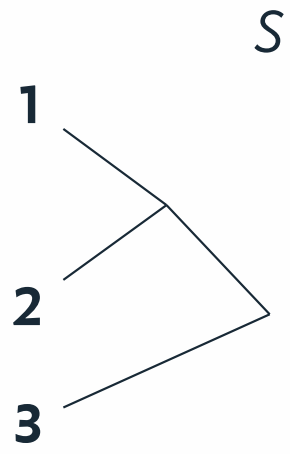
RECONSTRUCTION DE L'HISTOIRE ÉVOLUTIVE DE FAMILLES DE GÈNES EN SYNTÉNIE

Méthodes algorithmiques pour la réconciliation segmentale

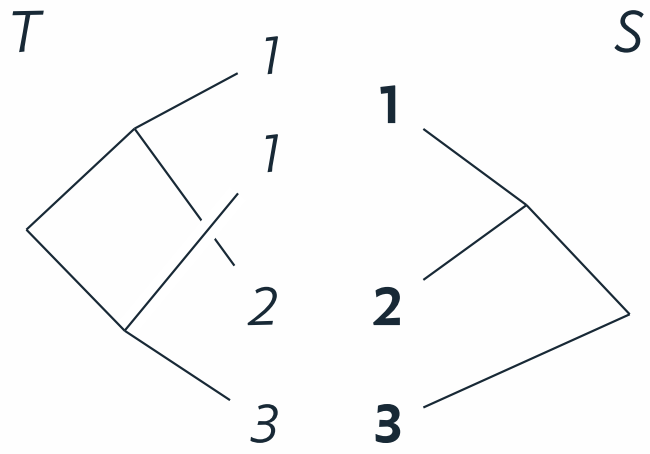
Mattéo Delabre

Laboratoire de biologie informatique et théorique
Université de Montréal
30 août 2022

PHYLOGÉNIE

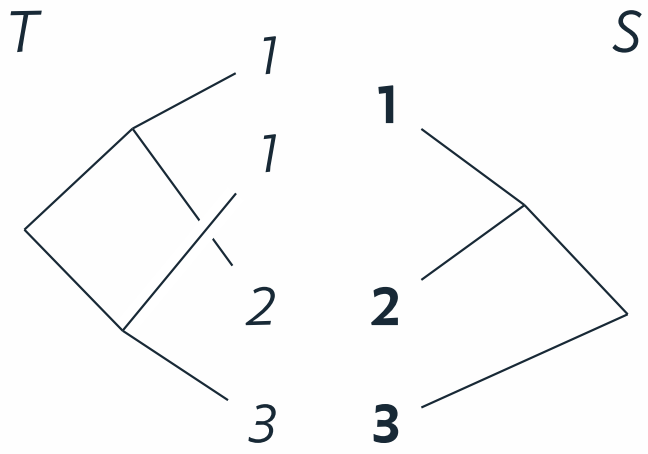


HIÉRARCHIE D'ÉVOLUTION

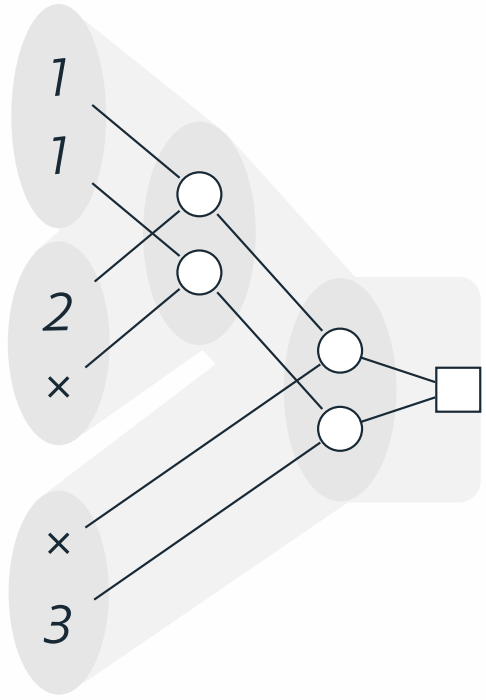


Dépendant	–	**Hôte**
Espèces	–	Habitats
Parasites	–	Hôtes
Gènes	–	Espèces
Domaines	–	Protéines

RÉCONCILIATION

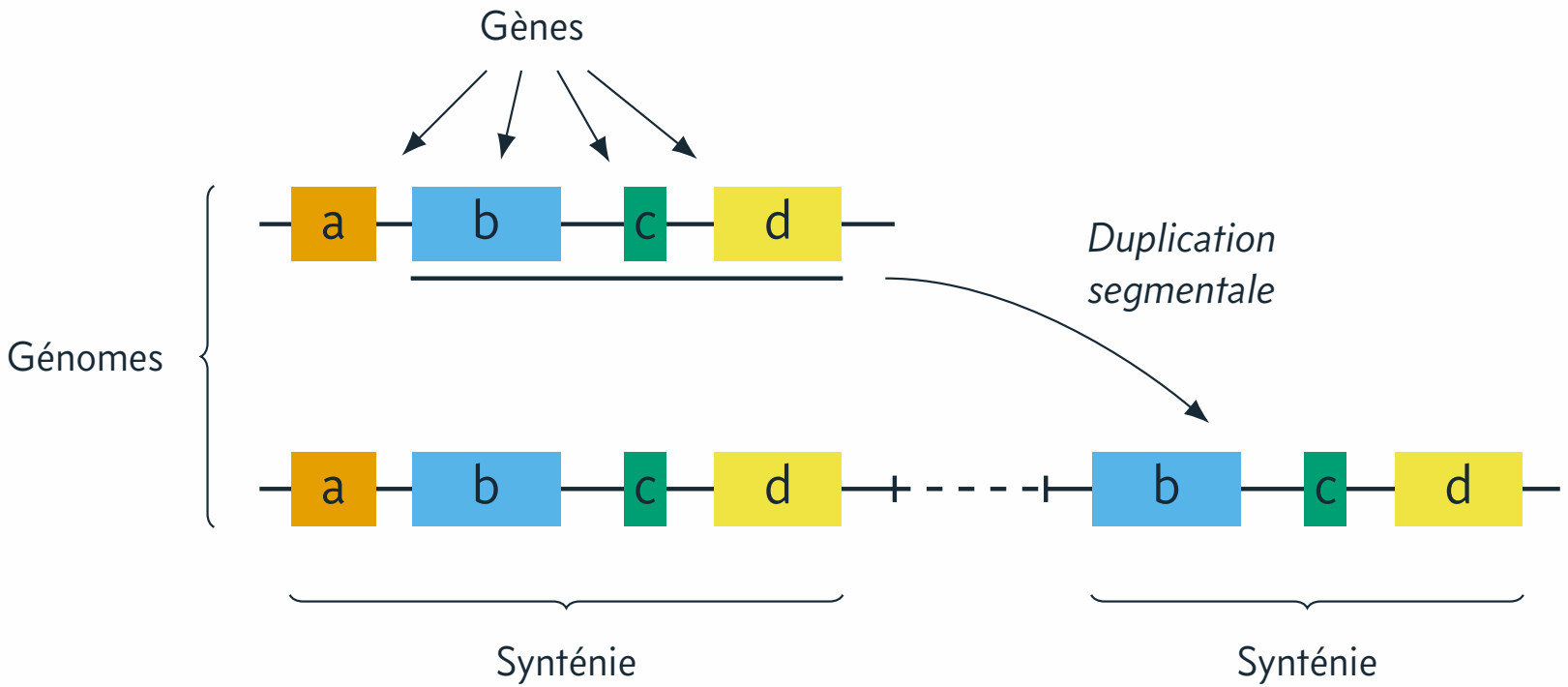


Dépendant	–	**Hôte**
Espèces	–	Habitats
Parasites	–	Hôtes
Gènes	–	Espèces
Domaines	–	Protéines

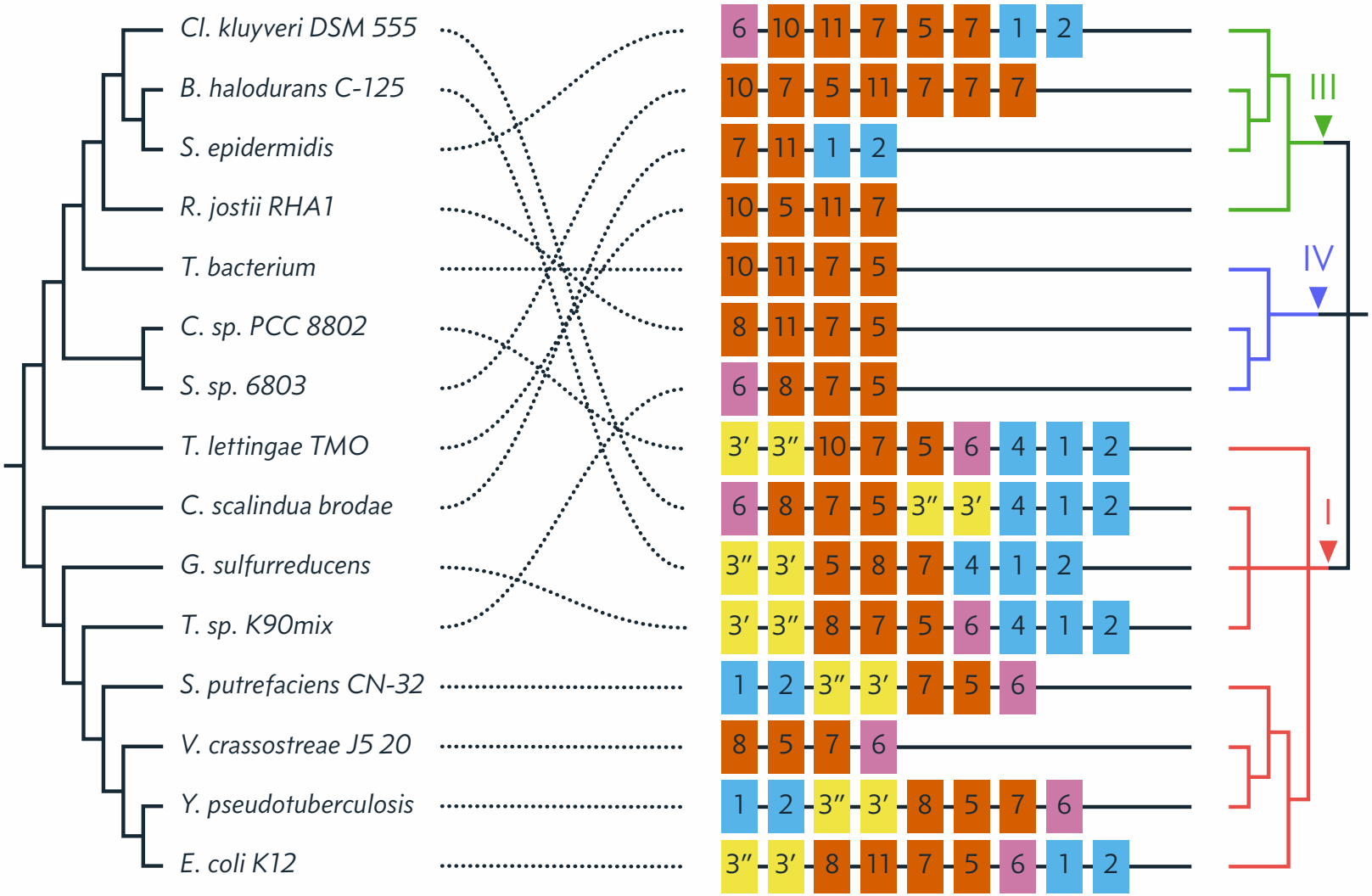


- **Réconciliation :** association de deux phylogénies expliquant leurs différences par des événements

ÉVOLUTION SEGMENTALE

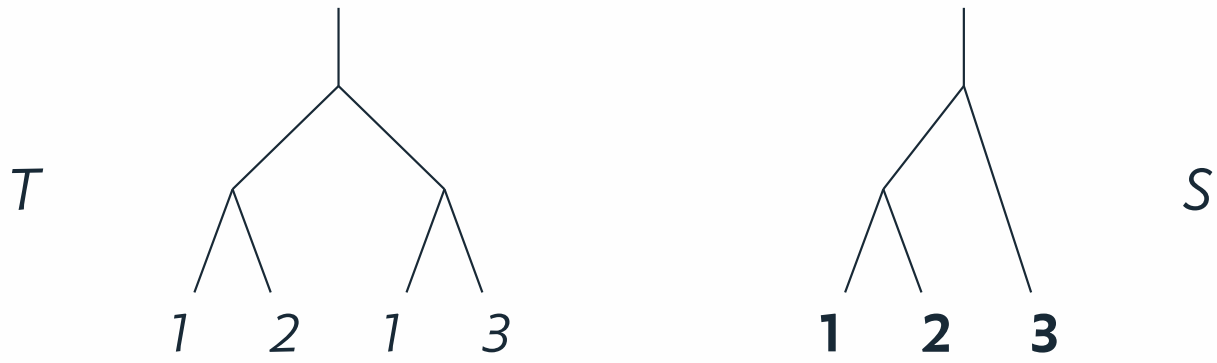


SYSTÈMES CRISPR-CAS



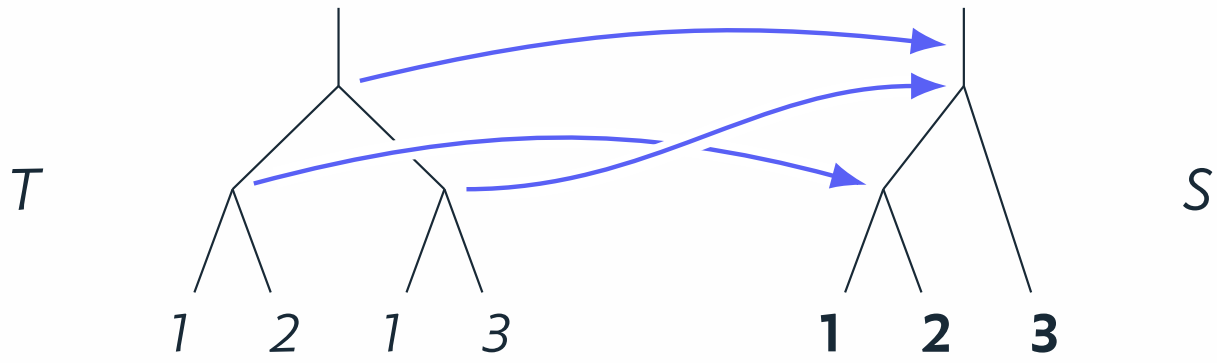
PLAN

PLONGEMENT



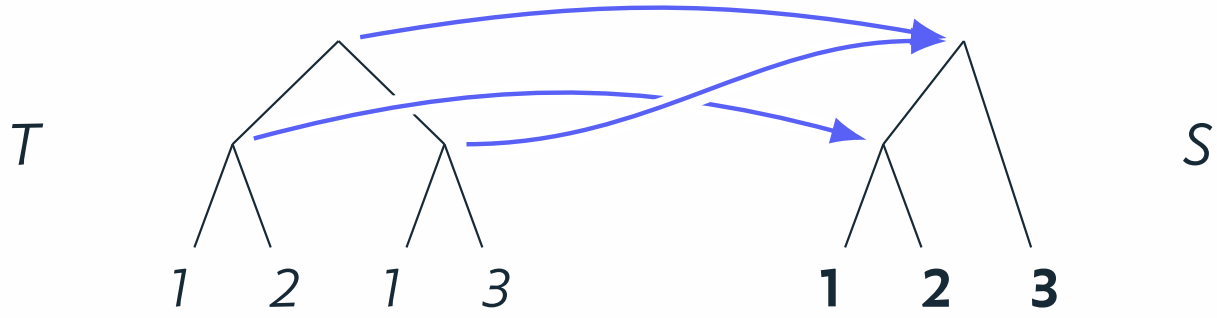
- Chaque embranchement de *T* est **associé** à un point dans *S*
- Sur un embranchement de *S* → *codivergence*
- Sur une branche de *S* → *événement indépendant*

PLONGEMENT



- Chaque embranchement de *T* est **associé** à un point dans *S*
- Sur un embranchement de *S* → *codivergence*
- Sur une branche de *S* → *événement indépendant*

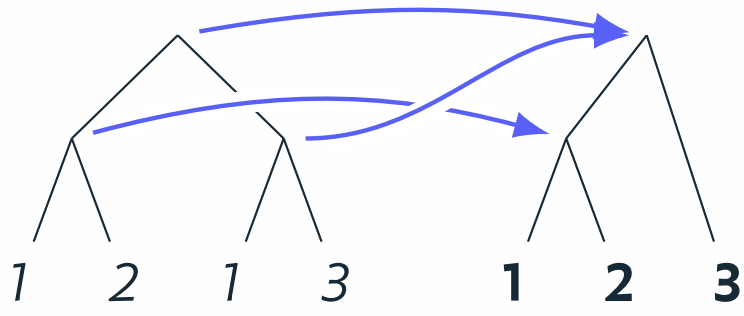
PLONGEMENT



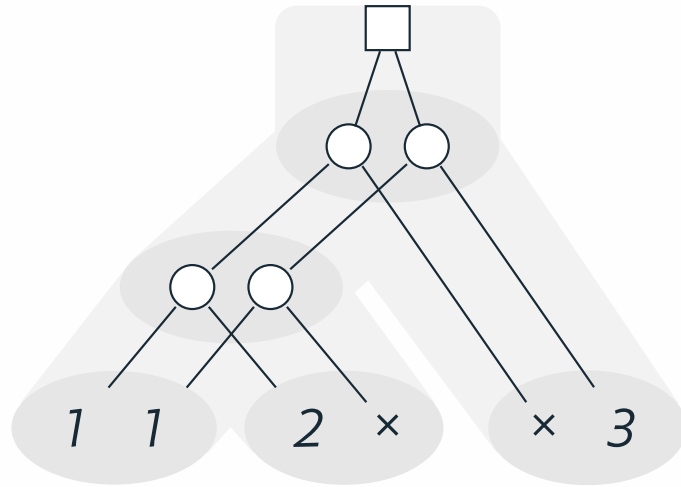
- Chaque embranchement de T est **associé** à un point dans S
- Sur un embranchement de S → *codivergence*
- Sur une branche de S → *événement indépendant*

MODÈLE DUPLICATION–PERTE

- **Modèle :** Types d'événements pouvant survenir



T S



Arbre réconcilié $\mathcal{R}$

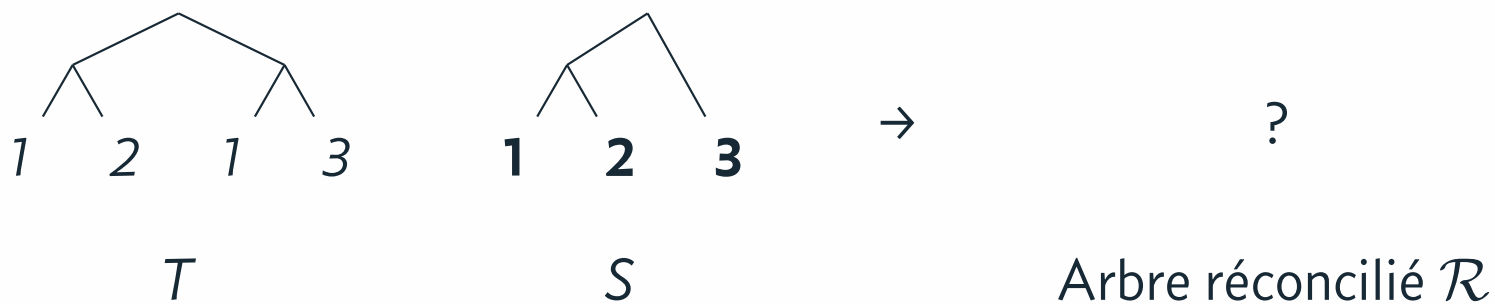
- **DL :** Spéciation ○ Duplication □ Perte ×

PARCIMONIE

- Attribution d'un **coût** δ pour chaque type d'événement
- Coût d'une réconciliation : somme des coûts des événements

Problème $\mathcal{M}$-Parcimonie

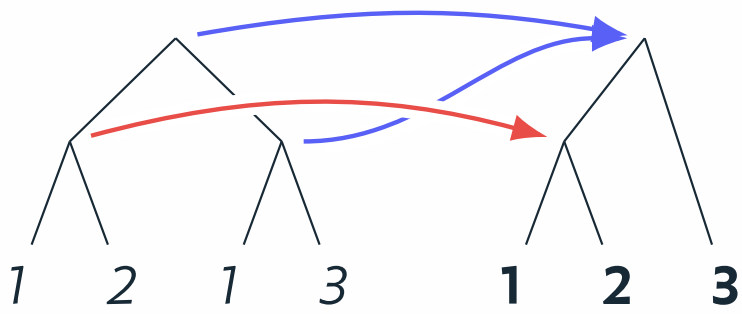
Entrée : Arbres phylogénétiques T et S
Sortie : Réconciliation sous $\mathcal{M}$ de coût minimum



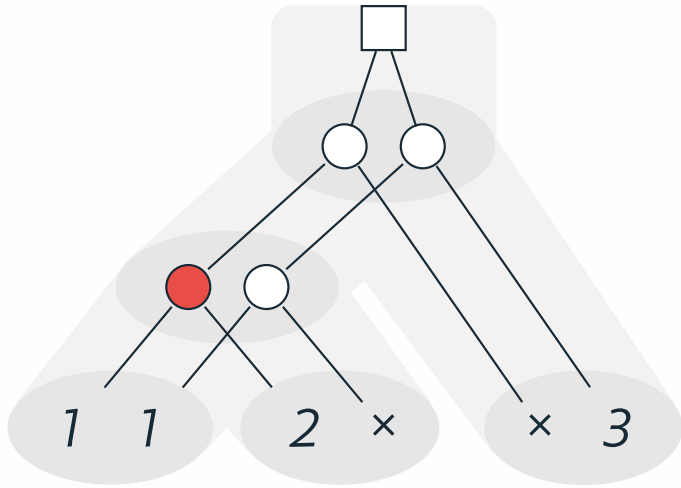
DL-PARCIMONIE

- **LCA** : Ancêtre commun le plus récent

1. Nœud de T ne peut être associé à un descendant du LCA
2. Associé à un ancêtre strict du LCA → réconciliation non-optimale



T S

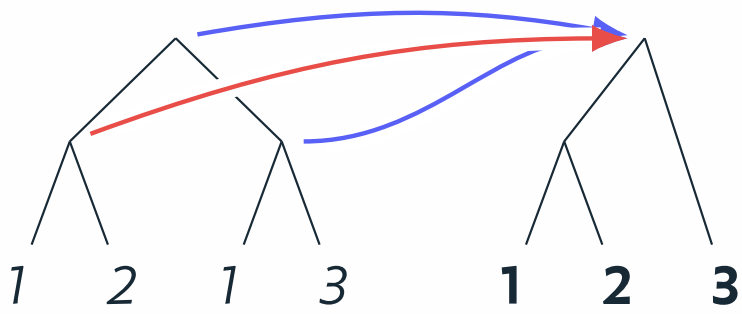


Arbre réconcilié $\mathcal{R}$

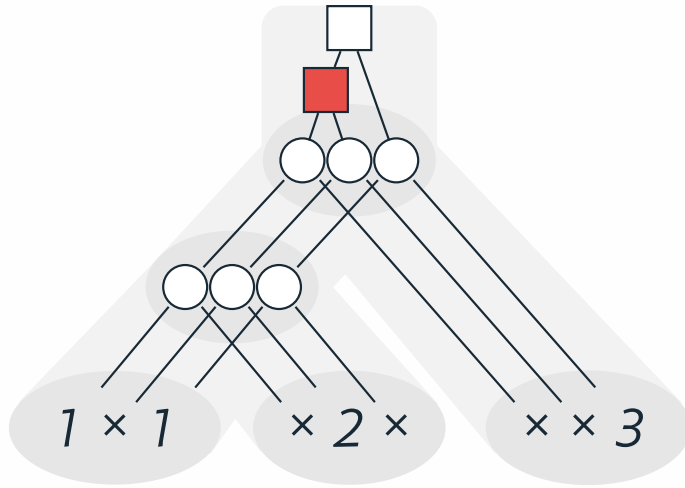
DL-PARCIMONIE

- **LCA** : Ancêtre commun le plus récent

1. Nœud de T ne peut être associé à un descendant du LCA
2. Associé à un ancêtre strict du LCA → réconciliation non-optimale



T S



Arbre réconcilié $\mathcal{R}$

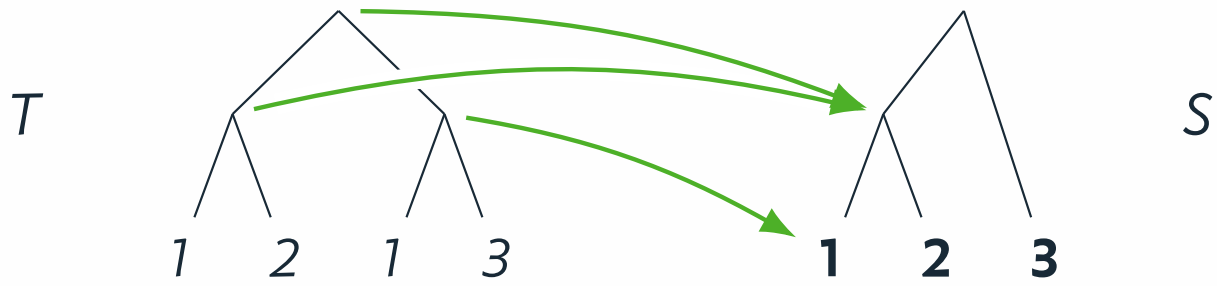
DL-PARCIMONIE

- **LCA** : Ancêtre commun le plus récent

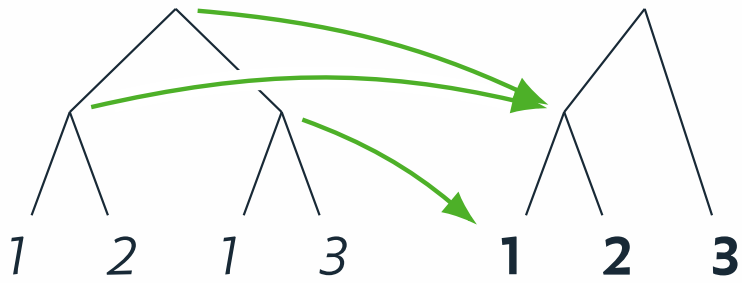
1. Nœud de T ne peut être associé à un descendant du LCA
2. Associé à un ancêtre strict du LCA → réconciliation non-optimale

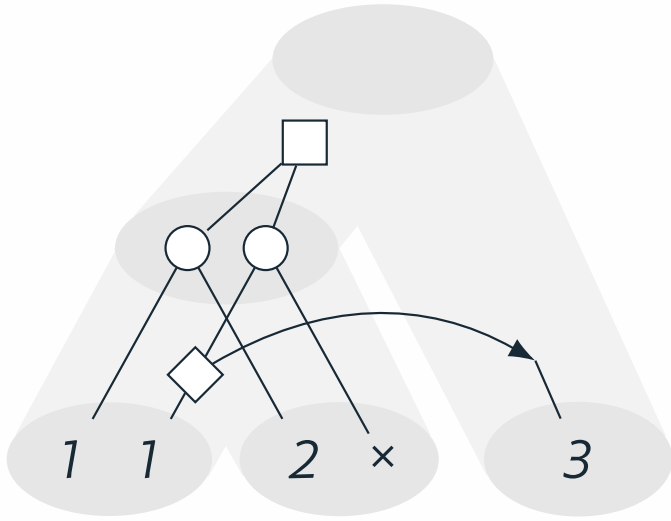
- Donc associer chaque nœud au LCA *(LCA-mapping)* est optimal
- Se calcule en **temps linéaire** en la taille de T

CHANGEMENT D'HÔTE



MODÈLE DUPLICATION–TRANSFERT–PERTE

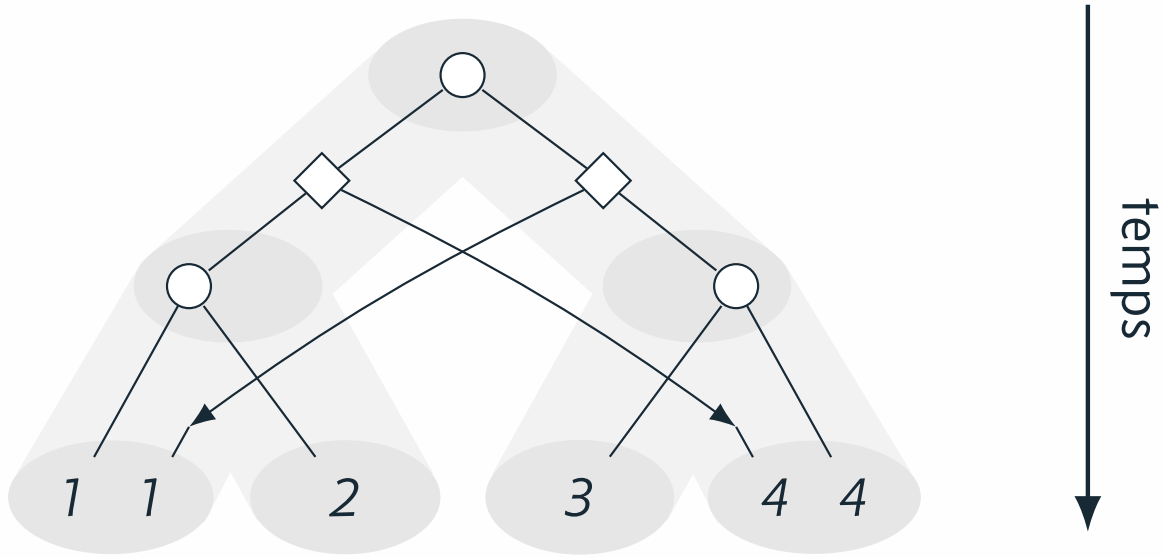




- **DTL :** Spéciation ○ Duplication □ Transfert ◇ Perte ×

CONTRAINTES D'ORDRE

- Transferts s'effectuent entre deux **points coexistants** de S
- Certaines combinaisons de transferts sont **irréalisables**



DTL-PARCIMONIE

- Réconciliation est **acyclique** si ses transferts sont réalisables

Problème DTL-Parcimonie

Entrée : Arbres phylogénétiques T et S
Sortie : Réconciliation DTL **acyclique** de coût minimum

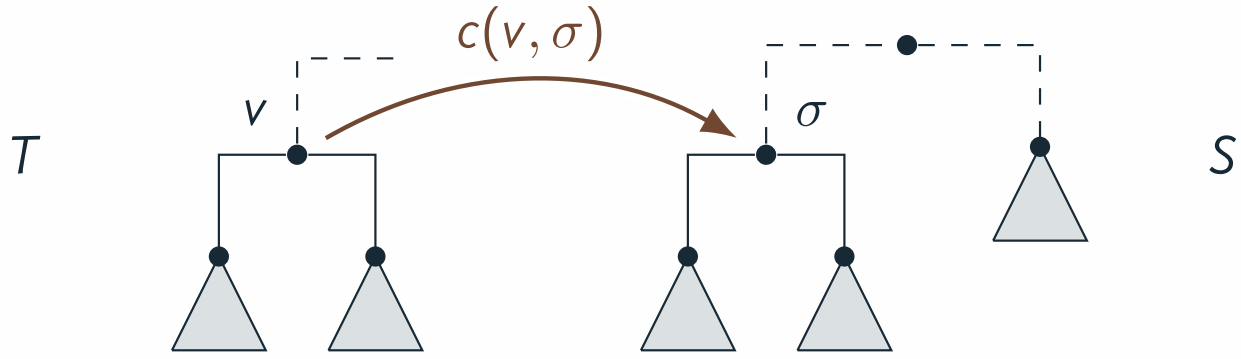
- **NP-complet** (problème du sous-graphe acyclique maximum)
- *(Vérification de l'acyclicité en temps polynomial)*
- **Relaxation** de la contrainte → problème DTL-parcimonie⋆

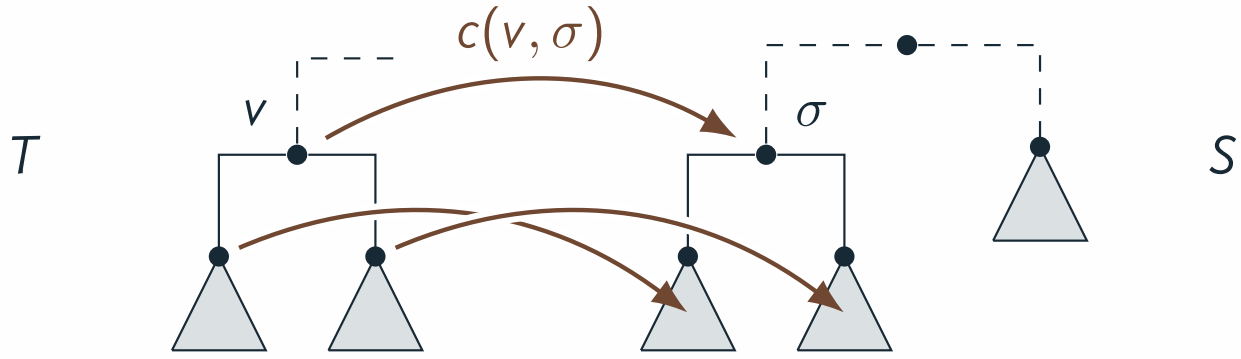
Problème DTL-Parcimonie⋆

Entrée : Arbres phylogénétiques T et S
Sortie : Réconciliation DTL ~~acyclique~~ de coût minimum

- LCA n'est pas toujours optimal
- **Plusieurs solutions** optimales en général
- Réconciliation optimale composée de sous-réconciliations optimales
- Sous-structure optimale → **programmation dynamique**

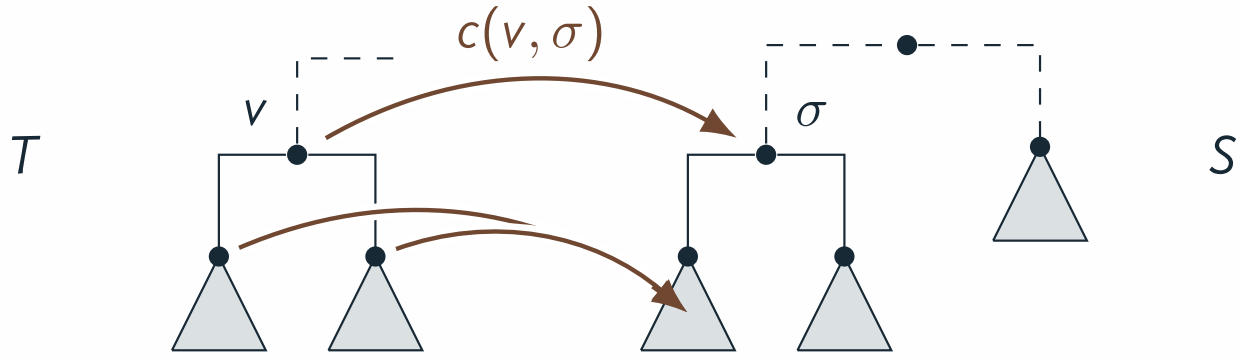
DTL-PARCIMONIE⋆





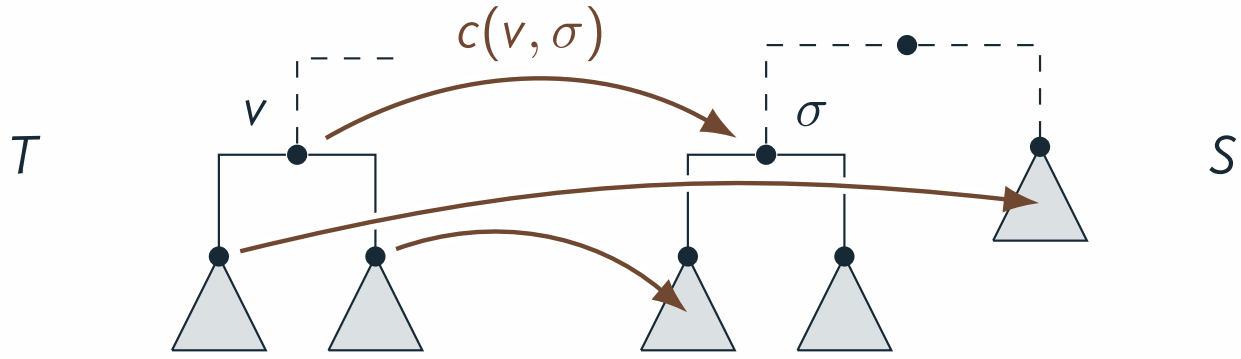
1 **Deux descendants, sous-arbre différent** Spéciation ◯

DTL-PARCIMONIE⋆



1. Deux descendants, sous-arbre différent — Spéciation ○
2. **Deux descendants, même sous-arbre** — Duplication □

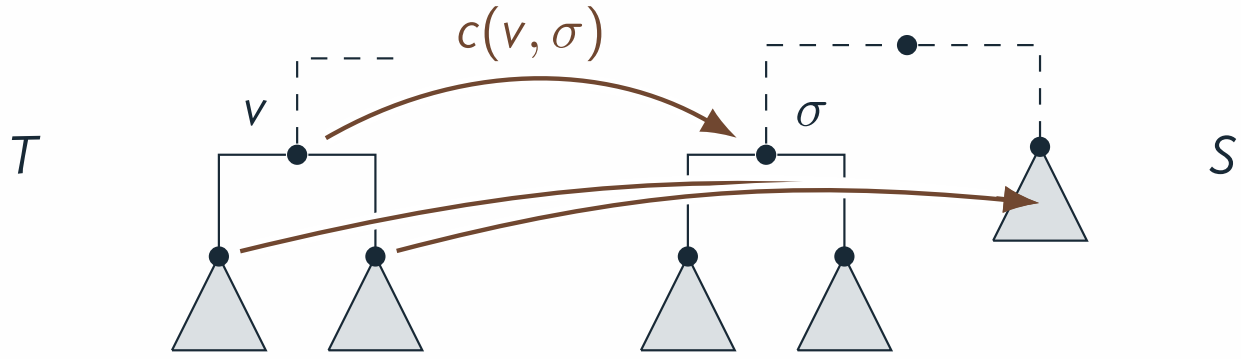
DTL-PARCIMONIE⋆



1. Deux descendants, sous-arbre différent
2. Deux descendants, même sous-arbre
3. **Un descendant, un séparé**

Spéciation ○
Duplication □
Transfert ◇

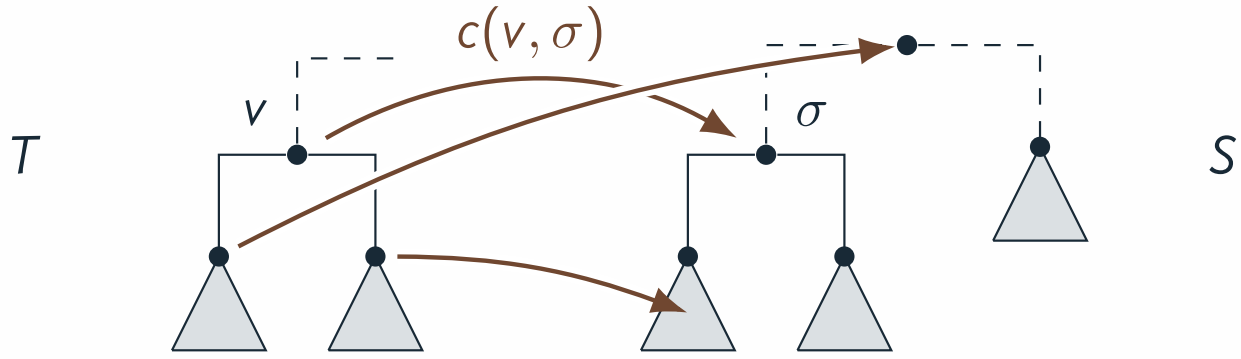
DTL-PARCIMONIE⋆



1. Deux descendants, sous-arbre différent
2. Deux descendants, même sous-arbre
3. Un descendant, un séparé
4. **Deux séparés**

Spéciation ○
Duplication □
Transfert ◇
2 Transferts ◇◇

DTL-PARCIMONIE⋆



1. Deux descendants, sous-arbre différent
2. Deux descendants, même sous-arbre
3. Un descendant, un séparé
4. Deux séparés
5. **Un ancêtre**

Spéciation ○
Duplication □
Transfert ◇
2 Transferts ◇◇
(invalide)

EXPLOITATION DES RÉCONCILIATIONS

- Analyse de l'espace des solutions
- Inférence d'arbres d'espèces
- Correction d'arbres de gènes
- Résolution de multifurcations dans les arbres

PLAN

1 Introduction

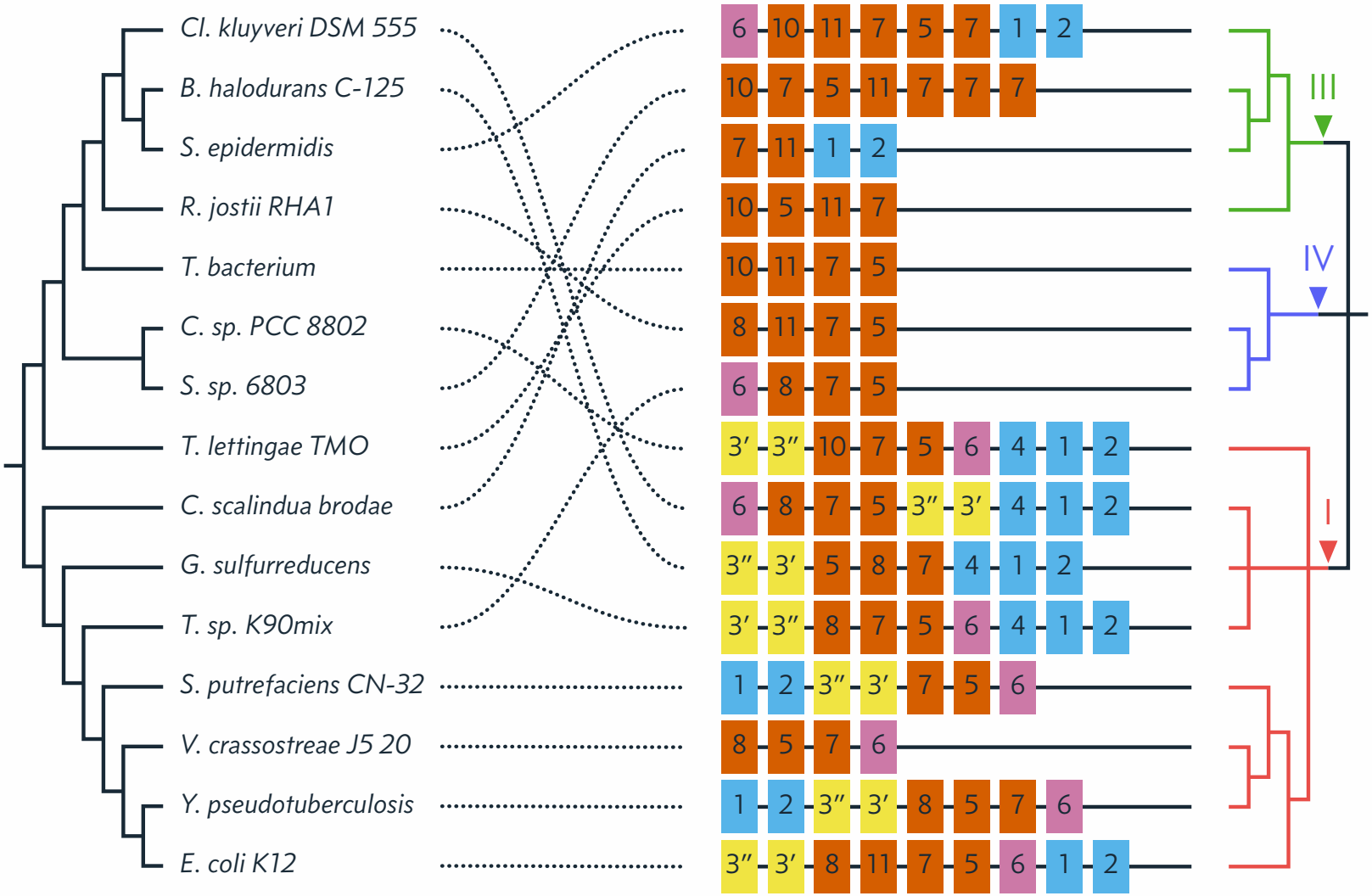
2 Réconciliation bilatérale

3 Réconciliation multiple

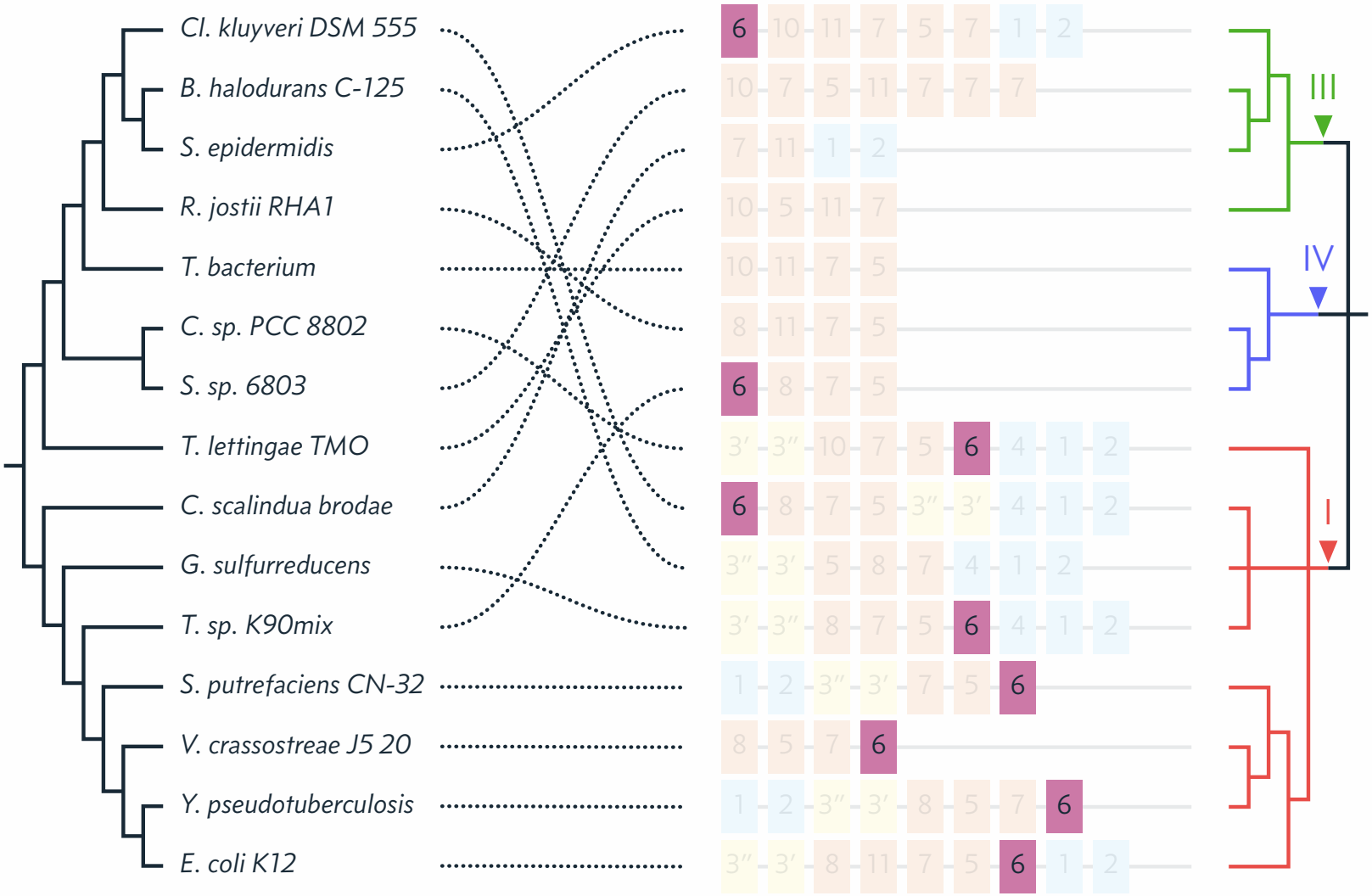
- Réconciliation d'adjacences (DeCo)
- Réconciliation segmentale (super-réconciliation)

4 Projet de recherche

SYSTÈMES CRISPR-CAS



SYSTÈMES CRISPR-CAS



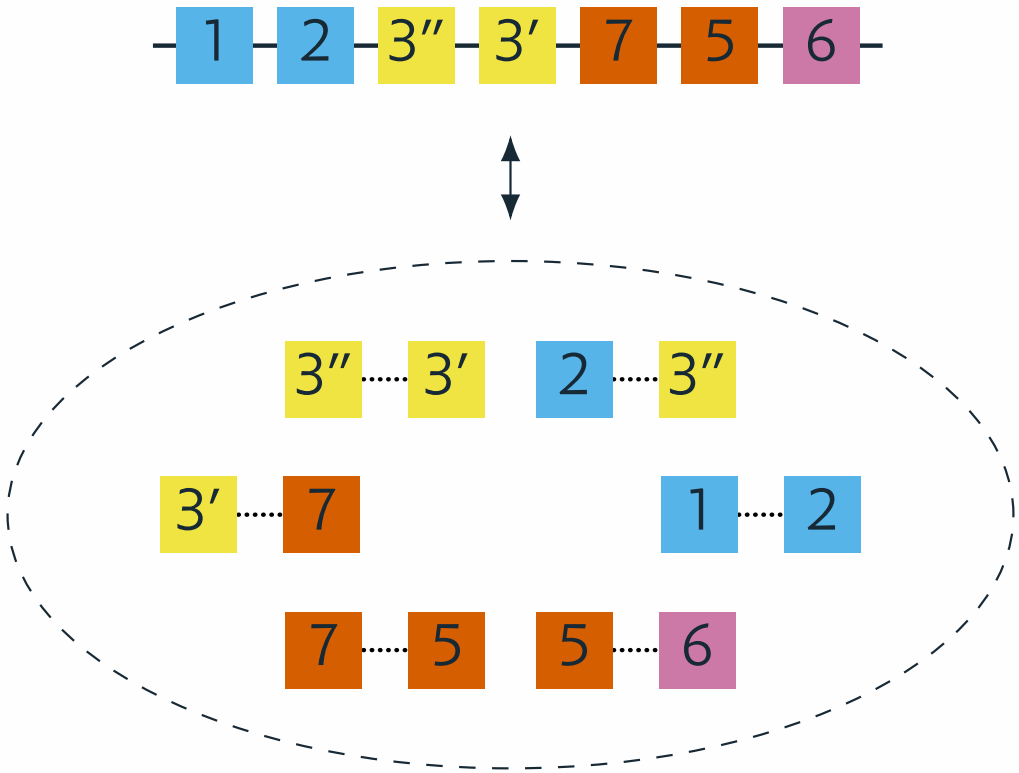
RÉCONCILIATION MULTIPLE

- Intégrer l'**évolution de l'organisation des gènes** dans la réconciliation
- Emprunte aux techniques de comparaison de génomes

- Ajout d'un « niveau intermédiaire » entre gènes et espèces
  - **Adjacences** → *DeCo*
  - **Segments** → *super-réconciliation*

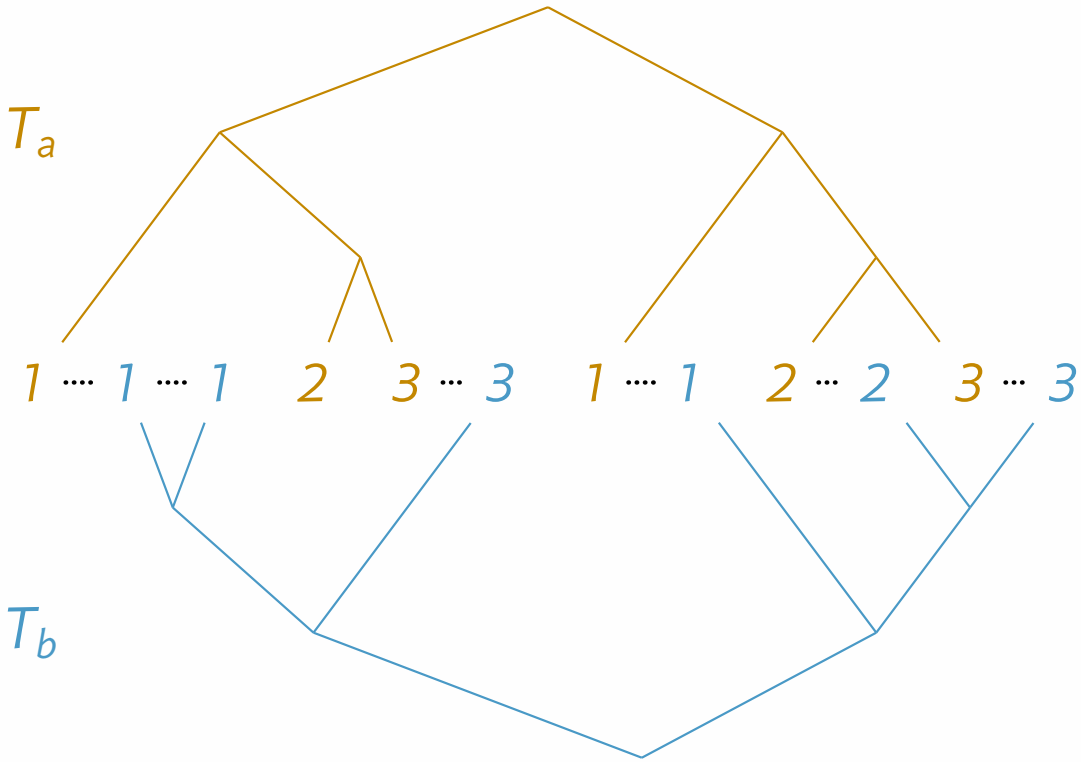
PLAN

ADJACENCES



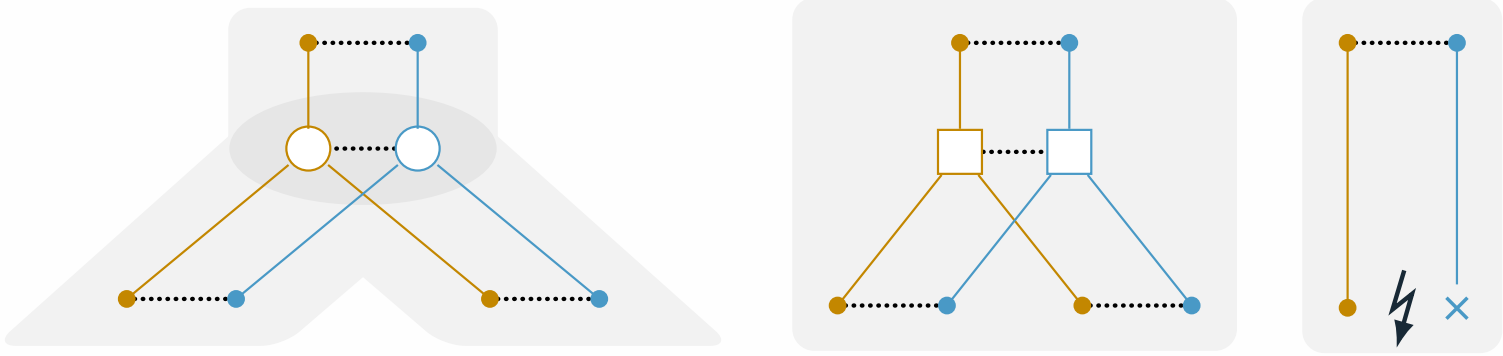
- **Adjacence :** Un gène précède l'autre dans le même génome

ADJACENCES ET PHYLOGÉNIES



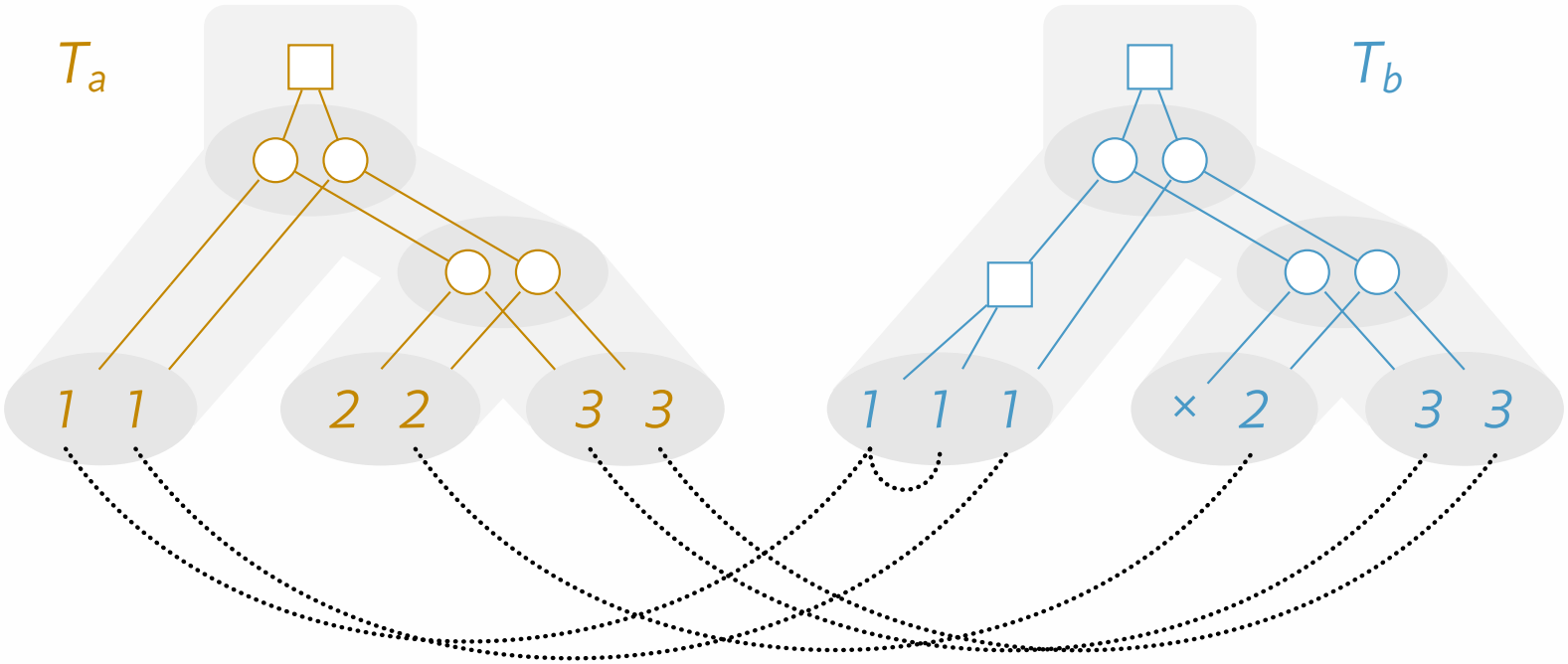
MODÈLE DECO

- Évolution d'adjacences menant aux adjacences observées
- Adjacences **évoluent indépendamment** les unes des autres
- Événement affecte une adjacence s'il affecte ses deux extrémités

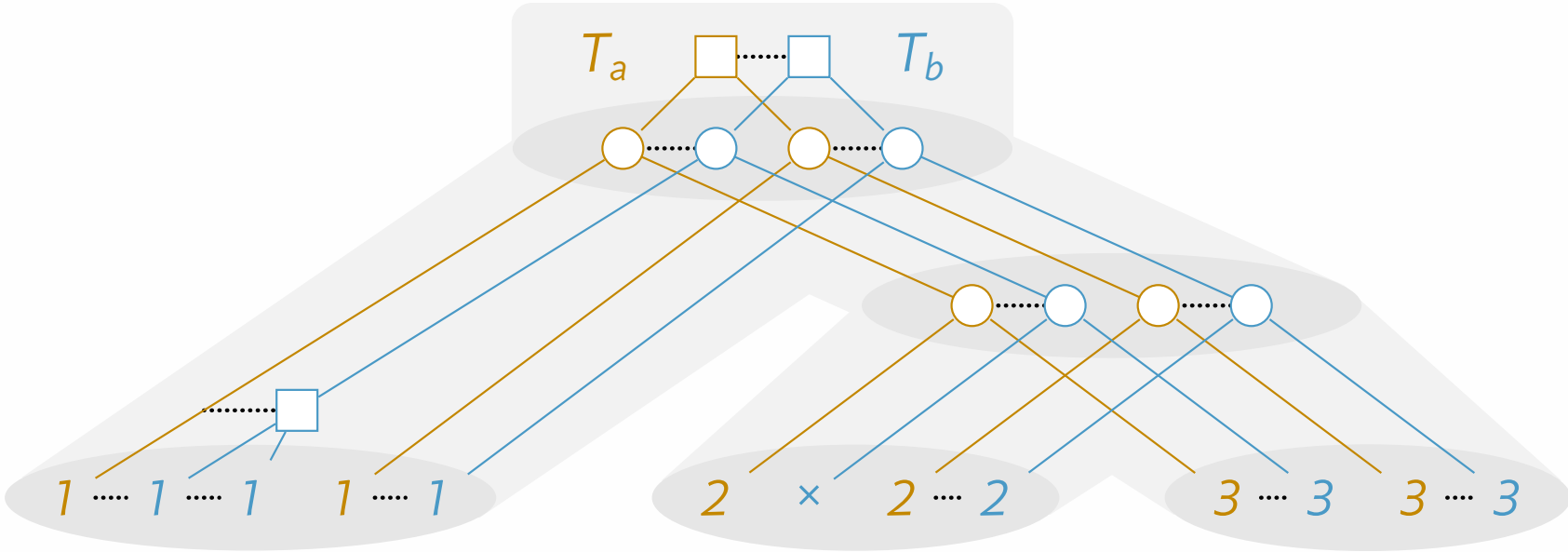


Spéciation Duplication Rupture

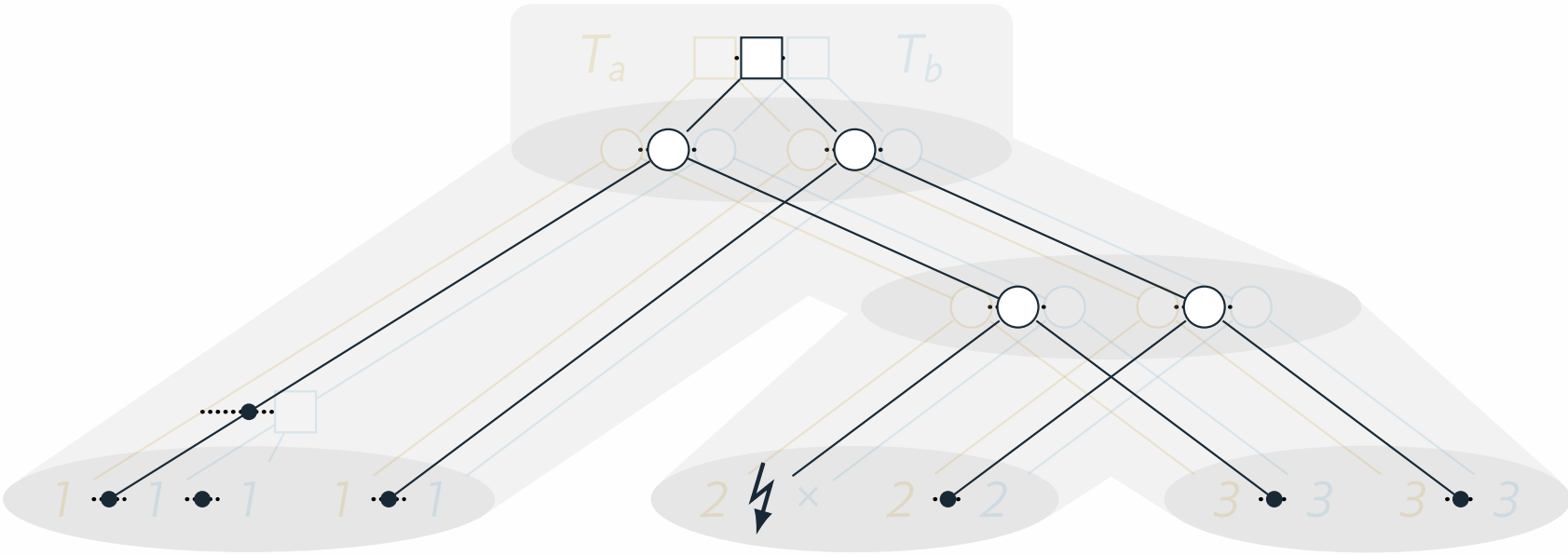
RÉCONCILIATION D'ADJACENCES (1/2)



RÉCONCILIATION D'ADJACENCES (2/2)



RÉCONCILIATION D'ADJACENCES (2/2)



- **Forêt d'adjacences :** Arbres pour chaque famille d'adjacence
- Chaque racine de la forêt est un **gain** d'adjacence
- Chaque adjacence peut être **rompue**

PROBLÈME ET ALGORITHME DECO

Problème Adjacence-Parcimonie

Entrée : Arbres phylogénétiques T_i et S ; adjacences sur les feuilles
Sortie : Réconciliation d'adjacences minimisant les gains et ruptures

- DeCo procède en deux étapes :

1. **Réconciliation individuelle** de chaque arbre selon DL, DTL, …
2. Arbre d'adjacence pour **chaque paire d'arbres** (indépendants)

PROBLÈME ET ALGORITHME DECO

Problème Adjacence-Parcimonie

Entrée : Arbres phylogénétiques T_i et S ; adjacences sur les feuilles
Sortie : Réconciliation d'adjacences minimisant les gains et ruptures

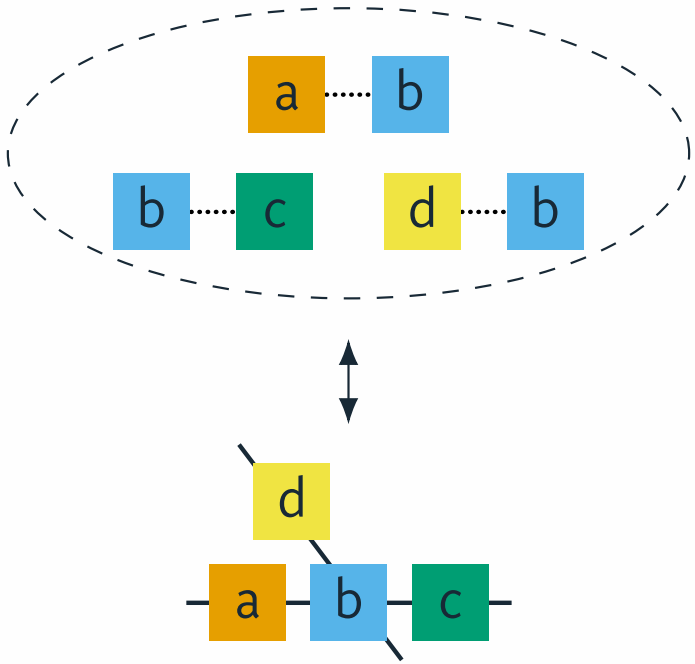
- DeCo procède en deux étapes :

1. **Réconciliation individuelle** de chaque arbre selon DL, DTL, …
2. Arbre d'adjacence pour **chaque paire d'arbres** (indépendants)

- N'est pas garanti optimal
- Complexité du problème général inconnue (NP-complet ?)

LINÉARITÉ

- **Adjacences linéaires :** aucun gène n'est dans plus de deux adjacences



- DeCo **ne garantit pas** la linéarité des adjacences inférées

PLAN

1 Introduction

2 Réconciliation bilatérale

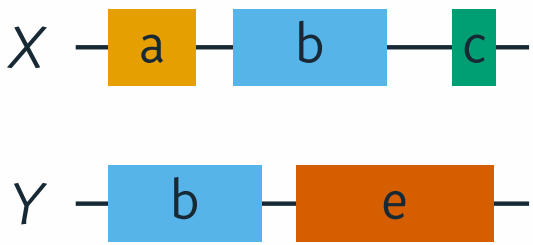
3 **Réconciliation multiple**
- Réconciliation d'adjacences (DeCo)
- **Réconciliation segmentale (super-réconciliation)**

4 Projet de recherche

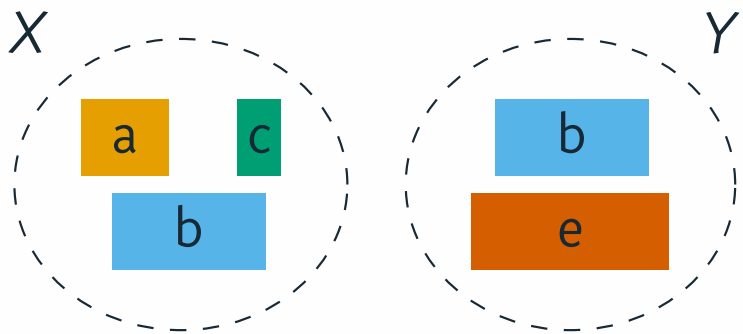
SYNTÉNIES

- Généralisation des adjacences à un nombre quelconque de gènes

1 Segments
(chaînes)

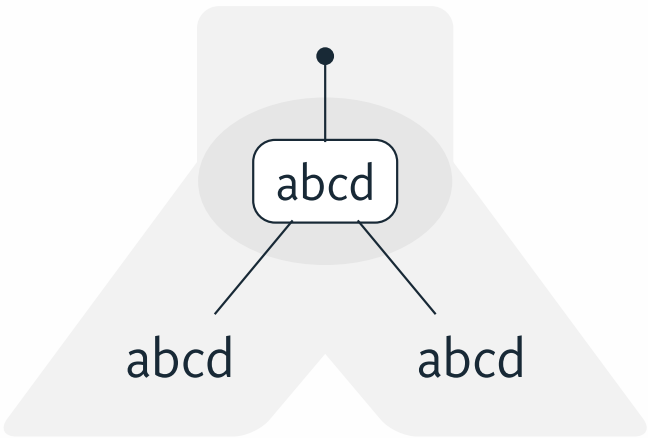


2 Groupes
(ensembles)

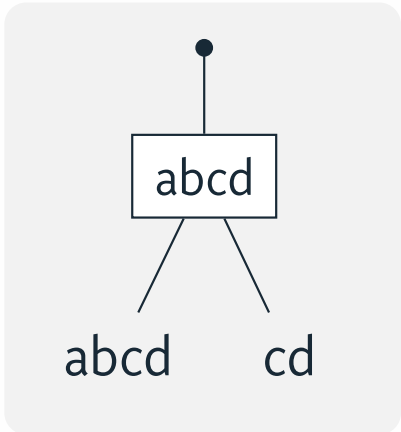


MODÈLE D'ÉVOLUTION DES SYNTÉNIES

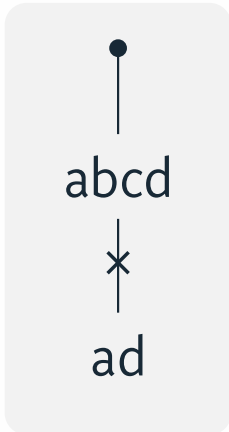
- Évolution de synténies menant aux synténies observées
- Événement affecte un **sous-segment**/**sous-ensemble** de la synténie



Spéciation

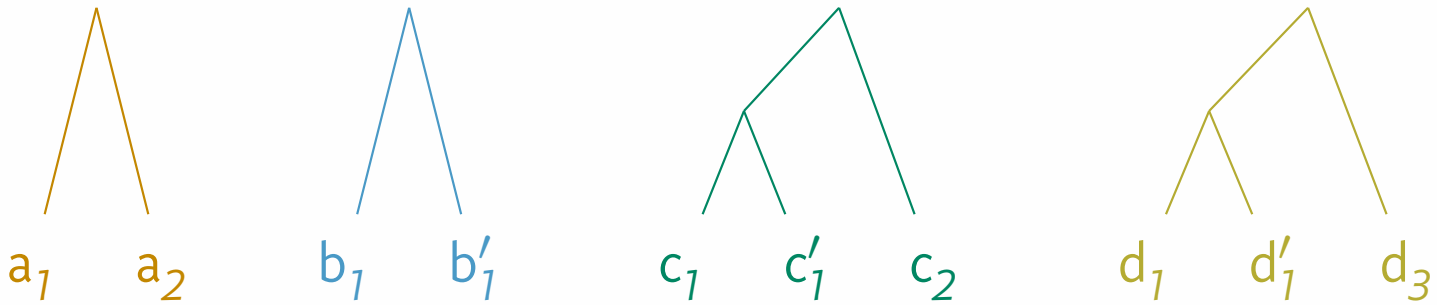


Duplication segmentale



Perte segmentale

SUPER-RÉCONCILIATION (1/2)



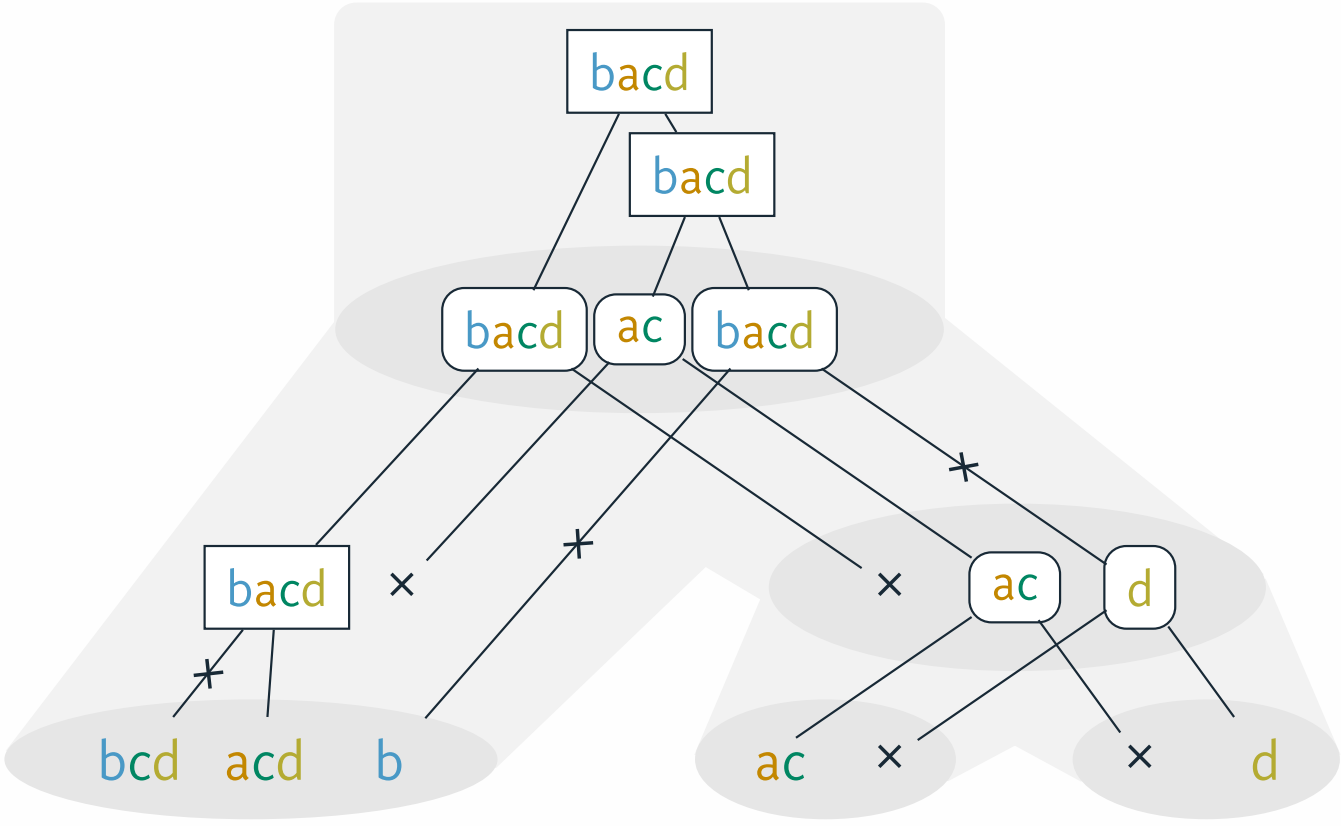
$b_1 c_1 d_1$ $a_1 c'_1 d'_1$

b'_1 $a_2 c_2$ d_3

Synténies

S

SUPER-RÉCONCILIATION (2/2)



- **Arbre de synténies :** Évolution des synténies
- *(Un seul arbre, pas une forêt, contrairement à DeCo)*

PROBLÈME SUPER-RÉCONCILIATION

Problème Super-Réconciliation

Entrée : Arbres phylogénétiques T_i et S ; synténies sur les feuilles

Sortie : Super-réconciliation de coût minimum

PROBLÈME SUPER-RÉCONCILIATION

Problème Super-Réconciliation

Entrée : Arbres phylogénétiques T_i et S ; synténies sur les feuilles
Sortie : Super-réconciliation de coût minimum

1. **Topologie** pour l'arbre de synténies
2. **Ordre** ancestral sur les gènes *(si besoin)*
3. **Réconciliation** de l'arbre de synténies
4. **Contenu** des synténies

} **NPC**

- Étape 3 réalisée séparément par *LCA-mapping*

PROBLÈME SUPER-RÉCONCILIATION

Problème Super-Réconciliation

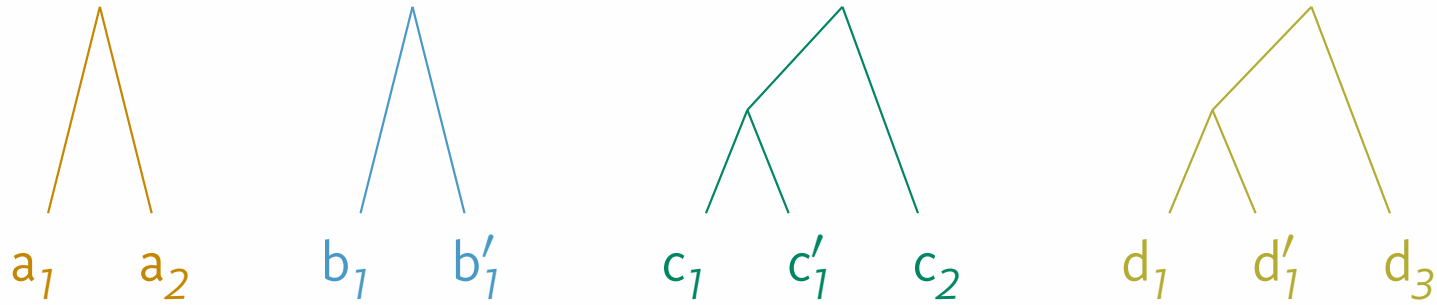
Entrée : Arbres phylogénétiques T_i et S ; synténies sur les feuilles
Sortie : Super-réconciliation de coût minimum

1. **Topologie** pour l'arbre de synténies
2. **Ordre** ancestral sur les gènes *(si besoin)*
3. **Réconciliation** de l'arbre de synténies
4. **Contenu** des synténies

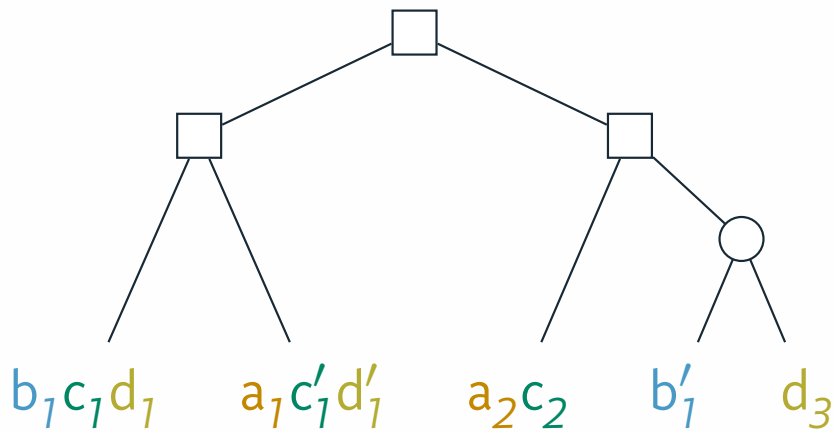
} ? ; } ? ; } ? ; } **NPC**

- Étape 3 réalisée séparément par *LCA-mapping*

TOPOLOGIE DE L'ARBRE DES SYNTÉNIES



- **Super-arbre** des arbres de gènes



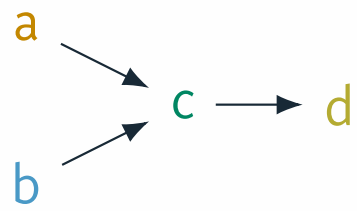
ORDRE ANCESTRAL SUR LES GÈNES

- Ordre relatif entre les gènes **conservé** à travers l'arbre
- **Tri topologique** sur graphe auxiliaire
- Dans le pire cas, $n!$ ordres possibles

bcd acd

b ac d

Synténies



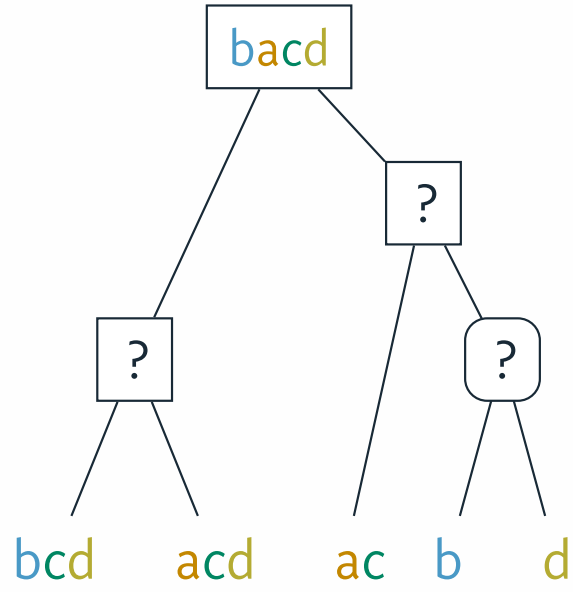
Graphe auxiliaire

abcd bacd

Ordres ancestraux possibles

CONTENU DES SYNTÉNIES : SEGMENTS

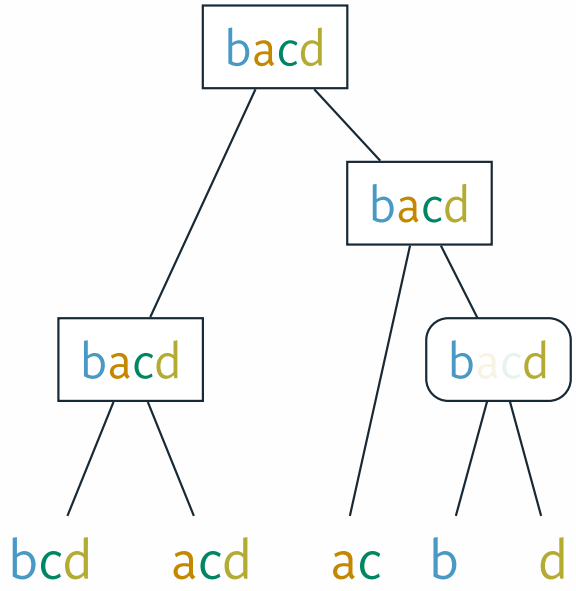
- Choisir synténies ancestrales pour **minimiser les pertes** segmentales



CONTENU DES SYNTÉNIES : SEGMENTS

- Choisir synténies ancestrales pour **minimiser les pertes** segmentales

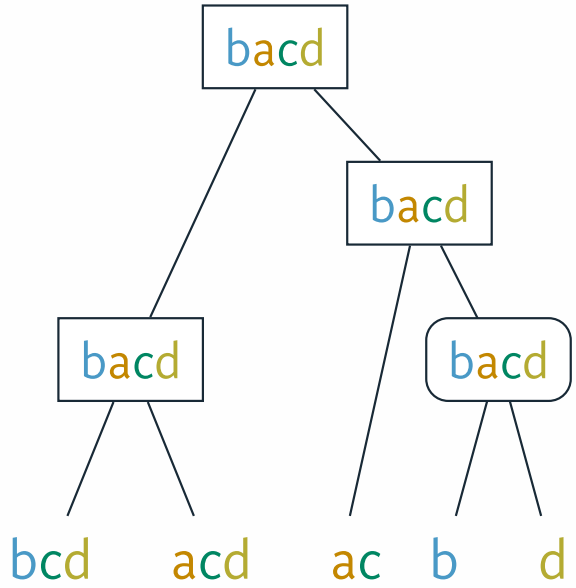
- Au moins les gènes du sous-arbre
- **Stratégie :** Ajouter des gènes supplémentaires de sorte à rassembler les segments de pertes



CONTENU DES SYNTÉNIES : SEGMENTS

- Choisir synténies ancestrales pour **minimiser les pertes** segmentales

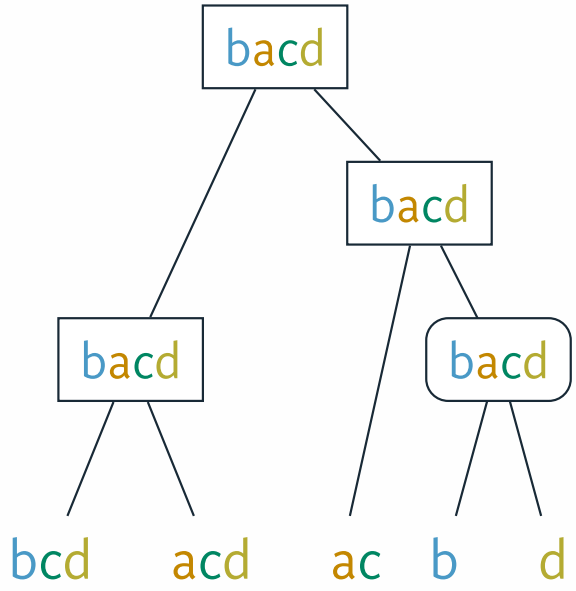
- Au moins les gènes du sous-arbre
- **Stratégie :** Ajouter des gènes supplémentaires de sorte à rassembler les segments de pertes



CONTENU DES SYNTÉNIES : SEGMENTS

- Choisir synténies ancestrales pour **minimiser les pertes** segmentales

- Au moins les gènes du sous-arbre
- **Stratégie :** Ajouter des gènes supplémentaires de sorte à rassembler les segments de pertes



- Au pire $\mathbf{2^n}$ **choix** pour chaque embranchement
- Faire mieux qu'explorer exhaustivement ? *(Question ouverte)*

CONTENU DES SYNTÉNIES : GROUPES

- Même problème pour les groupes
- **Au plus une perte par branche**
- Moins de choix de regroupement
- Algorithme linéaire en programmation dynamique

ADJACENCES

- ~~Événements segmentaux~~
- Réarrangements
- Duplications en tandem
- Fusions, fissions
- ~~Linéarité garantie~~
- Algorithme polynomial

SYNTÉNIES

- Événements segmentaux
- ~~Réarrangements~~
- ~~Duplications en tandem~~
- ~~Fusions, fissions~~
- Linéarité garantie
- Problèmes difficiles

PLAN

SUPER-RÉCONCILIATION AVEC TRANSFERTS

- Extension du modèle pour permettre les transferts segmentaux
- Réconciliation ne peut plus être calculée séparément des contenus
- Nouvel algorithme d'optimisation conjointe
- Présenté en mai à RECOMB-CG 2022

Y. Anselmetti, M. Delabre et N. El-Mabrouk. « Reconciliation with segmental duplication, transfer, loss and gain ». In : *Comparative Genomics*. Springer International Publishing, 2022, p. 124-145

COMPLEXITÉ DE SUPER-RÉCONCILIATION

Problème Super-Réconciliation

Entrée : Arbres phylogénétiques T_i et S ; synténies sur les feuilles
Sortie : Super-réconciliation de coût minimum

1. **Topologie** pour l'arbre de synténies
2. **Ordre** ancestral sur les gènes *(si besoin)*
3. **Réconciliation** de l'arbre de synténies
4. **Contenu** des synténies

} ? } ? } ? } **NPC**

- Étape 3 réalisée séparément par *LCA-mapping*

COMPLEXITÉ DE SUPER-RÉCONCILIATION

Problème Super-Réconciliation

Entrée : Arbres phylogénétiques T_i et S ; synténies sur les feuilles
Sortie : Super-réconciliation de coût minimum

1. **Topologie** pour l'arbre de synténies
2. **Ordre** ancestral sur les gènes *(si besoin)*
3. **Réconciliation** de l'arbre de synténies
4. **Contenu** des synténies

} P (3 et 4) ; } P (2 et 3) ; } NPC (1 et 2) ; } **NPC** (1 à 3)

- Étape 3 réalisée séparément par *LCA-mapping*

RÉDUCTION DE MNAE-3-SAT

MNAE-3-SAT (NP-complet)

Entrée : Variables $\{x_1, \ldots, x_n\}$; clauses à trois variables positives

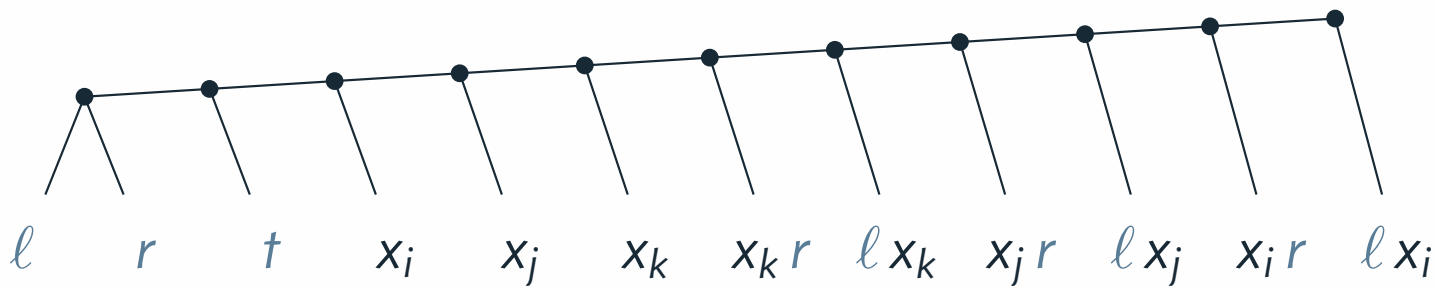
Sortie : Affectation qui rend vraies une ou deux variables par clause

RÉDUCTION DE MNAE-3-SAT

MNAE-3-SAT (NP-complet)

Entrée : Variables $\{x_1, \ldots, x_n\}$; clauses à trois variables positives
Sortie : Affectation qui rend vraies une ou deux variables par clause

- Un gène par variable et trois gènes spéciaux ℓ, t, r
- Ordre sur les gènes $\leftrightarrow$ Affectation des variables
- Clause $\{x_i, x_j, x_k\} \leftrightarrow$

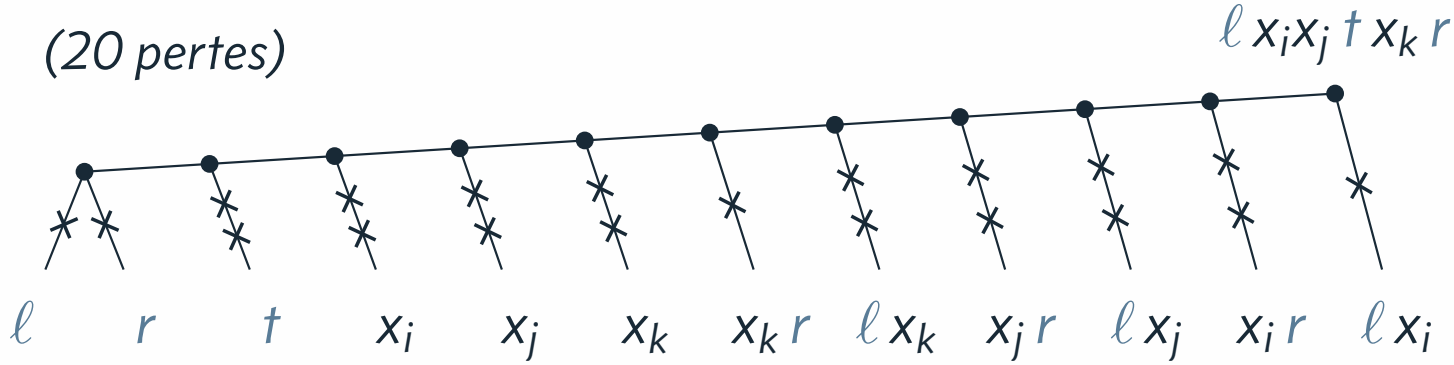


RÉDUCTION DE MNAE-3-SAT

MNAE-3-SAT (NP-complet)

Entrée : Variables $\{x_1, \ldots, x_n\}$; clauses à trois variables positives
Sortie : Affectation qui rend vraies une ou deux variables par clause

- Un gène par variable et trois gènes spéciaux ℓ, t, r
- Ordre sur les gènes $\leftrightarrow$ Affectation des variables
- Clause $\{x_i, x_j, x_k\} \leftrightarrow$

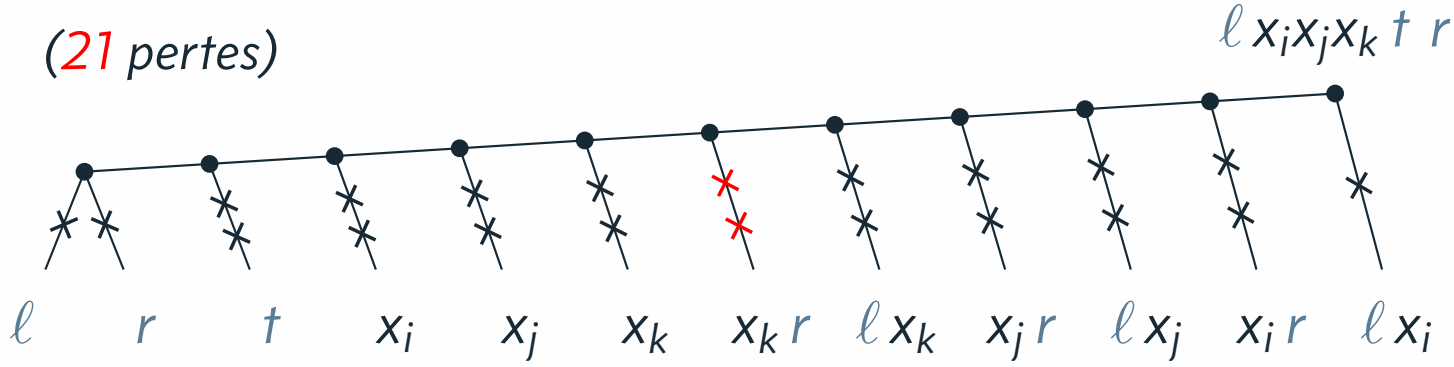


RÉDUCTION DE MNAE-3-SAT

MNAE-3-SAT (NP-complet)

Entrée : Variables $\{x_1, \ldots, x_n\}$; clauses à trois variables positives
Sortie : Affectation qui rend vraies une ou deux variables par clause

- Un gène par variable et trois gènes spéciaux ℓ, t, r
- Ordre sur les gènes $\leftrightarrow$ Affectation des variables
- Clause $\{x_i, x_j, x_k\} \leftrightarrow$



ALGORITHME POLYNOMIAL POUR ORDRE FIXE

- Ordre ancestral $x_1 x_2 \cdots x_n$ fixé
- Calcul de la réconciliation pour $x_1 \cdots x_i$ à partir de $x_1 \cdots x_{i-1}$
- Chaque embranchement a le choix de porter x_i ou non
- Si x_{i-1} était déjà omis, on peut omettre x_i « gratuitement »

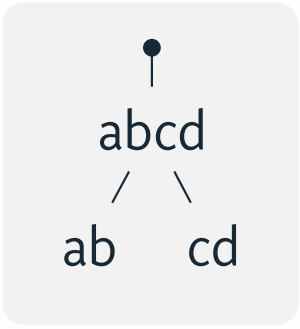
EXTENSIONS DU MODÈLE (1/2)

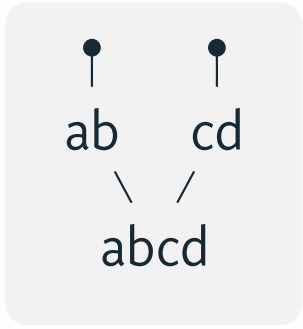
Modèle super-réconciliation plus limité que DeCo :

- Réarrangements ($abcd \rightsquigarrow ab\bar{d}\bar{c}$)
- Transferts s'insérant dans une synténie existante
- Duplications en tandem ($abcd \rightsquigarrow abcdcd$)
- Fission d'une synténie ($abcd \rightsquigarrow ab, cd$)

EXTENSIONS DU MODÈLE (2/2)

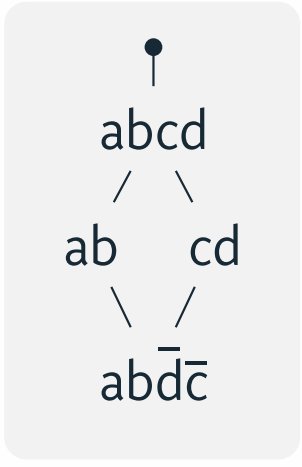
Fission

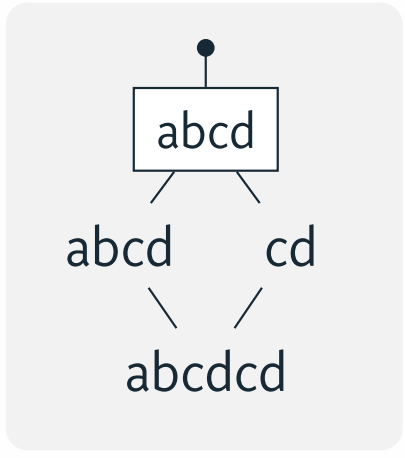


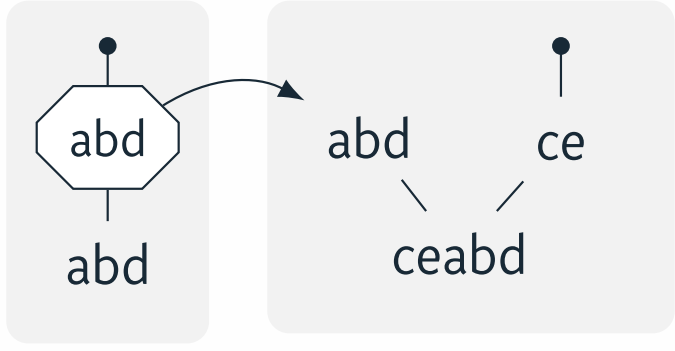


Fusion

- Ces deux événements supplémentaires suffisent (voir SCJ)







SIMULATIONS ET APPLICATIONS

- Validation des modèles par simulation
- Application aux synténies Cas

RÉFÉRENCES PRINCIPALES

- **Réconciliation :** Page. « Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas » (1994)
- **Transferts :** Tofigh, Hallett et Lagergren. « Simultaneous identification of duplications and lateral gene transfers » (mars 2011)
- **DeCo :** Duchemin, Anselmetti, Patterson, Ponty, Bérard, Chauve, Scornavacca, Daubin et Tannier. « DeCoSTAR : Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies » (mai 2017)
- **Super-réconciliation :** Delabre, El-Mabrouk, Huber, Lafond, Moulton, Noutahi et Castellanos. « Evolution through segmental duplications and losses : a super-reconciliation approach » (mai 2020)