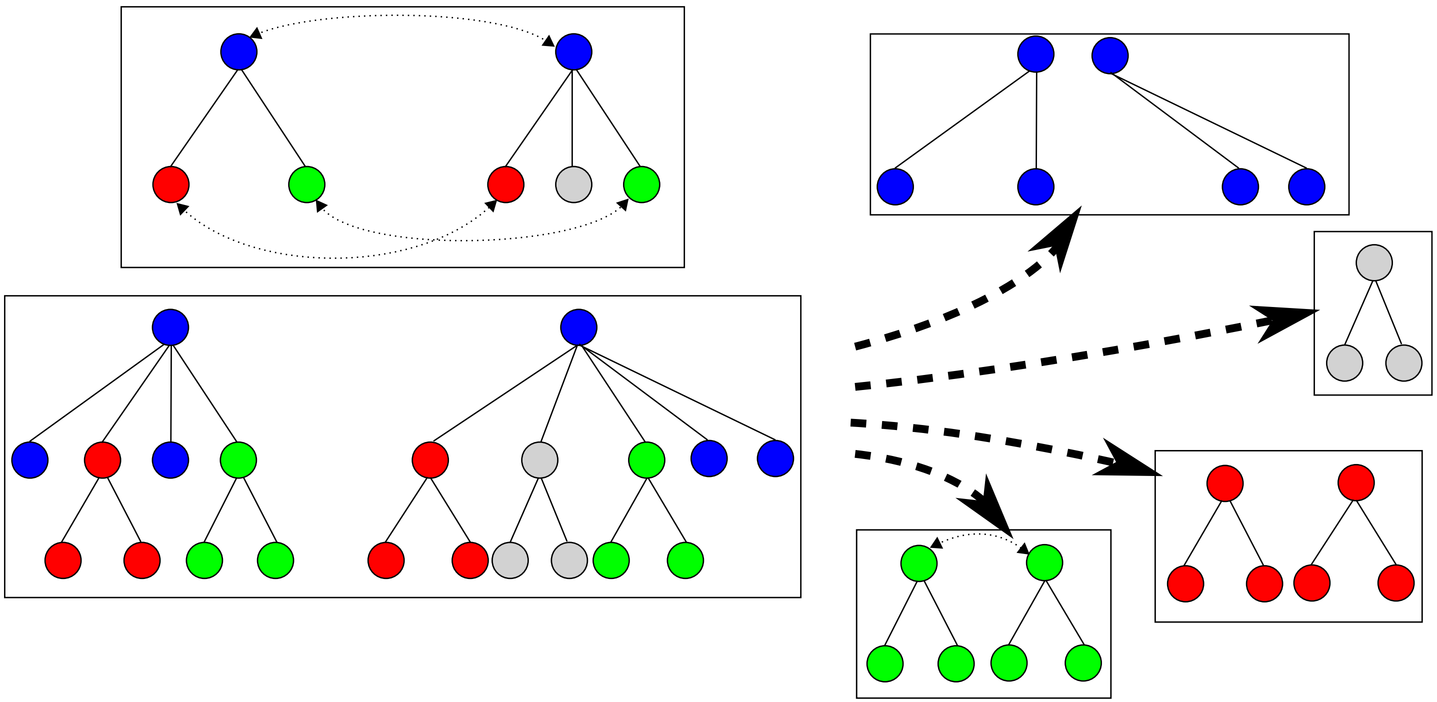
Séparation par couleur : problème



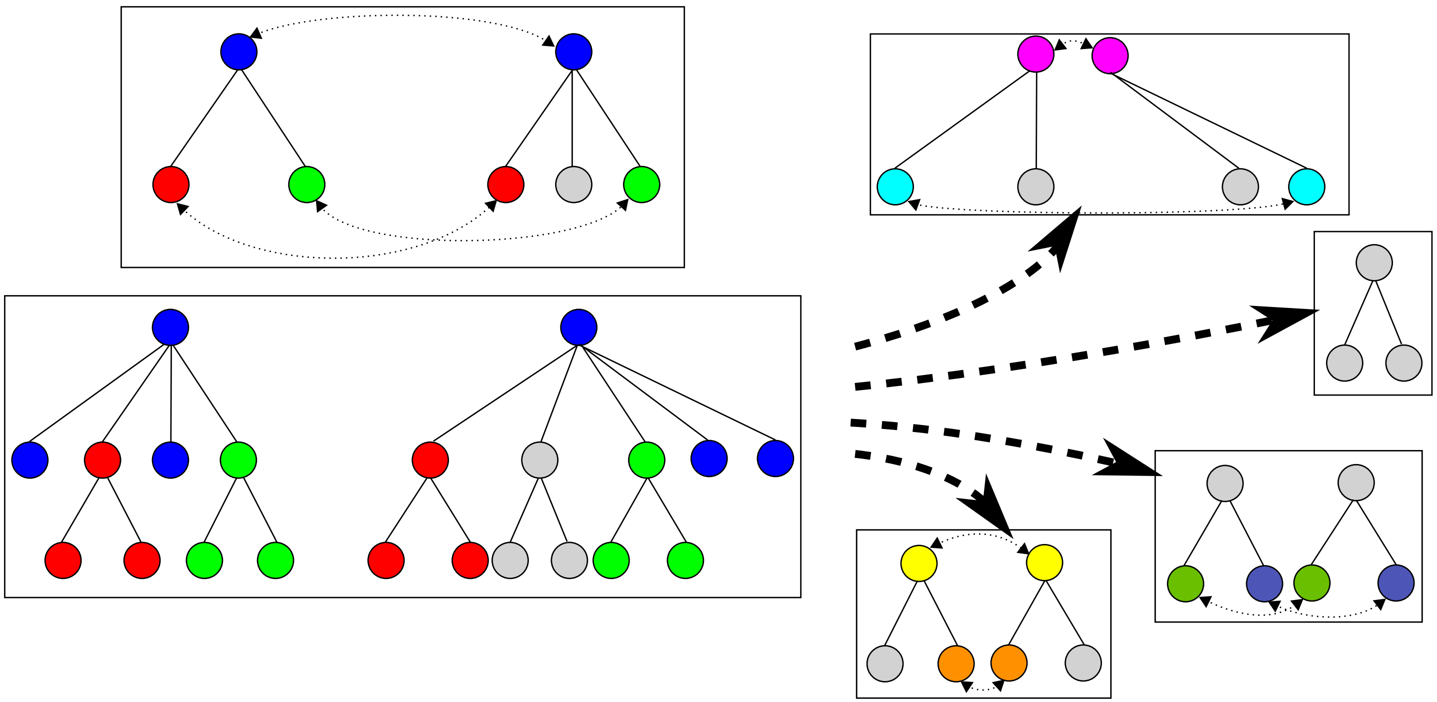
Séparation par couleur : problème



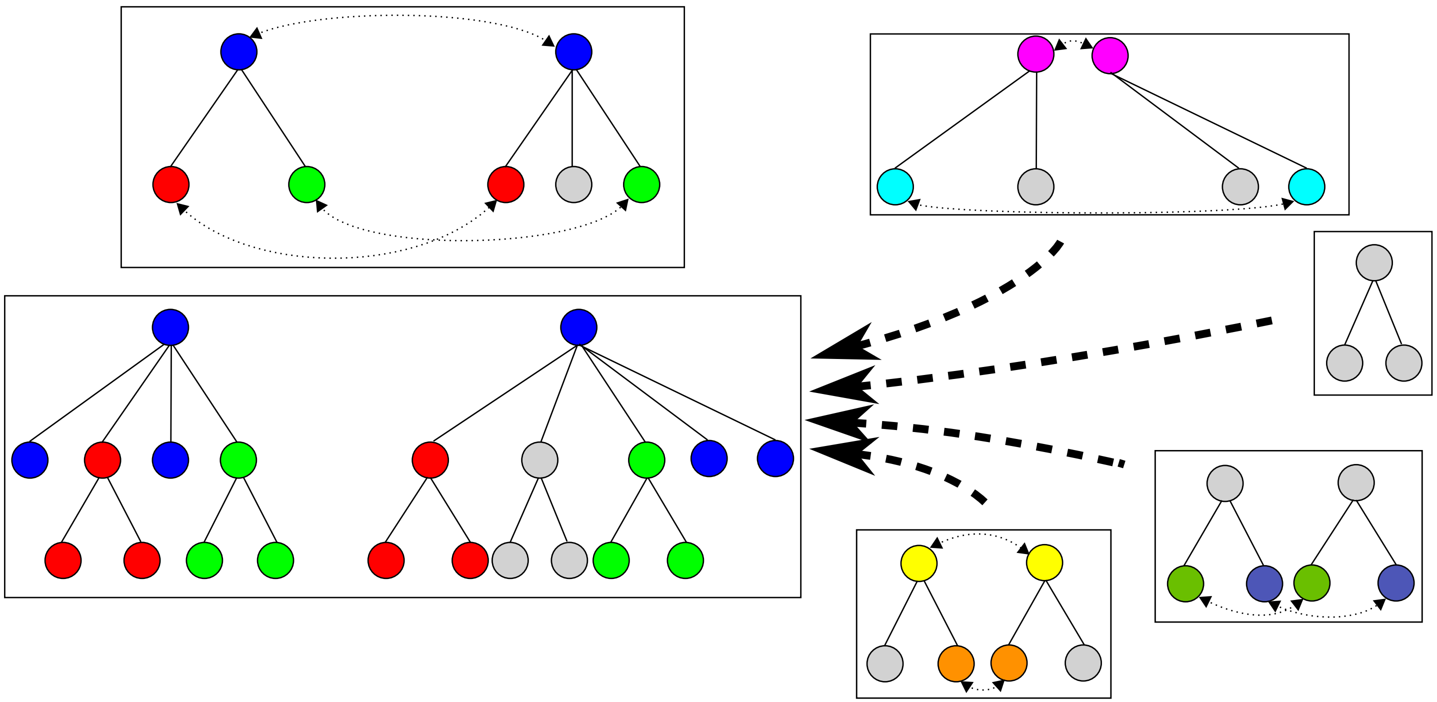
Séparation par couleur : problème



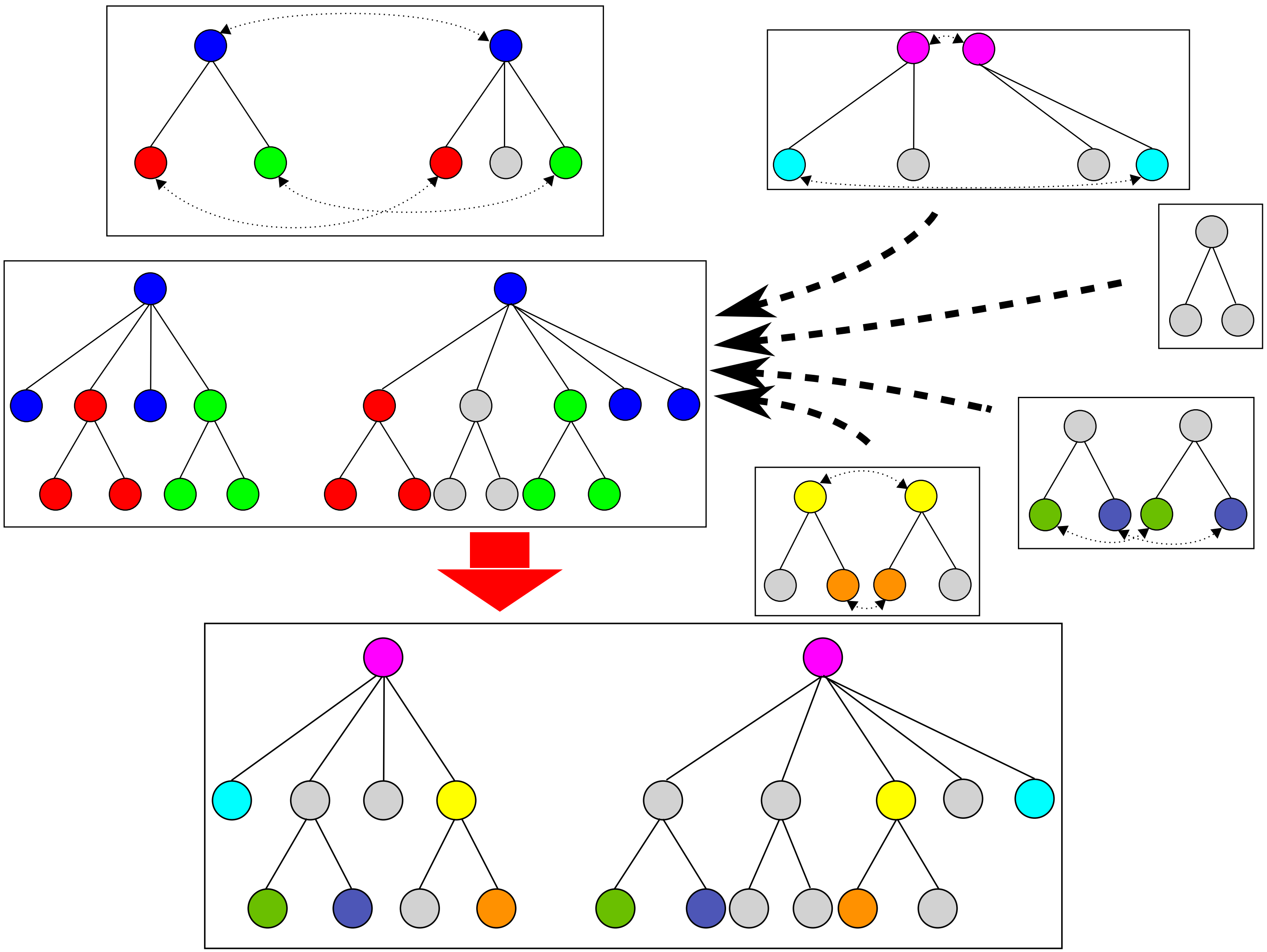
Séparation par couleur : problème



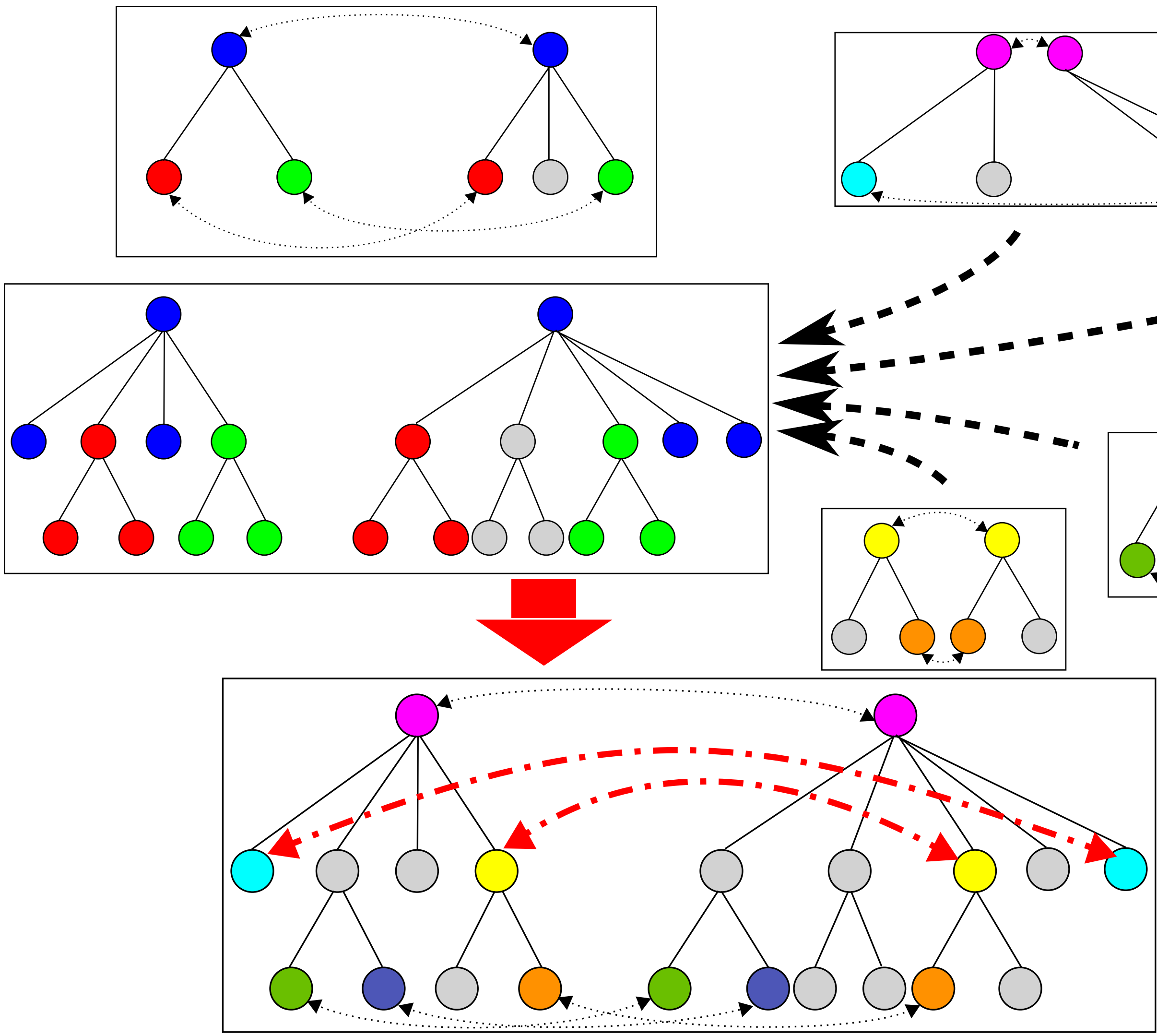
Séparation par couleur : problème



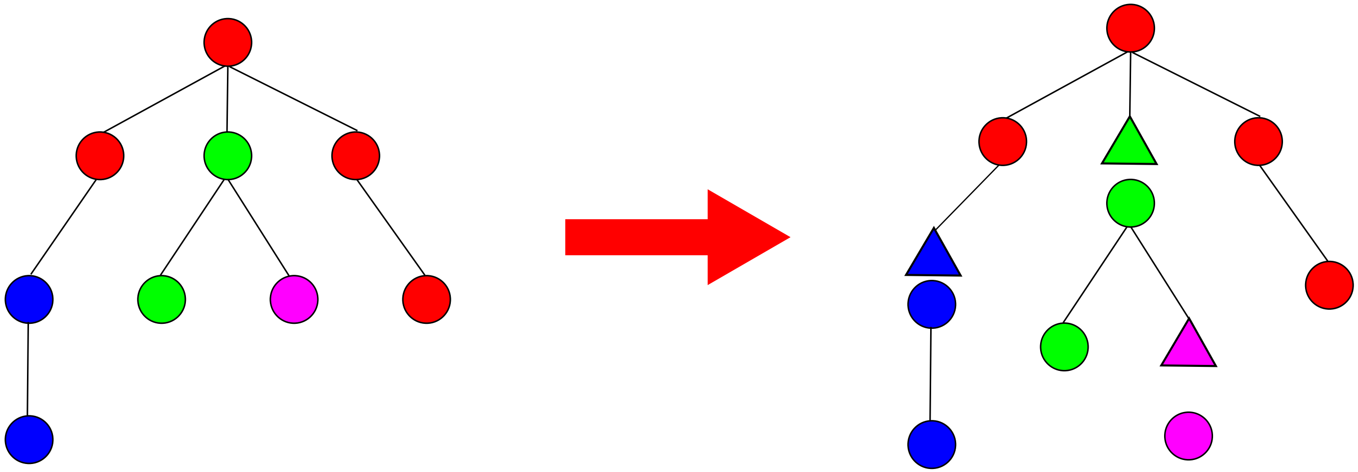
Séparation par couleur : problème



Séparation par couleur : problème

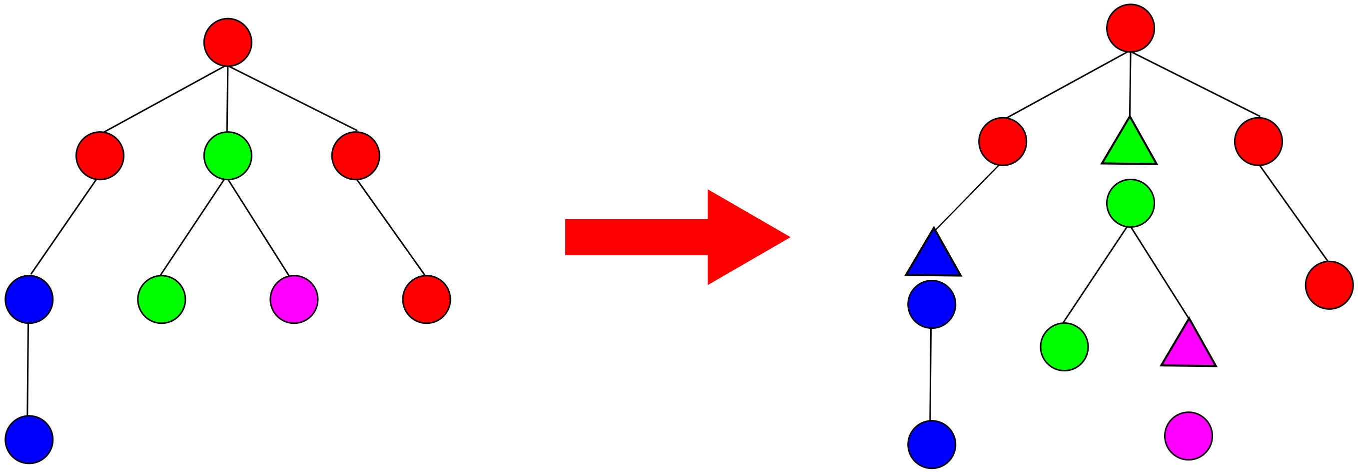


Algorithme : couleurs



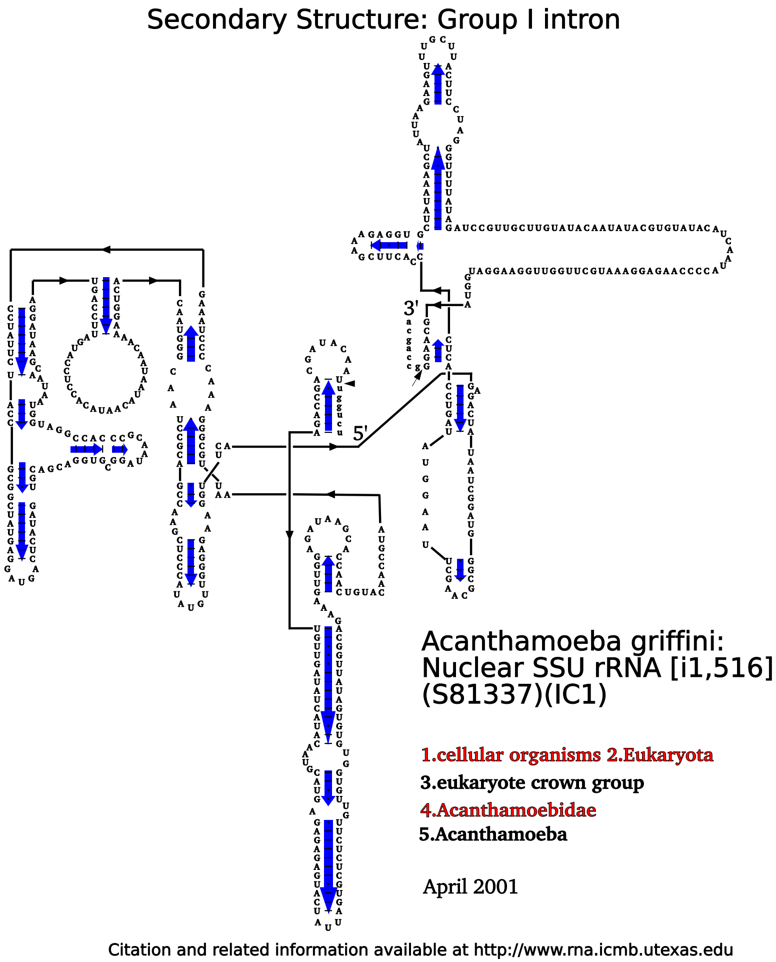
- Séparation de chaque sous-arbre de couleurs
- Ajout d'ancres
- Calcul des distances en fonction de la dépendance des couleurs (de bas en haut)

Algorithme : couleurs

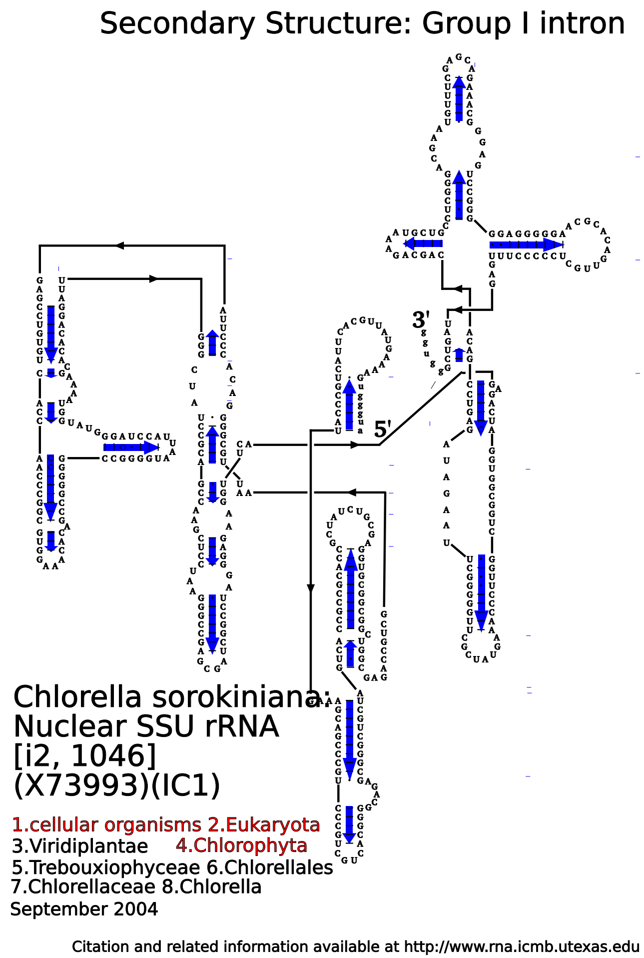


- Séparation de chaque sous-arbre de couleurs
- Ajout d'ancres
- Calcul des distances en fonction de la dépendance des couleurs (de bas en haut)
- $\Rightarrow$ Complexité dépendante de la taille du plus grand sous-arbre

Résultats

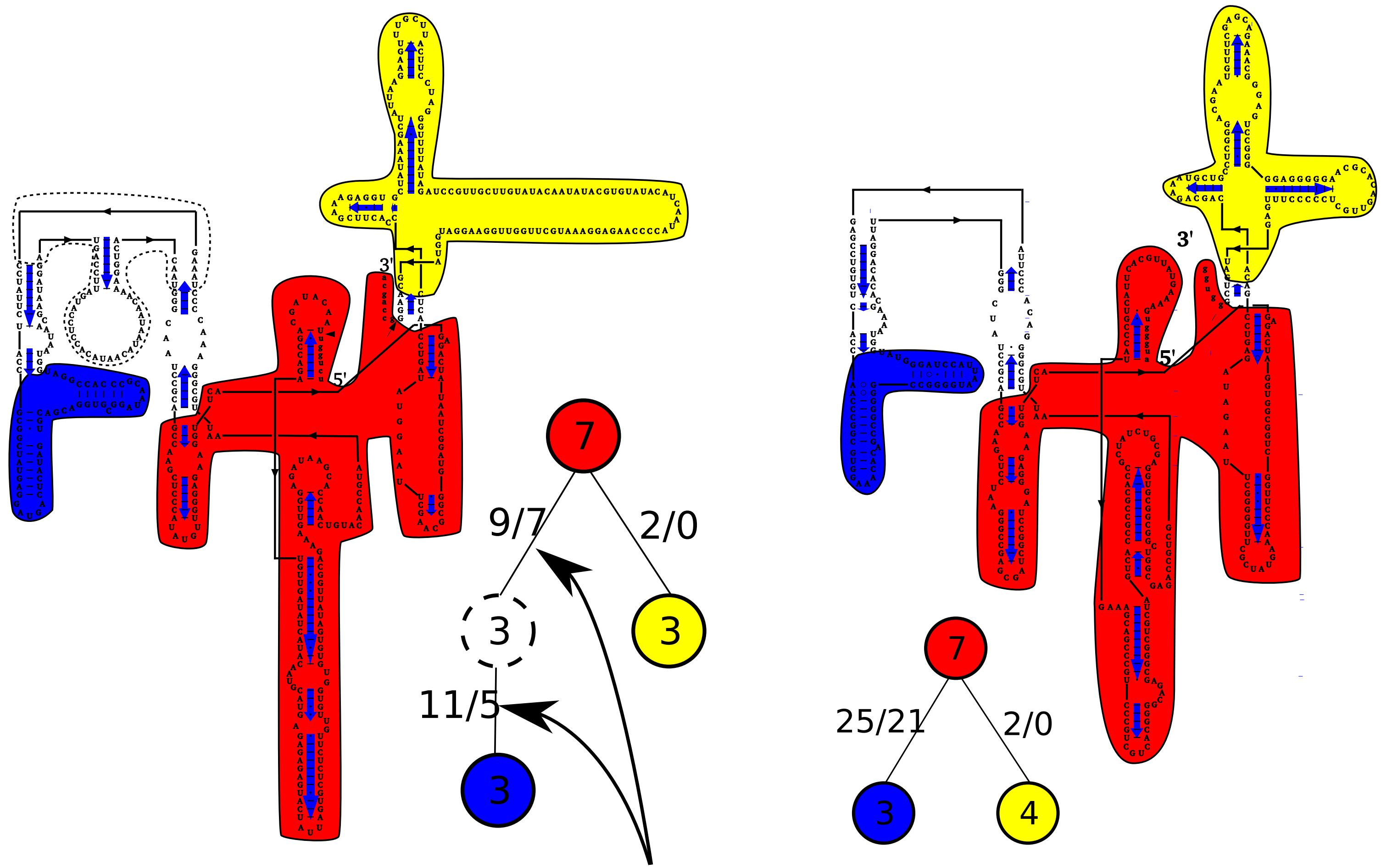


Secondary Structure: Group I intron
3'
5'
Acanthamoeba griffini:
Nuclear SSU rRNA [i1,516]
(S81337)(IC1)
1.cellular organisms 2.Eukaryota
3.eukaryote crown group
4.Acanthamoebidae
5.Acanthamoeba
April 2001
Citation and related information available at http://www.rna.icmb.utexas.edu

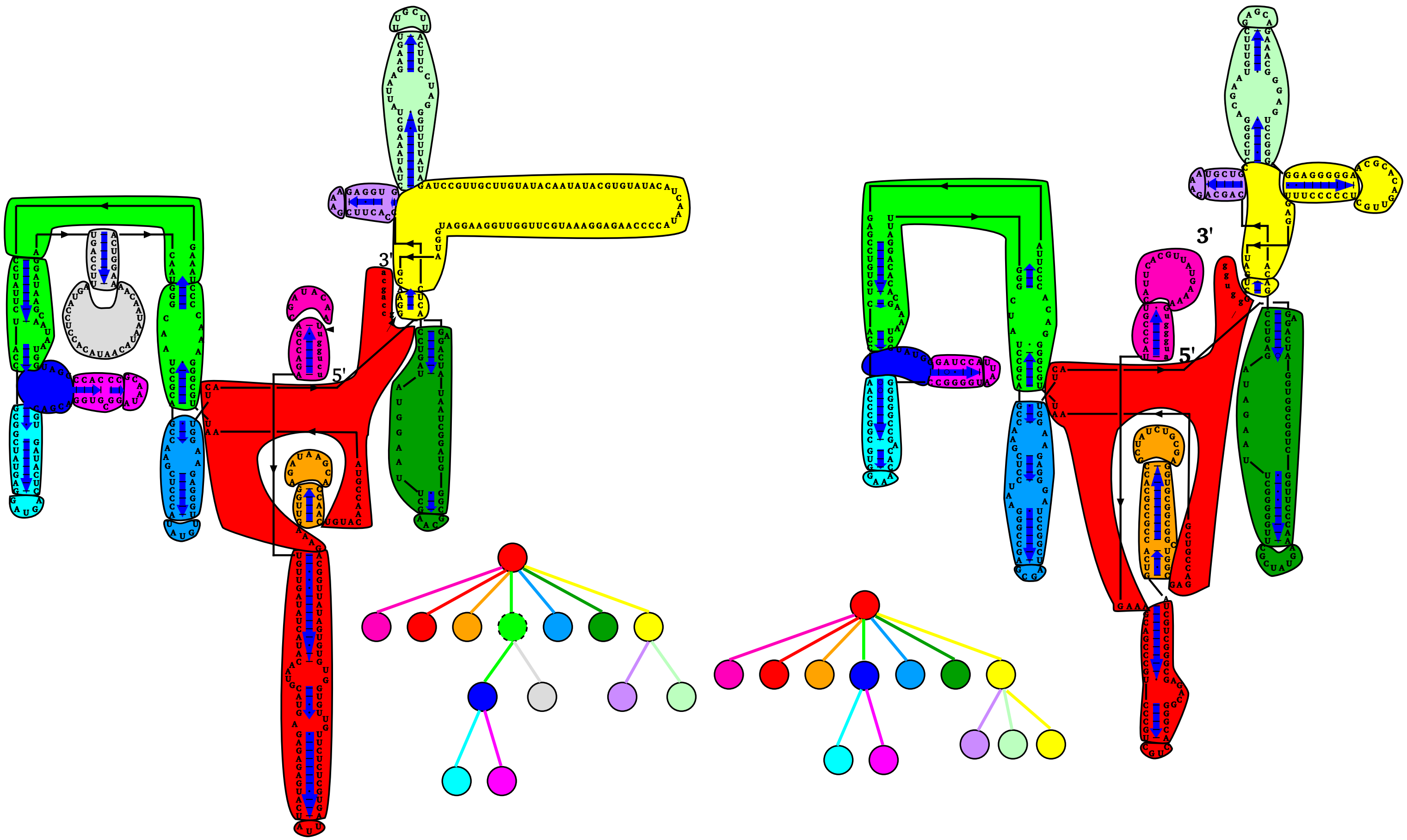


Secondary Structure: Group I intron
3'
5'
Chlorella sorokiniana:
Nuclear SSU rRNA
[i2, 1046]
(X73993)(IC1)
1.cellular organisms 2.Eukaryota
3.Viridiplantae 4.Chlorophyta
5.Trebouxiophyceae 6.Chlorellales
7.Chlorellaceae 8.Chlorella
September 2004
Citation and related information available at http://www.rna.icmb.utexas.edu

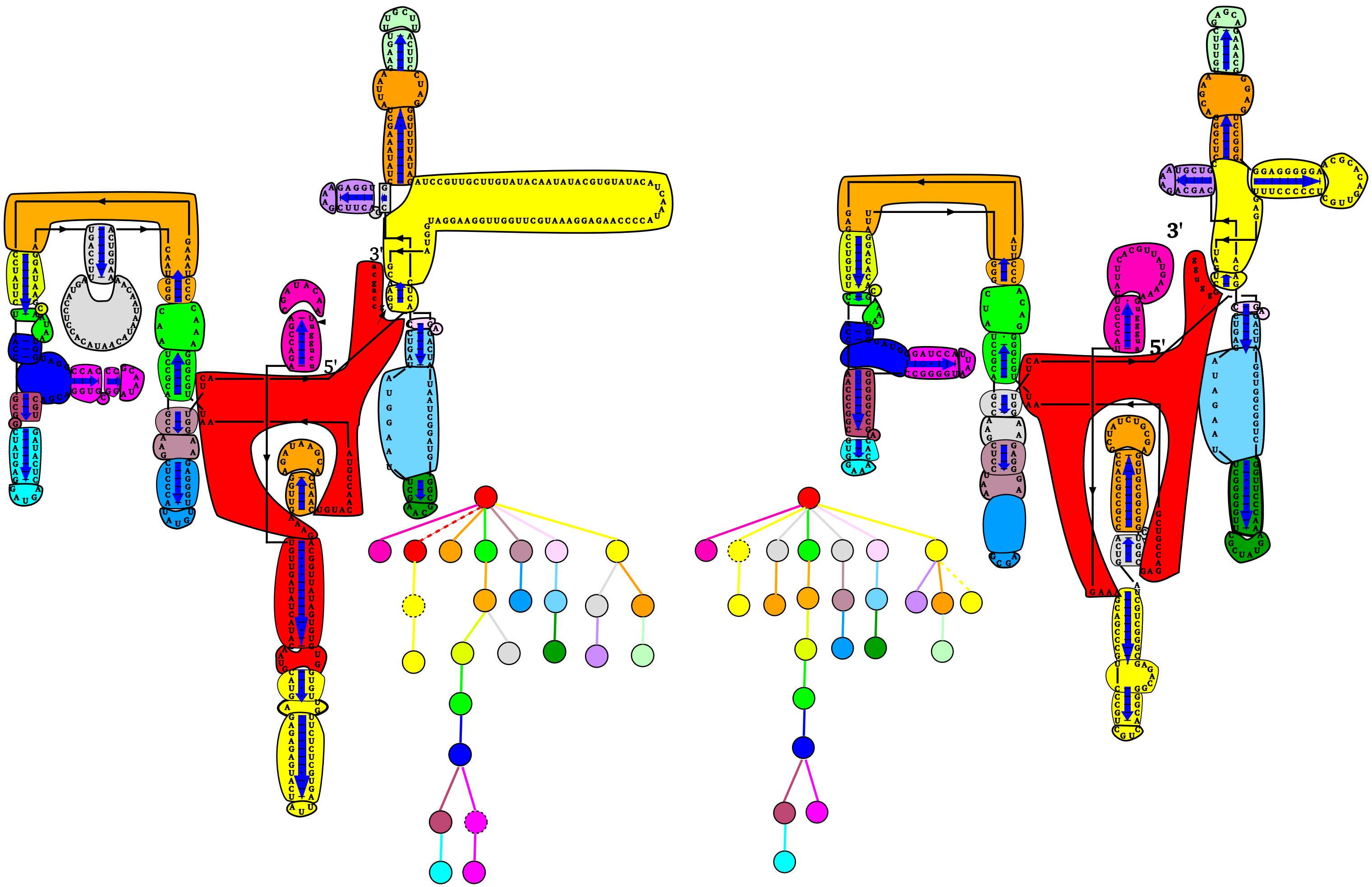
Résultats : Niveau 0



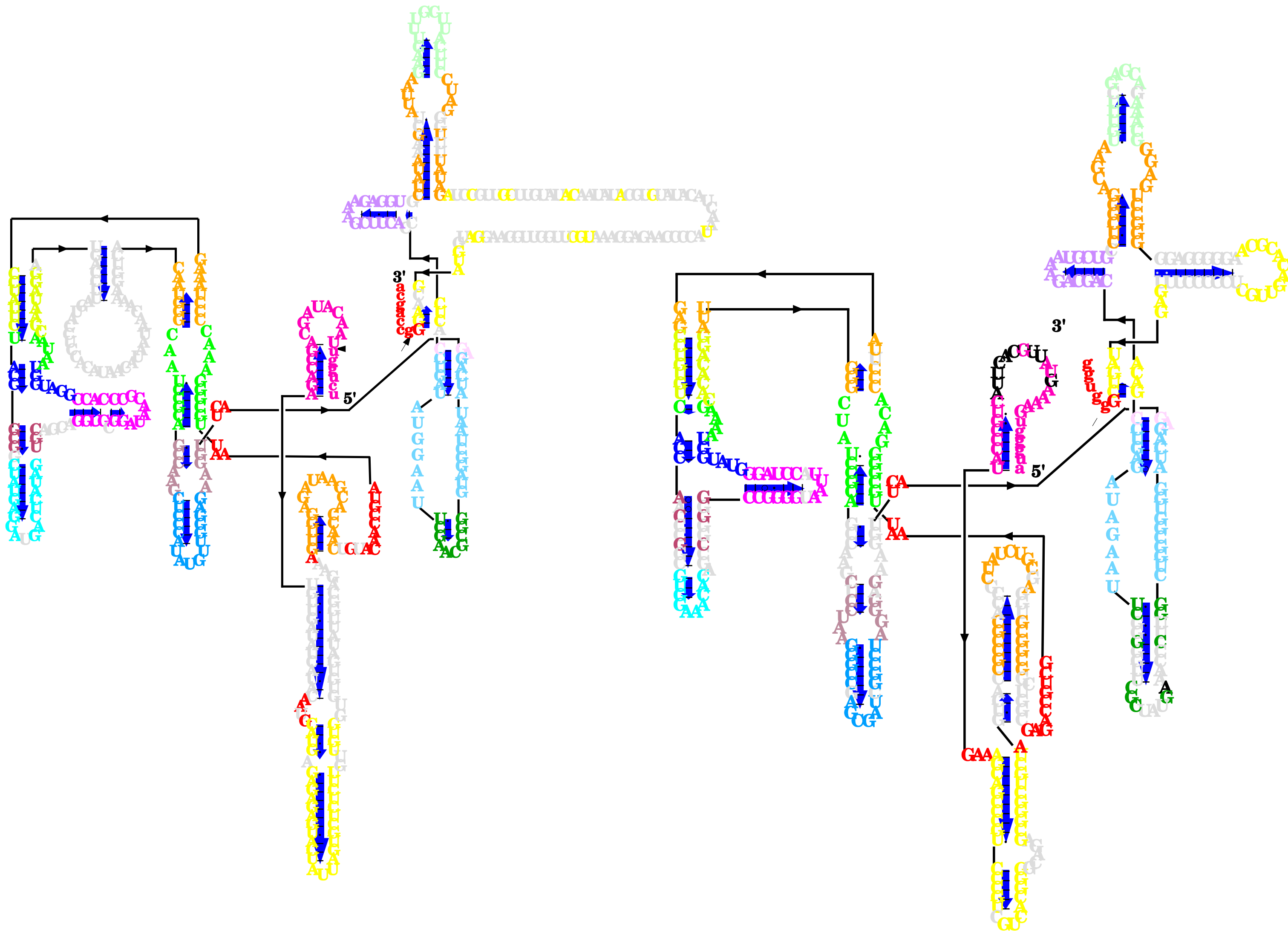
Résultats : Niveau 1



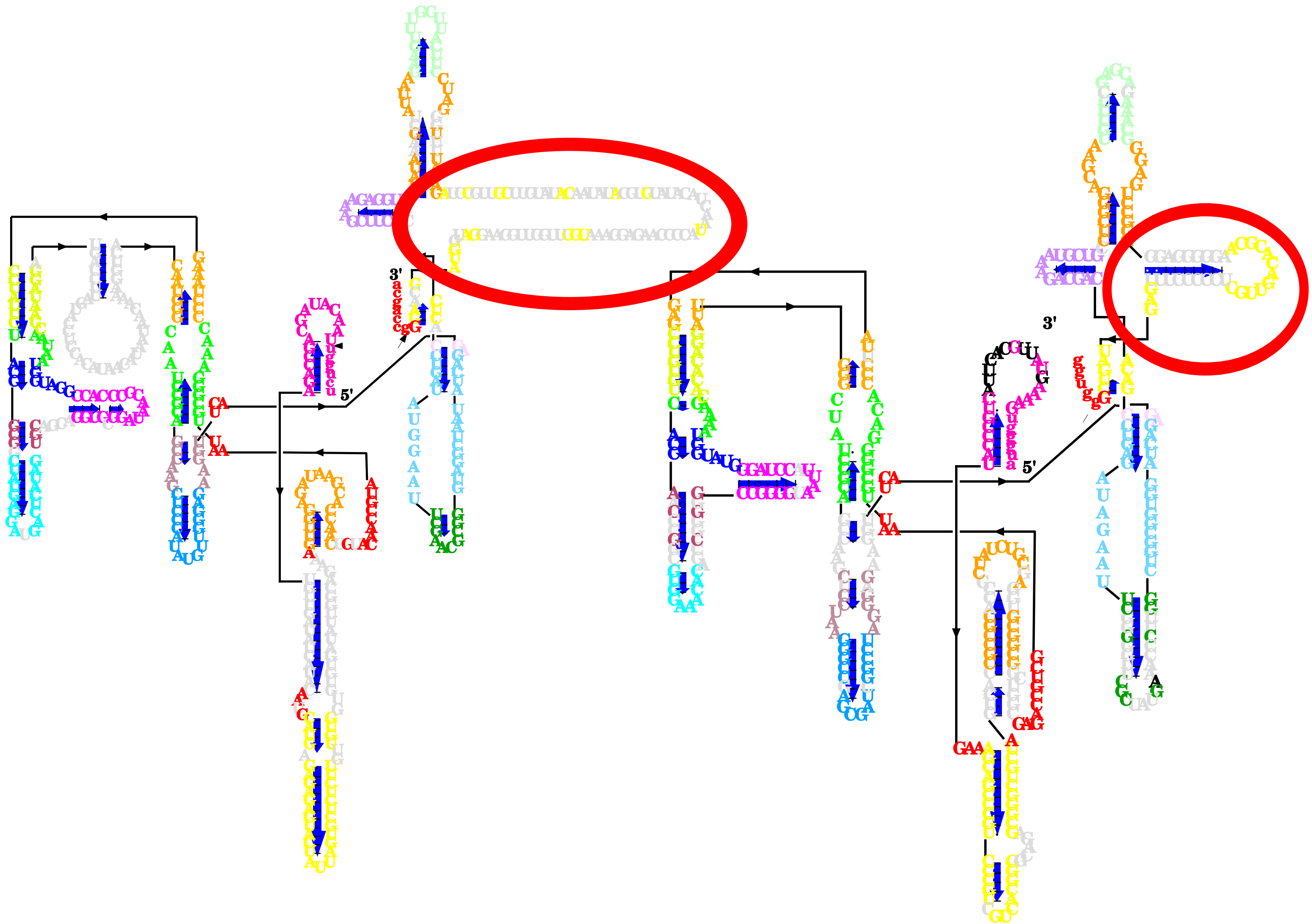
Résultats : Niveau 2



Résultats : Niveau 3



Résultats : Niveau 3



Conclusion

Conclusion(1)

Contribution :

- Nouvelle distance d'édition

Conclusion(1)

Contribution :

- Nouvelle distance d'édition
- MiGaL :

Conclusion(1)

Contribution :

- Nouvelle distance d'édition
- MiGaL :
  - Modélisation originale en adéquation avec la structure secondaire

Conclusion(1)

Contribution :

- Nouvelle distance d'édition
- MiGaL :
  - Modélisation originale en adéquation avec la structure secondaire
  - De nombreuses variantes possibles (étiquettes, nombre de niveau, . . .)

Conclusion(1)

Contribution :

- Nouvelle distance d'édition
- MiGaL :
  - Modélisation originale en adéquation avec la structure secondaire
  - De nombreuses variantes possibles (étiquettes, nombre de niveau, . . .)
- Algorithme de comparaison de deux MiGaLs

Conclusion(1)

Contribution :

- Nouvelle distance d'édition
- MiGaL :
  - Modélisation originale en adéquation avec la structure secondaire
  - De nombreuses variantes possibles (étiquettes, nombre de niveau, . . .)
- Algorithme de comparaison de deux MiGaLs
- Algorithme performant de comparaison de deux RNA-MiGaLs : séparation des couleurs

Conclusion(2)

Travaux à venir :

- Utilisation de RNA-MiGaL :

Conclusion(2)

Travaux à venir :

- Utilisation de RNA-MiGaL :
  - Inférence de modèle

Conclusion(2)

Travaux à venir :

- Utilisation de RNA-MiGaL :
  - Inférence de modèle
  - Repliement

Conclusion(2)

Travaux à venir :

- Utilisation de RNA-MiGaL :
  - Inférence de modèle
  - Repliement
- Distance entre deux RNA-MiGaL

Conclusion(2)

Travaux à venir :

- Utilisation de RNA-MiGaL :
  - Inférence de modèle
  - Repliement
- Distance entre deux RNA-MiGaL
- Extension aux structures tertiaires/protéines

Conclusion(2)

Travaux à venir :

- Utilisation de RNA-MiGaL :
  - Inférence de modèle
  - Repliement
- Distance entre deux RNA-MiGaL
- Extension aux structures tertiaires/protéines
- Comparaison multiple de structures secondaires