Algorithmique pour l'évolution des interactions géniques

DeCo et ARt-DeCo

Sèverine Bérard

ISE-M – Université de Montpellier

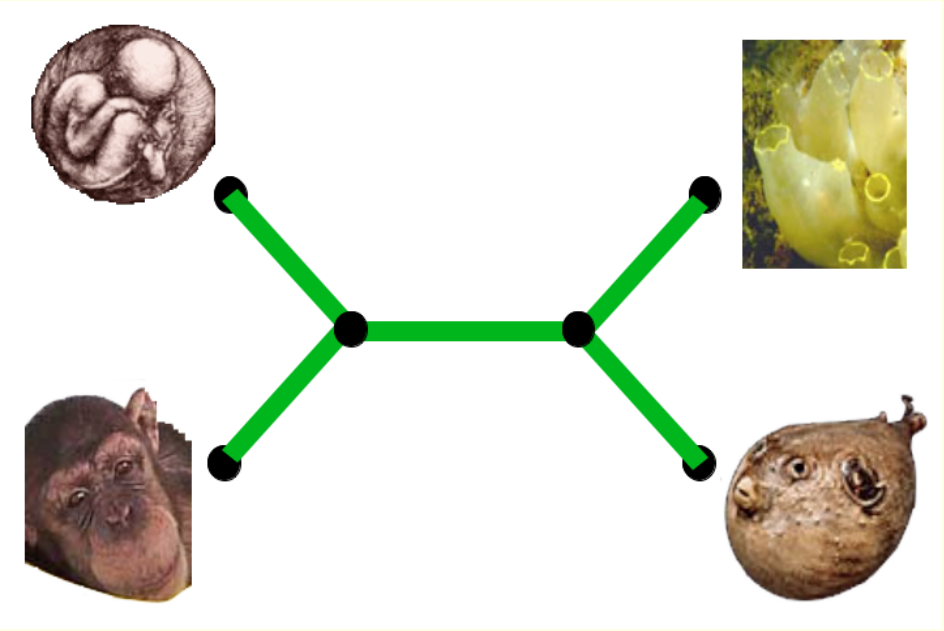
- Introduction
- Modèle
- Démarche
- Validation
- Utilisation de DeCo pour l'assemblage : ARt-DeCo

Inférence phylogénétique (1/3)

Homme	A**C**GTCGGTTGCCGAC
Chimpanzé	A**C**GTCGGTTGCCGAC
Fugu	A**T**GTCGGGTGCCGAC
Cione	A**A**GTCGCGTGCCGAC

$\approx$1980

$\Longrightarrow$

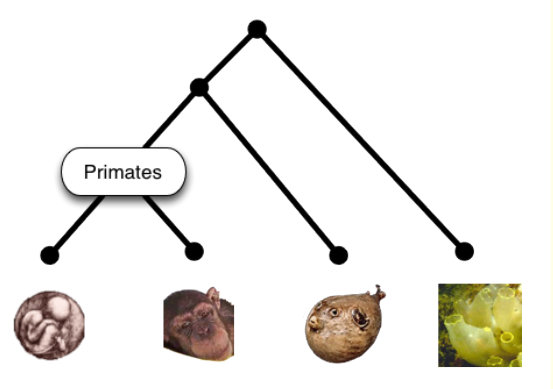


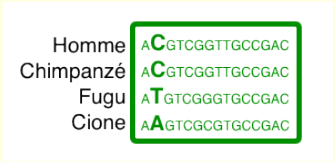
Événement : Substitutions

Inférence phylogénétique (2/3)

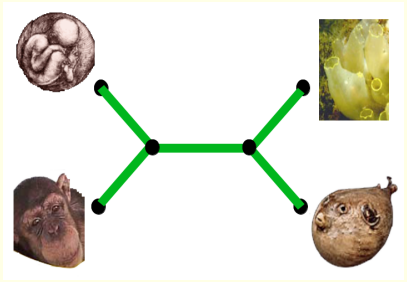
Données :

- Arbre des espèces
- Arbres de gènes racinés

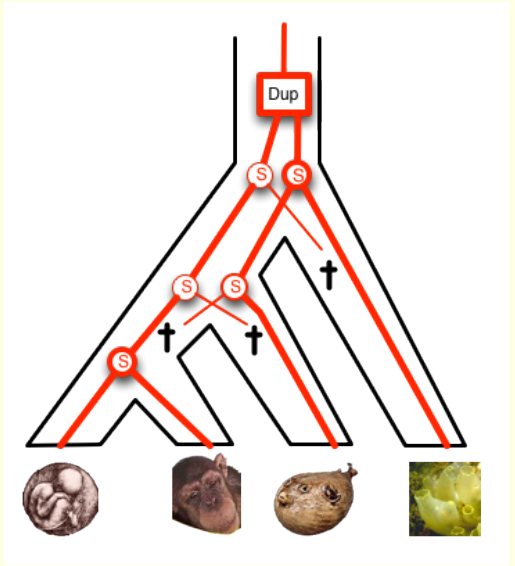




≈1980 ⟹



≈2000 ⟹

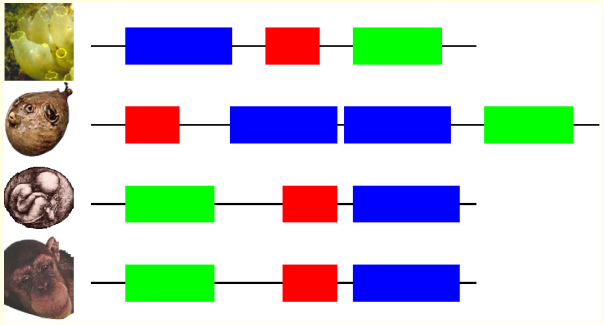


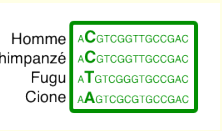
Événements : Substitutions

\+ duplications, pertes

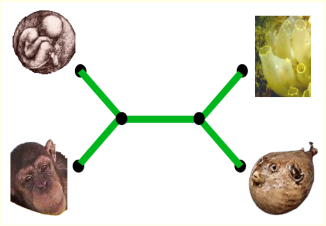
Inférence phylogénétique (3/3)

- Utilisation des adjacences de gènes
- Adjacence de gènes = 2 gènes voisins sur le génome

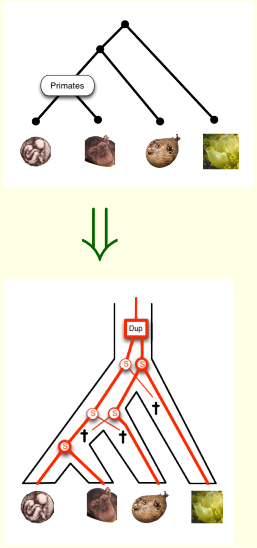




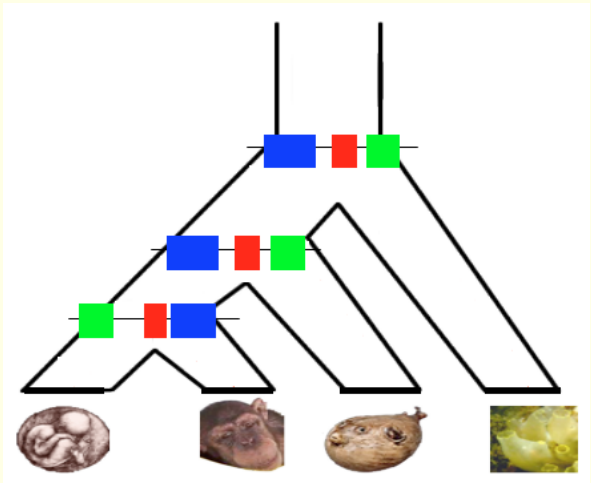
≈1980 ⇒



≈2000 ⇒



2012 ⟹



Événements :
Substitutions

\+ duplications, pertes

\+ réarrangements

Historique (1/2)

- Stage de M2R Informatique Coralie Gallien $jan \rightarrow sept\ 2011$
  donne lieu à `DeCo`
  Encadrement : Sèverine Bérard et Éric Tannier

- Preuve, codage et écriture de l'article $sept\ 2011 \rightarrow juin\ 2012$

- Parution de l'article $sept\ 2012$

BIOINFORMATICS

Vol. 28 ECCB 2012, pages i382–i388
doi:10.1093/bioinformatics/bts374

Evolution of gene neighborhoods within reconciled phylogenies

Sèverine Bérard[1,2,*], Coralie Gallien[1], Bastien Boussau[3,4], Gergely J. Szöllősi[3], Vincent Daubin[3] and Eric Tannier[3,5,*]

- Stage de M2R BCD Pierre-Antoine Jean $mars \rightarrow sept\ 2013$
  Algorithmique pour l'évolution des interactions géniques
  Encadrement : *Vincent Berry* et *Annie Chateau* (*ÉT*, *SB*)

Historique (2/2)

- Stage de M2R BCD Yoann Anselmetti $mars \rightarrow sept\ 2014$
  Reconstruction de l'histoire évolutive des génomes : réconciliation et reconstruction des relations ancestrales
  Encadrement : *SB*, *Vincent Berry*, *Annie Chateau* et *ÉT*

- Parution de l'article $oct\ 2015$


Anselmetti *et al. BMC Genomics* 2015, **16**(Suppl 10):S11
http://www.biomedcentral.com/1471-2164/16/S10/S11

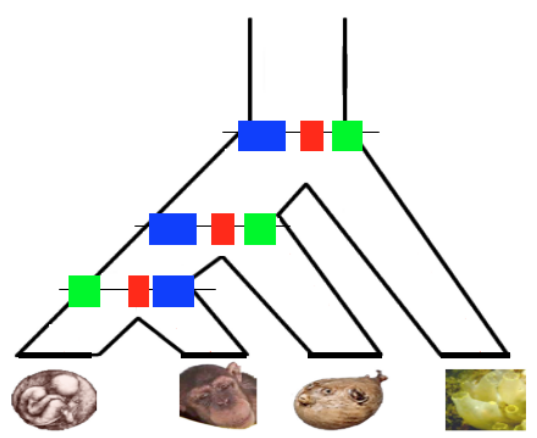
BMC Genomics

RESEARCH Open Access

Ancestral gene synteny reconstruction improves extant species scaffolding

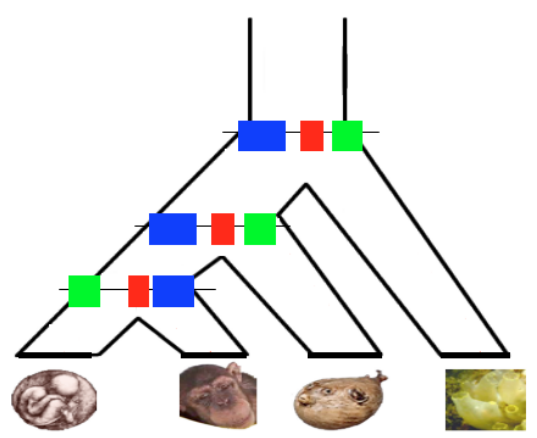
Yoann Anselmetti[1,3], Vincent Berry[2], Cedric Chauve[5], Annie Chateau[2], Eric Tannier[3,4], Sèverine Bérard[1,2*]

- Thèse Yoann Anselmetti (ISE-M - Montpellier) $oct\ 2014 \rightarrow 17$
  Assemblage et évolution de la structure de génomes anciens et actuels Encadrement : *Sèverine Bérard* et *Éric Tannier*



Objectifs

- Retracer l'histoire évolutive de relations entre gènes
- Adjacences : estimer les positions relatives des gènes ancestraux

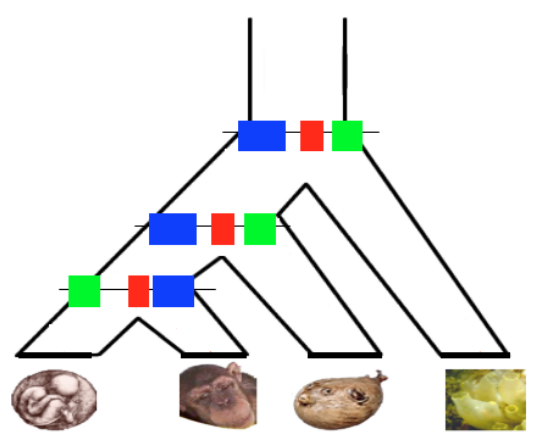


Objectifs

- Retracer l'histoire évolutive de relations entre gènes
- Adjacences : estimer les positions relatives des gènes ancestraux

Principes

- Parcimonie : min(#gains d'adjacence + #cassures d'adjacence)

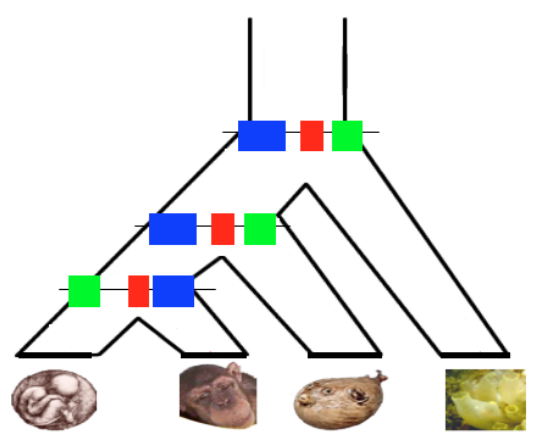


Objectifs

- Retracer l'histoire évolutive de relations entre gènes
- Adjacences : estimer les positions relatives des gènes ancestraux

Principes

- Parcimonie : min(#gains d'adjacence + #cassures d'adjacence)
- Algorithme exact



Objectifs

- Retracer l'histoire évolutive de relations entre gènes
- Adjacences : estimer les positions relatives des gènes ancestraux

Principes

- Parcimonie : min(#gains d'adjacence + #cassures d'adjacence)
- Algorithme exact
- Programmation dynamique, complexité polynomiale

- Introduction
- Modèle
- Démarche
- Validation
- Utilisation de DeCo pour l'assemblage : ARt-DeCo

- Adjacences : Deux gènes A_1 et A_2 sont adjacents s'ils se trouvent sur le même chromosome et qu'il n'y a pas de gène entre eux
  On note A_1A_2 ou A_2A_1 (symétrie) ou encore $A_1 \sim A_2$

- Adjacences : Deux gènes A_1 et A_2 sont adjacents s'ils se trouvent sur le même chromosome et qu'il n'y a pas de gène entre eux
  On note A_1A_2 ou A_2A_1 (symétrie) ou encore $A_1 \sim A_2$

- Arbre phylogénétique : graphe connexe non cyclique, orienté

- Adjacences : Deux gènes A_1 et A_2 sont adjacents s'ils se trouvent sur le même chromosome et qu'il n'y a pas de gène entre eux
  On note A_1A_2 ou A_2A_1 (symétrie) ou encore $A_1 \sim A_2$

- Arbre phylogénétique : graphe connexe non cyclique, orienté

  $\rightarrow$ Arbre d'espèces

  $\rightarrow$ Arbre de gènes

  $\rightarrow$ Arbre d'adjacences

- Adjacences : Deux gènes A_1 et A_2 sont adjacents s'ils se trouvent sur le même chromosome et qu'il n'y a pas de gène entre eux
  On note A_1A_2 ou A_2A_1 (symétrie) ou encore $A_1 \sim A_2$

- Arbre phylogénétique : graphe connexe non cyclique, orienté

  $\rightarrow$ Arbre d'espèces

  $\rightarrow$ Arbre de gènes

  $\rightarrow$ Arbre d'adjacences

- Forêt : ensemble d'arbres

- Adjacences : Deux gènes A_1 et A_2 sont adjacents s'ils se trouvent sur le même chromosome et qu'il n'y a pas de gène entre eux
  On note A_1A_2 ou A_2A_1 (symétrie) ou encore $A_1 \sim A_2$

- Arbre phylogénétique : graphe connexe non cyclique, orienté

  $\rightarrow$ Arbre d'espèces

  $\rightarrow$ Arbre de gènes

  $\rightarrow$ Arbre d'adjacences

- Forêt : ensemble d'arbres

- Pour un nœud n, $S(n)$ est son espèce, $E(n)$ l'événement associé

- Sur les espèces :

  $\rightarrow$ Spéciation ($Spec$) ●

Événements évolutifs

- Sur les espèces :

  → Spéciation ($Spec$) ●

- Sur les gènes :

  → Duplication de gène ($GDup$) ■

  → Perte de gène ($GLos$) ✖

Événements évolutifs

- Sur les espèces :
  - → Spéciation ($Spec$) ●
- Sur les gènes :
  - → Duplication de gène ($GDup$) ■
  - → Perte de gène ($GLos$) ✖
- Sur les adjacences
  - → Duplication d'adjacence ($ADup$) □
  - → Perte d'adjacence ($ALos$) ⛒
  - → Création d'adjacence ($Gain$) ▲ *(toute racine d'un arbre d'adjacence)*
  - → Cassure d'adjacence ($Break$) ⇄

Un arbre d'adjacences $\mathcal{A}$ est un arbre phylogénétique retraçant l'histoire évolutive d'adjacences

Arbre d'adjacences

Un arbre d'adjacences $\mathcal{A}$ est un arbre phylogénétique retraçant l'histoire évolutive d'adjacences

- Il est associé à 2 arbres de gènes, $\mathcal{G}_1$ et $\mathcal{G}_2$, et à une liste d'adjacences actuelles notée $\mathcal{L}$

Un arbre d'adjacences $\mathcal{A}$ est un arbre phylogénétique retraçant l'histoire évolutive d'adjacences

- Il est associé à 2 arbres de gènes, $\mathcal{G}_1$ et $\mathcal{G}_2$, et à une liste d'adjacences actuelles notée $\mathcal{L}$

- Tous ses nœuds sont étiquetés par des adjacences entre nœuds des arbres de gènes associés et/ou un événement évolutif
  $\rightarrow$ Évt feuilles : $\{GLos,\ ALos,\ Break,\ Extant\}$
  $\rightarrow$ Évt nœuds internes : $\{Spec,\ GDup,\ ADup\}$

Un arbre d'adjacences $\mathcal{A}$ est un arbre phylogénétique retraçant l'histoire évolutive d'adjacences

- Il est associé à 2 arbres de gènes, $\mathcal{G}_1$ et $\mathcal{G}_2$, et à une liste d'adjacences actuelles notée $\mathcal{L}$

- Tous ses nœuds sont étiquetés par des adjacences entre nœuds des arbres de gènes associés et/ou un événement évolutif
  $\rightarrow$ Évt feuilles : $\{GLos,\ ALos,\ Break,\ Extant\}$
  $\rightarrow$ Évt nœuds internes : $\{Spec,\ GDup,\ ADup\}$

- L'arbre respecte les règles de transmissions d'adjacences, et ses feuilles de type $Extant$ sont exactement les adjacences de $\mathcal{L}$

Un arbre d'adjacences $\mathcal{A}$ est un arbre phylogénétique retraçant l'histoire évolutive d'adjacences

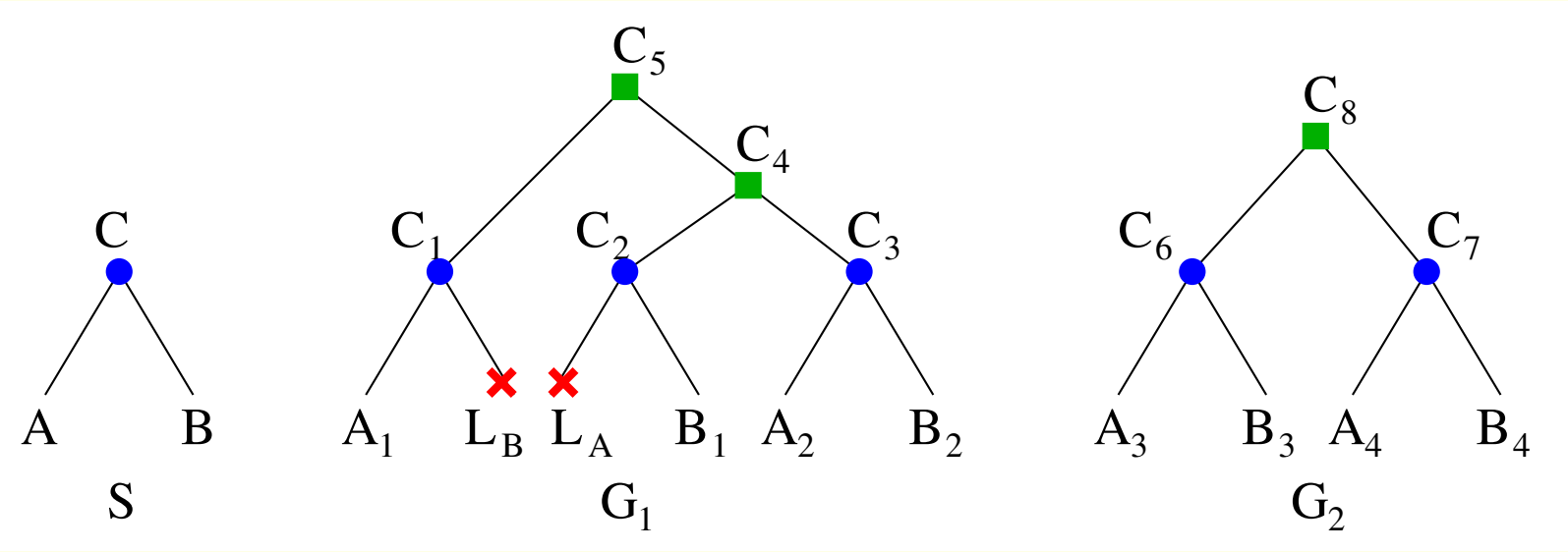
- Il est associé à 2 arbres de gènes, $\mathcal{G}_1$ et $\mathcal{G}_2$, et à une liste d'adjacences actuelles notée $\mathcal{L}$

- Tous ses nœuds sont étiquetés par des adjacences entre nœuds des arbres de gènes associés et/ou un événement évolutif
  $\rightarrow$ Évt feuilles : $\{GLos,\ ALos,\ Break,\ Extant\}$
  $\rightarrow$ Évt nœuds internes : $\{Spec,\ GDup,\ ADup\}$

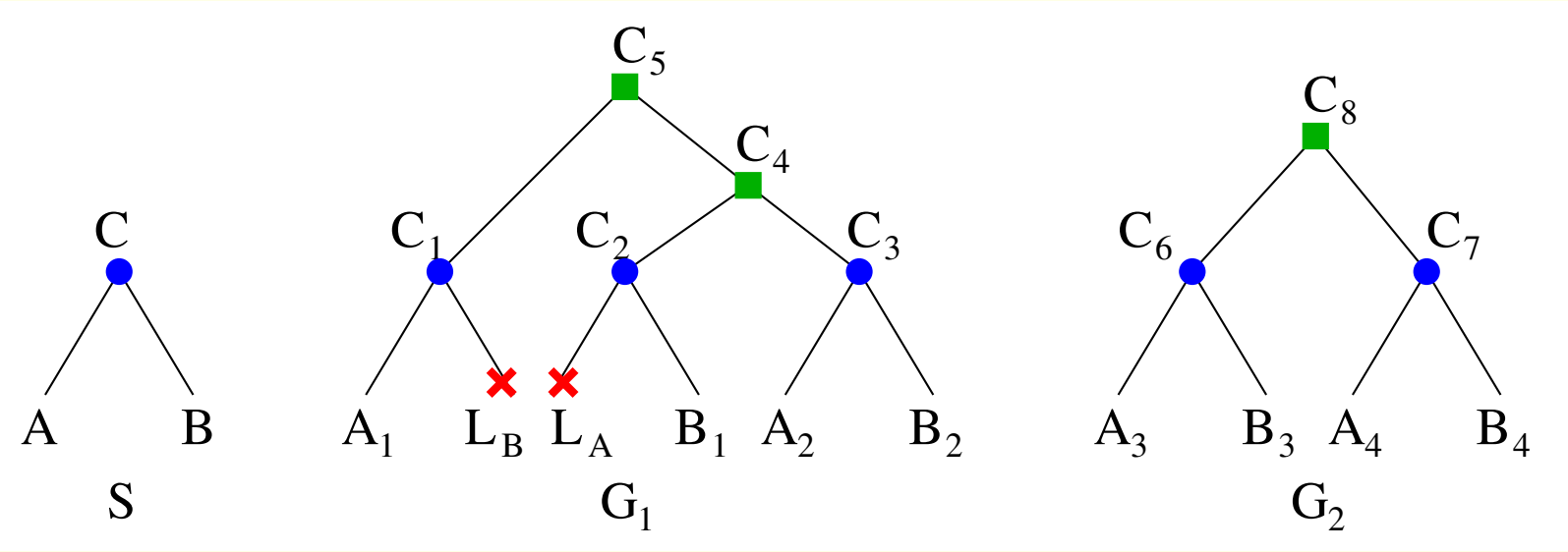
- L'arbre respecte les règles de transmissions d'adjacences, et ses feuilles de type $Extant$ sont exactement les adjacences de $\mathcal{L}$

- La racine correspond à la création d'une adjacence, ce nœud racine porte aussi une étiquette événement

Exemple d'arbres

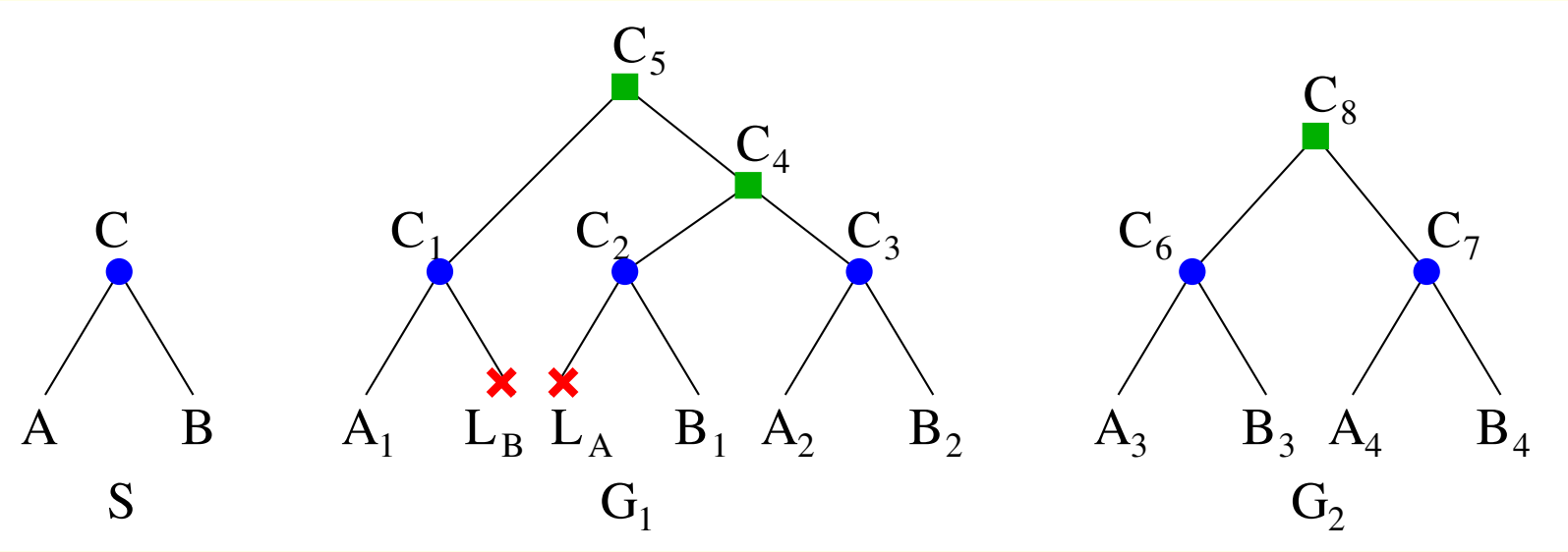


Exemple d'arbres

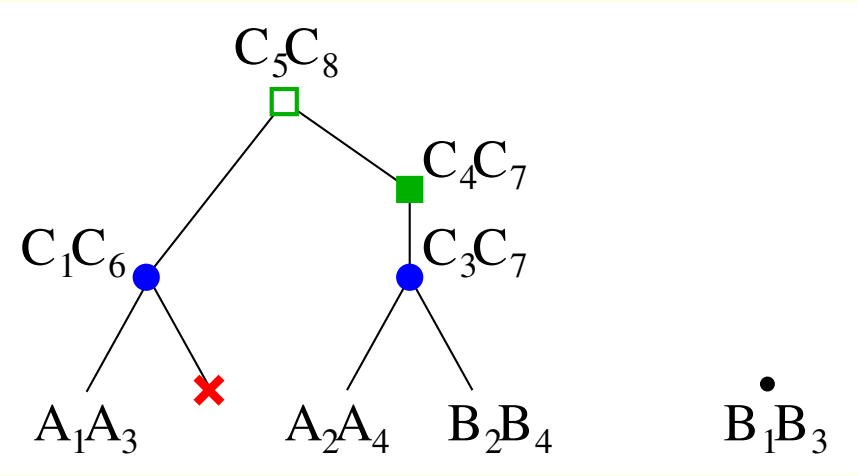


Liste d'adjacences associées $\mathcal{L} = \{A_1A_3, B_1B_3, A_2A_4, B_2B_4\}$

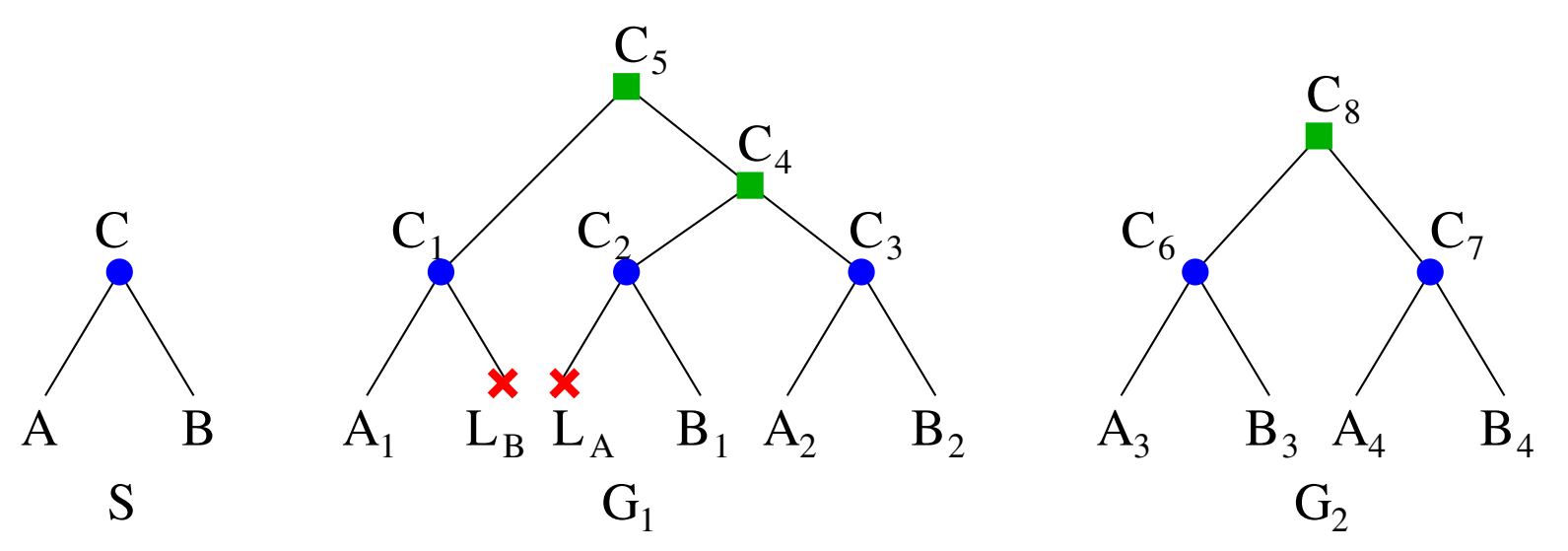
Exemple d'arbres



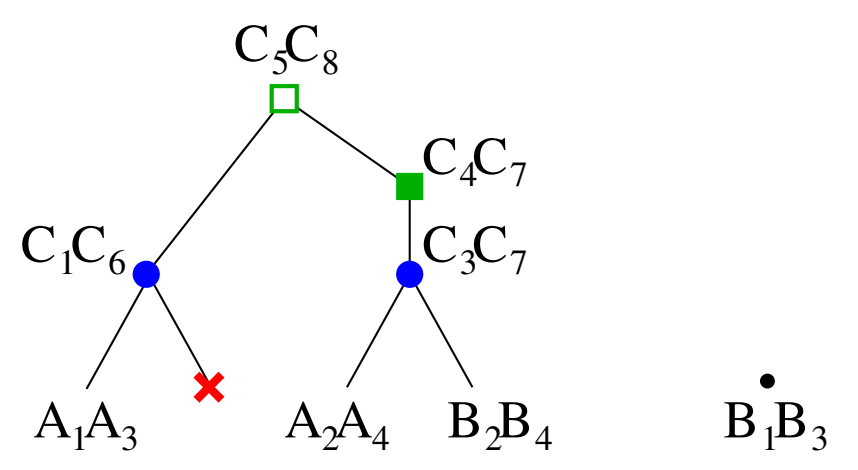
Liste d'adjacences associées $\mathcal{L} = \{A_1A_3, B_1B_3, A_2A_4, B_2B_4\}$



Exemple d'arbres



Liste d'adjacences associées $\mathcal{L} = \{A_1A_3, B_1B_3, A_2A_4, B_2B_4\}$



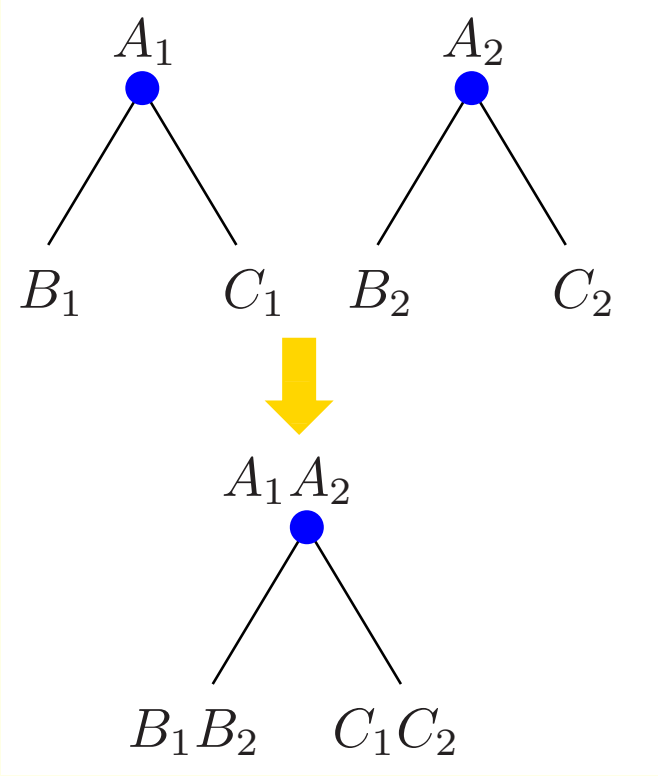
Remarque : Les arbres d'adjacences ne sont pas forcément binaires

Règles de transmissions d'adjacences (1/4)

Les nœuds d'un arbre d'adjacences $\mathcal{A}$ associé à 2 arbres de gènes $\mathcal{G}_1$ et $\mathcal{G}_2$ et à une liste d'adjacences $\mathcal{L}$ respectent les propriétés suivantes :

- A_1A_2 **Nœud de spéciation**
  Alors A_1 et A_2 doivent être 2 nœuds de spéciation dans leur arbre de gènes respectifs

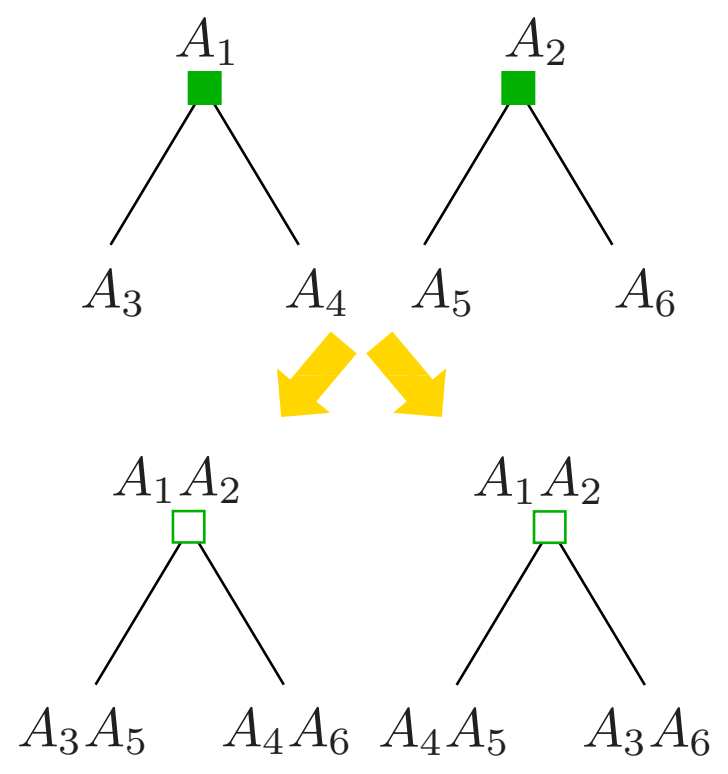
  Supposons que A_1 et A_2 aient 2 fils, respectivement B_1 et C_1 et B_2 et C_2 . Alors les fils de A_1A_2 dans $\mathcal{A}$ sont exactement B_1B_2 et C_1C_2



- A_1A_2 **Nœud de duplication d'adjacence**

Alors A_1 et A_2 doivent être 2 nœuds de duplication de gène

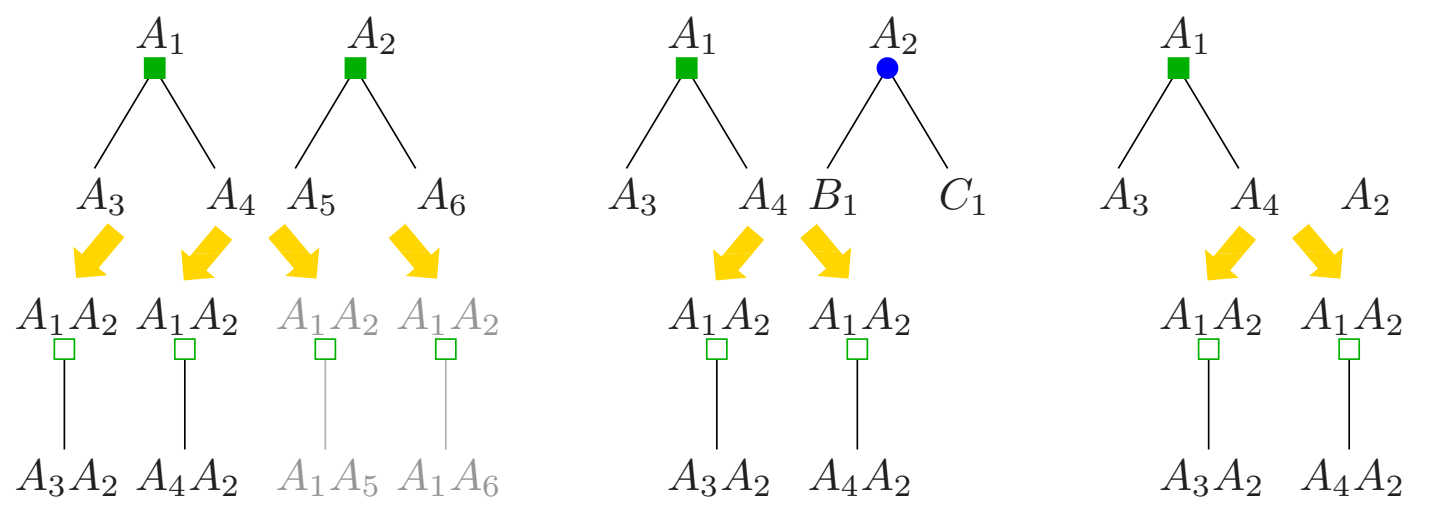
Si A_1 et A_2 ont 2 fils, respectivement A_3 et A_4 et A_5 et A_6 , alors A_1A_2 a exactement 2 fils dans $\mathcal{A}$ qui sont soit A_3A_5 et A_4A_6 soit A_4A_5 et A_3A_6



Règles de transmissions d'adjacences (3/4)

- A_1A_2 **Nœud de duplication de gène**
  Alors A_1 ou A_2 (ou les deux) doit être un nœud de duplication de gène dans son arbre de gènes

  Supposons que A_1 soit le nœud de duplication de gène avec 2 fils A_3 et A_4. Alors A_1A_2 a exactement 1 fils dans $\mathcal{A}$ qui est soit A_3A_2 soit A_4A_2



Règles de transmissions d'adjacences (4/4)

- A_1A_2 **Nœud de création d'adjacence**
  Alors A_1A_2 est la racine de $\mathcal{A}$ et A_1 et A_2 peuvent être de n'importe quel type *(sauf le couple $Spec/Extant$ qui n'existe pas)*

Règles de transmissions d'adjacences (4/4)

- $A_1 A_2$ **Nœud de création d'adjacence**
  Alors $A_1 A_2$ est la racine de $\mathcal{A}$ et A_1 et A_2 peuvent être de n'importe quel type *(sauf le couple $Spec/Extant$ qui n'existe pas)*

- $A_1 A_2$ **Feuille cassure d'adjacence**
  Alors A_1 et A_2 peuvent être de n'importe quel type

Règles de transmissions d'adjacences (4/4)

- A_1A_2 **Nœud de création d'adjacence**
  Alors A_1A_2 est la racine de $\mathcal{A}$ et A_1 et A_2 peuvent être de n'importe quel type *(sauf le couple $Spec/Extant$ qui n'existe pas)*

- A_1A_2 **Feuille cassure d'adjacence**
  Alors A_1 et A_2 peuvent être de n'importe quel type

- A_1A_2 **Feuille adjacence actuelle**
  Alors A_1 et A_2 doivent être 2 gènes actuels

Règles de transmissions d'adjacences (4/4)

- A_1A_2 **Nœud de création d'adjacence**
  Alors A_1A_2 est la racine de $\mathcal{A}$ et A_1 et A_2 peuvent être de n'importe quel type *(sauf le couple $Spec/Extant$ qui n'existe pas)*

- A_1A_2 **Feuille cassure d'adjacence**
  Alors A_1 et A_2 peuvent être de n'importe quel type

- A_1A_2 **Feuille adjacence actuelle**
  Alors A_1 et A_2 doivent être 2 gènes actuels

- A_1A_2 **Feuille perte d'adjacence**
  Alors A_1 et A_2 doivent être 2 perte de gènes

Règles de transmissions d'adjacences (4/4)

- A_1A_2 **Nœud de création d'adjacence**
  Alors A_1A_2 est la racine de $\mathcal{A}$ et A_1 et A_2 peuvent être de n'importe quel type *(sauf le couple $Spec/Extant$ qui n'existe pas)*

- A_1A_2 **Feuille cassure d'adjacence**
  Alors A_1 et A_2 peuvent être de n'importe quel type

- A_1A_2 **Feuille adjacence actuelle**
  Alors A_1 et A_2 doivent être 2 gènes actuels

- A_1A_2 **Feuille perte d'adjacence**
  Alors A_1 et A_2 doivent être 2 perte de gènes

- A_1A_2 **Feuille perte de gène**
  Alors A_1 ou bien A_2 doivent être une perte de gènes

Général Étant donné

- un arbre d'espèce,
- un ensemble d'arbres de gènes,
- une liste d'adjacences actuelles,

trouver l'histoire évolutive des adjacences minimisant le nombre de création et de cassure d'adjacences

Général Étant donné

- un arbre d'espèce,
- un ensemble d'arbres de gènes,
- une liste d'adjacences actuelles,

trouver l'histoire évolutive des adjacences minimisant le nombre de création et de cassure d'adjacences

Sous-problème Étant donné

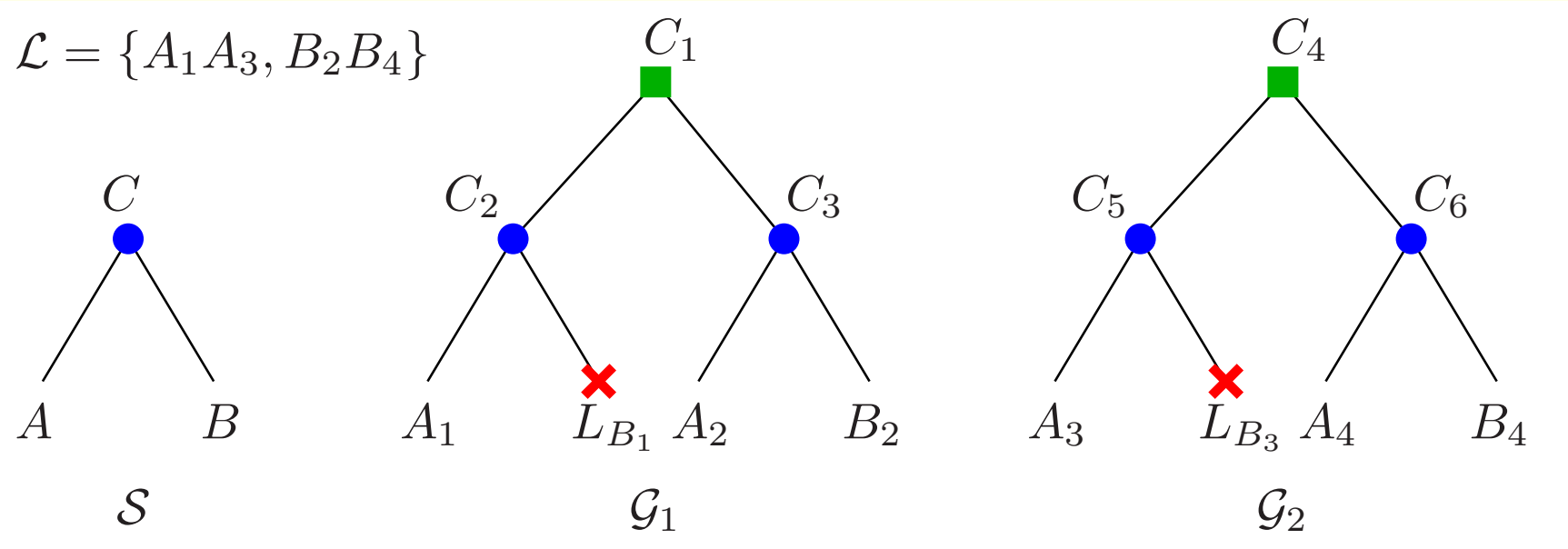
- un arbre d'espèce,
- 2 arbres de gènes réconciliés,
- une liste $\mathcal{L}$ d'adj. actuelles entre feuilles de ces arbres,

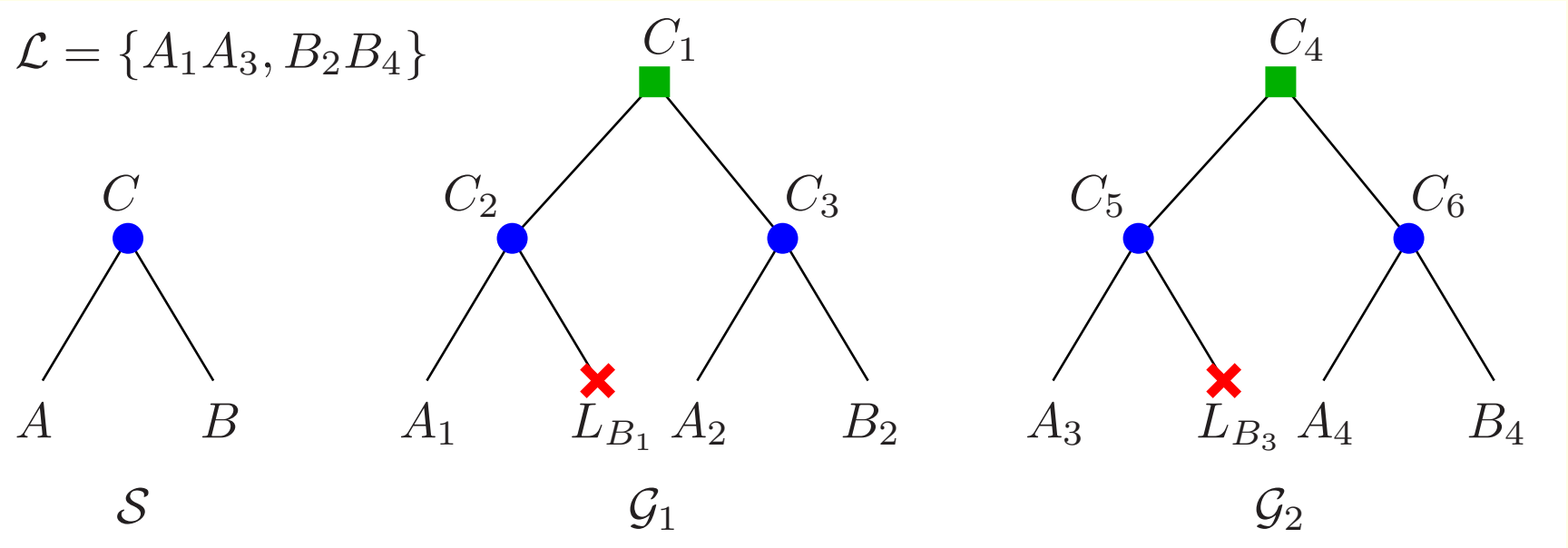
trouver l'histoire évolutive des adjacences de $\mathcal{L}$ minimisant le nombre de création et de cassure d'adjacences

$\Rightarrow$ Résolu par l'algorithme DeCo

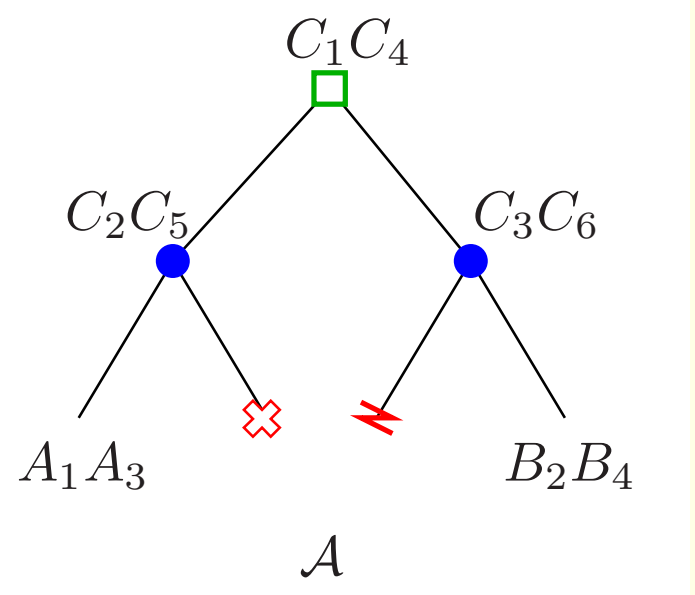
- Introduction
- Modèle
- Démarche
  - → Étape 1 : Réconciliation
  - → Étape 2 : Création des classes d'adjacences
  - → Étape 3 : Algorithme DeCo
  - → Étape 4 : Synthèse
- Validation
- Utilisation de DeCo pour l'assemblage : ARt-DeCo

- Introduction
- Modèle
- Démarche
  - → Étape 1 : Réconciliation
  - → Étape 2 : Création des classes d'adjacences
  - → Étape 3 : Algorithme DeCo
  - → Étape 4 : Synthèse
- Validation
- Utilisation de DeCo pour l'assemblage : ARt-DeCo

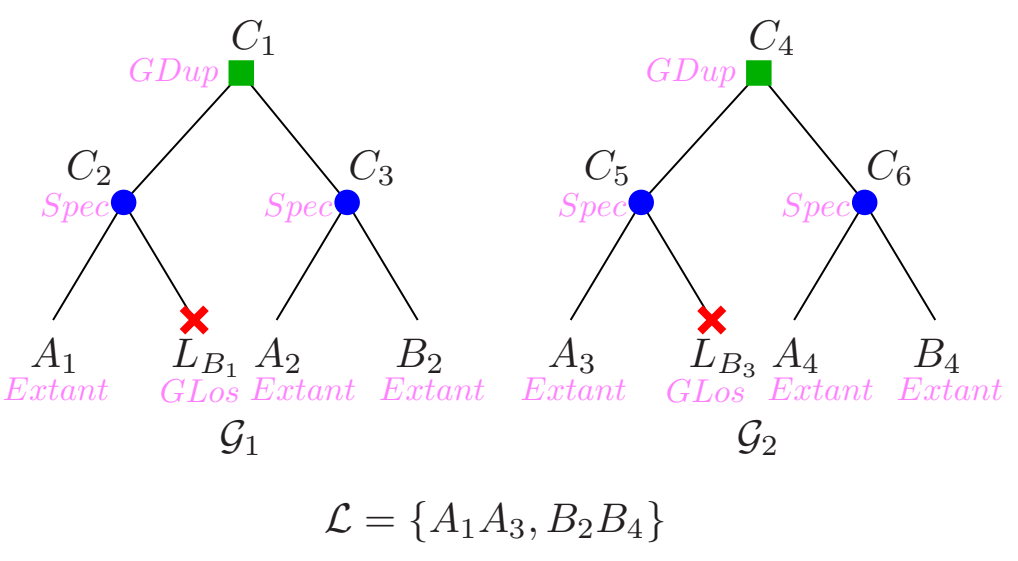

$\mathcal{L} = \{A_1 A_3, B_2 B_4\}$
C
A
B
$\mathcal{S}$
C_1
C_2
C_3
A_1
L_{B_1}
A_2
B_2
$\mathcal{G}_1$
C_4
C_5
C_6
A_3
L_{B_3}
A_4
B_4
$\mathcal{G}_2$



1. Calcul d'une matrice de coût par programmation dynamique
2. Phase de *backtracking*



Calcul des coûts



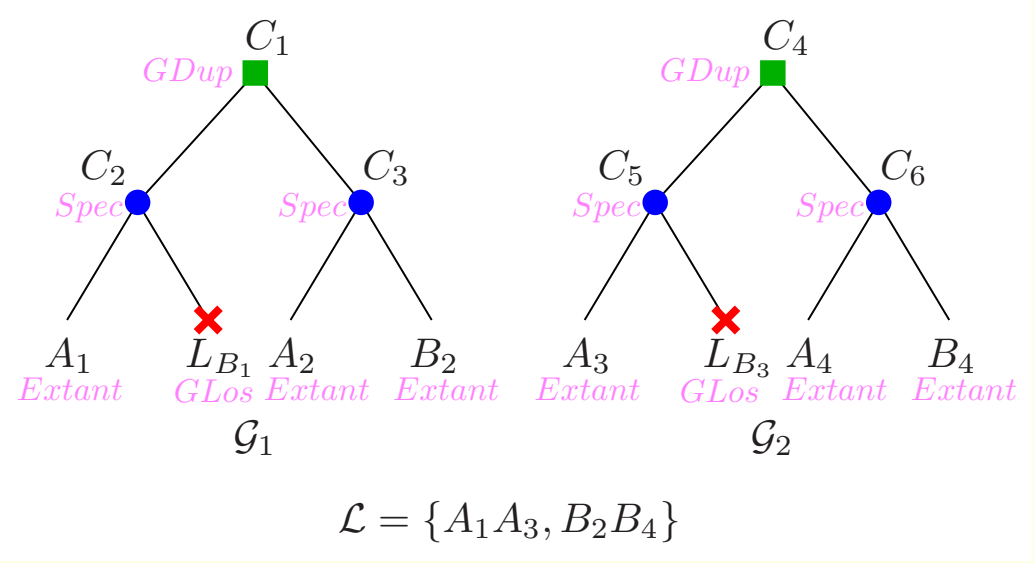
c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1							
L_{B_1}							
A_2							
B_2							
C_2							
C_3							
C_1							

- Matrice de programmation dynamique, coûts c_0 et c_1 :

$\rightarrow$ $c_0(n_1, n_2)$ coût minimum d'une histoire évolutive où n_1 et n_2 ne sont pas adjacents

$\rightarrow$ $c_1(n_1, n_2)$ coût minimum d'une histoire évolutive où n_1 et n_2 sont adjacents

Calcul des coûts



c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1		X		X	X	X	X
L_{B_1}	X		X		X	X	X
A_2		X		X	X	X	X
B_2	X		X		X	X	X
C_2	X	X	X	X			
C_3	X	X	X	X			
C_1	X	X	X	X			

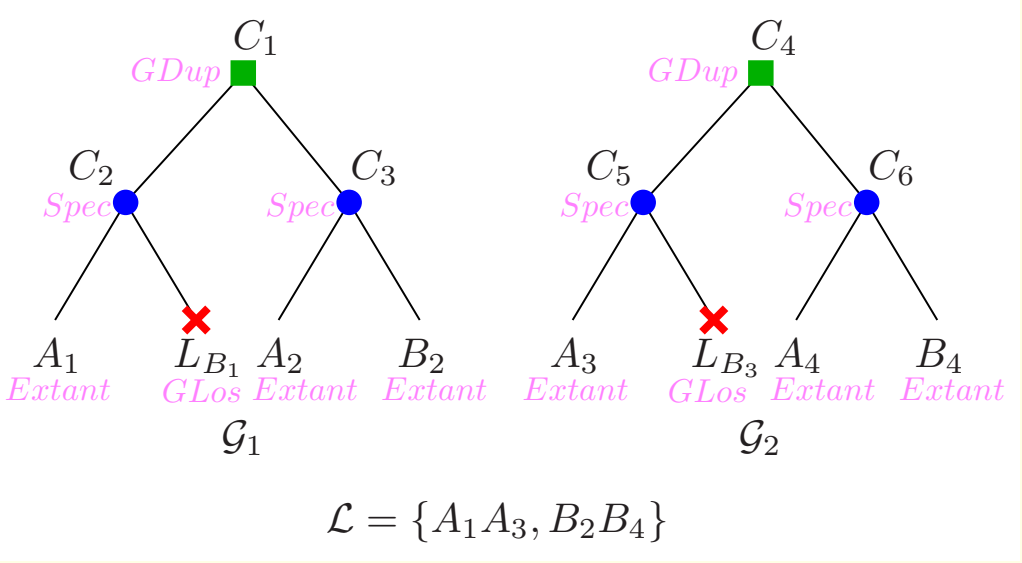
- Matrice de programmation dynamique, coûts c_0 et c_1 :

$\rightarrow$ $c_0(n_1, n_2)$ coût minimum d'une histoire évolutive où n_1 et n_2 ne sont pas adjacents

$\rightarrow$ $c_1(n_1, n_2)$ coût minimum d'une histoire évolutive où n_1 et n_2 sont adjacents

- Calcul des coûts entre nœuds de même espèce

Calcul des coûts



c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1		×		×	×	×	×
L_{B_1}	×		×		×	×	×
A_2		×		×	×	×	×
B_2	×		×		×	×	×
C_2	×	×	×	×			
C_3	×	×	×	×			
C_1	×	×	×	×			

- Matrice de programmation dynamique, coûts c_0 et c_1 :

→ $c_0(n_1, n_2)$ coût minimum d'une histoire évolutive où n_1 et n_2 ne sont pas adjacents

→ $c_1(n_1, n_2)$ coût minimum d'une histoire évolutive où n_1 et n_2 sont adjacents

- Calcul des coûts entre nœuds de même espèce

- Calcul des coûts selon les événements associés aux nœuds ⇒ plusieurs cas

À chaque cas est associé une formule de récurrence :

1. $\underline{E(n_1) = Extant \text{ et } E(n_2) = Extant}$
   Si $n_1 n_2 \in \mathcal{L}$ alors $c_1(n_1, n_2) = 0$ et $c_0(n_1, n_2) = \infty$
   sinon $c_1(n_1, n_2) = \infty$ et $c_0(n_1, n_2) = 0$

À chaque cas est associé une formule de récurrence :

1. $\underline{E(n_1) = Extant \text{ et } E(n_2) = Extant}$
   Si $n_1 n_2 \in \mathcal{L}$ alors $c_1(n_1, n_2) = 0$ et $c_0(n_1, n_2) = \infty$
   sinon $c_1(n_1, n_2) = \infty$ et $c_0(n_1, n_2) = 0$

2. $\underline{E(n_1) = GLos \text{ et } E(n_2) \neq GLos}$
   Alors $c_1(n_1, n_2) = 0$ et $c_0(n_1, n_2) = 0$

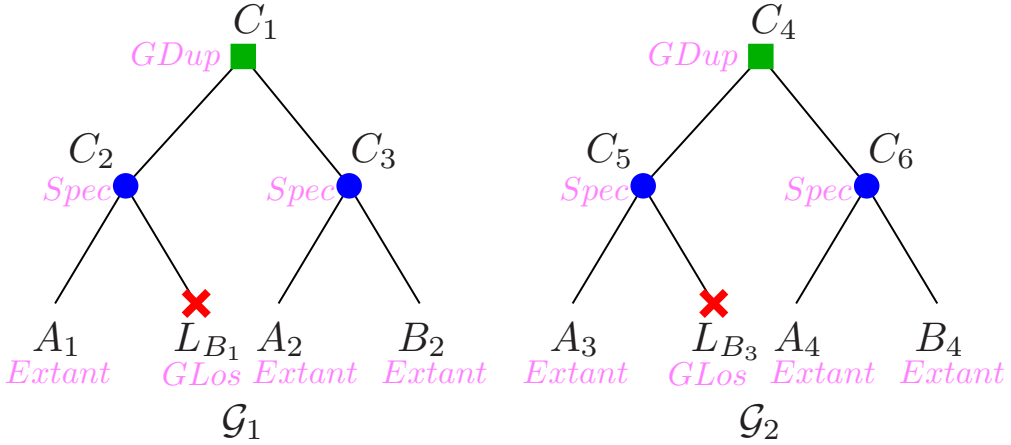
À chaque cas est associé une formule de récurrence :

1. $\underline{E(n_1) = Extant \text{ et } E(n_2) = Extant}$
   Si $n_1 n_2 \in \mathcal{L}$ alors $c_1(n_1, n_2) = 0$ et $c_0(n_1, n_2) = \infty$
   sinon $c_1(n_1, n_2) = \infty$ et $c_0(n_1, n_2) = 0$

2. $\underline{E(n_1) = GLos \text{ et } E(n_2) \neq GLos}$
   Alors $c_1(n_1, n_2) = 0$ et $c_0(n_1, n_2) = 0$

3. $\underline{E(n_1) = GLos \text{ et } E(n_2) = GLos}$
   Alors $c_1(n_1, n_2) = 0$ et $c_0(n_1, n_2) = 0$

Exemple des cas des feuilles

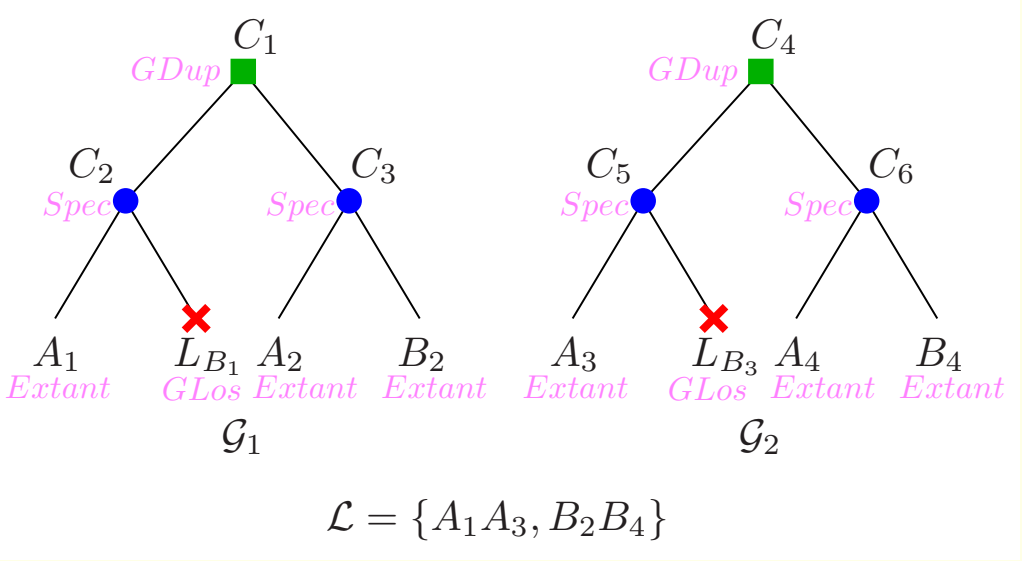


$\mathcal{L} = \{A_1A_3, B_2B_4\}$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1		X		X	X	X	X
L_{B_1}	X		X		X	X	X
A_2		X		X	X	X	X
B_2	X		X		X	X	X
C_2	X	X	X	X			
C_3	X	X	X	X			
C_1	X	X	X	X			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$

Exemple des cas des feuilles

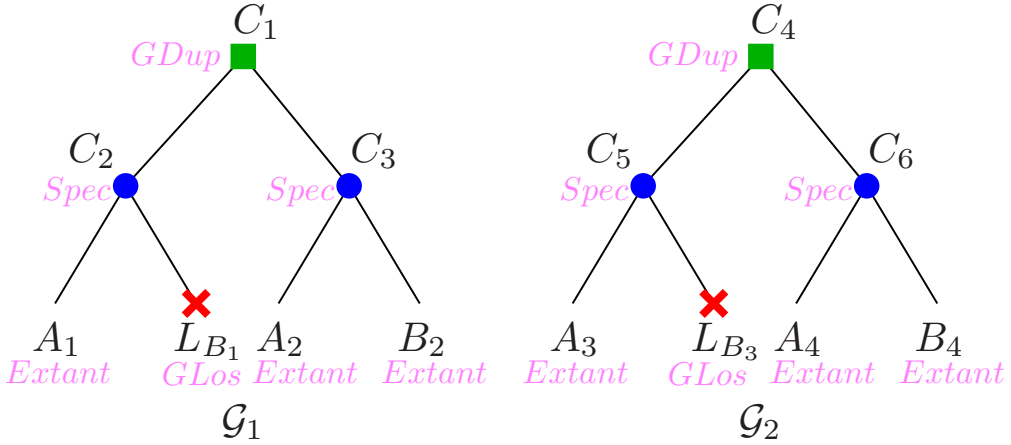


$\mathcal{L} = \{A_1A_3, B_2B_4\}$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X		X	X	X	X
L_{B_1}	X		X		X	X	X
A_2		X		X	X	X	X
B_2	X		X		X	X	X
C_2	X	X	X	X			
C_3	X	X	X	X			
C_1	X	X	X	X			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$

Exemple des cas des feuilles

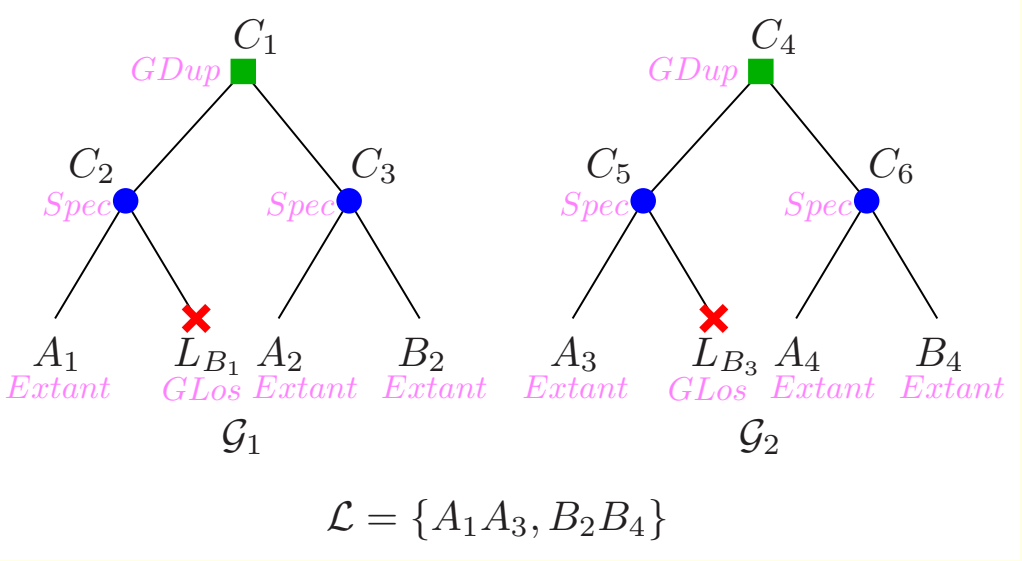


$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X		X	X	X	X
L_{B_1}	X		X		X	X	X
A_2		X		X	X	X	X
B_2	X		X		X	X	X
C_2	X	X	X	X			
C_3	X	X	X	X			
C_1	X	X	X	X			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$
- A_1-A_4 : cas 1. $Extant/Extant$, $A_1A_4 \notin \mathcal{L}$

Exemple des cas des feuilles

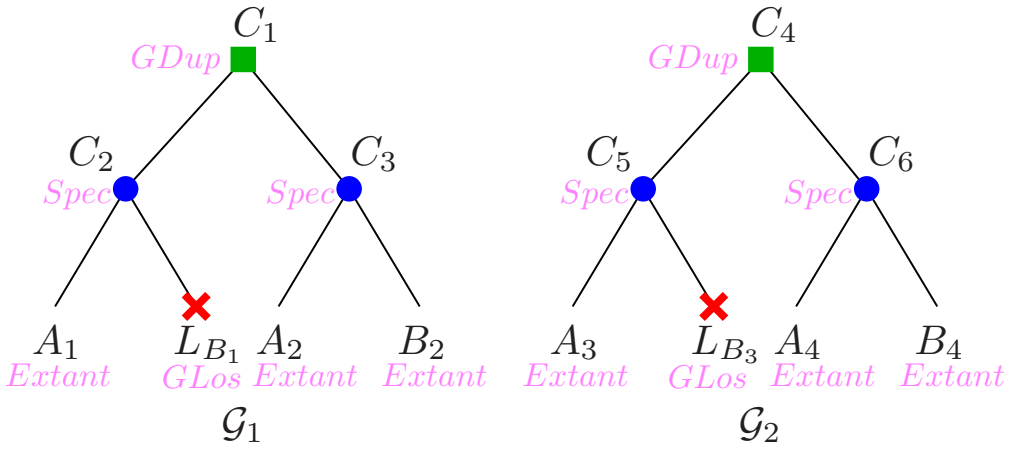


$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X		X		X	X	X
A_2		X		X	X	X	X
B_2	X		X		X	X	X
C_2	X	X	X	X			
C_3	X	X	X	X			
C_1	X	X	X	X			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$
- A_1-A_4 : cas 1. $Extant/Extant$, $A_1A_4 \notin \mathcal{L}$

Exemple des cas des feuilles

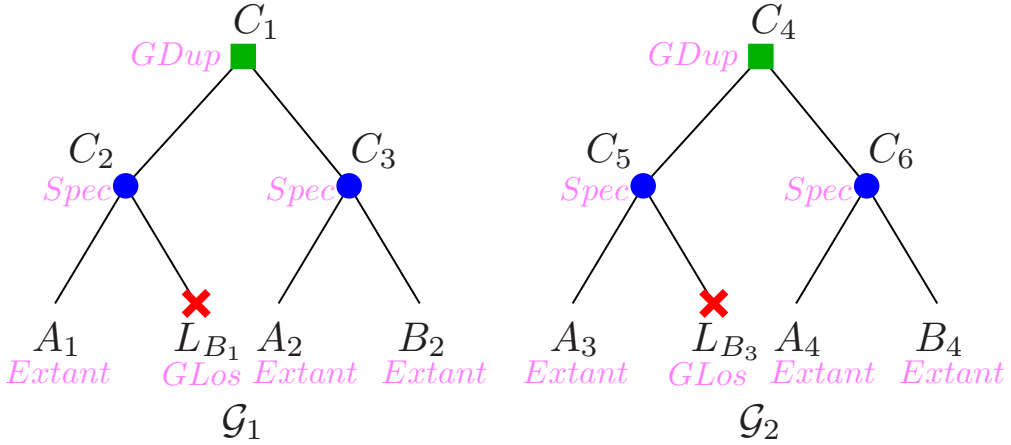


$\mathcal{L} = \{A_1A_3, B_2B_4\}$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗		✗		✗	✗	✗
A_2		✗		✗	✗	✗	✗
B_2	✗		✗		✗	✗	✗
C_2	✗	✗	✗	✗			
C_3	✗	✗	✗	✗			
C_1	✗	✗	✗	✗			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$
- A_1-A_4 : cas 1. $Extant/Extant$, $A_1A_4 \notin \mathcal{L}$
- L_{B_1}-L_{B_3} : cas 3. $GLos/GLos$

Exemple des cas des feuilles

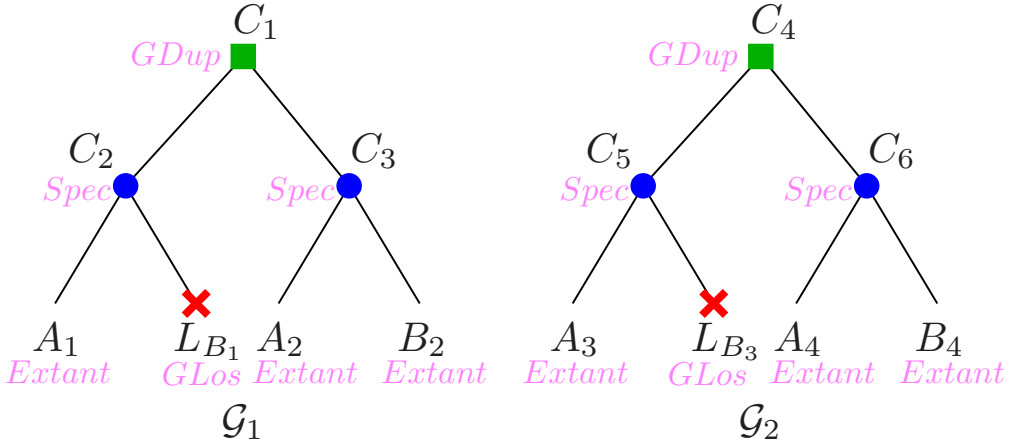


$\mathcal{L} = \{A_1A_3, B_2B_4\}$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗	$0/0$	✗		✗	✗	✗
A_2		✗		✗	✗	✗	✗
B_2	✗		✗		✗	✗	✗
C_2	✗	✗	✗	✗			
C_3	✗	✗	✗	✗			
C_1	✗	✗	✗	✗			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$
- A_1-A_4 : cas 1. $Extant/Extant$, $A_1A_4 \notin \mathcal{L}$
- L_{B_1}-L_{B_3} : cas 3. $GLos/GLos$

Exemple des cas des feuilles

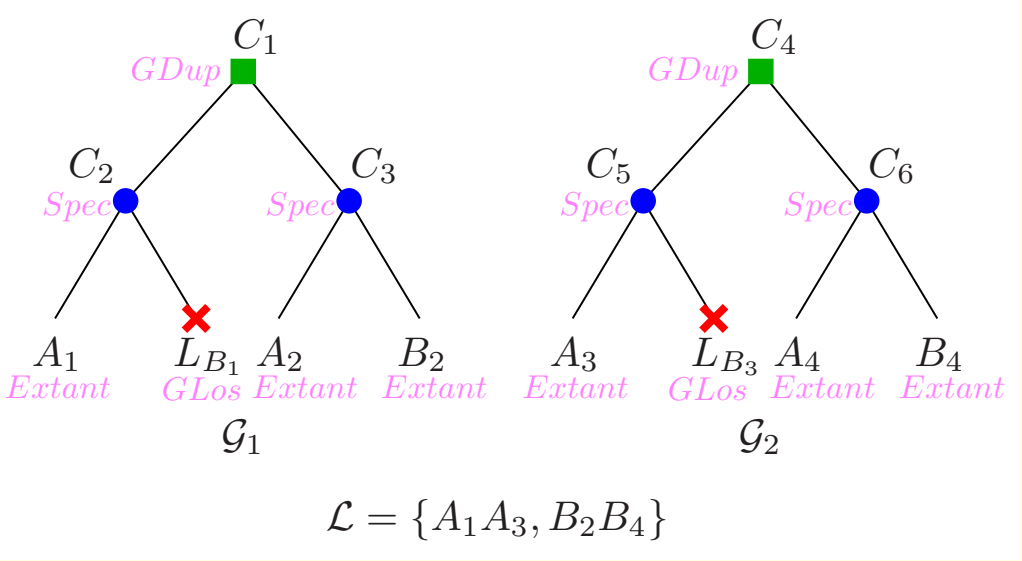


$\mathcal{L} = \{A_1A_3, B_2B_4\}$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗	$0/0$	✗		✗	✗	✗
A_2		✗		✗	✗	✗	✗
B_2	✗		✗		✗	✗	✗
C_2	✗	✗	✗	✗			
C_3	✗	✗	✗	✗			
C_1	✗	✗	✗	✗			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$
- A_1-A_4 : cas 1. $Extant/Extant$, $A_1A_4 \notin \mathcal{L}$
- L_{B_1}-L_{B_3} : cas 3. $GLos/GLos$
- L_{B_1}-B_4 : cas 2. $GLos/Any$

Exemple des cas des feuilles

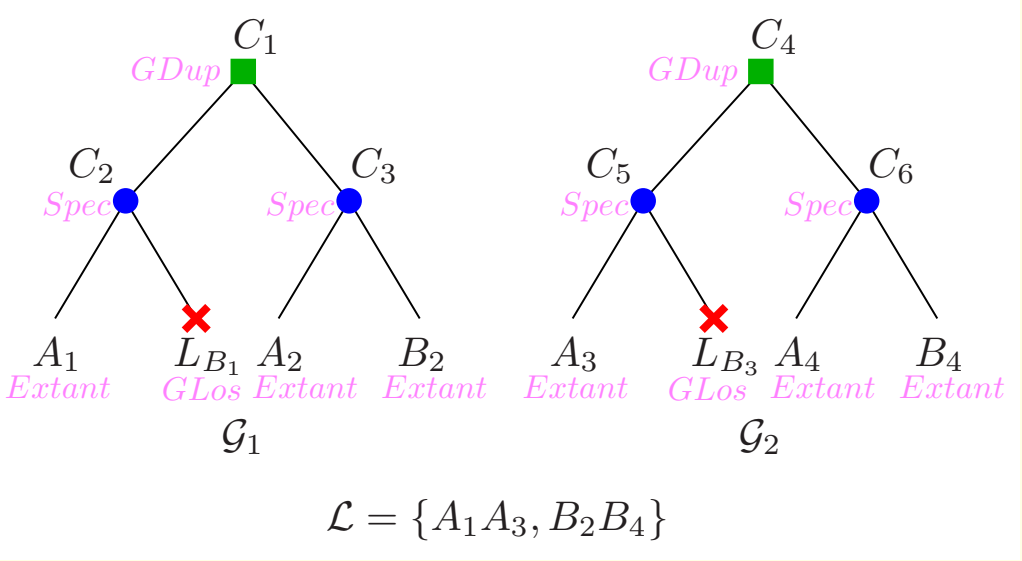


$\mathcal{L} = \{A_1A_3, B_2B_4\}$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗	$0/0$	✗	$0/0$	✗	✗	✗
A_2		✗		✗	✗	✗	✗
B_2	✗		✗		✗	✗	✗
C_2	✗	✗	✗	✗			
C_3	✗	✗	✗	✗			
C_1	✗	✗	✗	✗			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$
- A_1-A_4 : cas 1. $Extant/Extant$, $A_1A_4 \notin \mathcal{L}$
- L_{B_1}-L_{B_3} : cas 3. $GLos/GLos$
- L_{B_1}-B_4 : cas 2. $GLos/Any$

Exemple des cas des feuilles

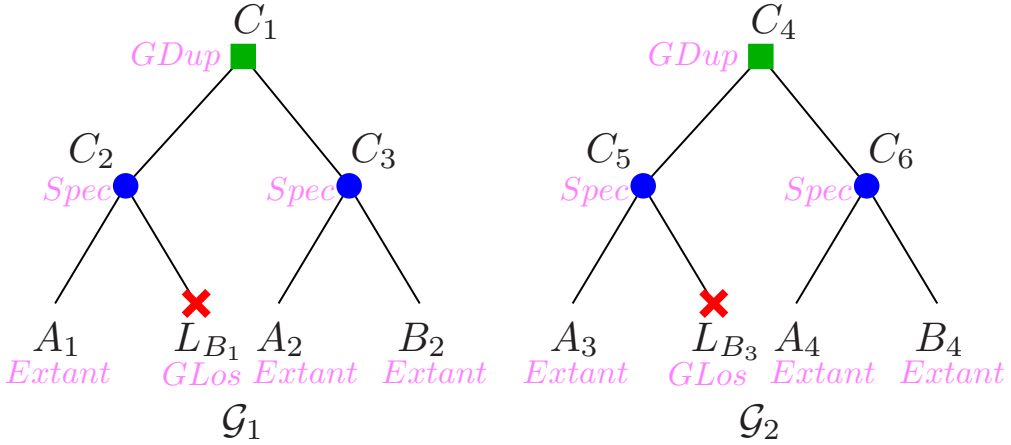


$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗	$0/0$	✗	$0/0$	✗	✗	✗
A_2		✗		✗	✗	✗	✗
B_2	✗		✗		✗	✗	✗
C_2	✗	✗	✗	✗			
C_3	✗	✗	✗	✗			
C_1	✗	✗	✗	✗			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$
- A_1-A_4 : cas 1. $Extant/Extant$, $A_1A_4 \notin \mathcal{L}$
- L_{B_1}-L_{B_3} : cas 3. $GLos/GLos$
- L_{B_1}-B_4 : cas 2. $GLos/Any$
- ...

Exemple des cas des feuilles



$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗	$0/0$	✗	$0/0$	✗	✗	✗
A_2	$0/\infty$	✗	$0/\infty$	✗	✗	✗	✗
B_2	✗	$0/0$	✗	$\infty/0$	✗	✗	✗
C_2	✗	✗	✗	✗			
C_3	✗	✗	✗	✗			
C_1	✗	✗	✗	✗			

- A_1-A_3 : cas 1. $Extant/Extant$, $A_1A_3 \in \mathcal{L}$
- A_1-A_4 : cas 1. $Extant/Extant$, $A_1A_4 \notin \mathcal{L}$
- L_{B_1}-L_{B_3} : cas 3. $GLos/GLos$
- L_{B_1}-B_4 : cas 2. $GLos/Any$
- ...

4. $E(n_1) \in \{Extant, Spec\}$ et $E(n_2) = GDup$

$$c_1(n_1, n_2) = \min \begin{cases} c_1(n_1, fg(n_2)) + c_0(n_1, fd(n_2)) \\ c_0(n_1, fg(n_2)) + c_1(n_1, fd(n_2)) \\ c_1(n_1, fg(n_2)) + c_1(n_1, fd(n_2)) + C(Gain) \\ c_0(n_1, fg(n_2)) + c_0(n_1, fd(n_2)) + C(Break) \end{cases}$$

4. $E(n_1) \in \{Extant, Spec\}$ et $E(n_2) = GDup$

$$c_1(n_1, n_2) = \min \begin{cases} c_1(n_1, fg(n_2)) + c_0(n_1, fd(n_2)) \\ c_0(n_1, fg(n_2)) + c_1(n_1, fd(n_2)) \\ c_1(n_1, fg(n_2)) + c_1(n_1, fd(n_2)) + C(Gain) \\ c_0(n_1, fg(n_2)) + c_0(n_1, fd(n_2)) + C(Break) \end{cases}$$

$$c_0(n_1, n_2) = \min \begin{cases} c_0(n_1, fg(n_2)) + c_0(n_1, fd(n_2)) \\ c_0(n_1, fg(n_2)) + c_1(n_1, fd(n_2)) + C(Gain) \\ c_1(n_1, fg(n_2)) + c_0(n_1, fd(n_2)) + C(Gain) \\ c_1(n_1, fg(n_2)) + c_1(n_1, fd(n_2)) + 2 * C(Gain) \end{cases}$$

5. $E(n_1) = Spec$ et $E(n_2) = Spec$

(on suppose que $S(fg(n_1)) = S(fg(n_2))$ et $S(fd(n_1)) = S(fd(n_2))$)

$$c_1(n_1, n_2) = \min \begin{cases} c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) \\ c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + C(Break) \\ c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + C(Break) \\ c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + 2 * C(Break) \end{cases}$$

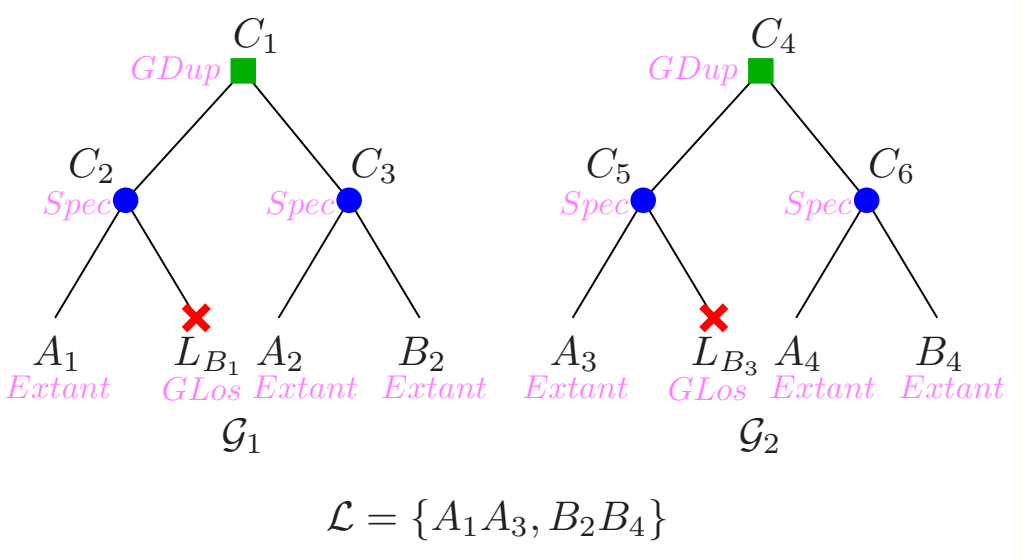
5. $\underline{E(n_1) = Spec \text{ et } E(n_2) = Spec}$

(on suppose que $S(fg(n_1)) = S(fg(n_2))$ et $S(fd(n_1)) = S(fd(n_2))$)

$$c_1(n_1, n_2) = \min \begin{cases} c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) \\ c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + C(Break) \\ c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + C(Break) \\ c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + 2 * C(Break) \end{cases}$$

$$c_0(n_1, n_2) = \min \begin{cases} c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) \\ c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + C(Gain) \\ c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + C(Gain) \\ c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + 2 * C(Gain) \end{cases}$$

Exemple des cas *Spec/Spec* et *Spec/GDup*

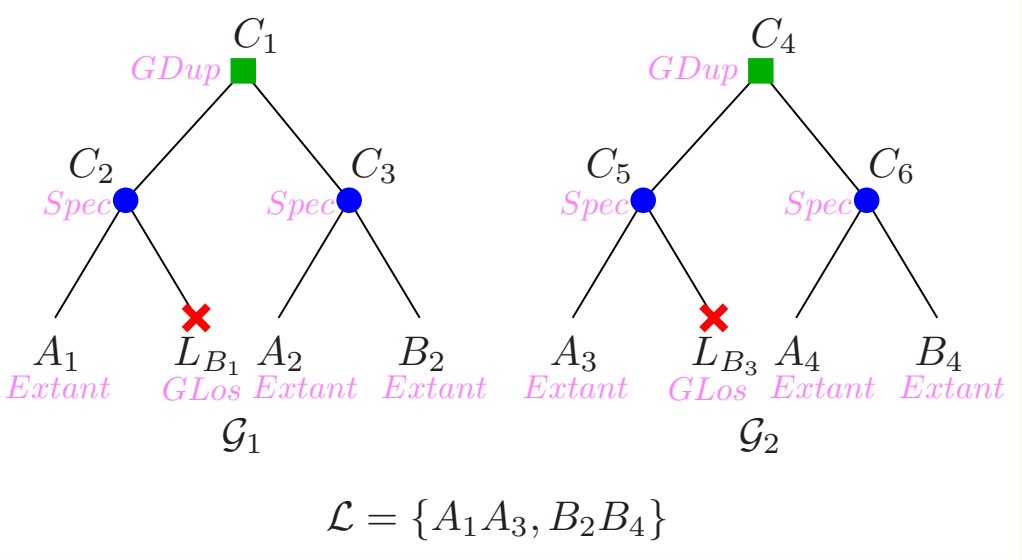


c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X			
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_5 : cas 5. *Spec/Spec*

$$c_1(C_2, C_5) = \min \begin{cases} c_1(A_1, A_3) + c_1(L_{B_1}, L_{B_3}) \\ c_1(A_1, A_3) + c_0(L_{B_1}, L_{B_3}) + C(Break) \\ c_0(A_1, A_3) + c_1(L_{B_1}, L_{B_3}) + C(Break) \\ c_0(A_1, A_3) + c_0(L_{B_1}, L_{B_3}) + 2 * C(Break) \end{cases}$$

Exemple des cas *Spec/Spec* et *Spec/GDup*

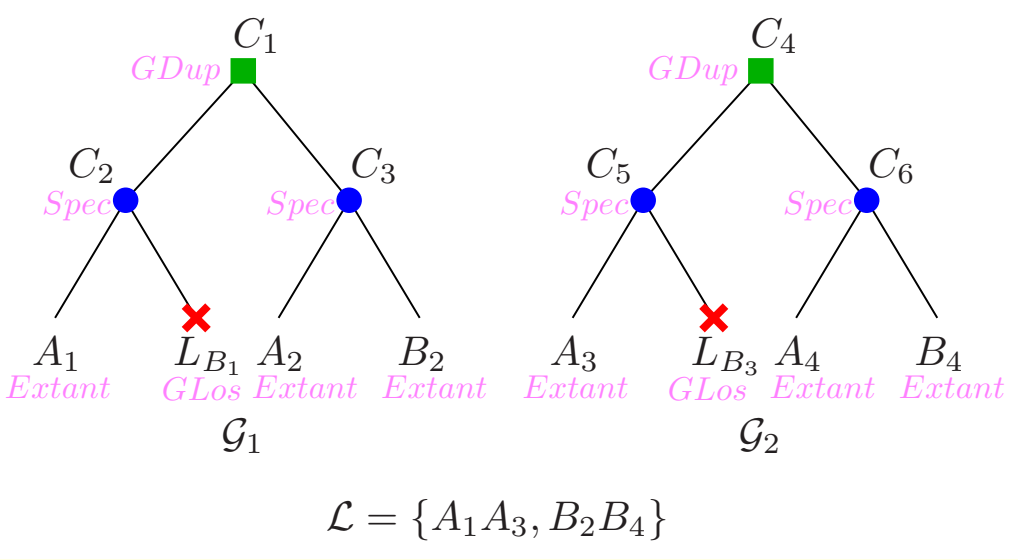


c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗	$0/0$	✗	$0/0$	✗	✗	✗
A_2	$0/\infty$	✗	$0/\infty$	✗	✗	✗	✗
B_2	✗	$0/0$	✗	$\infty/0$	✗	✗	✗
C_2	✗	✗	✗	✗			
C_3	✗	✗	✗	✗			
C_1	✗	✗	✗	✗			

- C_2-C_5 : cas 5. *Spec/Spec*

$$c_1(C_2, C_5) = \min \begin{cases} 0 + 0 \\ 0 + 0 + 1 \\ \infty + 0 + 1 \\ \infty + 0 + 2 \end{cases}$$

Exemple des cas *Spec/Spec* et *Spec/GDup*



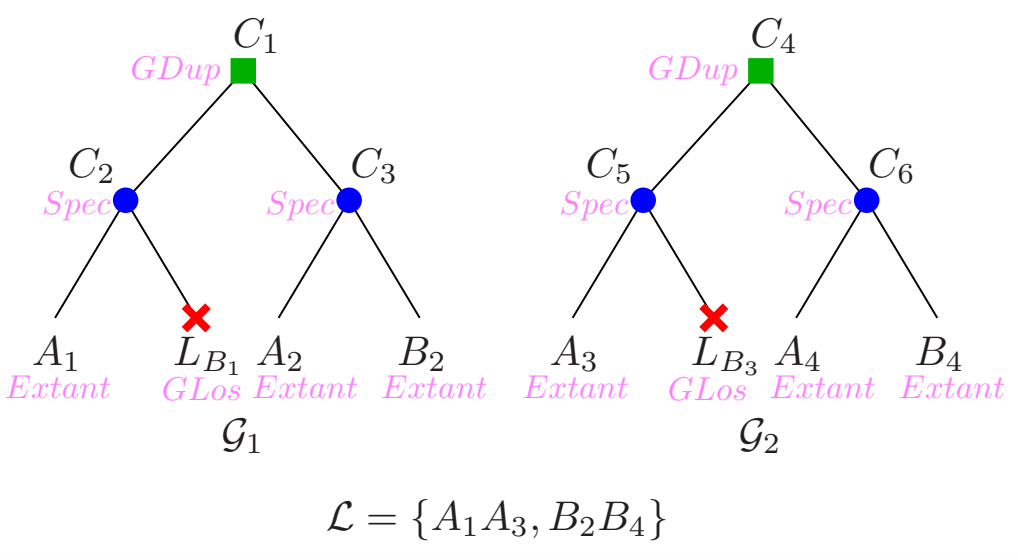
$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$/0$		
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_5 : cas 5. *Spec/Spec*

$$c_1(C_2, C_5) = \min \begin{cases} 0 & + & 0 & & \\ 0 & + & 0 & + & 1 \\ \infty & + & 0 & + & 1 \\ \infty & + & 0 & + & 2 \end{cases}$$

Exemple des cas *Spec/Spec* et *Spec/GDup*

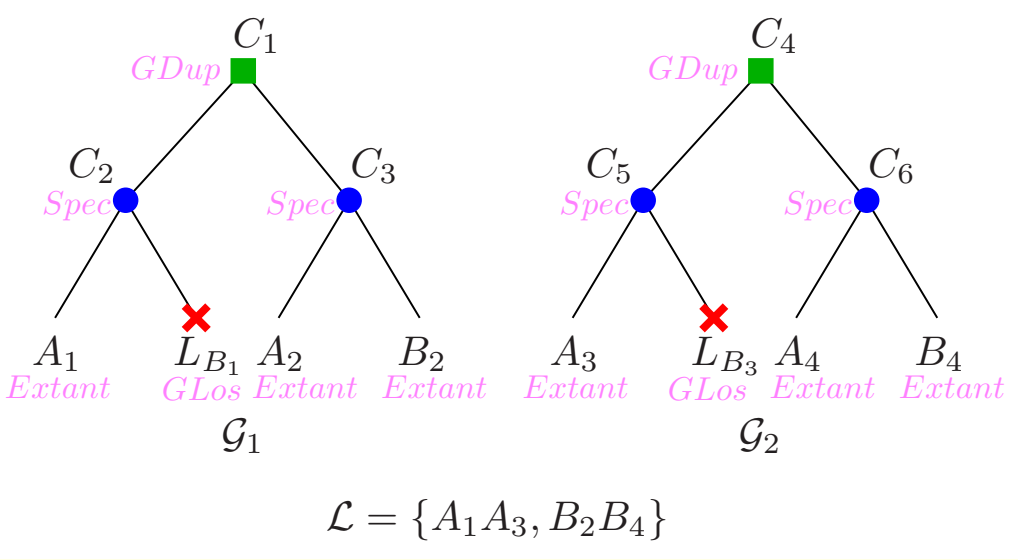


c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$/0$		
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_5 : cas 5. *Spec/Spec*

$$c_0(C_2, C_5) = \min \begin{cases} c_0(A_1, A_3) + c_0(L_{B_1}, L_{B_3}) \\ c_1(A_1, A_3) + c_0(L_{B_1}, L_{B_3}) + C(Gain) \\ c_0(A_1, A_3) + c_1(L_{B_1}, L_{B_3}) + C(Gain) \\ c_1(A_1, A_3) + c_1(L_{B_1}, L_{B_3}) + 2 * C(Gain) \end{cases}$$

Exemple des cas *Spec/Spec* et *Spec/GDup*



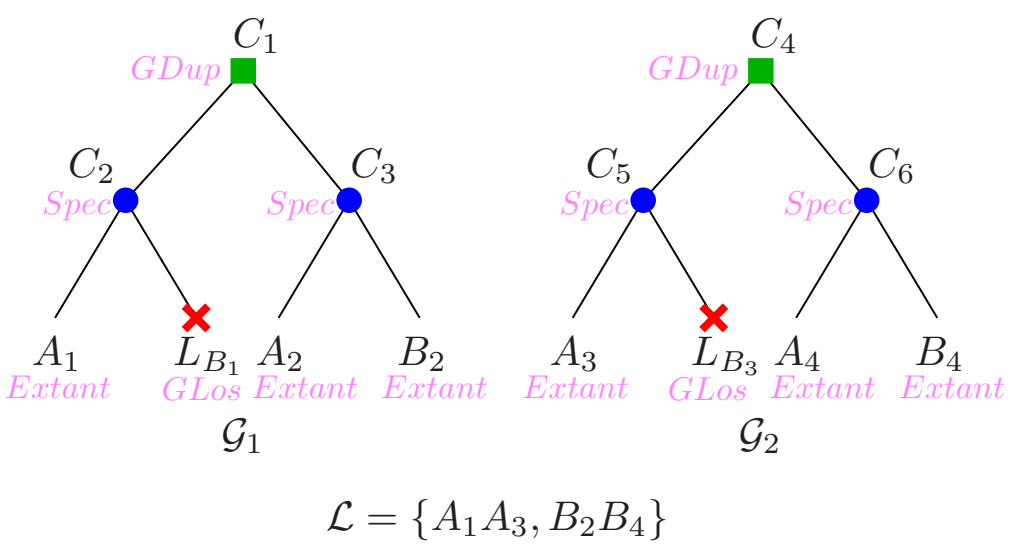
$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$/0$		
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_5 : cas 5. *Spec/Spec*

$$c_0(C_2, C_5) = \min \begin{cases} \infty + 0 \\ 0 + 0 + 1 \\ \infty + 0 + 1 \\ 0 + 0 + 2 \end{cases}$$

Exemple des cas *Spec/Spec* et *Spec/GDup*



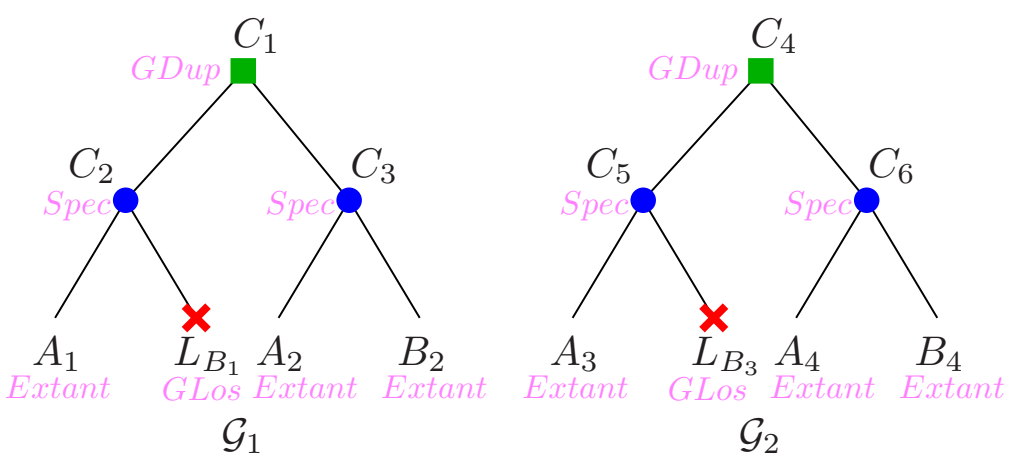
$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$		
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_5 : cas 5. *Spec/Spec*

$$c_0(C_2, C_5) = \min \begin{cases} \infty + 0 \\ 0 + 0 + 1 \\ \infty + 0 + 1 \\ 0 + 0 + 2 \end{cases}$$

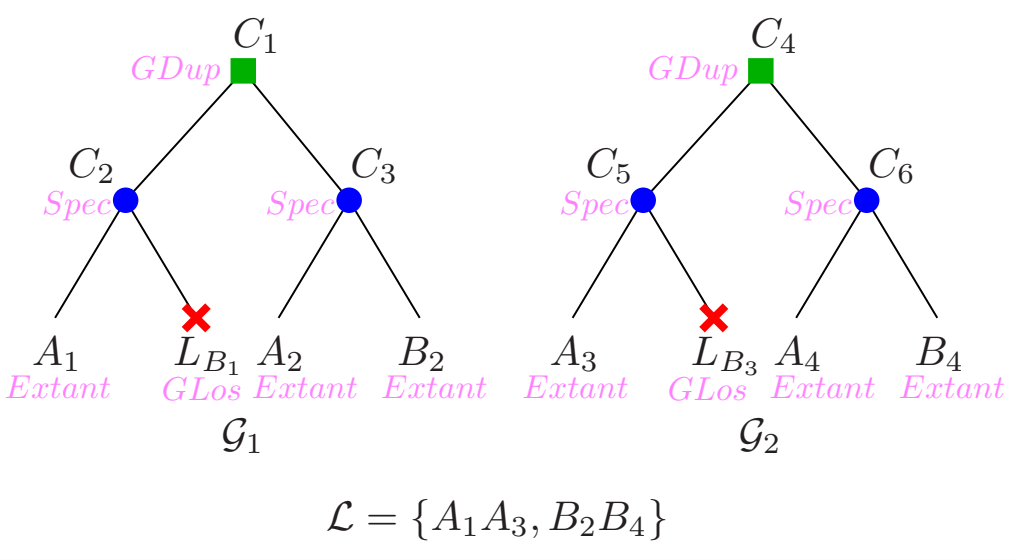
Exemple des cas *Spec/Spec* et *Spec/GDup*



c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$		
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_6 : cas 5. *Spec/Spec*

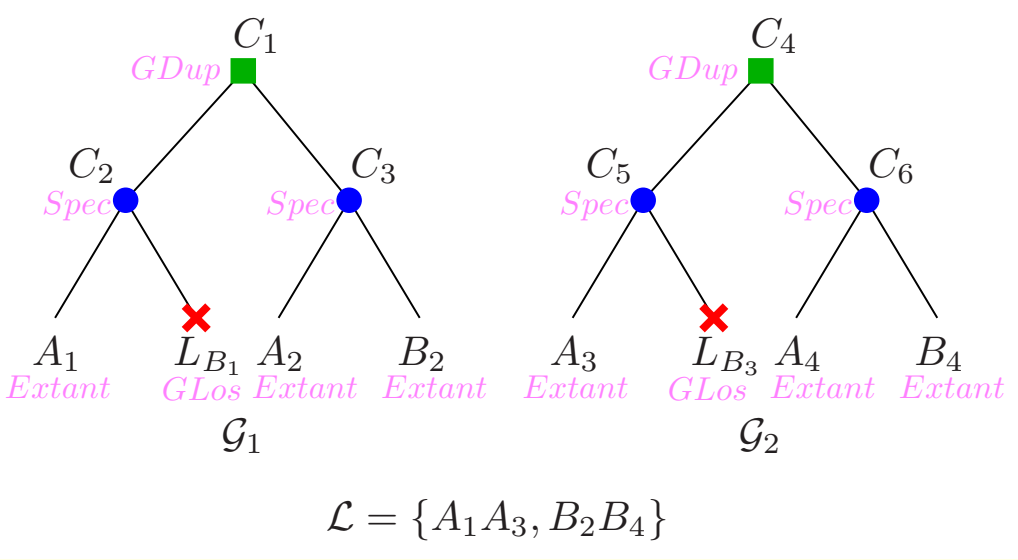
Exemple des cas *Spec/Spec* et *Spec/GDup*



c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$	$0/1$	
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_6 : cas 5. *Spec/Spec*

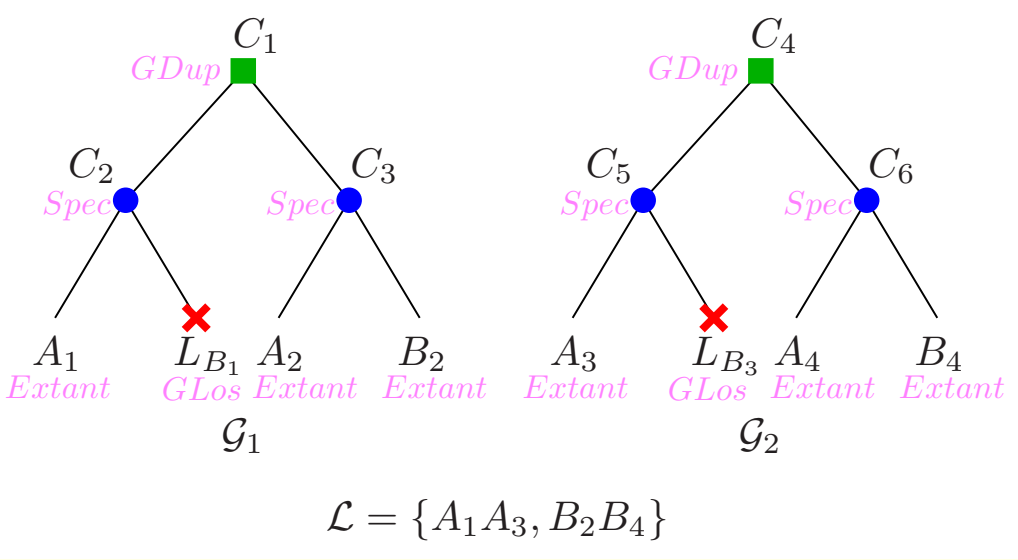
Exemple des cas *Spec/Spec* et *Spec/GDup*



c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$	$0/1$	
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_6 : cas 5. *Spec/Spec*
- C_2-C_4 : cas 4. *Spec/GDup*

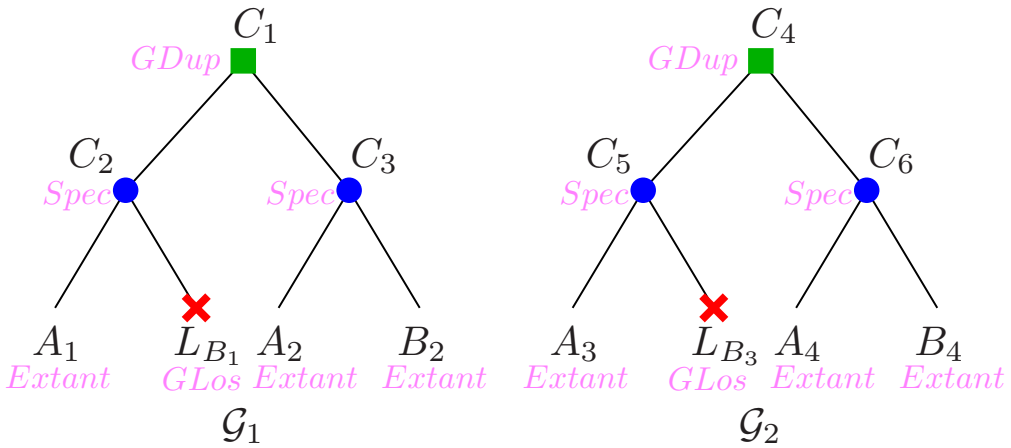
Exemple des cas *Spec/Spec* et *Spec/GDup*



c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$	$0/1$	$1/0$
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_6 : cas 5. *Spec/Spec*
- C_2-C_4 : cas 4. *Spec/GDup*

Exemple des cas *Spec/Spec* et *Spec/GDup*

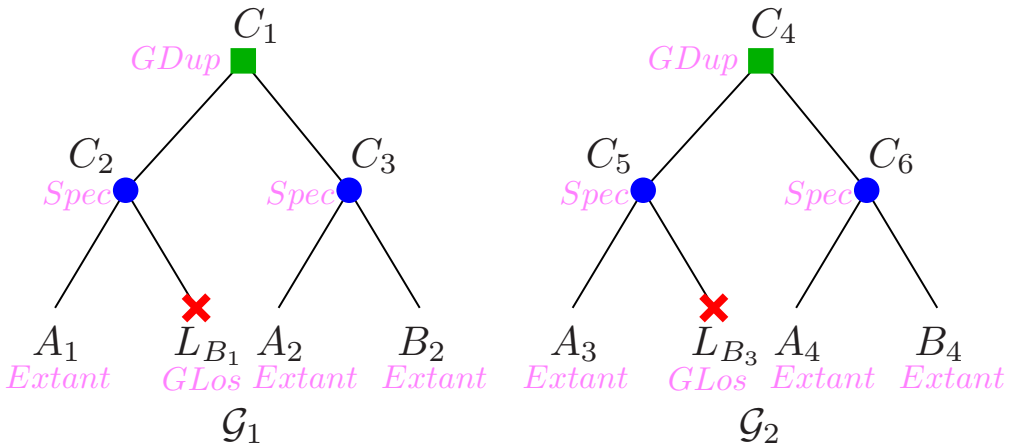


$\mathcal{L} = \{A_1A_3, B_2B_4\}$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$	$0/1$	$1/0$
C_3	X	X	X	X			
C_1	X	X	X	X			

- C_2-C_6 : cas 5. *Spec/Spec*
- C_2-C_4 : cas 4. *Spec/GDup*
- C_3-C_5 & C_3-C_6 : cas 5. *Spec/Spec*

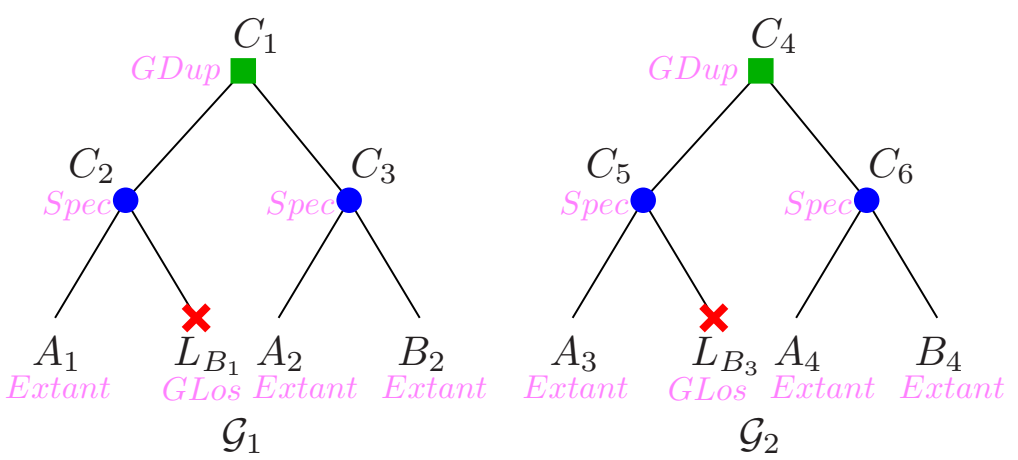
Exemple des cas *Spec/Spec* et *Spec/GDup*



c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$	$0/1$	$1/0$
C_3	X	X	X	X	$0/1$	$1/1$	
C_1	X	X	X	X			

- C_2-C_6 : cas 5. *Spec/Spec*
- C_2-C_4 : cas 4. *Spec/GDup*
- C_3-C_5 & C_3-C_6 : cas 5. *Spec/Spec*

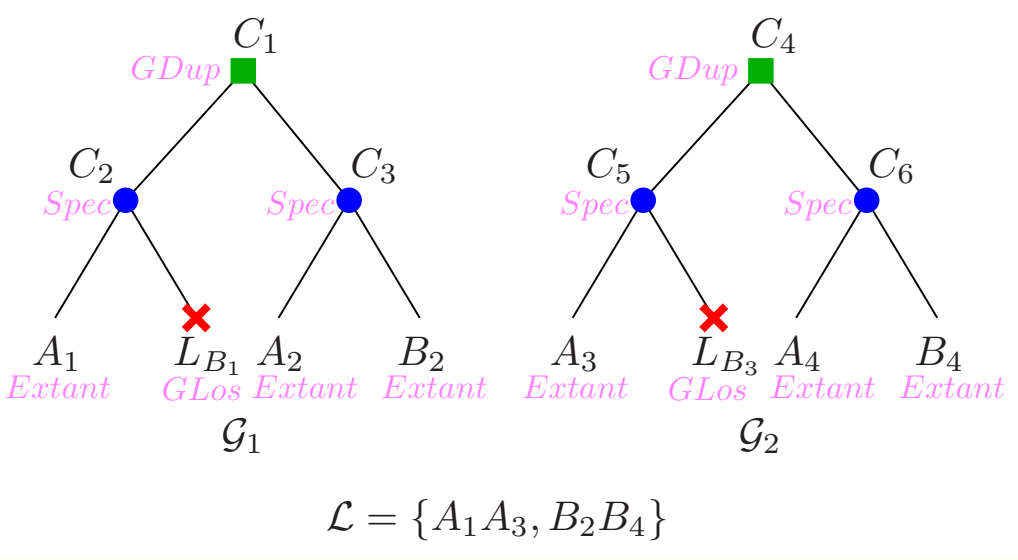
Exemple des cas *Spec/Spec* et *Spec/GDup*



c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$	$0/1$	$1/0$
C_3	X	X	X	X	$0/1$	$1/1$	
C_1	X	X	X	X			

- C_2-C_6 : cas 5. *Spec/Spec*
- C_2-C_4 : cas 4. *Spec/GDup*
- C_3-C_5 & C_3-C_6 : cas 5. *Spec/Spec*
- C_3-C_4, C_1-C_5 & C_1-C_6 : cas 4. *Spec/GDup*

Exemple des cas *Spec/Spec* et *Spec/GDup*



c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$	$0/1$	$1/0$
C_3	X	X	X	X	$0/1$	$1/1$	$1/1$
C_1	X	X	X	X	$1/0$	$1/1$	

- C_2-C_6 : cas 5. *Spec/Spec*
- C_2-C_4 : cas 4. *Spec/GDup*
- C_3-C_5 & C_3-C_6 : cas 5. *Spec/Spec*
- C_3-C_4, C_1-C_5 & C_1-C_6 : cas 4. *Spec/GDup*

Le dernier cas

Remarque : On n'examine pas le cas $E(n_1) = Extant$ et $E(n_2) = Spec$ car $S(n_1) \neq S(n_2)$

6. $E(n_1) = GDup$ and $E(n_2) = GDup$

$$c_1(n_1, n_2) = \min(D1, D2, D12)$$

Le dernier cas

Remarque : On n'examine pas le cas $E(n_1) = Extant$ et $E(n_2) = Spec$ car $S(n_1) \neq S(n_2)$

6. $E(n_1) = GDup$ and $E(n_2) = GDup$

$$c_1(n_1, n_2) = \min(D1, D2, D12)$$

D1 : cas où n_1 se duplique avant n_2

$$D1 = \min \begin{cases} c_1(fg(n_1), n_2) + c_0(fd(n_1), n_2) \\ c_0(fg(n_1), n_2) + c_1(fd(n_1), n_2) \\ c_1(fg(n_1), n_2) + c_1(fd(n_1), n_2) + C(Gain) \\ c_0(fg(n_1), n_2) + c_0(fd(n_1), n_2) + C(Break) \end{cases}$$

Le dernier cas

Remarque : On n'examine pas le cas $E(n_1) = Extant$ et $E(n_2) = Spec$ car $S(n_1) \neq S(n_2)$

6. $E(n_1) = GDup$ and $E(n_2) = GDup$

$$c_1(n_1, n_2) = \min(D1, D2, D12)$$

D1 : cas où n_1 se duplique avant n_2

$$D1 = \min \begin{cases} c_1(fg(n_1), n_2) + c_0(fd(n_1), n_2) \\ c_0(fg(n_1), n_2) + c_1(fd(n_1), n_2) \\ c_1(fg(n_1), n_2) + c_1(fd(n_1), n_2) + C(Gain) \\ c_0(fg(n_1), n_2) + c_0(fd(n_1), n_2) + C(Break) \end{cases}$$

D2 : cas où n_2 se duplique avant n_1

$$D2 = \min \begin{cases} c_1(n_1, fg(n_2)) + c_0(n_1, fd(n_2)) \\ c_0(n_1, fg(n_2)) + c_1(n_1, fd(n_2)) \\ c_1(n_1, fg(n_2)) + c_1(n_1, fd(n_2)) + C(Gain) \\ c_0(n_1, fg(n_2)) + c_0(n_1, fd(n_2)) + C(Break) \end{cases}$$

Le dernier cas (suite)

D12 : cas où n_1 et n_2 se dupliquent en même temps (16 cas)

$$D12 = \min \begin{cases}
c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) \\
c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) \\
c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Gain) \\
c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + 2 * C(Gain) \\
c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Break) \\
c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) + C(Break) \\
c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Gain) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Gain) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) \\
c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) \\
c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) \\
c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + 2 * C(Break)
\end{cases}$$

Le dernier cas (suite)

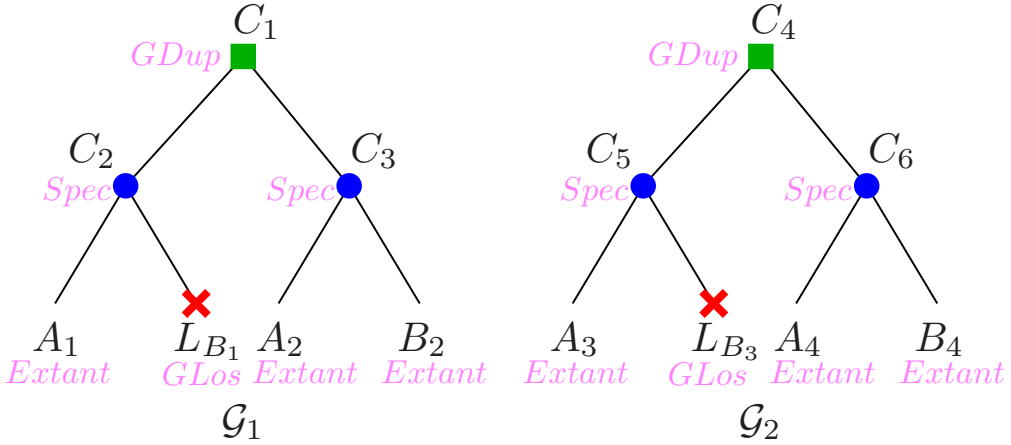
D12 : cas où n_1 et n_2 se dupliquent en même temps (16 cas)

$$D12 = \min \begin{cases}
c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) \\
c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) \\
c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Gain) \\
c_1(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + 2 * C(Gain) \\
c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Break) \\
c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) + C(Break) \\
c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Gain) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Gain) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) \\
c_0(fg(n_1), fg(n_2)) + c_1(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) \\
c_1(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Gain) \\
c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_1(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_1(fd(n_1), fg(n_2)) + C(Break) \\
c_0(fg(n_1), fg(n_2)) + c_0(fd(n_1), fd(n_2)) + c_0(fg(n_1), fd(n_2)) + c_0(fd(n_1), fg(n_2)) + 2 * C(Break)
\end{cases}$$

Coût c_0 pour le cas $GDup$-$GDup$:

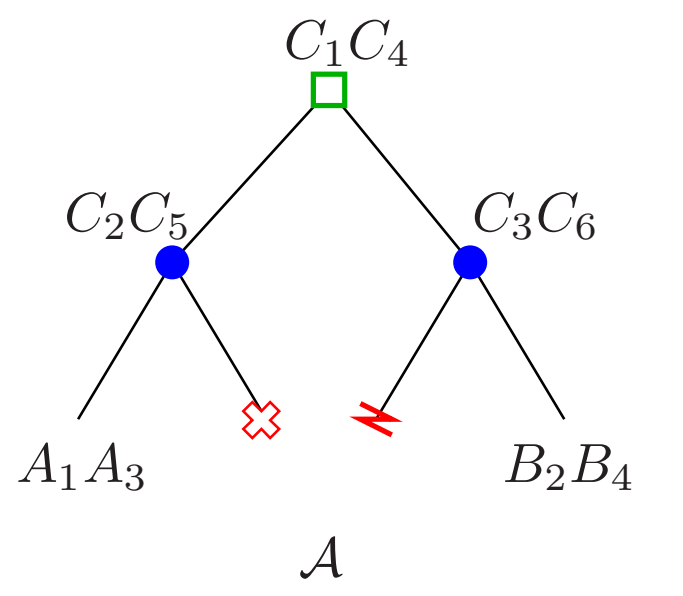
$$c_0(n_1, n_2) \sim \min(D1_0, D2_0)$$

Fin de l'exemple : le backtracking

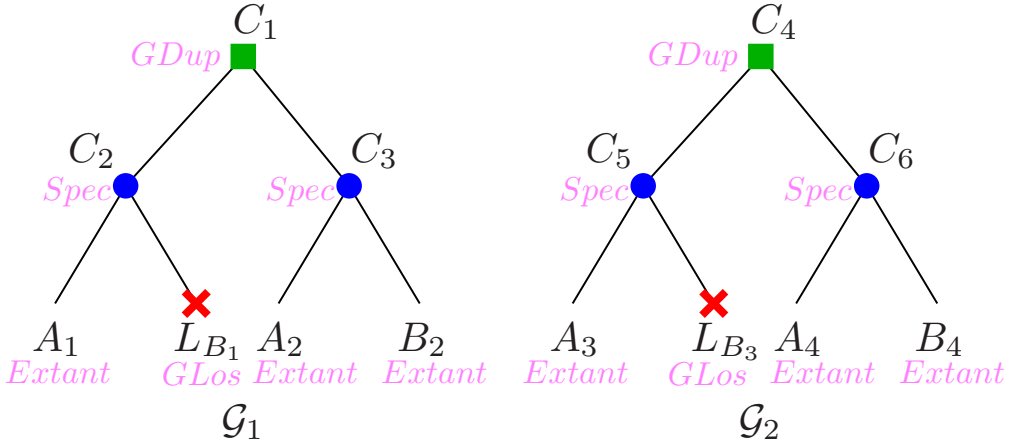


$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗	$0/0$	✗	$0/0$	✗	✗	✗
A_2	$0/\infty$	✗	$0/\infty$	✗	✗	✗	✗
B_2	✗	$0/0$	✗	$\infty/0$	✗	✗	✗
C_2	✗	✗	✗	✗	$1/0$	$0/1$	$1/0$
C_3	✗	✗	✗	✗	$0/1$	$1/1$	$1/1$
C_1	✗	✗	✗	✗	$1/0$	$1/1$	$2/1$

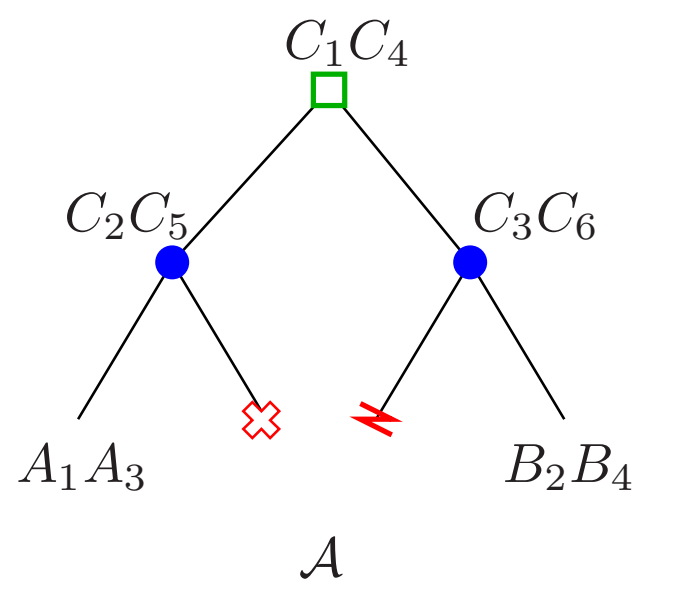


Fin de l'exemple : le backtracking

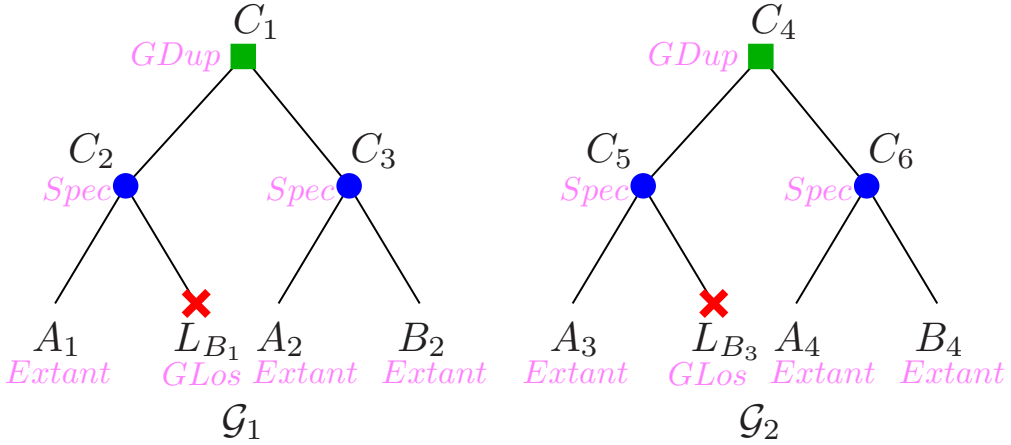


$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗	$0/0$	✗	$0/0$	✗	✗	✗
A_2	$0/\infty$	✗	$0/\infty$	✗	✗	✗	✗
B_2	✗	$0/0$	✗	$\infty/0$	✗	✗	✗
C_2	✗	✗	✗	✗	$1/0$	$0/1$	$1/0$
C_3	✗	✗	✗	✗	$0/1$	$1/1$	$1/1$
C_1	✗	✗	✗	✗	$1/0$	$1/1$	$2/1$

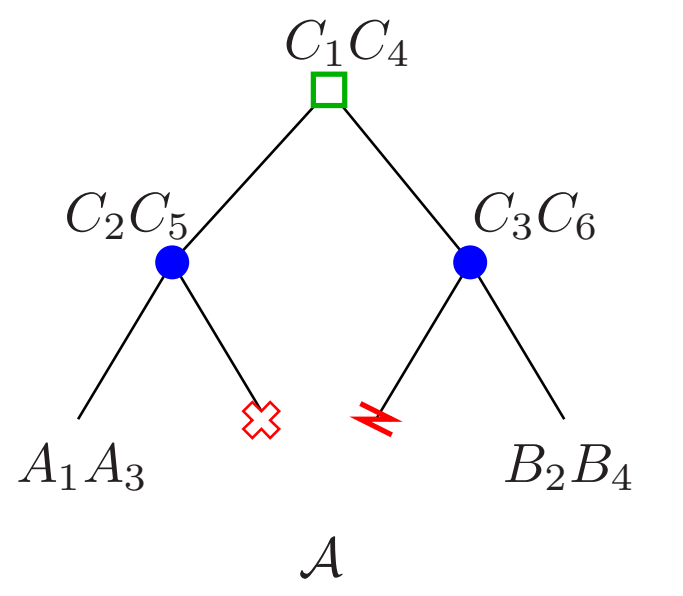


Fin de l'exemple : le backtracking

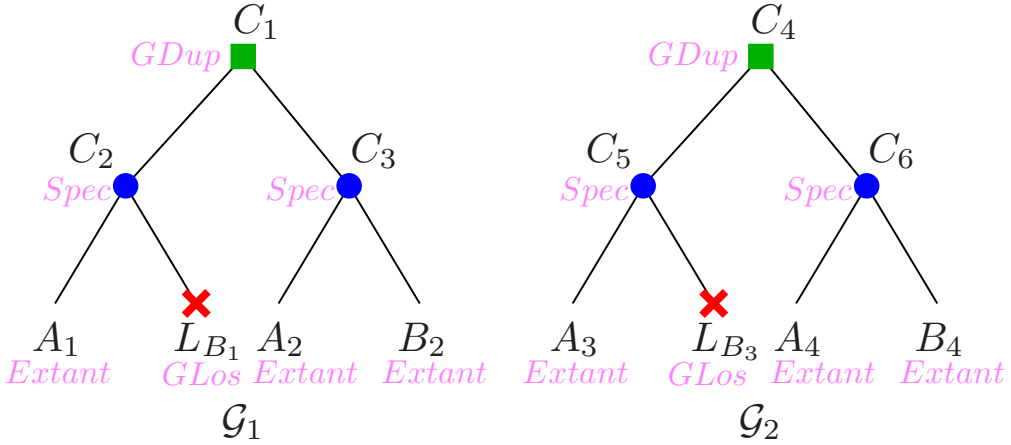


$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	✗	$0/\infty$	✗	✗	✗	✗
L_{B_1}	✗	$0/0$	✗	$0/0$	✗	✗	✗
A_2	$0/\infty$	✗	$0/\infty$	✗	✗	✗	✗
B_2	✗	$0/0$	✗	$\infty/0$	✗	✗	✗
C_2	✗	✗	✗	✗	$1/0$	$0/1$	$1/0$
C_3	✗	✗	✗	✗	$0/1$	$1/1$	$1/1$
C_1	✗	✗	✗	✗	$1/0$	$1/1$	$2/1$

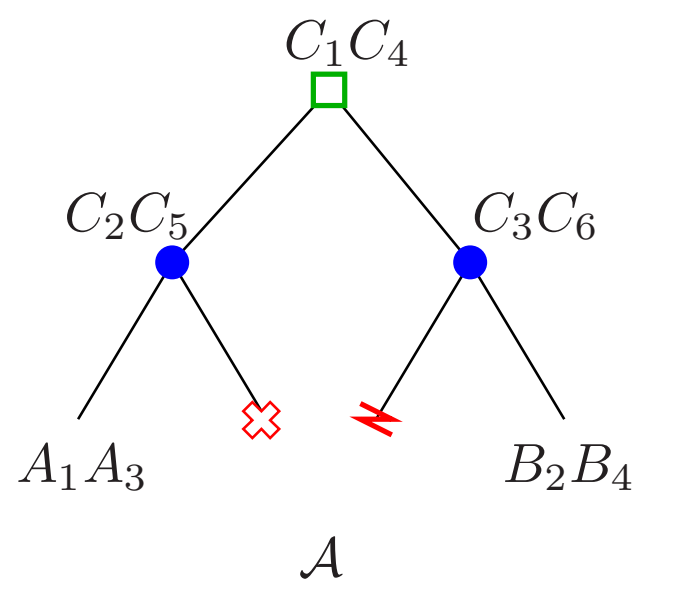


Fin de l'exemple : le backtracking

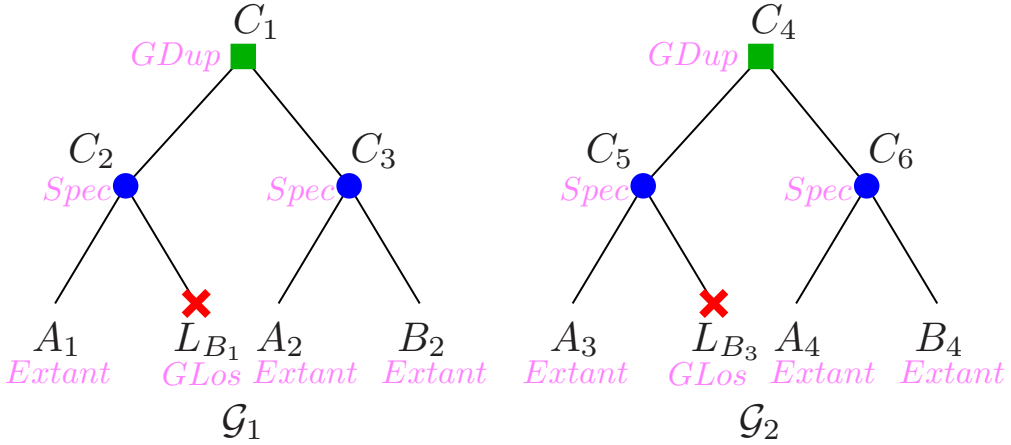


$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$	$0/1$	$1/0$
C_3	X	X	X	X	$0/1$	$1/1$	$1/1$
C_1	X	X	X	X	$1/0$	$1/1$	$2/1$

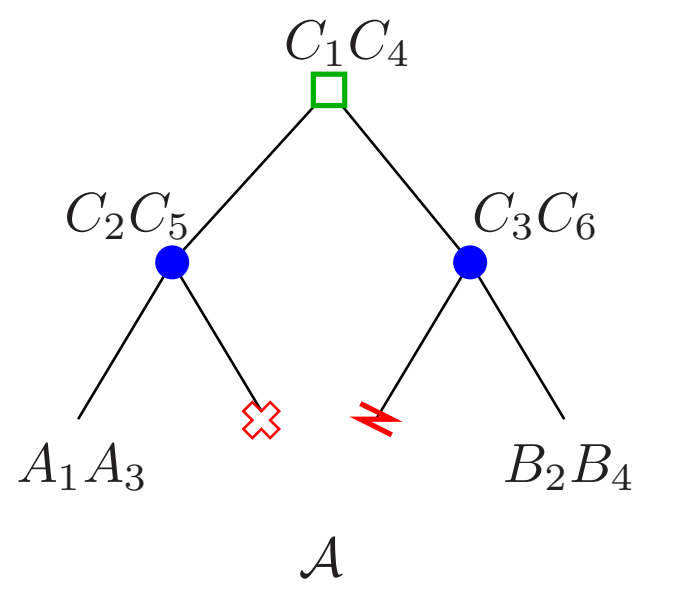


Fin de l'exemple : le backtracking



$$\mathcal{L} = \{A_1A_3, B_2B_4\}$$

c_0/c_1	A_3	L_{B_3}	A_4	B_4	C_5	C_6	C_4
A_1	$\infty/0$	X	$0/\infty$	X	X	X	X
L_{B_1}	X	$0/0$	X	$0/0$	X	X	X
A_2	$0/\infty$	X	$0/\infty$	X	X	X	X
B_2	X	$0/0$	X	$\infty/0$	X	X	X
C_2	X	X	X	X	$1/0$	$0/1$	$1/0$
C_3	X	X	X	X	$0/1$	$1/1$	$1/1$
C_1	X	X	X	X	$1/0$	$1/1$	$2/1$



- DeCo gère les pertes et les duplications de gènes (sans besoin de les ordonner), ainsi que des familles de gènes sans restriction d'un seul représentant pas espèce

Caractéristiques de DeCo

- DeCo gère les pertes et les duplications de gènes (sans besoin de les ordonner), ainsi que des familles de gènes sans restriction d'un seul représentant pas espèce

- Algorithme exact : on minimise le nombre de cassures et de créations d'adjacence

Caractéristiques de DeCo

- DeCo gère les pertes et les duplications de gènes (sans besoin de les ordonner), ainsi que des familles de gènes sans restriction d'un seul représentant pas espèce
- Algorithme exact : on minimise le nombre de cassures et de créations d'adjacence
- Généralisation des algorithmes de parcimonie de Fitch-Sankoff sur un alphabet binaire

Caractéristiques de DeCo

- DeCo gère les pertes et les duplications de gènes (sans besoin de les ordonner), ainsi que des familles de gènes sans restriction d'un seul représentant pas espèce

- Algorithme exact : on minimise le nombre de cassures et de créations d'adjacence

- Généralisation des algorithmes de parcimonie de Fitch-Sankoff sur un alphabet binaire

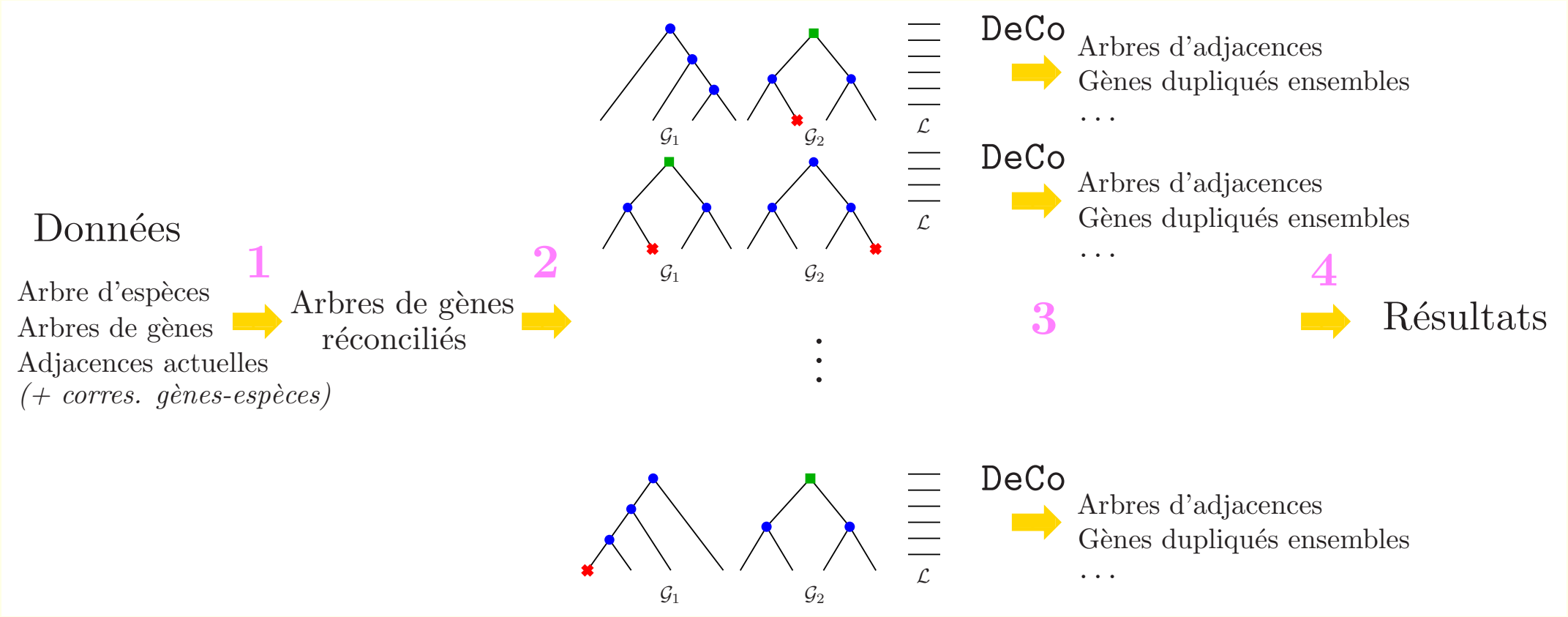
- Hypothèse : les adjacences évoluent indépendemment
  $\Rightarrow$ utilisation du principe de programmation dynamique

- DeCo gère les pertes et les duplications de gènes (sans besoin de les ordonner), ainsi que des familles de gènes sans restriction d'un seul représentant pas espèce
- Algorithme exact : on minimise le nombre de cassures et de créations d'adjacence
- Généralisation des algorithmes de parcimonie de Fitch-Sankoff sur un alphabet binaire
- Hypothèse : les adjacences évoluent indépendemment
  $\Rightarrow$ utilisation du principe de programmation dynamique
- Complexité en temps et en espace en $O(n^2)$ où n est le nombre de nœuds d'un arbre de gènes

Caractéristiques de DeCo

- DeCo gère les pertes et les duplications de gènes (sans besoin de les ordonner), ainsi que des familles de gènes sans restriction d'un seul représentant pas espèce
- Algorithme exact : on minimise le nombre de cassures et de créations d'adjacence
- Généralisation des algorithmes de parcimonie de Fitch-Sankoff sur un alphabet binaire
- Hypothèse : les adjacences évoluent indépendemment
  $\Rightarrow$ utilisation du principe de programmation dynamique
- Complexité en temps et en espace en $O(n^2)$ où n est le nombre de nœuds d'un arbre de gènes
- Inclus dans un "pipeline" pour traiter le problème général

Pipeline DeCo

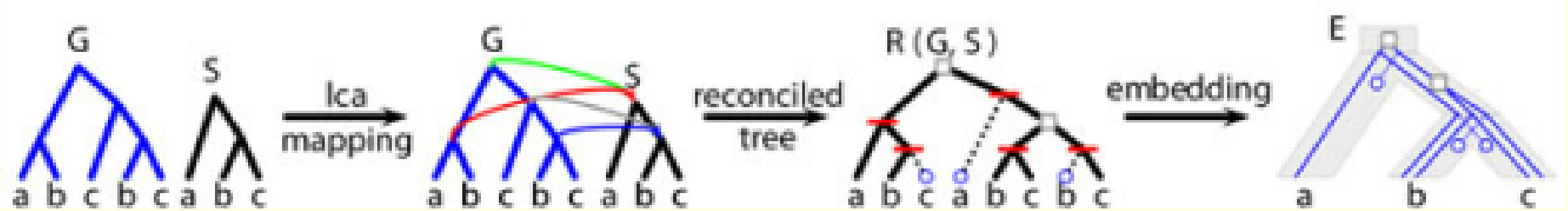


1. Réconciliation
2. Création des classes
3. DeCo
4. Synthèse

Programmé en C++/Bio++, disponible sur http://pbil.univ-lyon1.fr/software/DeCo/

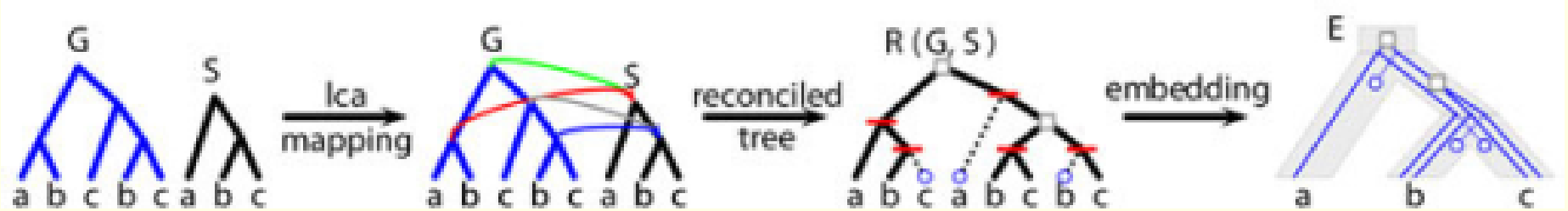
- Introduction
- Modèle
- Démarche
  - → Étape 1 : Réconciliation
  - → Étape 2 : Création des classes d'adjacences
  - → Étape 3 : Algorithme DeCo
  - → Étape 4 : Synthèse
- Validation
- Utilisation de DeCo pour l'assemblage : ARt-DeCo

- Réconciliation LCA (Least Common Ancestor)



- Solution unique
- Événements géniques pris en compte : duplications et pertes

- Réconciliation LCA (Least Common Ancestor)



- Solution unique
- Événements géniques pris en compte : duplications et pertes
- Constitue le "pré-traitement" : tous les arbres de gènes du jeu de donnée sont réconciliés avec l'arbre d'espèce $\mathcal{S}$
  $\rightarrow$ certains sont éliminés (non binaire par exemple)
  $\rightarrow$ d'autres effeuillés (gènes d'espèces $\notin \mathcal{S}$)

- Introduction
- Modèle
- Démarche
  - → Étape 1 : Réconciliation
  - → Étape 2 : Création des classes d'adjacences
  - → Étape 3 : Algorithme DeCo
  - → Étape 4 : Synthèse
- Validation
- Utilisation de DeCo pour l'assemblage : ARt-DeCo

- On groupe ensemble les adjacences qui peuvent avoir une origine évolutive commune

- On groupe ensemble les adjacences qui peuvent avoir une origine évolutive commune

- Analogie avec les familles de gènes :
  $\rightarrow$ une classe produit une forêt d'arbres d'adjacences
  $\rightarrow$ un arbre d'adjacences contient des adjacences homologues
  $\Rightarrow$ famille d'adjacences

Les classes d'adjacences

- On groupe ensemble les adjacences qui peuvent avoir une origine évolutive commune

- Analogie avec les familles de gènes :
  $\rightarrow$ une classe produit une forêt d'arbres d'adjacences
  $\rightarrow$ un arbre d'adjacences contient des adjacences homologues
  $\Rightarrow$ famille d'adjacences

- Chaque classe est associée à exactement 2 arbres de gènes, ceux-ci peuvent-être des sous-arbres d'arbres passés en entrée

- On groupe ensemble les adjacences qui peuvent avoir une origine évolutive commune

- Analogie avec les familles de gènes :
  $\rightarrow$ une classe produit une forêt d'arbres d'adjacences
  $\rightarrow$ un arbre d'adjacences contient des adjacences homologues
  $\Rightarrow$ famille d'adjacences

- Chaque classe est associée à exactement 2 arbres de gènes, ceux-ci peuvent-être des sous-arbres d'arbres passés en entrée

- Les classes d'adjacences forment des classes d'équivalence (au sens mathématique)

Relation d'équivalence entre adjacences

Deux adjacences AB et CD sont dans la même classe d'équivalence si

- A et C sont dans le même arbre de gènes, noté $\mathcal{G}_1$, ainsi que B et D, dans un arbre noté $\mathcal{G}_2$

Relation d'équivalence entre adjacences

Deux adjacences AB et CD sont dans la même classe d'équivalence si

- A et C sont dans le même arbre de gènes, noté $\mathcal{G}_1$, ainsi que B et D, dans un arbre noté $\mathcal{G}_2$

$\rightarrow$ si A et B sont dans le même arbre de gènes ($\mathcal{G}_1 = \mathcal{G}_2$), alors $LCA(A, B) = LCA(C, D) = n_d$ (et c'est un nœud de duplication)

Relation d'équivalence entre adjacences

Deux adjacences AB et CD sont dans la même classe d'équivalence si

- A et C sont dans le même arbre de gènes, noté $\mathcal{G}_1$, ainsi que B et D, dans un arbre noté $\mathcal{G}_2$

$\rightarrow$ si A et B sont dans le même arbre de gènes ($\mathcal{G}_1 = \mathcal{G}_2$), alors $LCA(A, B) = LCA(C, D) = n_d$ (et c'est un nœud de duplication)

$\rightarrow$ sinon, il existe 2 nœuds $n_1 \in \mathcal{G}_1, n_2 \in \mathcal{G}_2$ tels que

— $S(n_1) = S(n_2)$

— A et C sont des descendants de n_1

— B et D sont des descendants de n_2

Relation d'équivalence entre adjacences

Deux adjacences AB et CD sont dans la même classe d'équivalence si

- A et C sont dans le même arbre de gènes, noté $\mathcal{G}_1$, ainsi que B et D, dans un arbre noté $\mathcal{G}_2$

$\rightarrow$ si A et B sont dans le même arbre de gènes ($\mathcal{G}_1 = \mathcal{G}_2$), alors $LCA(A, B) = LCA(C, D) = n_d$ (et c'est un nœud de duplication)

$\rightarrow$ sinon, il existe 2 nœuds $n_1 \in \mathcal{G}_1, n_2 \in \mathcal{G}_2$ tels que
- $S(n_1) = S(n_2)$
- A et C sont des descendants de n_1
- B et D sont des descendants de n_2

Les 2 arbres de gènes associés à cette classe sont soit

- les 2 sous-arbres de $\mathcal{G}_1 = \mathcal{G}_2$ enracinés aux fils de n_d
- le sous-arbre de $\mathcal{G}_1$ enraciné à n_1 et le sous-arbre de $\mathcal{G}_2$ enraciné à n_2

- Introduction
- Modèle
- Démarche
  - → Étape 1 : Réconciliation
  - → Étape 2 : Création des classes d'adjacences
  - → Étape 3 : Algorithme DeCo
  - → Étape 4 : Synthèse
- Validation
- Utilisation de DeCo pour l'assemblage : ARt-DeCo

- À l'issue de tous les appels à `DeCo` on récupère plusieurs informations :

  → Les adjacences ancestrales par espèces

  → Les gènes qui se sont dupliqués ou perdus ensemble

  → Les arbres d'adjacences

  → Les degrés des gènes ancestraux (nombre de voisins)

  → Taille des arbres d'adjacences, nb d'arbres /classes, . . .

- Introduction
- Modèle
- Démarche
- Validation
- Utilisation de DeCo pour l'assemblage : ARt-DeCo

- Deux jeux de données principaux : mammifères et plantes

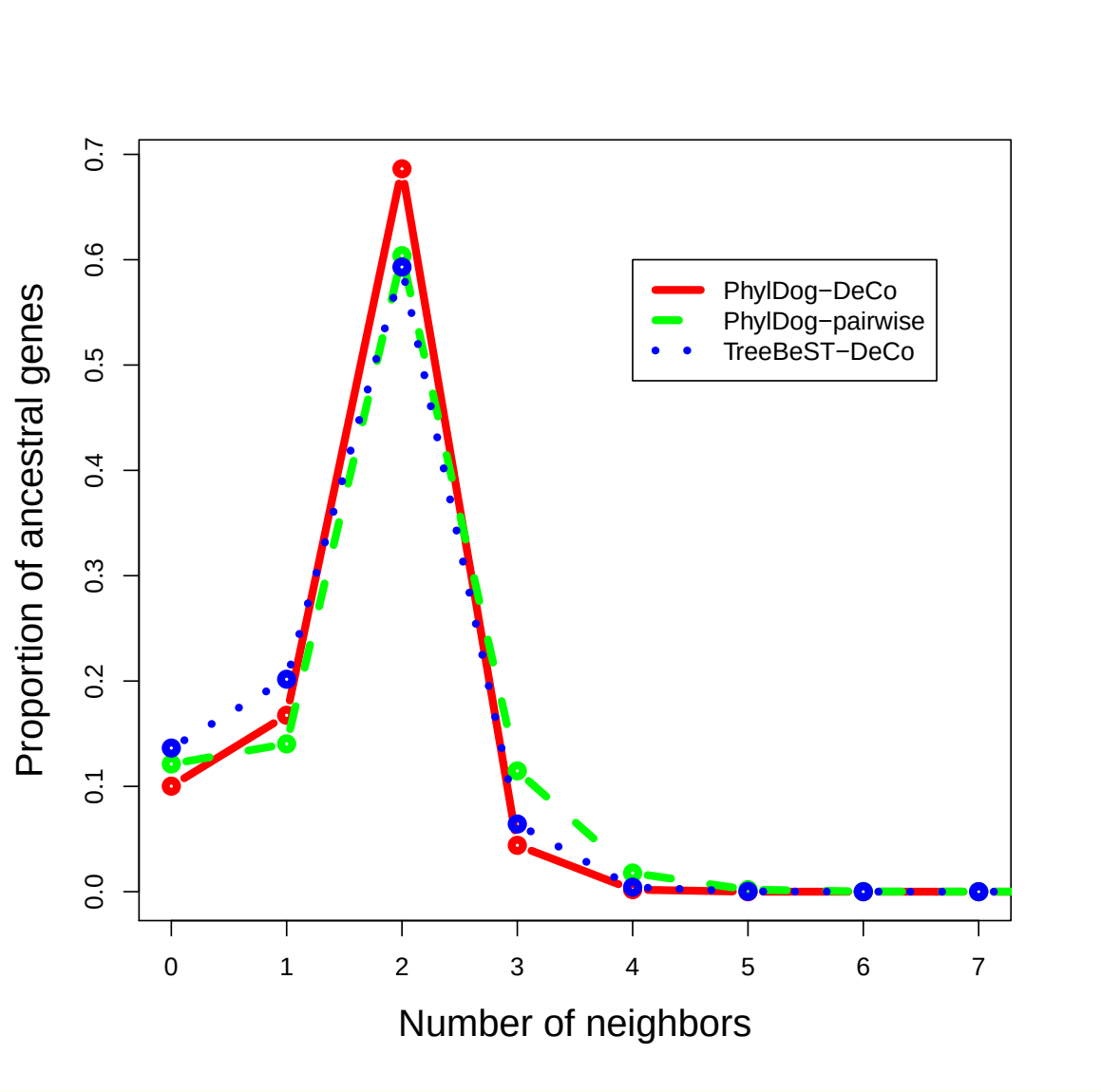
  → Mammifères : 5039 arbres de gènes (Ensembl 57)
  11 espèces dont les génomes étaient assemblés,
  $\sim 107\,000$ gènes

  → Plantes (angiospermes) : 35 182 arbres (EnsemblPlant 12)
  9 espèces dont les génomes étaient assemblés

- Gros jeux de données plantes $\sim 50\,000$ arbres, $\sim 615\,000$ gènes

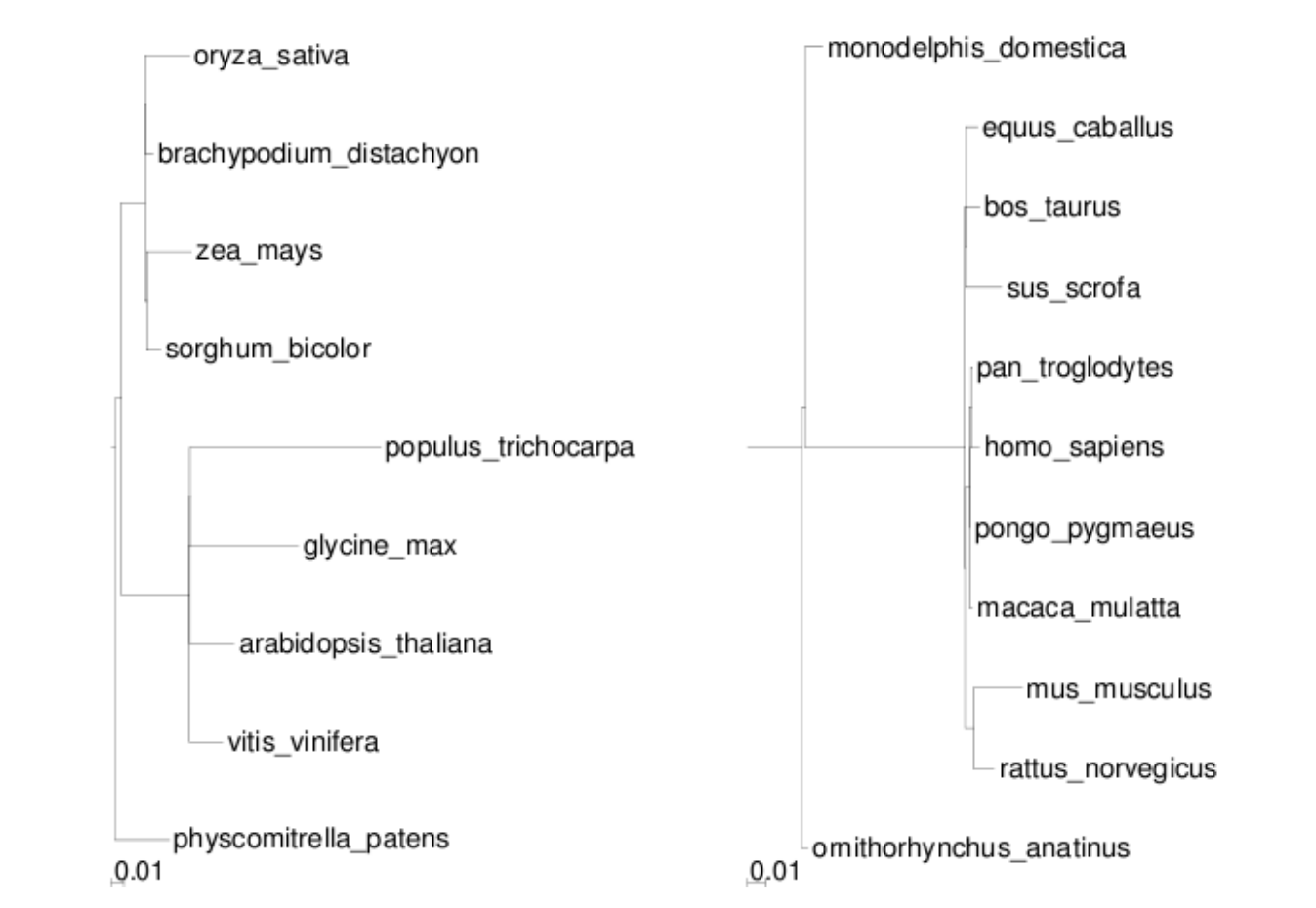
- Temps de calcul moyen d'une dizaine de minutes (réconciliation comprise)

Linéarité : nombres de voisins



Pour avoir de meilleurs génomes ancestraux il faut à la fois de bons arbres et de bonnes méthodes d'inférences d'adjacences

Phylogénies en fonction du nb de dupli en bloc



- Clades des angiospermes et des mammifères : temps de divergence et nb d'espèces assemblées similaires
- Longueur de branches = # $ADup$ / # nb gènes ancestraux sur cette branche
- En moyenne les lg de branches sont 3 fois plus longues chez les plantes

Première méthode pour reconstruire l'évolution de relations entres gènes

Première méthode pour reconstruire l'évolution de relations entres gènes

- Modèle minimaliste mais algorithme exact et rapide

Première méthode pour reconstruire l'évolution de relations entres gènes

- Modèle minimaliste mais algorithme exact et rapide
- Appliquée ici à la relation d'adjacence $\Rightarrow$ donne accès à de l'information sur les génomes ancestraux

Première méthode pour reconstruire l'évolution de relations entres gènes

- Modèle minimaliste mais algorithme exact et rapide
- Appliquée ici à la relation d'adjacence $\Rightarrow$ donne accès à de l'information sur les génomes ancestraux
- Peut facilement être étendu à d'autres type de relations : interactions protéines-protéines, présence dans une même voie métabolique, co-expression, ...

- DeCoLT Lateral gene transfer, rearrangement, reconciliation *Murray Patterson*, *Gergely Szöllősi*, *Vincent Daubin* et *Eric Tannier*. BMC Bioinformatics 2013, 14(Suppl 15):S4

- DeCoLT Lateral gene transfer, rearrangement, reconciliation *Murray Patterson*, *Gergely Szöllősi*, *Vincent Daubin* et *Eric Tannier*. BMC Bioinformatics 2013, 14(Suppl 15):S4

- Thèse Magali Semeria (LBBE - Lyon) $sept\,2012 \rightarrow dec\,2015$
  Modèle probabiliste pour l'évolution des relations entre gènes
  Encadrement : *Laurent Guéguen* et *Éric Tannier*

- DeCoLT Lateral gene transfer, rearrangement, reconciliation *Murray Patterson*, *Gergely Szöllősi*, *Vincent Daubin* et *Eric Tannier*. BMC Bioinformatics 2013, 14(Suppl 15):S4

- Thèse Magali Semeria (LBBE - Lyon) $sept\ 2012 \rightarrow dec\ 2015$
  Modèle probabiliste pour l'évolution des relations entre gènes
  Encadrement : *Laurent Guéguen* et *Éric Tannier*

- Thèse Yoann Anselmetti (ISE-M - Montpellier) $oct\ 2014 \rightarrow 17$
  Assemblage et évolution de la structure de génomes anciens et actuels
  Encadrement : *Sèverine Bérard* et *Éric Tannier*

- DeCoLT Lateral gene transfer, rearrangement, reconciliation *Murray Patterson*, *Gergely Szöllősi*, *Vincent Daubin* et *Eric Tannier*. BMC Bioinformatics 2013, 14(Suppl 15):S4

- Thèse Magali Semeria (LBBE - Lyon) $sept\ 2012 \rightarrow dec\ 2015$
  Modèle probabiliste pour l'évolution des relations entre gènes
  Encadrement : *Laurent Guéguen* et *Éric Tannier*

- Thèse Yoann Anselmetti (ISE-M - Montpellier) $oct\ 2014 \rightarrow 17$
  Assemblage et évolution de la structure de génomes anciens et actuels
  Encadrement : *Sèverine Bérard* et *Éric Tannier*

  $\rightarrow$ Première idée : Prise en compte de l'état d'assemblage des génomes pour moins pénaliser les cassures d'adjacences aux feuilles des génomes mal/peu assemblés

- `DeCoLT` Lateral gene transfer, rearrangement, reconciliation *Murray Patterson*, *Gergely Szöllősi*, *Vincent Daubin* et *Eric Tannier*. BMC Bioinformatics 2013, 14(Suppl 15):S4

- Thèse Magali Semeria (LBBE - Lyon) $sept\ 2012 \rightarrow dec\ 2015$
  Modèle probabiliste pour l'évolution des relations entre gènes
  Encadrement : *Laurent Guéguen* et *Éric Tannier*

- Thèse Yoann Anselmetti (ISE-M - Montpellier) $oct\ 2014 \rightarrow 17$
  Assemblage et évolution de la structure de génomes anciens et actuels
  Encadrement : *Sèverine Bérard* et *Éric Tannier*

  $\rightarrow$ Première idée : Prise en compte de l'état d'assemblage des génomes pour moins pénaliser les cassures d'adjacences aux feuilles des génomes mal/peu assemblés

  $\Rightarrow$ `ARt-DeCo`

- Introduction
- Modèle
- Démarche
- Validation
- Utilisation de `DeCo` pour l'assemblage : `ARt-DeCo`

Des génomes très fragmentés

Caenorhabditis_elegans 7
Drosophila_melanogaster 11
Monodelphis_domestica 11
Saccharomyces_cerevisiae 17
Gorilla_gorilla 27
Tetraodon_nigroviridis 27
Homo_sapiens 39
Mus_musculus 40
Bos_taurus 43
Pongo_abelii 53
Rattus_norvegicus 69
Taeniopygia_guttata 69
Equus_caballus 102
Chlorocebus_sabaeus 115
Felis_catus 213
Canis_familiaris 227
Pan_troglodytes 245
Ciona_savignyi 332
Papio_anubis 367
Cavia_porcellus 369
Danio_rerio 448
Lepisosteus_oculatus 494
Otolemur_garnettii 523
Gasterosteus_aculeatus 550
Loxodonta_africana 583
Ciona_intestinalis 614
Meleagris_gallopavo 631
Ovis_aries 638
Callithrix_jacchus 653
Nomascus_leucogenys 657
Macaca_mulatta 692
Mustela_putorius_furo 731
Ficedula_albicollis 757
Ictidomys_tridecemlineatus 761
Gallus_gallus 766
Oryzias_latipes 807
Oryctolagus_cuniculus 1043
Oreochromis_niloticus 1075
Sus_scrofa 1332
Poecilia_formosa 1504
Takifugu_rubripes 1923
Pelodiscus_sinensis 2162
Xenopus_tropicalis 2232
Anas_platyrhynchos 2338
Anolis_carolinensis 2376
Myotis_lucifugus 2393
Xiphophorus_maculatus 2445
Astyanax_mexicanus 2557
Ailuropoda_melanoleuca 2599
Sarcophilus_harrisii 3505
Latimeria_chalumnae 3948
Dasypus_novemcinctus 3958
Vicugna_pacos 4951
Gadus_morhua 5336
Petromyzon_marinus 5530
Tursiops_truncatus 5624
Pteropus_vampyrus 6328
Microcebus_murinus 6837
Ochotona_princeps 7324
Tupaia_belangeri 8121
Dipodomys_ordii 9720
Sorex_araneus 9740
Ornithorhynchus_anatinus 10158
Procavia_capensis 10465
Choloepus_hoffmanni 10770
Tarsius_syrichta 11528
Erinaceus_europaeus 11536
Echinops_telfairi 12301
Macropus_eugenii 12704

Une solution

- S'aider des informations phylogénétiques et des adjacences de gènes pour faire le "scaffolding" des espèces actuelles et ancestrales en même temps

Une solution

- S'aider des informations phylogénétiques et des adjacences de gènes pour faire le "scaffolding" des espèces actuelles et ancestrales en même temps

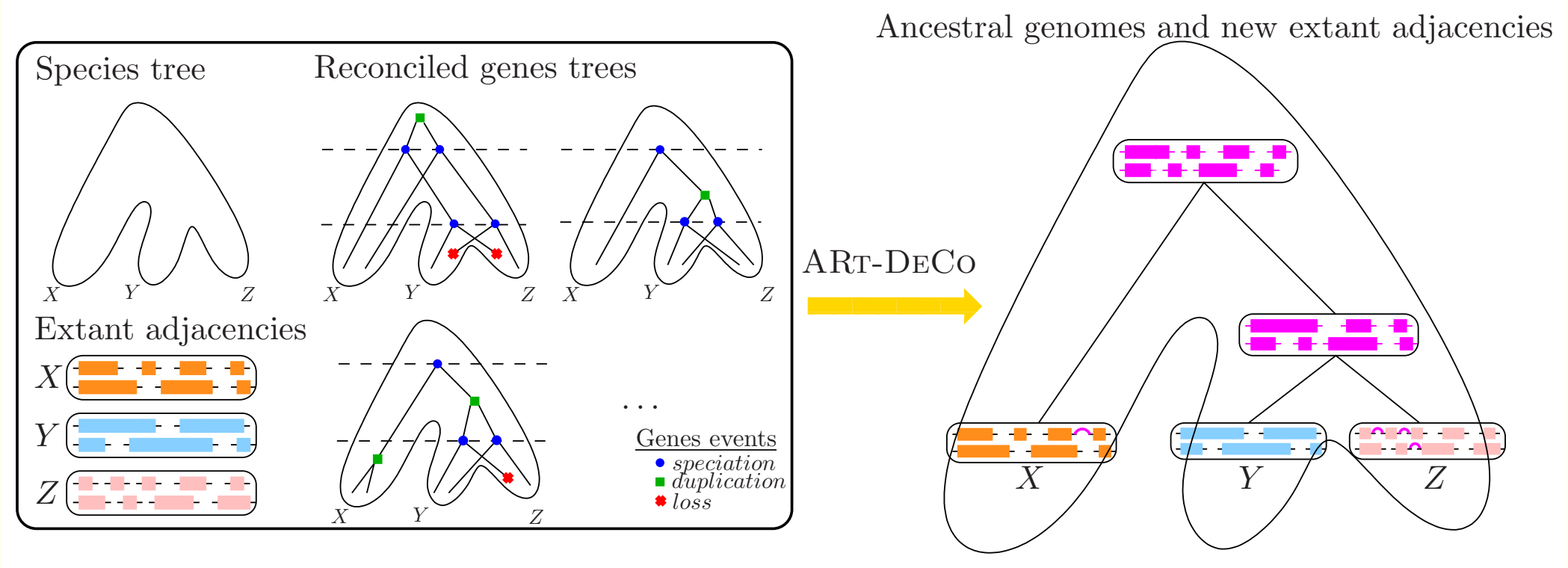
- C'est l'idée d'`ARt-DeCo` (`Assembly Recovery through DeCo`)

Une solution

- S'aider des informations phylogénétiques et des adjacences de gènes pour faire le "scaffolding" des espèces actuelles et ancestrales en même temps

- C'est l'idée d'`ARt-DeCo` (`Assembly Recovery through DeCo`)

- Travaux antérieurs apparentés :
  - → treecat (Husemann & Stoye, Algorithm for Molecular Biology, 2010)
  - → RACA (Kim *et al.*, PNAS, 2013)
  - → Aganezov *et al.*, Computational Biology and Chemistry, 2015


Species tree
Reconciled genes trees
X
Y
Z
Extant adjacencies
X
Y
Z
...
Genes events
speciation
duplication
loss
ARt-DeCo
Ancestral genomes and new extant adjacencies
X
Y
Z

- Fragmentation simulée sur 7 tétrapodes, sans duplications
  - $\rightarrow$ Famille de gènes universels uni-copie
  - $\rightarrow$ On retrouve $84 - 93$ % des cassures simulées
  - $\rightarrow$ Faible taux de faux positifs (~ 1.5 %)

- Fragmentation simulée sur 7 tétrapodes, sans duplications
  - $\rightarrow$ Famille de gènes universels uni-copie
  - $\rightarrow$ On retrouve $84 - 93$ % des cassures simulées
  - $\rightarrow$ Faible taux de faux positifs (~ 1.5 %)
- Réduction de la fragmentation (69 eucaryotes)

- Fragmentation simulée sur 7 tétrapodes, sans duplications
  - → Famille de gènes universels uni-copie
  - → On retrouve $84 - 93\ \%$ des cassures simulées
  - → Faible taux de faux positifs ($\sim 1.5\ \%$)

- Réduction de la fragmentation (69 eucaryotes)
  - → Chez les génomes actuels :

    Macropus_eugenii 12,704 contigs → 3,877 new adj

    Microcebus_murinus 6,837 contigs → 1,945 new adj

    Ailuropoda_melanoleuca 2,599 contigs → 747 new adj

    Loxodonta_africana 583 contigs → 23 new adj

    Ciona_savignyi 332 contigs → 0 new adj

    Mus_musculus 40 contigs → 1 new adj

- Fragmentation simulée sur 7 tétrapodes, sans duplications
  - → Famille de gènes universels uni-copie
  - → On retrouve $84 - 93$ % des cassures simulées
  - → Faible taux de faux positifs (~ 1.5 %)

- Réduction de la fragmentation (69 eucaryotes)
  - → Chez les génomes actuels :

    Macropus_eugenii 12,704 contigs → 3,877 new adj

    Microcebus_murinus 6,837 contigs → 1,945 new adj

    Ailuropoda_melanoleuca 2,599 contigs → 747 new adj

    Loxodonta_africana 583 contigs → 23 new adj

    Ciona_savignyi 332 contigs → 0 new adj

    Mus_musculus 40 contigs → 1 new adj
  - → Analyse de potentielles nouvelles adjacences

- Fonction objectif : minimiser #(▲,⚡)

$\rightarrow$ $c_0(g_1, g_2)$: Minimum cost where genes g_1 and g_2 are not adjacents

$\rightarrow$ $c_1(g_1, g_2)$: Minimum cost where genes g_1 and g_2 are adjacents

- Fonction objectif : minimiser #(▲,⚡)

→ $c_0(g_1, g_2)$: Minimum cost where genes g_1 and g_2 are not adjacents

→ $c_1(g_1, g_2)$: Minimum cost where genes g_1 and g_2 are adjacents

- Équations de programmation dynamique

Cas 1 $E(g_1) =$ Extant and $E(g_2) =$ Extant

Cas 2 $E(g_1) =$ ✚ and $E(g_2) \neq$ ✚

Cas 3 $E(g_1) =$ ✚ and $E(g_2) =$ ✚

Cas 4 $E(g_1) =$ {Extant, ●} and $E(g_2) =$ ■

Cas 5 $E(g_1) =$ ● and $E(g_2) =$ ●

Cas 6 $E(g_1) =$ ■ and $E(g_2) =$ ■

Evolutionary events :

- ● Speciation
- ■ Gene duplication
- ✚ Gene loss
- ⚡ Adjacency break
- ▲ Adjacency gain

- Modification **Cas 1**

- Modification **Cas 1**

$\rightarrow$ DeCo : $\begin{cases} \text{si } g_1 \sim g_2 & c_1(g_1, g_2) = 0 \quad \text{et } c_0(g_1, g_2) = \infty \\ \text{sinon} & c_0(g_1, g_2) = 0 \quad \text{et } c_1(g_1, g_2) = \infty \end{cases}$

- Modification **Cas 1**

$\rightarrow$ DeCo : $\begin{cases} \text{si } g_1 \sim g_2 & c_1(g_1, g_2) = 0 \quad \text{et } c_0(g_1, g_2) = \infty \\ \text{sinon} & c_0(g_1, g_2) = 0 \quad \text{et } c_1(g_1, g_2) = \infty \end{cases}$

$\rightarrow$ ARt-DeCo : $\begin{cases} c_1(g_1, g_2) = -log_b(P(g_1 \sim g_2)) \\ c_0(g_1, g_2) = -log_b(1 - P(g_1 \sim g_2)) \end{cases}$

où $P(g_1 \sim g_2)$ est la probabilité que $g_1 \sim g_2$

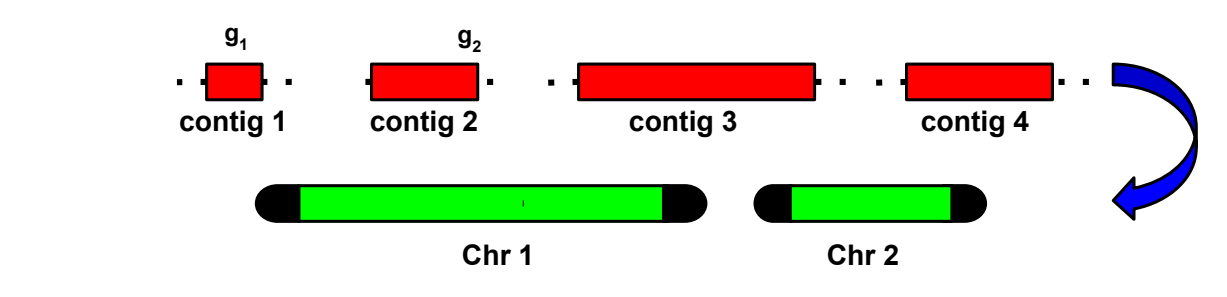
- Modification **Cas 1**

→ DeCo : $\begin{cases} \text{si } g_1 \sim g_2 & c_1(g_1, g_2) = 0 \quad \text{et } c_0(g_1, g_2) = \infty \\ \text{sinon} & c_0(g_1, g_2) = 0 \quad \text{et } c_1(g_1, g_2) = \infty \end{cases}$

→ ARt-DeCo : $\begin{cases} c_1(g_1, g_2) = -log_b(P(g_1 \sim g_2)) \\ c_0(g_1, g_2) = -log_b(1 - P(g_1 \sim g_2)) \end{cases}$

où $P(g_1 \sim g_2)$ est la probabilité que $g_1 \sim g_2$

- Calcul de $P(g_1 \sim g_2)$



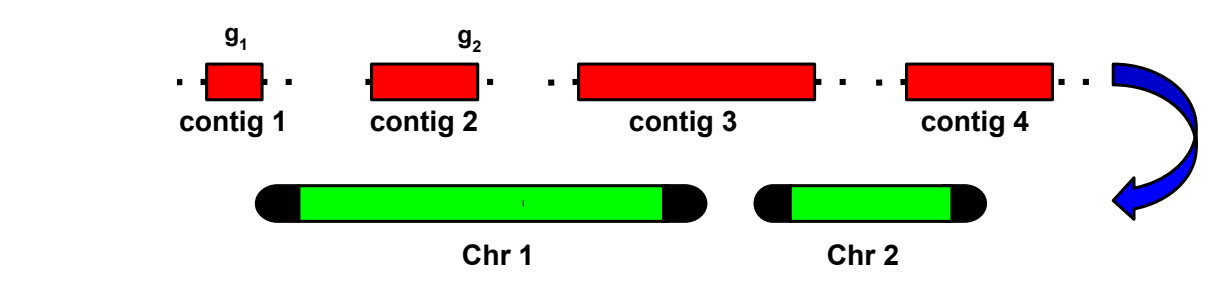
- Modification **Cas 1**

$\rightarrow$ DeCo : $\begin{cases} \text{si } g_1 \sim g_2 & c_1(g_1, g_2) = 0 \quad \text{et } c_0(g_1, g_2) = \infty \\ \text{sinon} & c_0(g_1, g_2) = 0 \quad \text{et } c_1(g_1, g_2) = \infty \end{cases}$

$\rightarrow$ ARt-DeCo : $\begin{cases} c_1(g_1, g_2) = -log_b(P(g_1 \sim g_2)) \\ c_0(g_1, g_2) = -log_b(1 - P(g_1 \sim g_2)) \end{cases}$

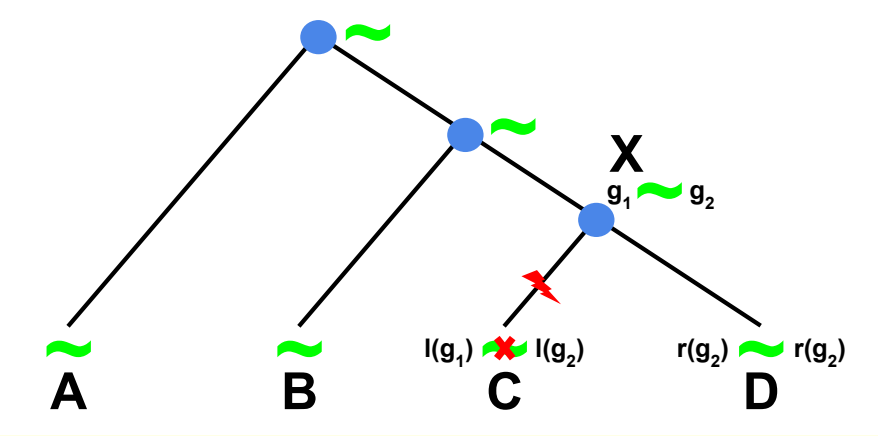
où $P(g_1 \sim g_2)$ est la probabilité que $g_1 \sim g_2$

- Calcul de $P(g_1 \sim g_2)$



$$P(g_1 \sim g_2) = \frac{\text{\# of complete scaffolding where } g_1 \sim g_2}{\text{Total \# of complete scaffolding}} = \frac{24}{144} = \frac{1}{6}$$

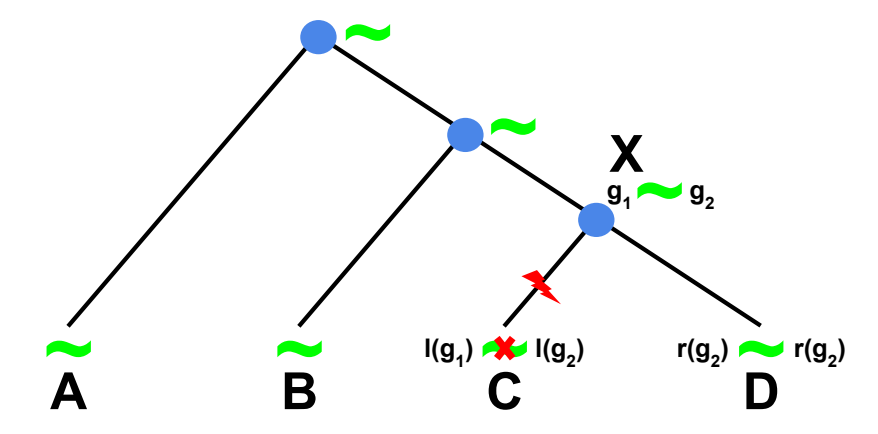
Choix de la base b du logarithme



$$Br = C(\text{⚡})$$

$$p = P(l(g_1) \sim l(g_2))$$

Choix de la base b du logarithme

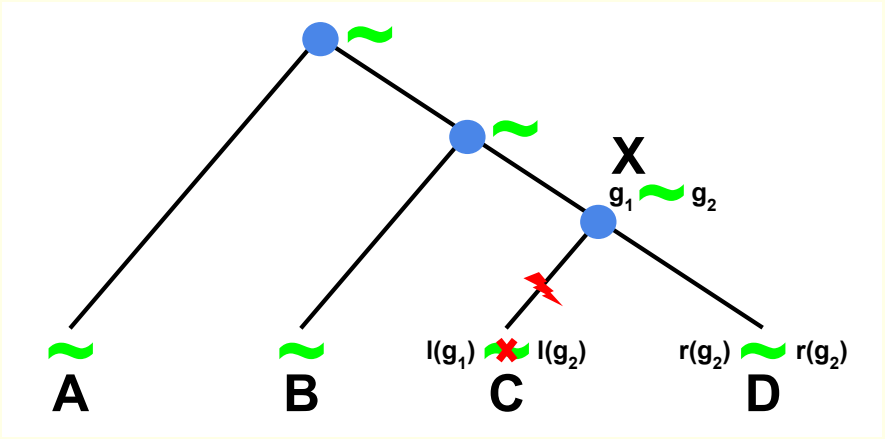


$$Br = C(\lightning)$$
$$p = P(l(g_1) \sim l(g_2))$$

Pour privilégier la prédiction d'une adjacence dans l'espèce **C** :

$$c_1(g_1, g_2) = min \begin{cases} (1) & c_1(l(g_1), l(g_2)) + c_1(r(g_1), r(g_2)) \\ (2) & c_1(l(g_1), l(g_2)) + c_0(r(g_1), r(g_2)) + C(\lightning) \\ (3) & c_0(l(g_1), l(g_2)) + c_1(r(g_1), r(g_2)) + C(\lightning) \\ (4) & c_0(l(g_1), l(g_2)) + c_0(r(g_1), r(g_2)) + 2 \times C(\lightning) \end{cases}$$

Choix de la base b du logarithme



$$Br = C(\lightning)$$
$$p = P(l(g_1) \sim l(g_2))$$

Pour privilégier la prédiction d'une adjacence dans l'espèce **C** :

$$c_1(g_1, g_2) = min \begin{cases} (1) & c_1(l(g_1), l(g_2)) + c_1(r(g_1), r(g_2)) \\ (2) & c_1(l(g_1), l(g_2)) + c_0(r(g_1), r(g_2)) + C(\lightning) \\ (3) & c_0(l(g_1), l(g_2)) + c_1(r(g_1), r(g_2)) + C(\lightning) \\ (4) & c_0(l(g_1), l(g_2)) + c_0(r(g_1), r(g_2)) + 2 \times C(\lightning) \end{cases}$$

$$(1) < (3) \quad \Rightarrow \quad b > \left(\frac{1-p}{p}\right)^{\frac{1}{Br}} \quad \Rightarrow \quad b = \left\lceil 1.05 * \left(\frac{1-p}{p}\right)^{\frac{1}{Br}} \right\rceil$$

Résultats : jeu de donnée des 7 tétrapodes

- Jeu de donnée proche de celui d'Aganezov *et al.*, 2015
  - → 7 tetrapods : Human, Chimpanzee, Macaque, Mouse, Rat, Dog and Chicken
  - → 8,818 universal unicopy gene families
  - → Run in $< 1\,min$ on computer desk

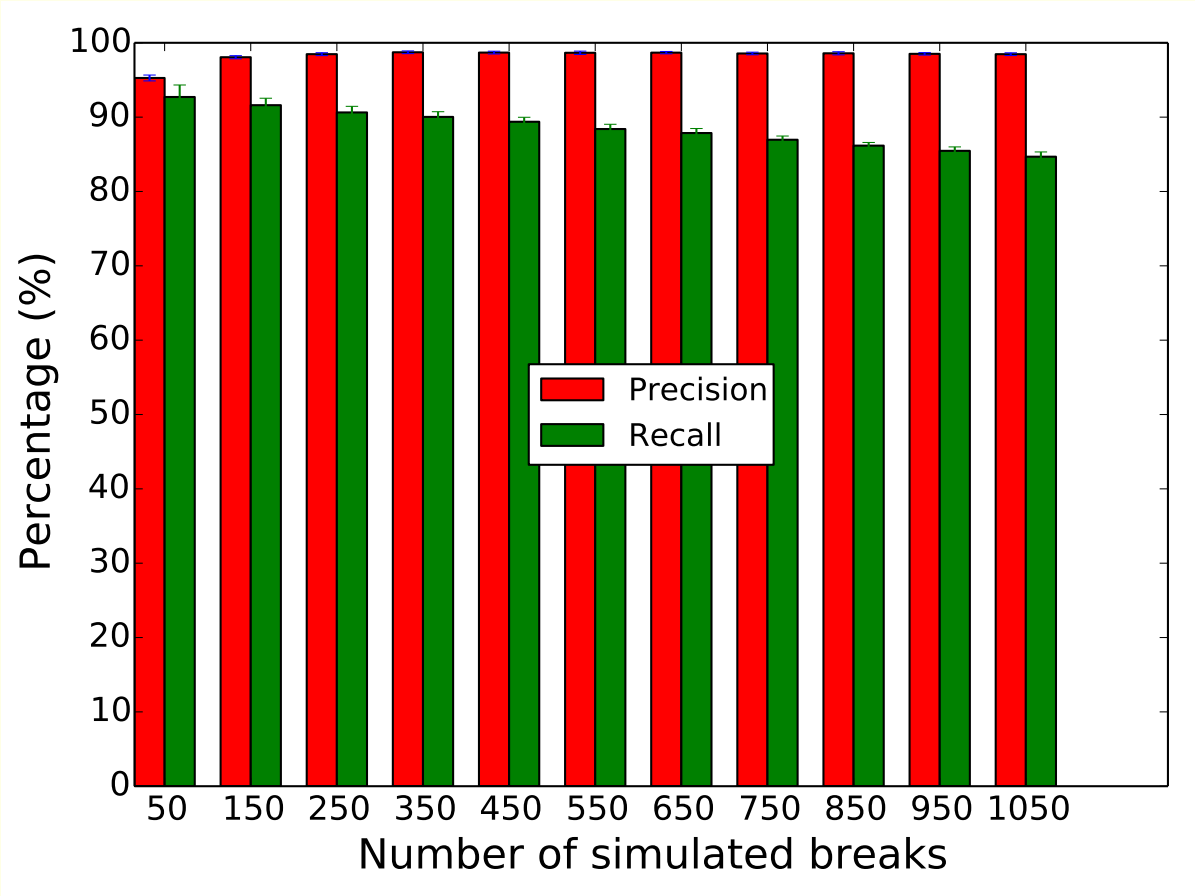
- Jeu de donnée proche de celui d'Aganezov *et al.*, 2015
  - → 7 tetrapods : Human, Chimpanzee, Macaque, Mouse, Rat, Dog and Chicken
  - → 8,818 universal unicopy gene families
  - → Run in $< 1\,min$ on computer desk

- Experiment
  - → Simulate breaks in each species (except on the outgroup species : Chicken)
  - → Different number of breaks from 50 to 1050, by step of 100

Résultats : jeu de donnée des 7 tétrapodes

$$Precision = \frac{TP}{TP+FP} = \frac{\text{\# of true inferred adj}}{\text{\# of inferred adj}}$$

$$Recall = \frac{TP}{TP+FN} = \frac{\text{\# of true inferred adj}}{\text{\# of simulated adj breaks}}$$



Résultats : jeu de donnée des 69 eucaryotes

- Data set on Ensembl 79
  - $\rightarrow$ 69 eucaryotes
  - $\rightarrow$ $20,279$ protein coding gene trees
  - $\rightarrow$ $1,222,543$ protein coding genes
  - $\rightarrow$ Run in $\sim$ $18\,h$ on computer desk

Résultats : jeu de donnée des 69 eucaryotes

- Data set on Ensembl 79
  - → 69 eucaryotes
  - → $20,279$ protein coding gene trees
  - → $1,222,543$ protein coding genes
  - → Run in $\sim$ $18\,h$ on computer desk

- Experiment
  - → Run `ARt-DeCo` to predict new adjacencies in extant species
  - → Analyze percentage of fragmentation reduction
  - → Case study of potential new extant adjacencies

Reducing genome fragmentation on 69 eukaryotes

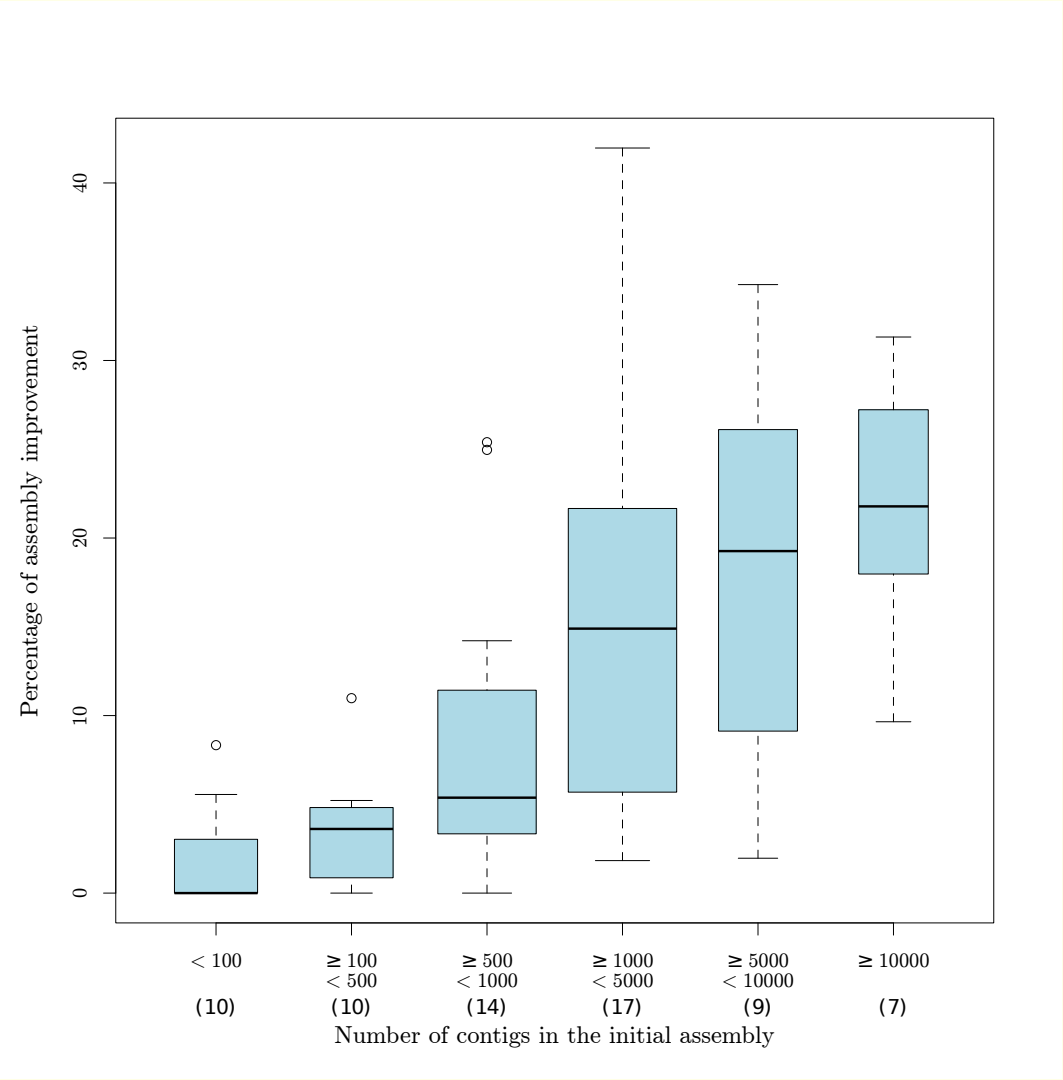
$C_I =$ Initial nb of contigs

$C_N =$ Final nb of scaffolds

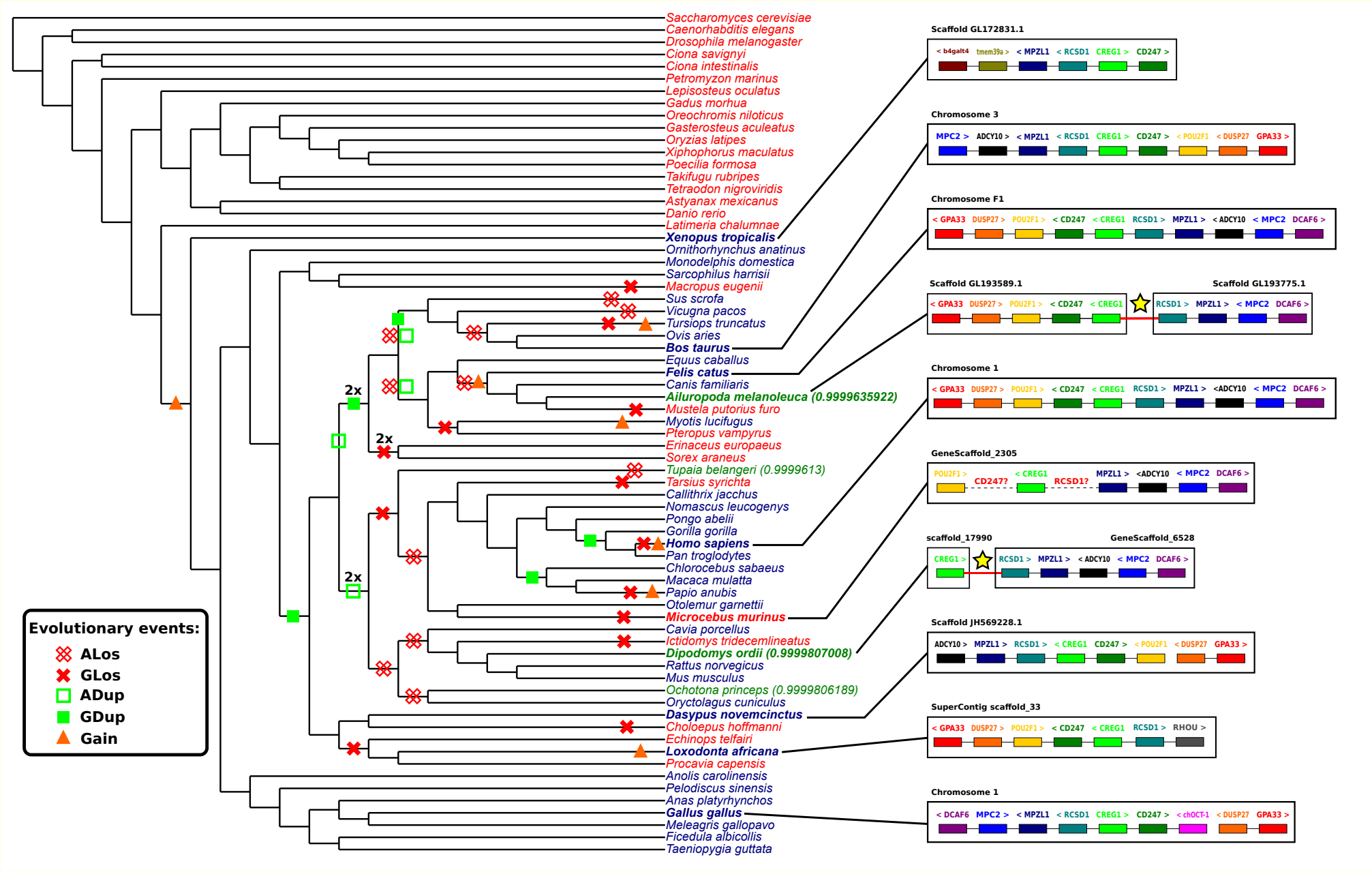
$p =$ Expected nb of chromosomes

% of assembly improvement :

$$\left(\frac{C_I - C_N}{C_I - p}\right) * 100$$



Revealing potential new extant adjacencies



- Conclusion
  - → Joint phylogenetic scaffolding of extant and ancestral genomes
  - → Polynomial algorithm $O(n^2)$ per adjacency class
  - → Gene order is still not completely resolved
- Perspectives
  - → Joint sequence-based phylogenetic scaffolding of extant and ancestral species)

DeCo →

ARt-DeCo →

- Des parties de cette présentation sont issues des rapports de stage de Coralie Gallien et Pierre-Antoine Jean et de la présentation à RECOMB'CG 15 de Yoann Anselmetti
- Image réconciliation issue de http://openi.nlm.nih.gov/

Institut des Sciences de l'Evolution-Montpellier