21/02/2011
Séminaire (A)CRO – LIF – Marseille

Problèmes d'optimisation combinatoire sur les réseaux phylogénétiques

Philippe Gambette

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Reconstruction de réseaux à partir de triplets
- Reconstruction de réseaux à partir de clades
- Sélection des données
- Visualisation de réseaux phylogénétiques
- Perspectives

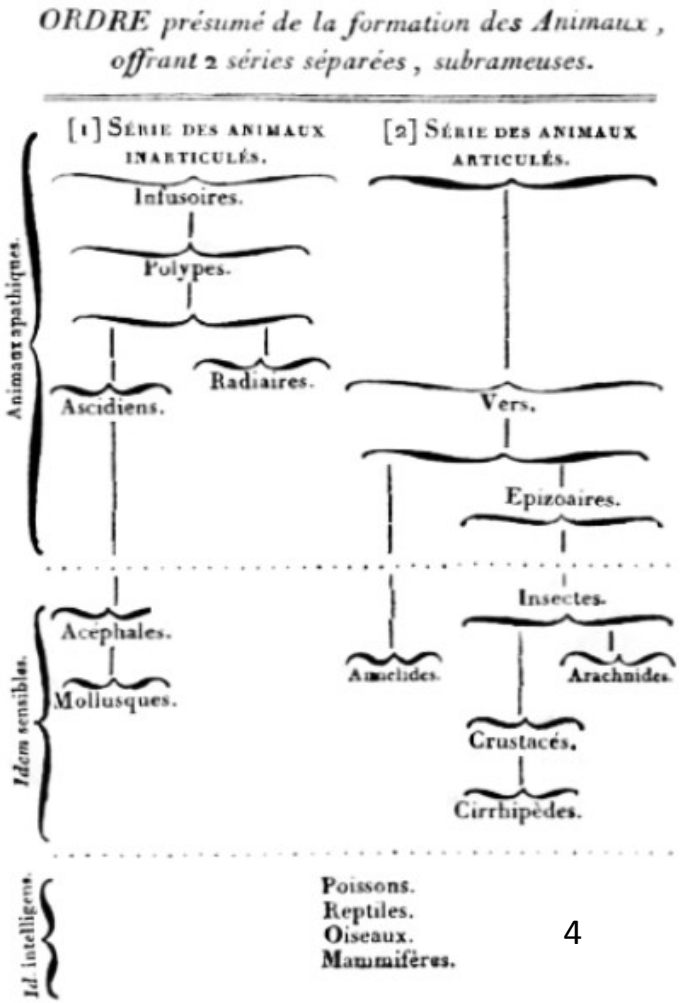
Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Reconstruction de réseaux à partir de triplets
- Reconstruction de réseaux à partir de clades
- Sélection des données
- Visualisation de réseaux phylogénétiques
- Perspectives

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

- Les **organiser** en fonction de caractères communs **classification**
- Décrire leur évolution



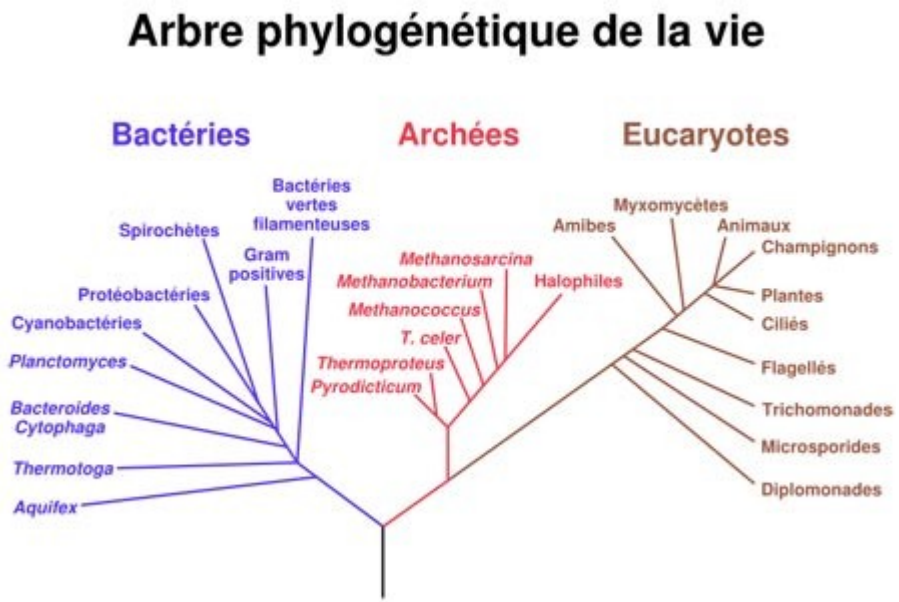
D'après Lamarck : Histoire naturelle des animaux sans vertèbres (1815)

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

- Les organiser en fonction de caractères communs
- Décrire leur **évolution**

modélisation



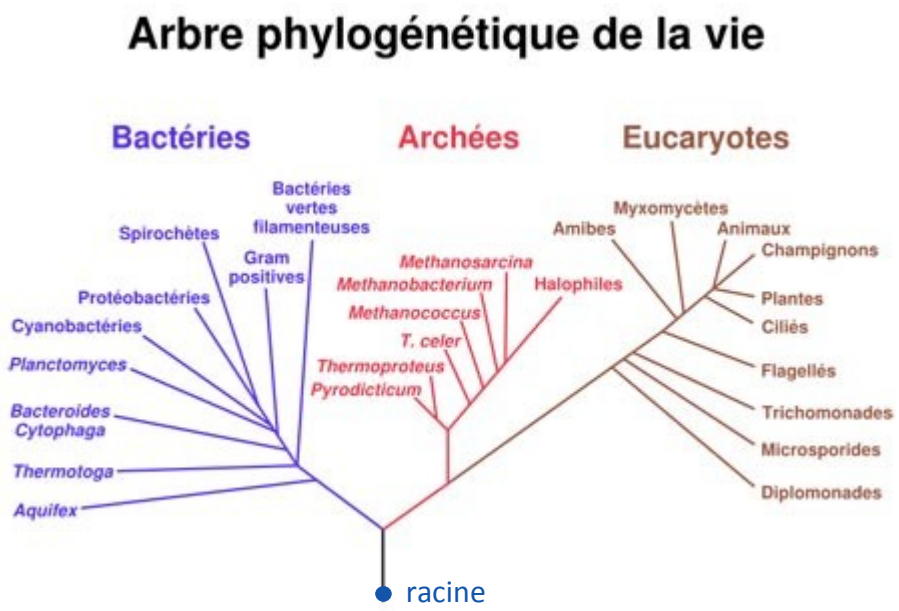
D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

- Les organiser en fonction de caractères communs
- Décrire leur **évolution**

modélisation



D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

Les arbres phylogénétiques

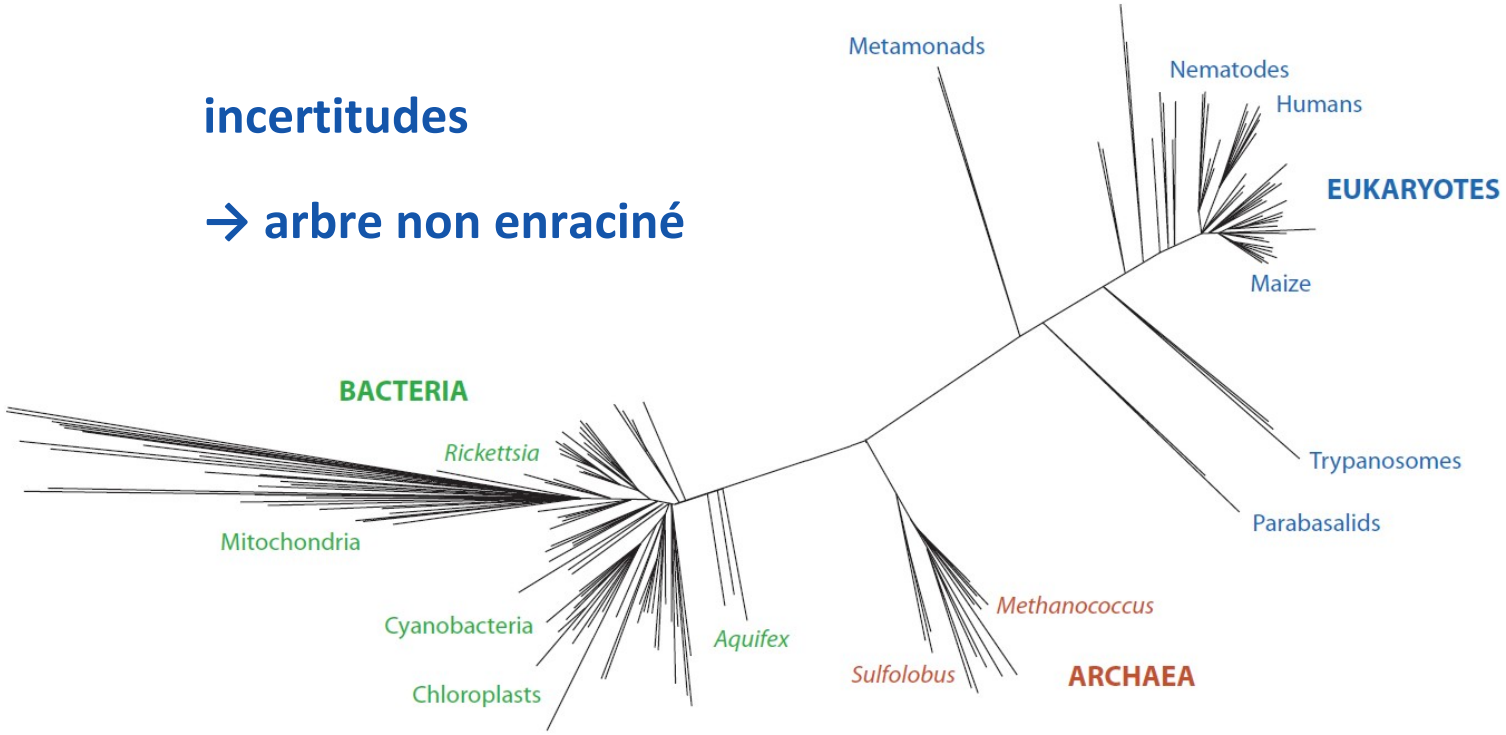
Arbre phylogénétique d'un ensemble d'espèces :

- Les organiser en fonction de caractères communs
- Décrire leur **évolution**

modélisation

incertitudes

→ arbre non enraciné

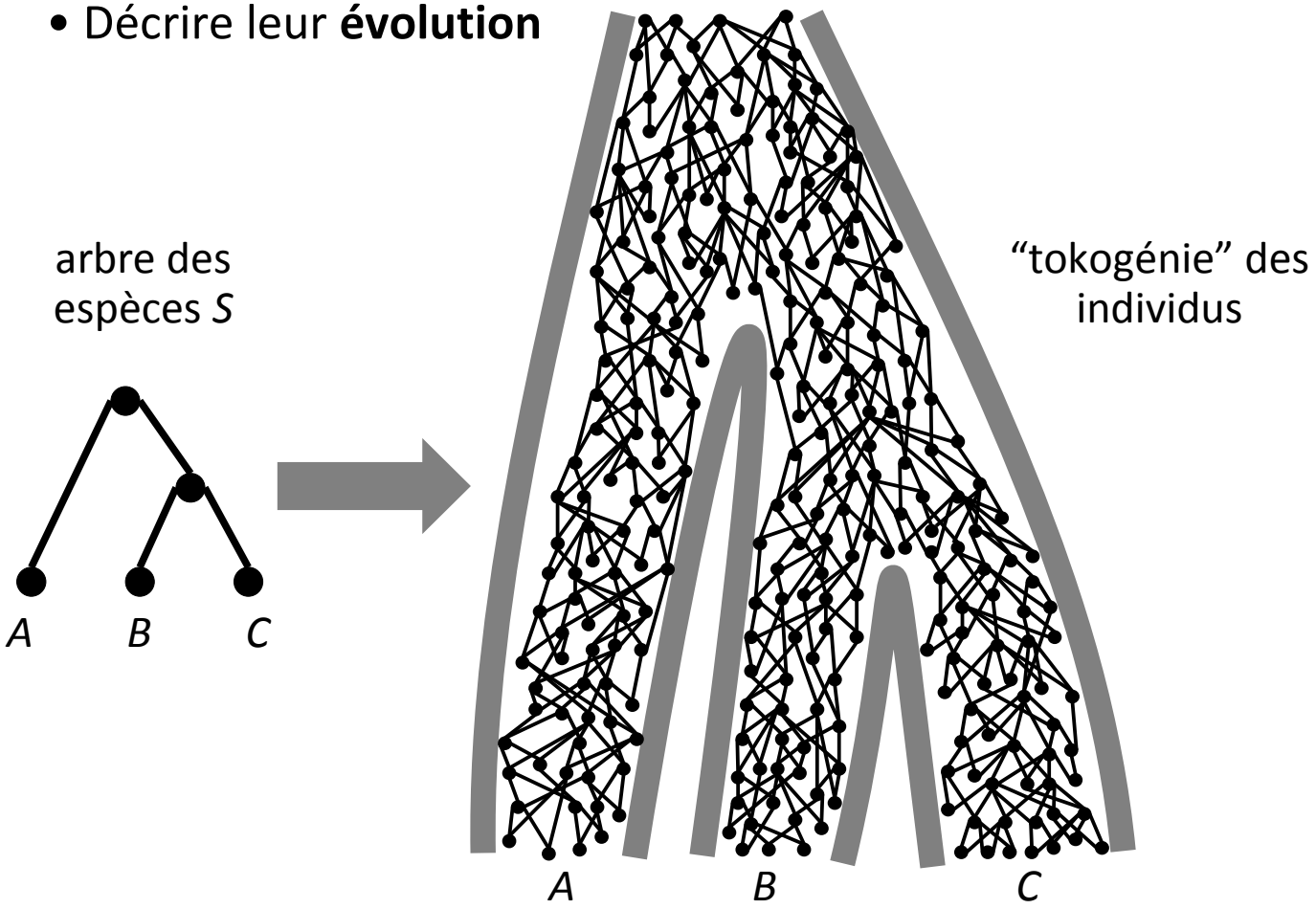


D'après Christophe Blumrich, David S. Spencer,
cité dans Doolittle : Uprooting the Tree of Life, Scientific American (Fév. 2000)

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

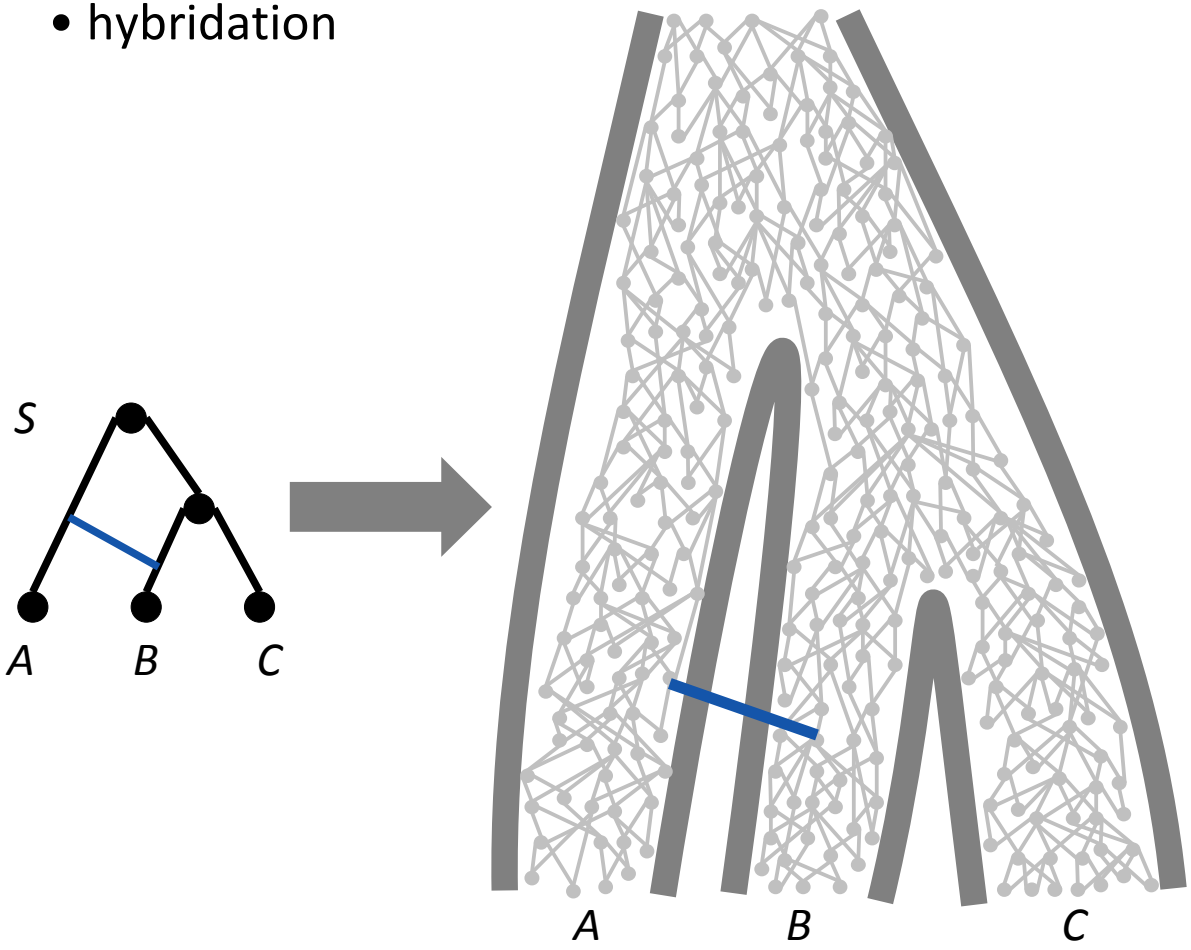
- Les organiser en fonction de caractères communs
- Décrire leur **évolution**



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

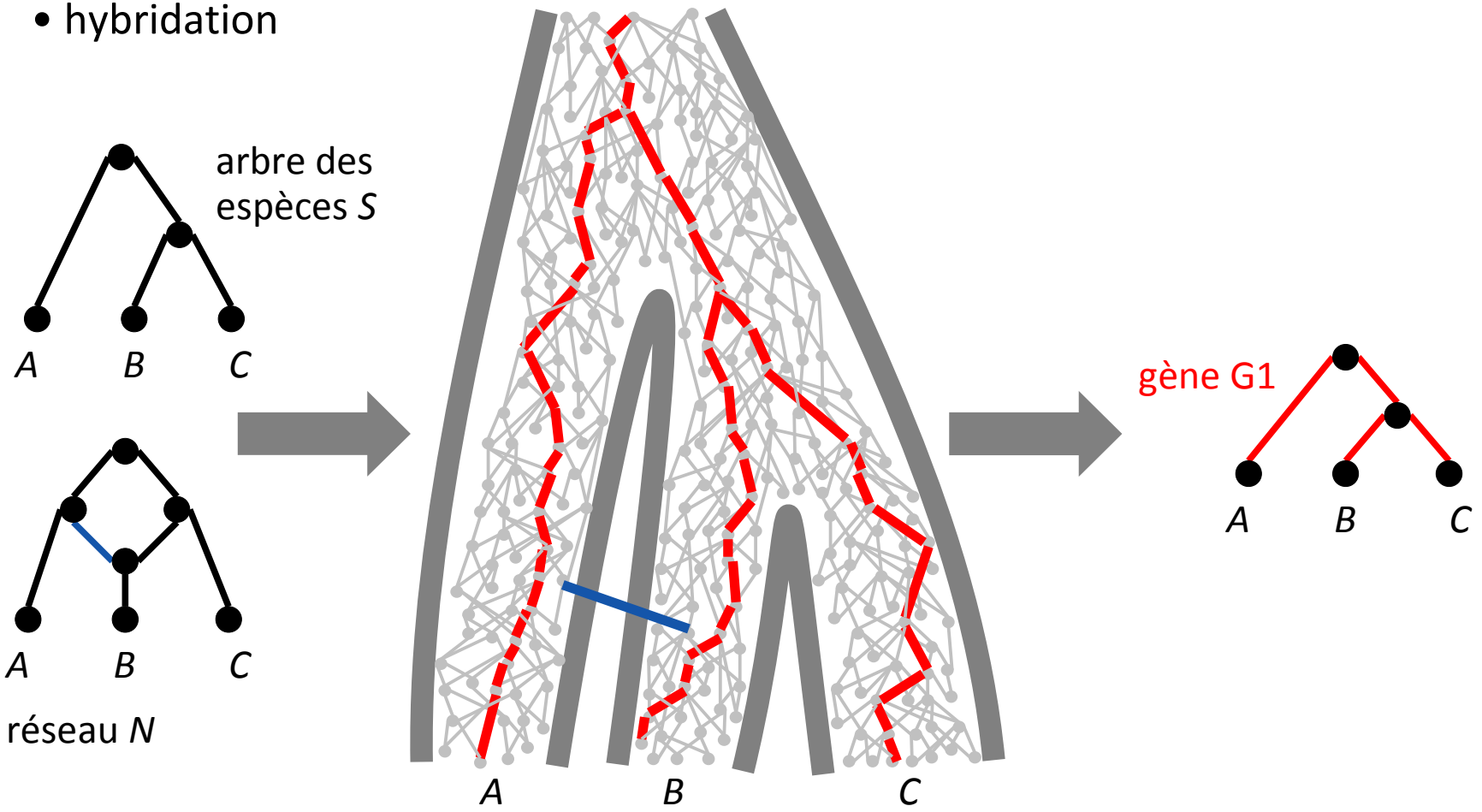
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

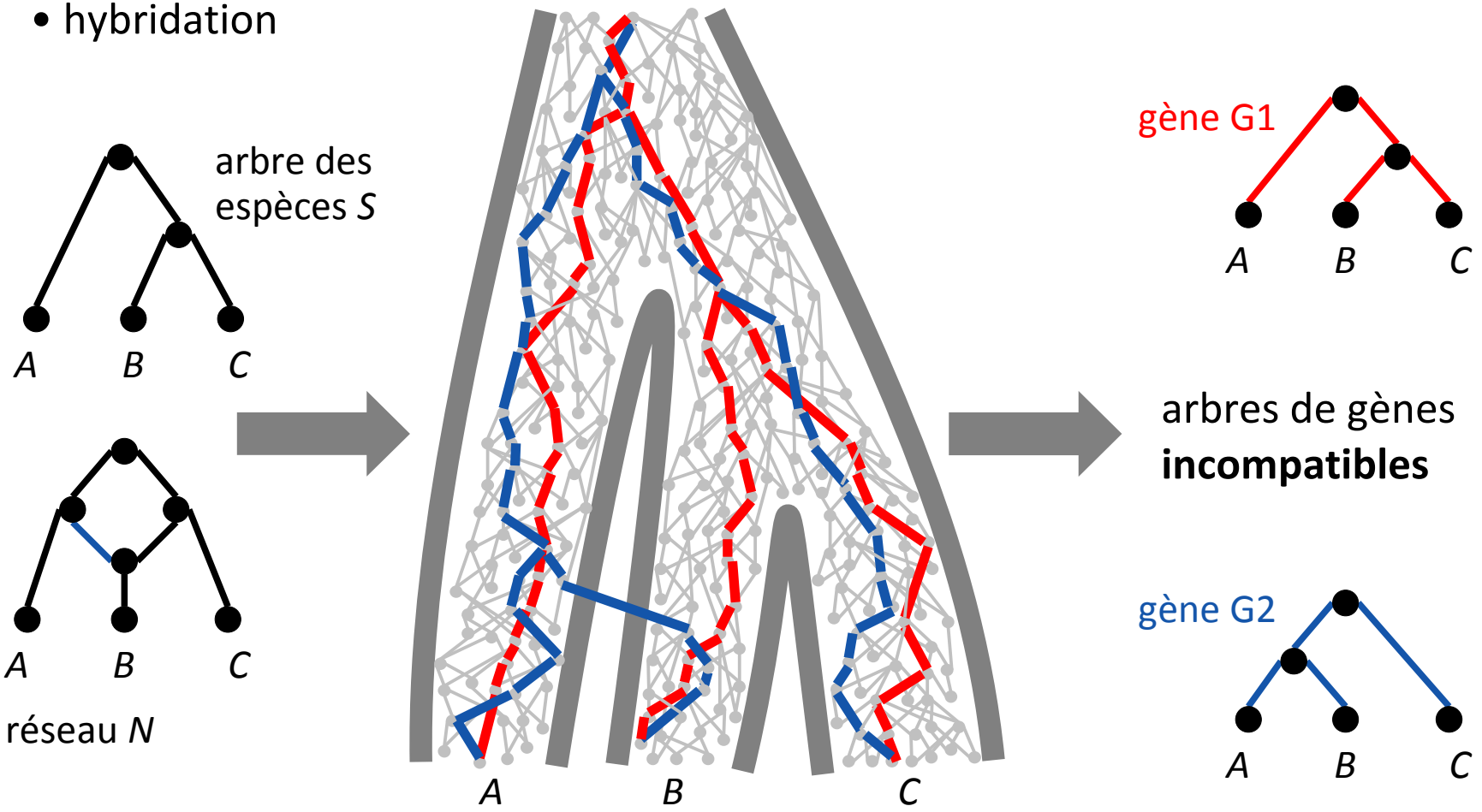
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

- transfert horizontal
- hybridation

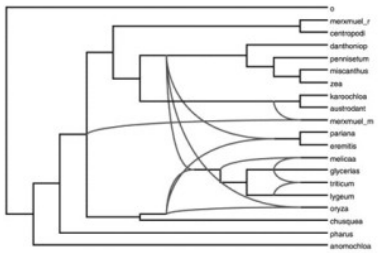


Les réseaux phylogénétiques

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **explicites**

modélisation de l'évolution



réseau à une couche de réticulation

Dendroscope

Simplistic

réseau de niveau 2

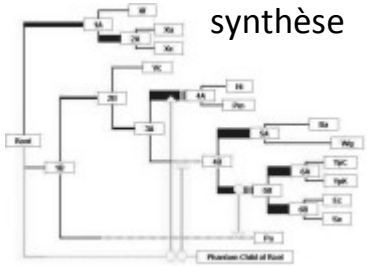
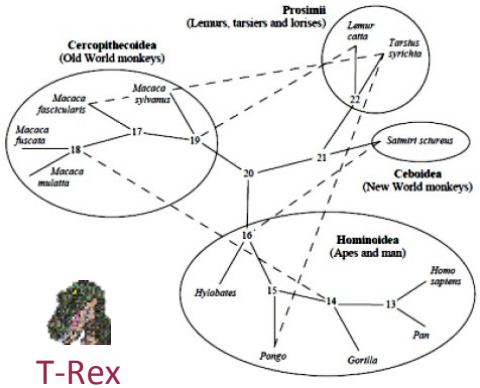


diagramme de synthèse

HorizStory



T-Rex

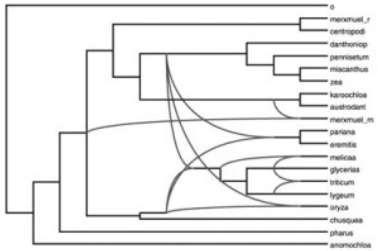
réticulogramme

Les réseaux phylogénétiques

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **explicites**

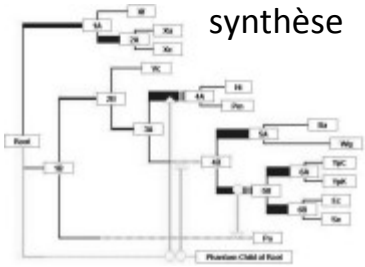
modélisation de l'évolution



réseau à une couche de réticulation

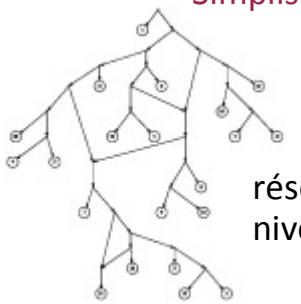
Dendroscope

diagramme de synthèse



HorizStory

Simplistic

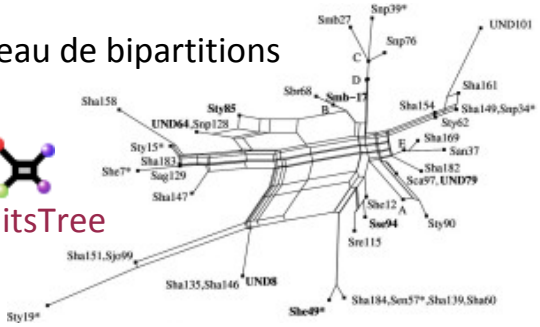


réseau de niveau 2

- réseaux phylogénétiques **abstraits**

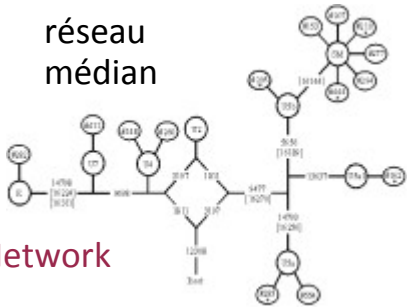
classification, visualisation de données

réseau de bipartitions



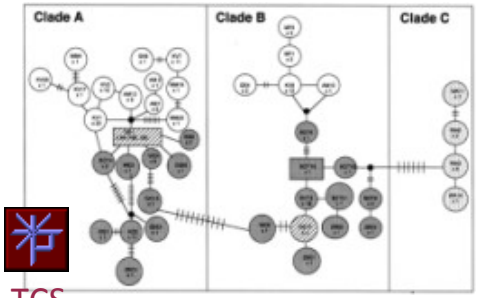
SplitsTree

réseau médian



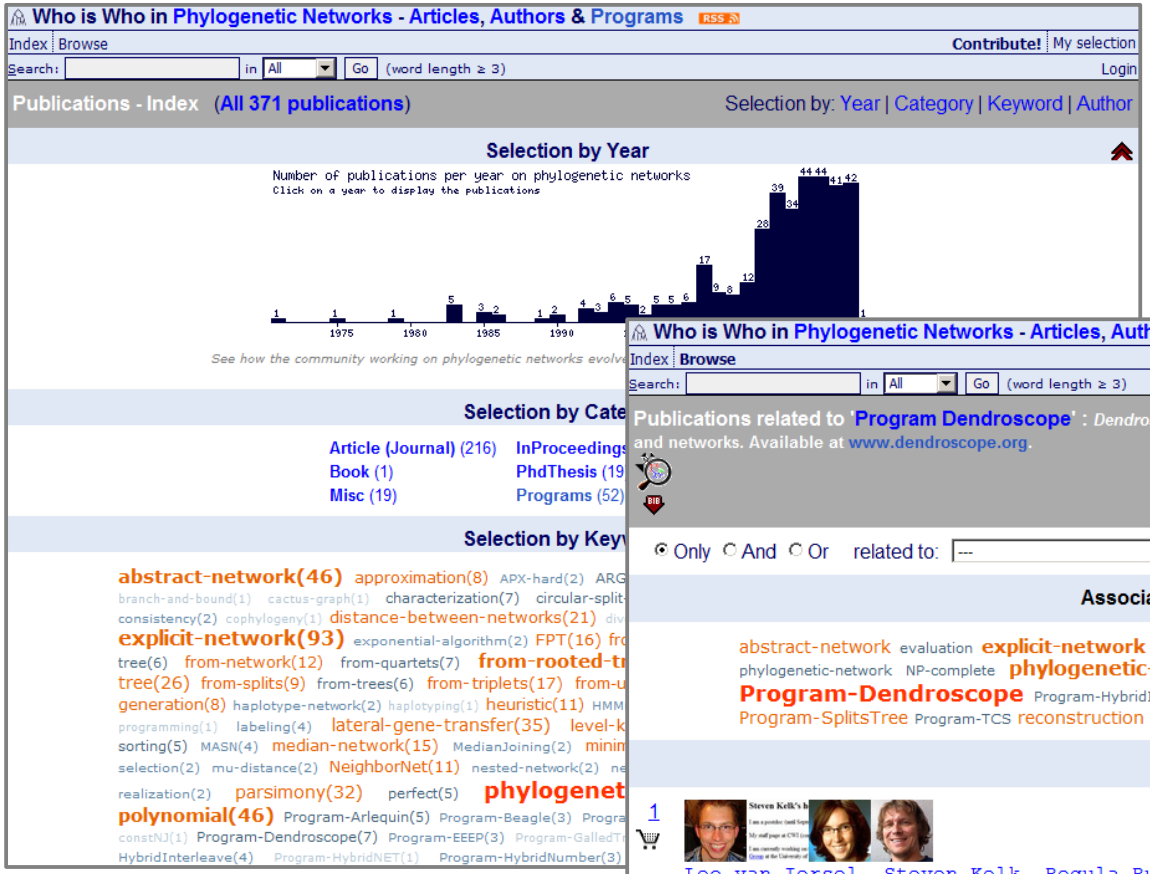
Network

réseau couvrant minimum



TCS

Plate-forme bibliographique



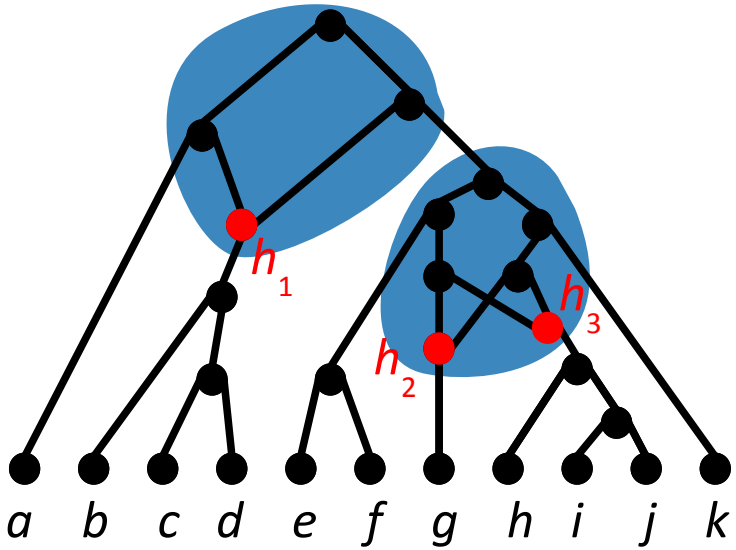
Who is Who in Phylogenetic Networks, Articles, Authors & Programs

Basé sur BibAdmin
par Sergiu Chelcea
+ nuages de mots, histogramme des dates, liste des journaux, graphes de co-auteurs, définition des mots-clés.

http://www.atgc-montpellier.fr/phylnet

Les réseaux phylogénétiques

Réseau phylogénétique explicite enraciné :



Sommets à plus d'un parent :
réticulations

Partie **non arborée** : ***blob***.

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Reconstruction de réseaux à partir de triplets
- Reconstruction de réseaux à partir de clades
- Sélection des données
- Visualisation de réseaux phylogénétiques
- Perspectives

Reconstruction de réseaux phylogénétiques

```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

méthodes de distance

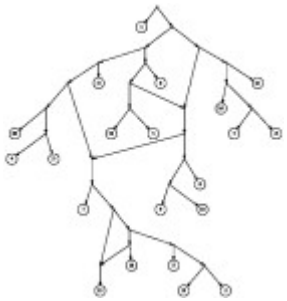
Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002

méthodes de parcimonie

Hein 1990 - Kececioglu & Gusfield 1994 - Jin, Nakhleh, Snir, Tuller 2009

méthodes de vraisemblance

Snir & Tuller 2009 - Jin, Nakhleh, Snir, Tuller 2009 - Velasco & Sober 2009



réseau *N*

Reconstruction de réseaux phylogénétiques

Problème : méthodes généralement lentes, explosion du nombre de séquences.

```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

méthodes de distance

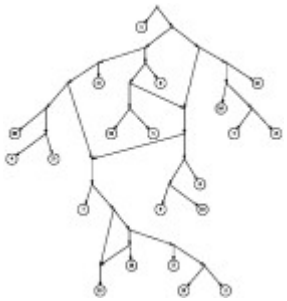
Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002

méthodes de parcimonie

Hein 1990 - Kececioglu & Gusfield 1994 - Jin, Nakhleh, Snir, Tuller 2009

méthodes de vraisemblance

Snir & Tuller 2009 - Jin, Nakhleh, Snir, Tuller 2009 - Velasco & Sober 2009



réseau *N*

Reconstruction de réseaux phylogénétiques

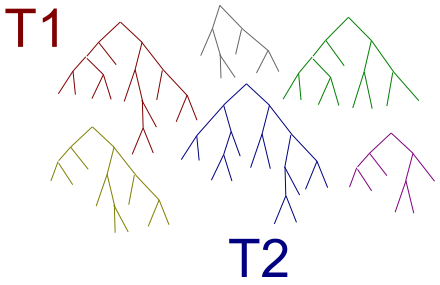
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

Reconstruction d'un arbre pour chaque gène présent chez plusieurs espèces

Guindon & Gascuel, SB, 2003



{arbres}

Base HOGENOM

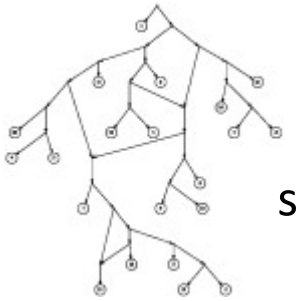
Phyl-ARIANE

ANR

Dufayard, Duret, Penel, Gouy, Rechenmann & Perrière, BioInf, 2005

Réconciliation ou consensus d'arbres

réseau explicite



super-réseau optimal *N*

Reconstruction de réseaux phylogénétiques

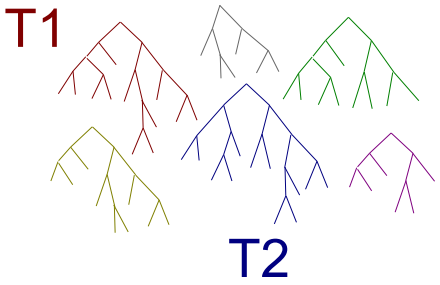
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

Reconstruction d'un arbre pour chaque gène présent chez plusieurs espèces

Guindon & Gascuel, SB, 2003



{arbres}

Base HOGENOM

Phyl-ARIANE

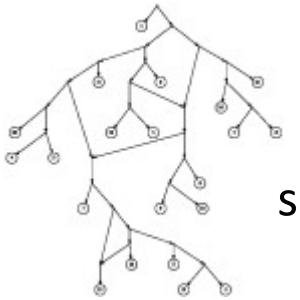
ANR

Dufayard, Duret, Penel, Gouy, Rechenmann & Perrière, BioInf, 2005

> 500 espèces, >70 000 arbres

Réconciliation ou consensus d'arbres

réseau explicite



super-réseau optimal *N*

Problème : la réconciliation d'arbres est un problème difficile

(NP-complet pour 2 arbres avec le minimum d'hybridations)

Bordewich & Semple, DAM, 2007

Triplets et quadruplets, clades et bipartitions

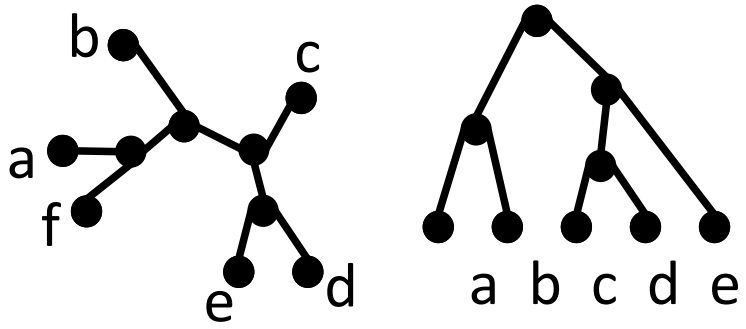
Problème :
Reconstruire le **super-réseau** d'un ensemble d'arbres est

difficile.

Idée :
reconstuire un réseau contenant tous les :

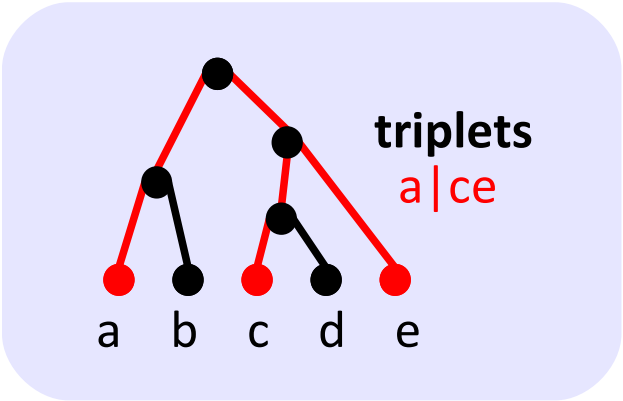
triplets
quadruplets
clades
bipartitions

des arbres en entrée ?



Triplets et quadruplets, clades et bipartitions

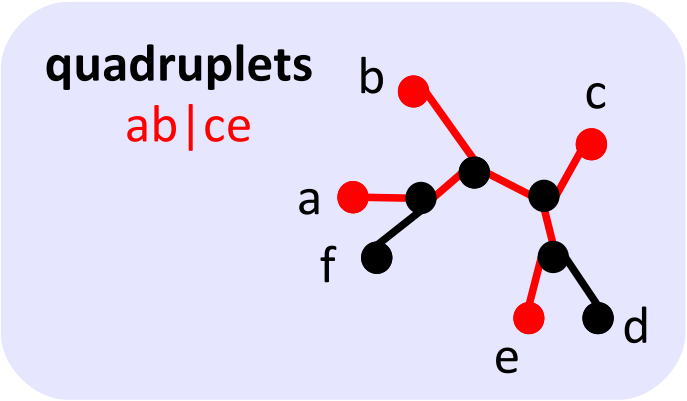
Idée :
reconstuire un réseau contenant tous les :

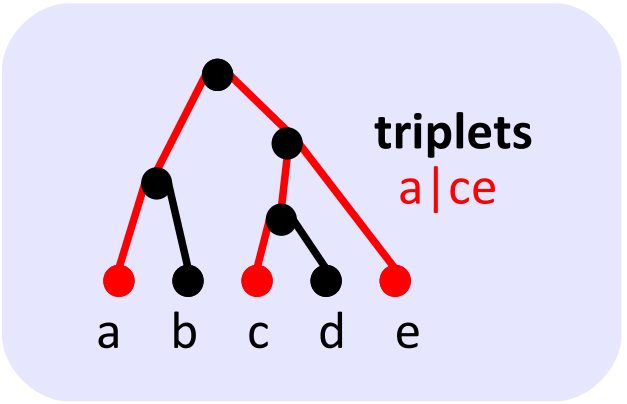


des arbres en entrée ?

Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :



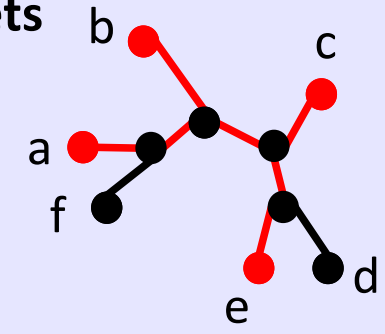


des arbres en entrée ?

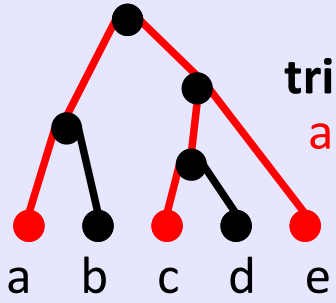
Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :

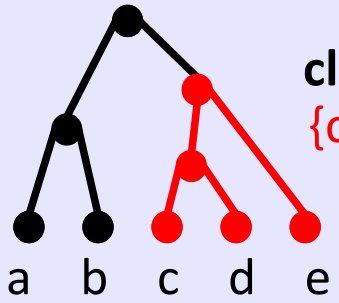
quadruplets
ab|ce



triplets
a|ce



clades
{c,d,e}

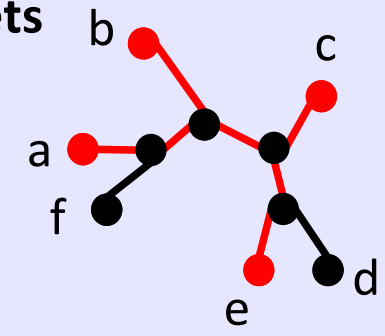


des arbres en entrée ?

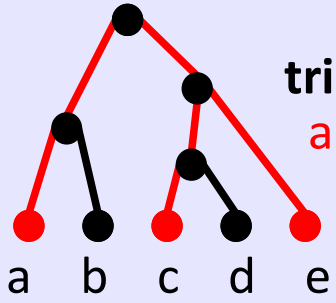
Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :

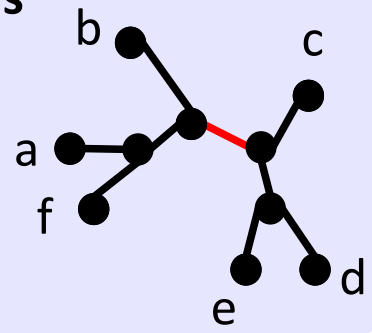
quadruplets
ab|ce



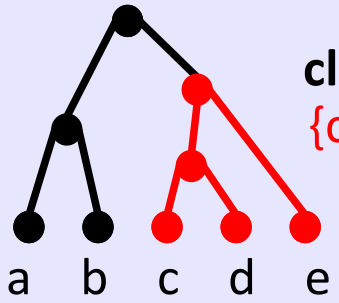
triplets
a|ce



bipartitions
{a,b,f}
{c,d,e}



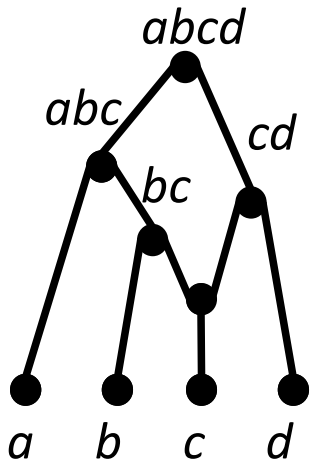
clades
{c,d,e}



des arbres en entrée ?

Clades stricts et souples

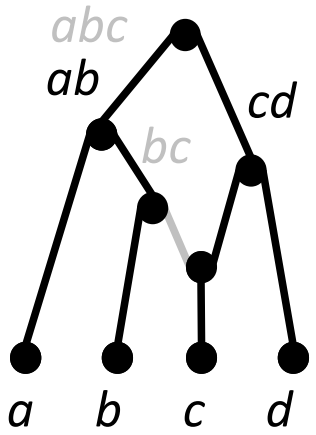
Clade “strict” : ensemble des feuilles sous un noeud du réseau



Clades stricts et souples

Clade “souple” : clade d'un arbre inclus dans le réseau

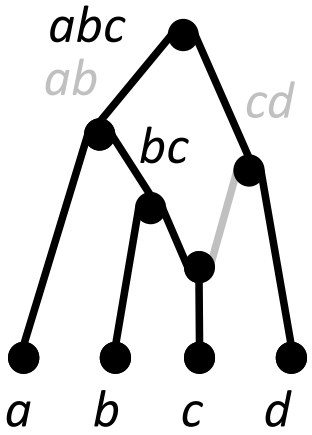
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



Clades stricts et souples

Clade “souple” : clade d'un arbre inclus dans le réseau

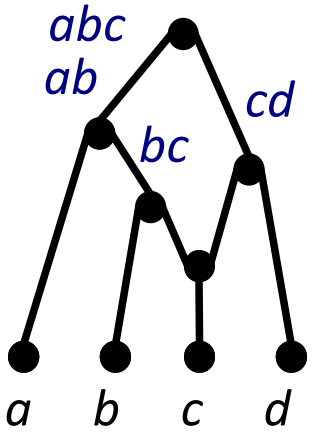
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



Clades stricts et souples

Modèle de **transmission arborée** des gènes
(gène transmis intégralement)

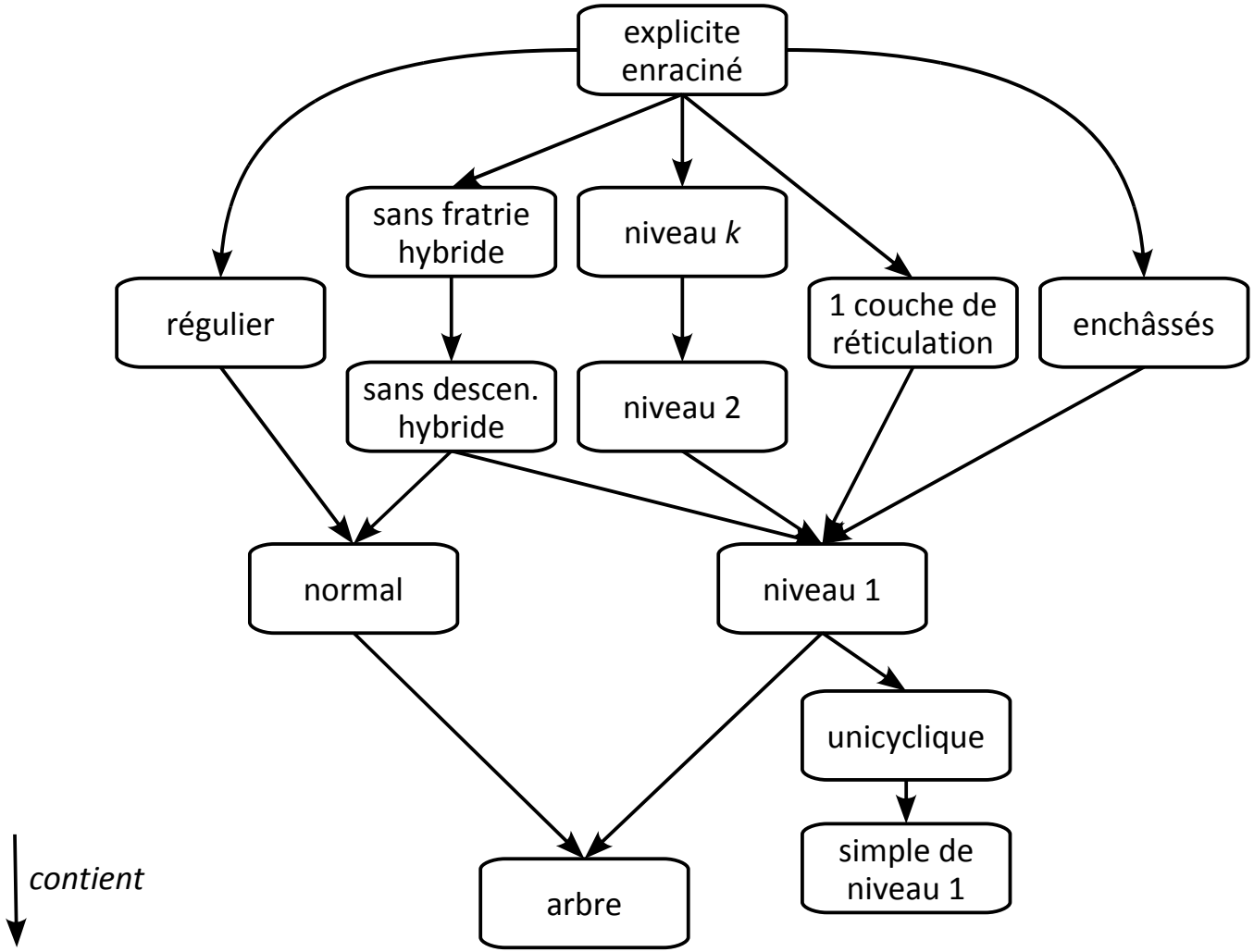
Clade “souple” : clade d'un arbre inclus dans le réseau



L'ensemble *S*(*N*) de **tous les clades souplement compatibles** avec *N* peut être de taille **exponentielle.**
Tester si un **clade souple** appartient à un réseau : **NP-complet.**

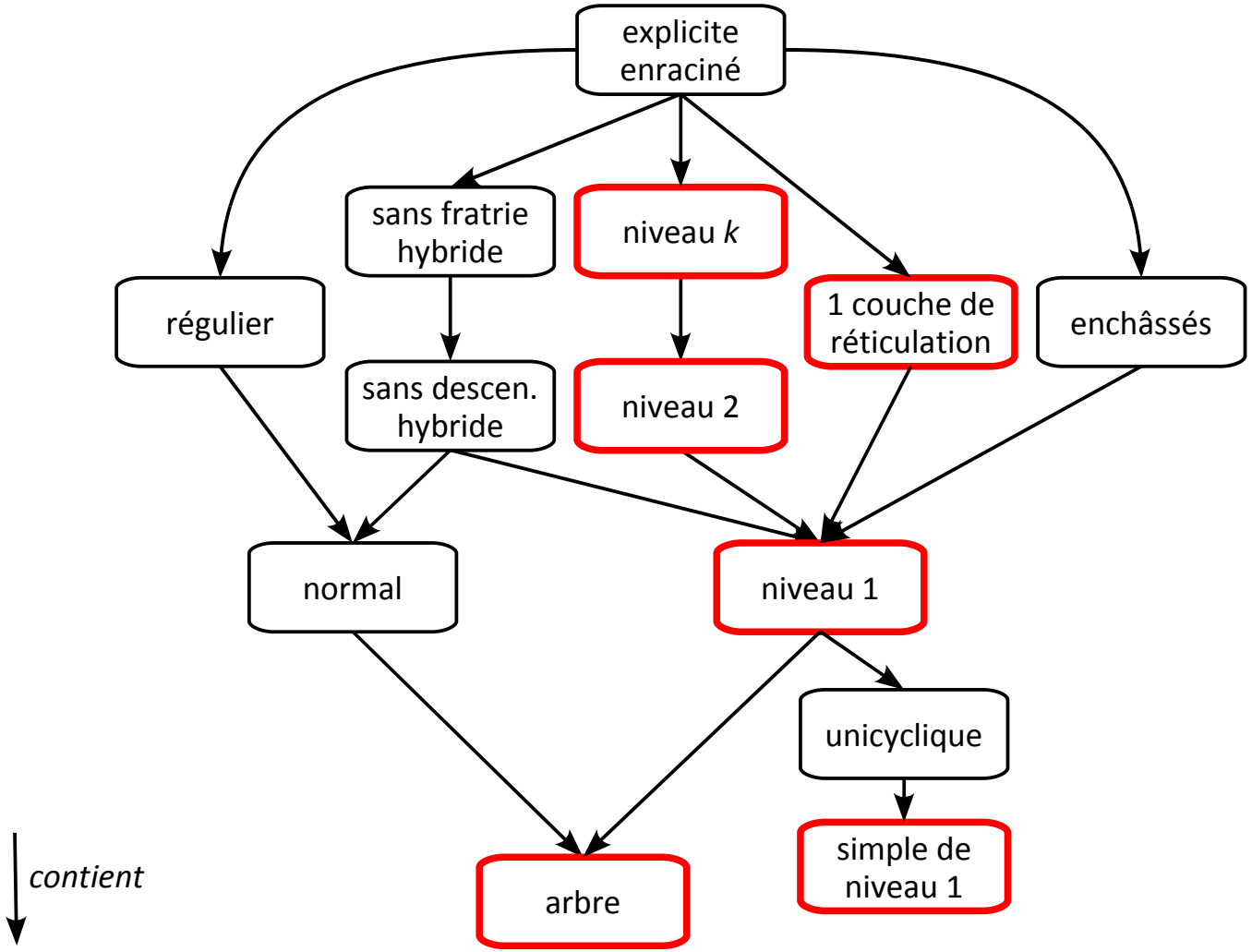
Kanj, Nakhleh, Than & Xia, TCS, 2008

Hiérarchie de sous-classes de réseaux



http://www.lirmm.fr/~gambette/RePhylogeneticNetworks.php

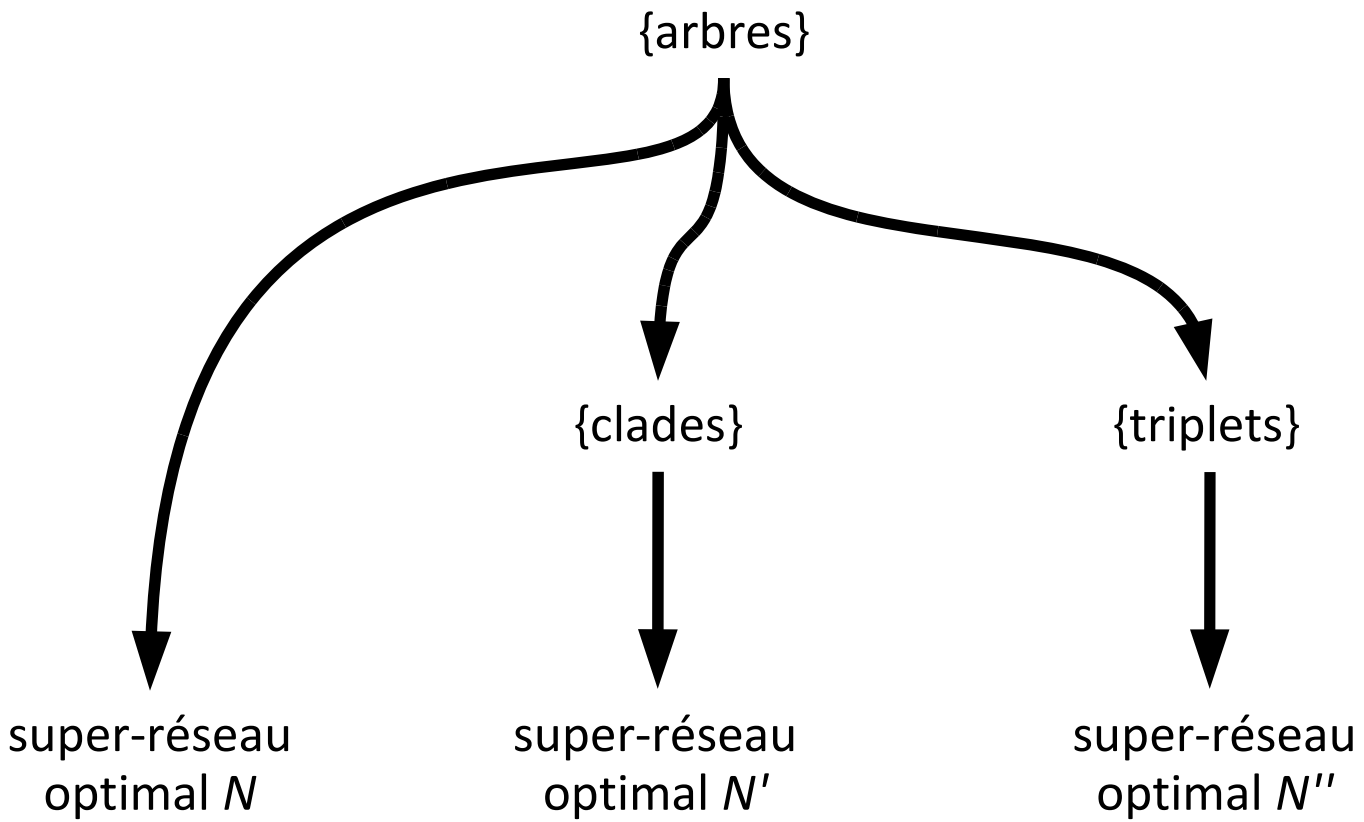
Hiérarchie de sous-classes de réseaux



http://www.lirmm.fr/~gambette/RePhylogeneticNetworks.php

Reconstruction combinatoire de réseaux phylogénétiques

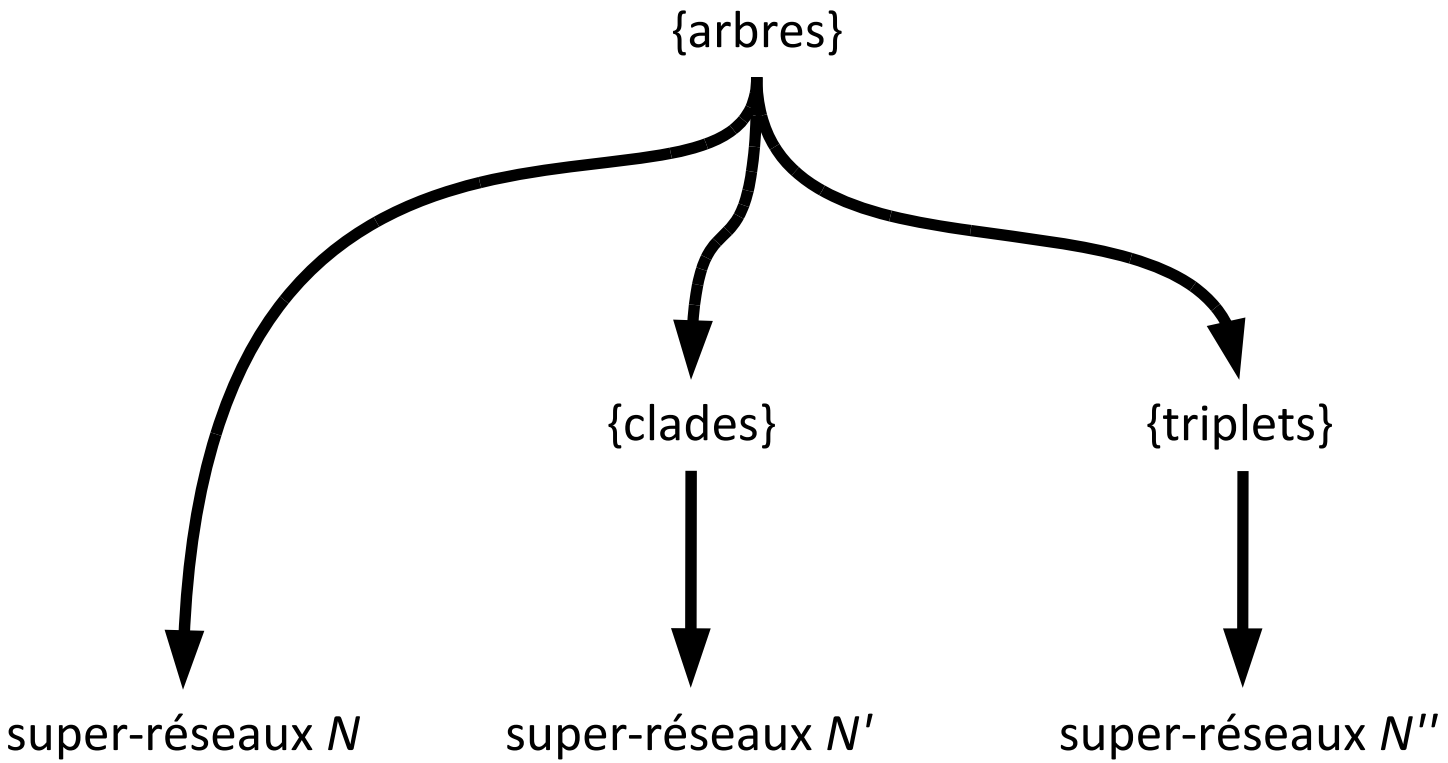
Idée :
modifier le type de données à traiter



$$N = N' = N'' \ ?$$

Reconstruction combinatoire de réseaux phylogénétiques

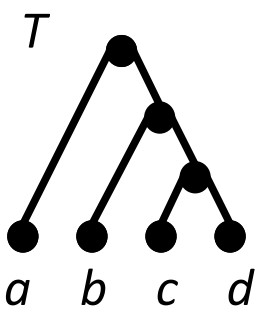
Idée :
modifier le type de données à traiter

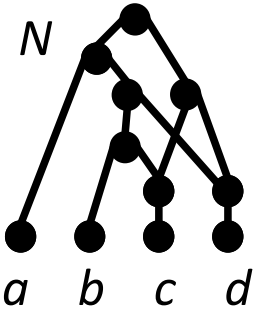


$$\{N\} \subseteq \{N'\} \subseteq \{N''\}$$

Reconstruction combinatoire de réseaux phylogénétiques

Un réseau qui contient l'ensemble de **tous les triplets ou clades d'un arbre *T*** ne contient **pas forcément *T***.



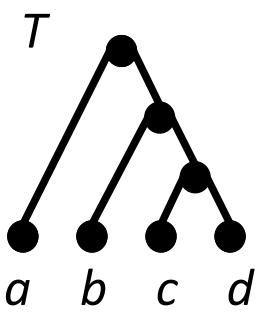


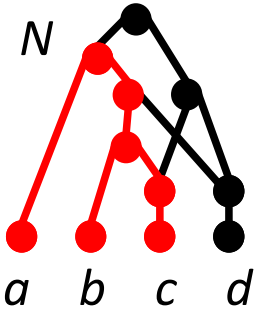
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

contient {*abcd*,*bcd*,*cd*,*a*,*b*,*c*,*d*}
mais pas *T*

Reconstruction combinatoire de réseaux phylogénétiques

Un réseau qui contient l'ensemble de **tous les triplets ou clades d'un arbre *T*** ne contient **pas forcément *T***.



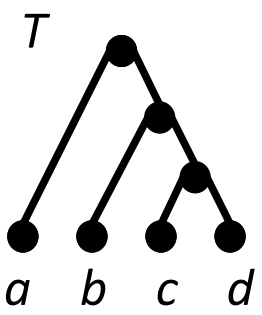


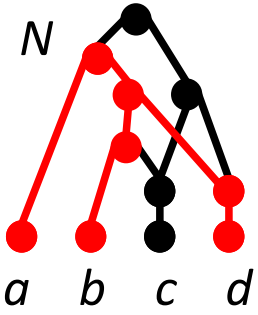
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

contient {*abcd*,*bcd*,*cd*,*a*,*b*,*c*,*d*}
mais pas *T*

Reconstruction combinatoire de réseaux phylogénétiques

Un réseau qui contient l'ensemble de **tous les triplets ou clades d'un arbre *T*** ne contient **pas forcément *T***.



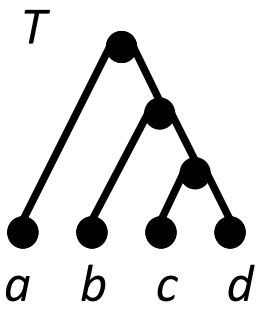


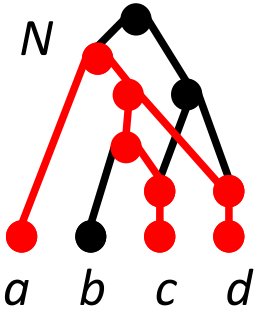
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

contient {*abcd*,*bcd*,*cd*,*a*,*b*,*c*,*d*}
mais pas *T*

Reconstruction combinatoire de réseaux phylogénétiques

Un réseau qui contient l'ensemble de **tous les triplets ou clades d'un arbre *T*** ne contient **pas forcément *T***.



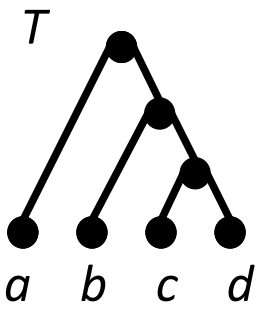


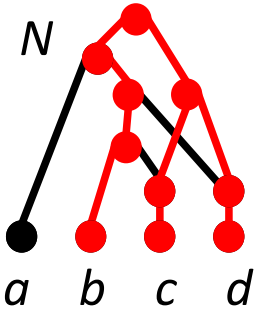
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

contient {*abcd*,*bcd*,*cd*,*a*,*b*,*c*,*d*}
mais pas *T*

Reconstruction combinatoire de réseaux phylogénétiques

Un réseau qui contient l'ensemble de **tous les triplets ou clades d'un arbre *T*** ne contient **pas forcément *T***.



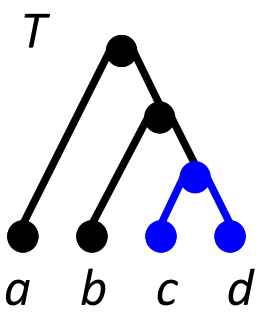


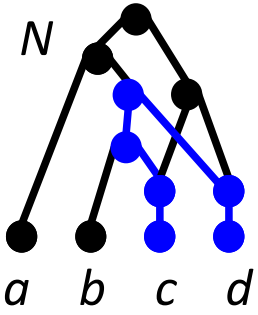
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

contient {*abcd*,*bcd*,*cd*,*a*,*b*,*c*,*d*}
mais pas *T*

Reconstruction combinatoire de réseaux phylogénétiques

Un réseau qui contient l'ensemble de **tous les triplets ou clades d'un arbre *T*** ne contient **pas forcément *T***.



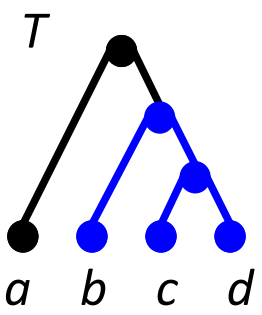


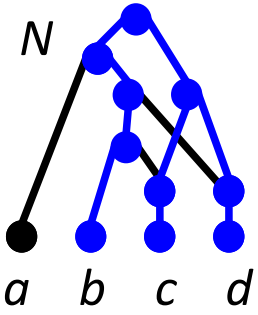
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

contient {*abcd*,*bcd*,*cd*,*a*,*b*,*c*,*d*}
mais pas *T*

Reconstruction combinatoire de réseaux phylogénétiques

Un réseau qui contient l'ensemble de **tous les triplets ou clades d'un arbre *T*** ne contient **pas forcément *T***.



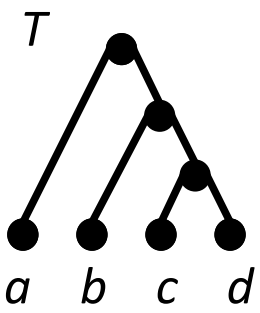


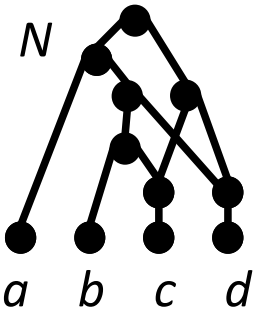
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

contient {*abcd*,*bcd*,*cd*,*a*,*b*,*c*,*d*}
mais pas *T*

Reconstruction combinatoire de réseaux phylogénétiques

Un réseau qui contient l'ensemble de **tous les triplets ou clades d'un arbre *T*** ne contient **pas forcément *T***.

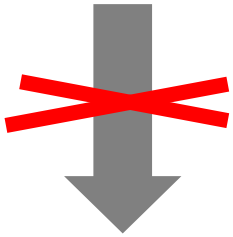




contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

contient {*abcd*,*bcd*,*cd*,*a*,*b*,*c*,*d*}
mais pas *T*

contient les clades / triplets d'un arbre *T*

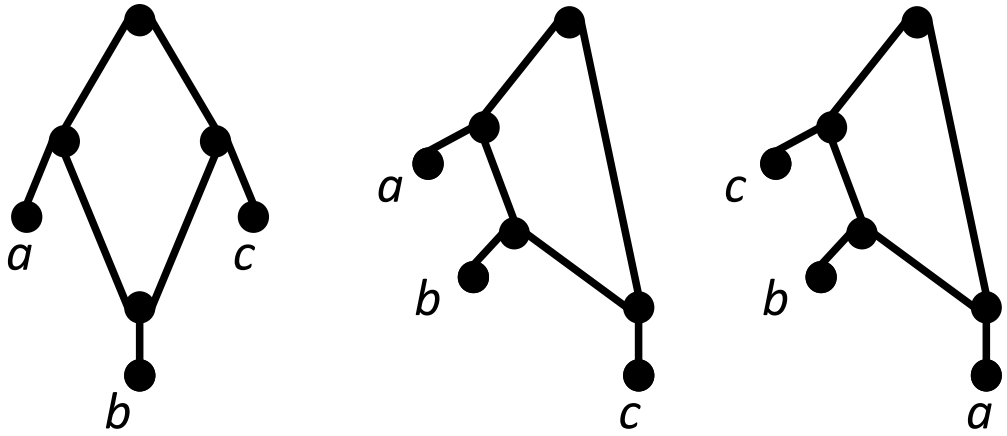


contient *T*.

Ambiguïté des solutions

Ambiguïté de la reconstruction, même à partir de données complètes et correctes

Plusieurs réseaux minimaux **distincts** ont exactement
le **même ensemble** d'arbres, de triplets, de clades.



Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 somments)

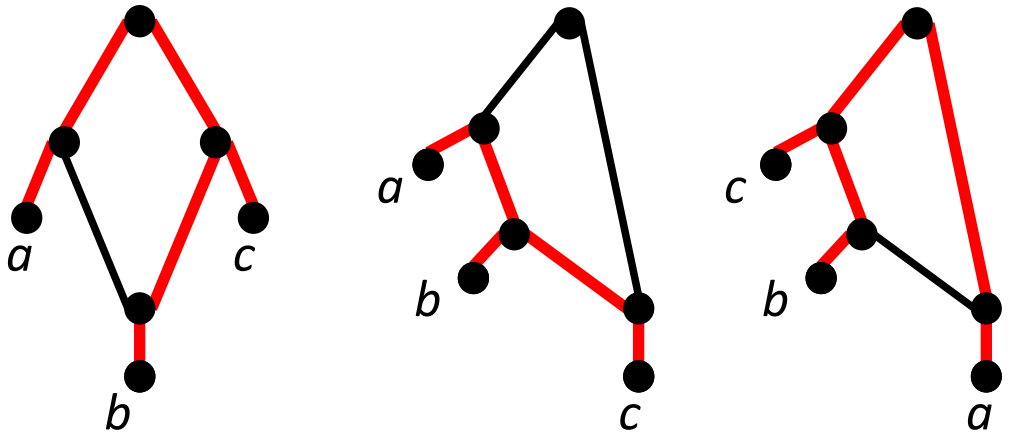
Gambette & Huber, 2010

Ambiguïté des solutions

Ambiguïté de la reconstruction, même à partir de données complètes et correctes

Plusieurs réseaux minimaux **distincts** ont exactement
le **même ensemble** d'arbres, de triplets, de clades.

a|bc



Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 somments)

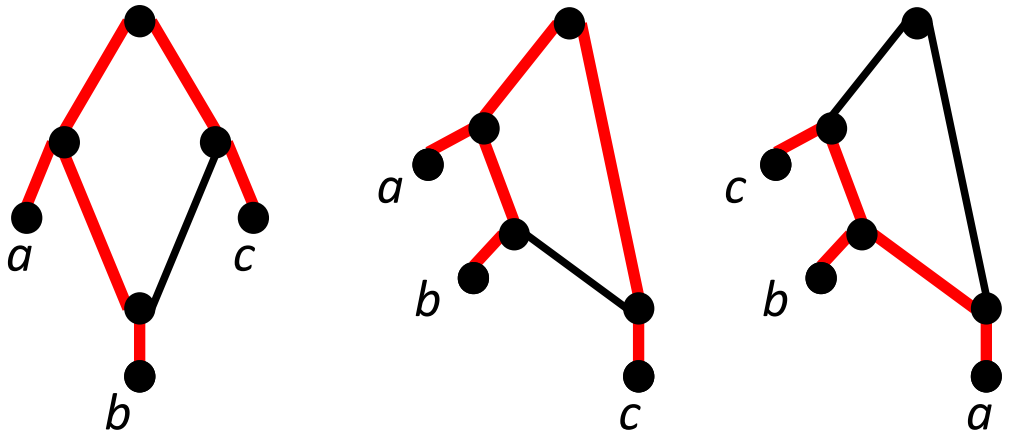
Gambette & Huber, 2010

Ambiguïté des solutions

Ambiguïté de la reconstruction, même à partir de données complètes et correctes

Plusieurs réseaux minimaux **distincts** ont exactement
le **même ensemble** d'arbres, de triplets, de clades.

c|ab



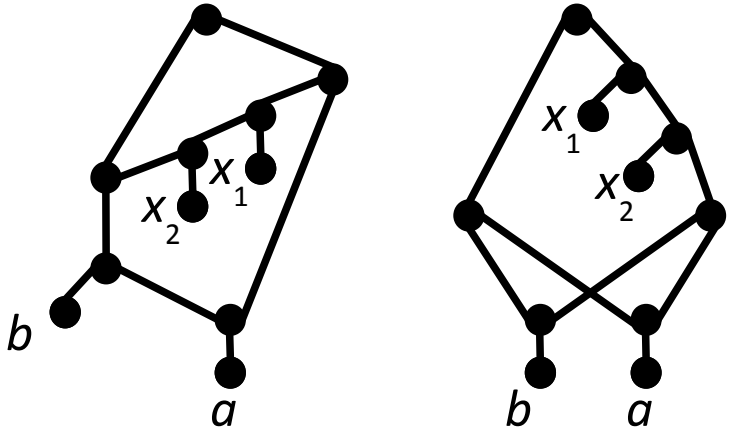
Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 somments)

Gambette & Huber, 2010

Ambiguïté des solutions

Ambiguïté de la reconstruction, même à partir de données complètes et correctes

Plusieurs réseaux minimaux **distincts** ont exactement
le **même ensemble** d'arbres, de triplets, de clades.



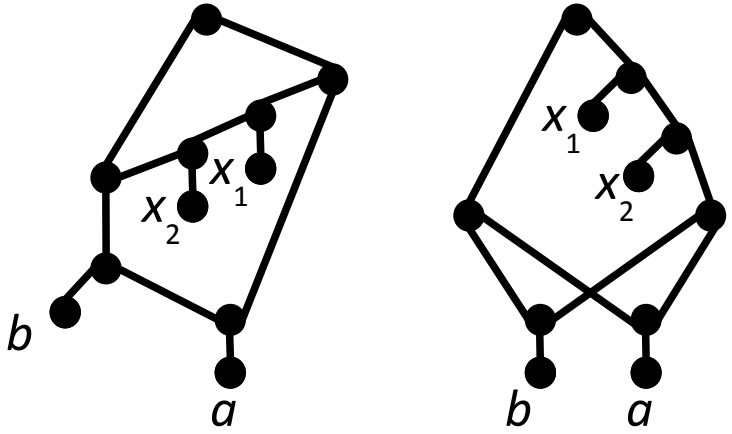
2 réseaux de niveau 2 avec le même ensemble de triplets

Gambette & Huber, 2010

Ambiguïté des solutions

Ambiguïté de la reconstruction, même à partir de données complètes et correctes

Plusieurs réseaux minimaux **distincts** ont exactement
le **même ensemble** d'arbres, de triplets, de clades.



2 réseaux de niveau 2 avec le même ensemble de triplets

Même avec des données de départ **complètes** et **correctes**,
impossible de choisir entre les formes ambiguës !

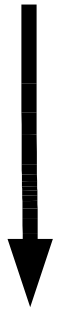
Gambette & Huber, 2010

Plan

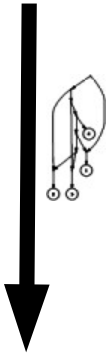
- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Reconstruction de réseaux à partir de triplets
- Reconstruction de réseaux à partir de clades
- Sélection des données
- Visualisation de réseaux phylogénétiques
- Perspectives

Reconstruction depuis les triplets

{arbres}



{triplets}



N'
réseau
de niveau *k*

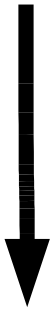
Reconstruction d'un réseau de **niveau** *k* à partir d'un ensemble de **triplets**

Jansson, Nguyen & Sung, JOC, 2006 : NP-complet pour niveau 1,
Van Iersel, Kelk & Mnich, JBCB, 2009 : NP-complet pour niveau *k*

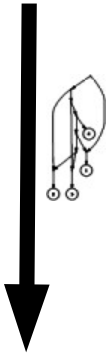
niveau =
mesure de “complexité”, d'éloignement par rapport à une structure d'arbre.

Reconstruction depuis les triplets

{arbres}



{triplets}

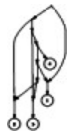


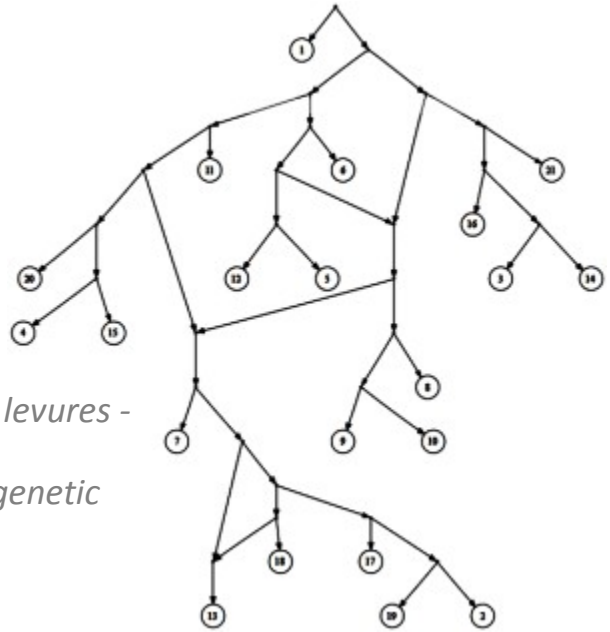
N'
réseau
de niveau *k*

Méthodes exactes rapides pour reconstruire un **réseau de niveau 1 et 2** (s'il en existe un) à partir d'un ensemble **dense** de **triplets**

Jansson, Nguyen & Sung, SODA'05 : O(n^3) pour niveau 1,
van Iersel, Kelk & al, RECOMB'08 : O(n^8) pour niveau 2,
To & Habib, CPM'09 : $O(n^{5k+4})$ pour niveau k
Van Iersel & Kelk, J. Theor. Biol., 2011 : NP-complet de trouver le niveau minimal

dense =
sur chaque ensemble de 3 feuilles, au moins 1 triplet existe dans *T*.

Programme Simplistic 

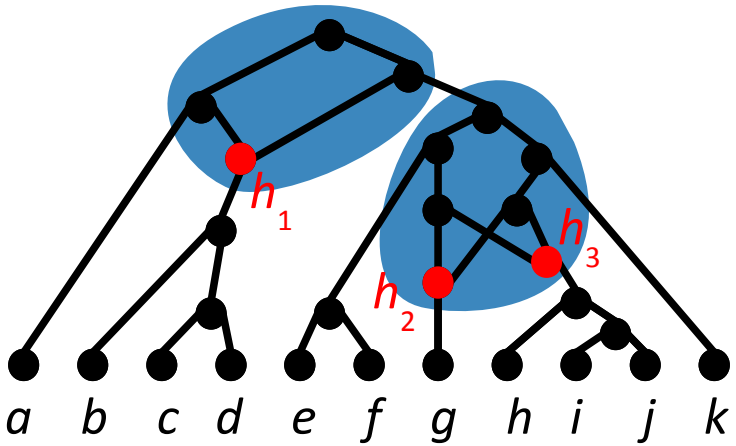


Réseau phylogénétique de levures - Van Iersel et al. : Constructing level-2 phylogenetic networks from triplets. RECOMB 2008

http://homepages.cwi.nl/~kelk/simplistic.html

Réseaux phylogénétiques de niveau *k*

Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



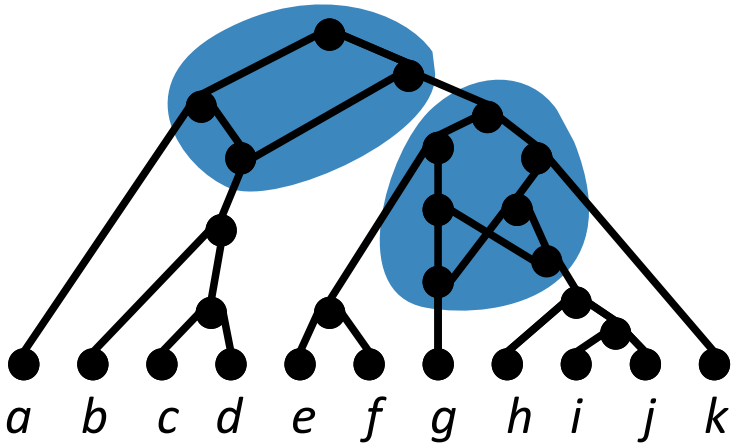
réseau de niveau 2

niveau =
nombre maximum d'hybridations par partie non arborée (*blob*).

Choy, Jansson, Sadakane & Sung, TCS, 2005

Réseaux phylogénétiques de niveau *k*

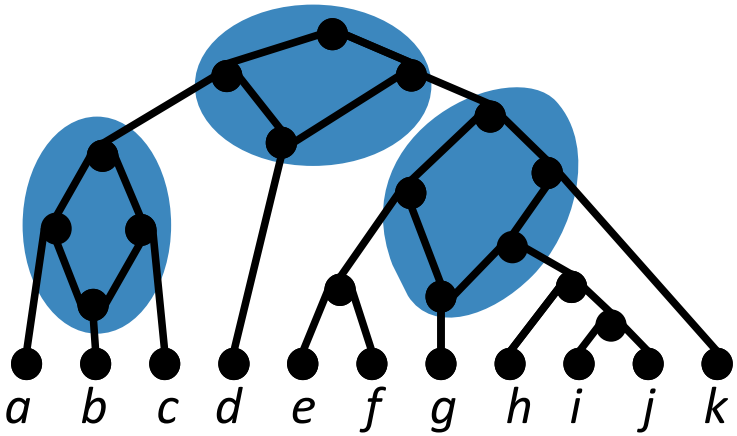
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau de niveau 2

niveau =
nombre maximum d'hybridations
par partie non arborée (*blob*).

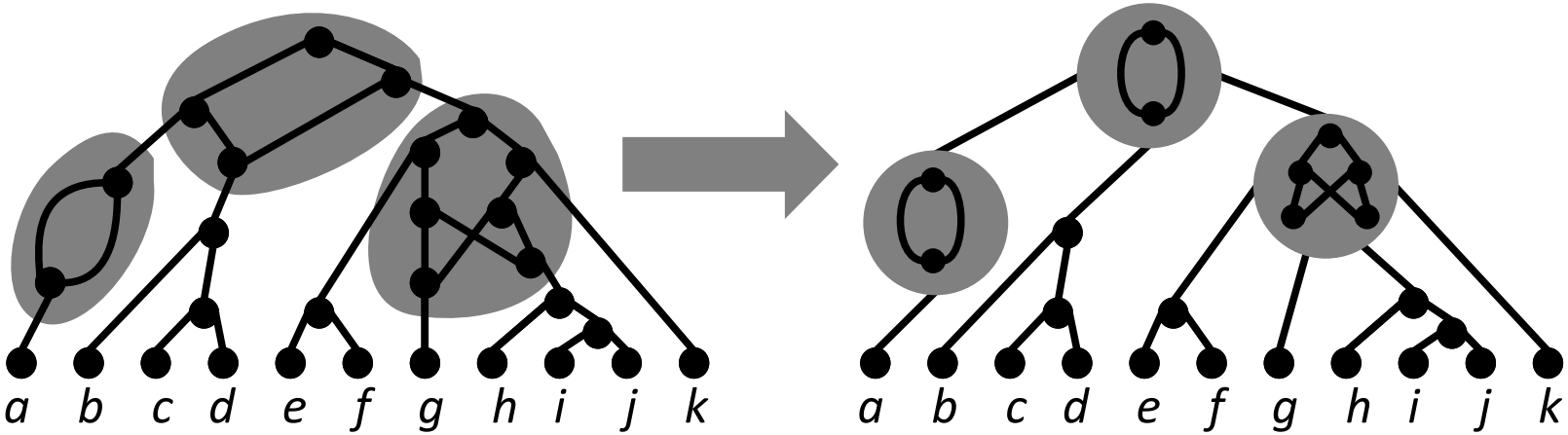
réseau de niveau 1
(“*galled tree*”)



Gusfield, Eddhu & Langley, IJOC, 2004

Décomposition des réseaux de niveau *k*

Décomposition en blobs :



N, réseau de niveau *k*.

décomposition **arborée** de *N* en **générateurs**.

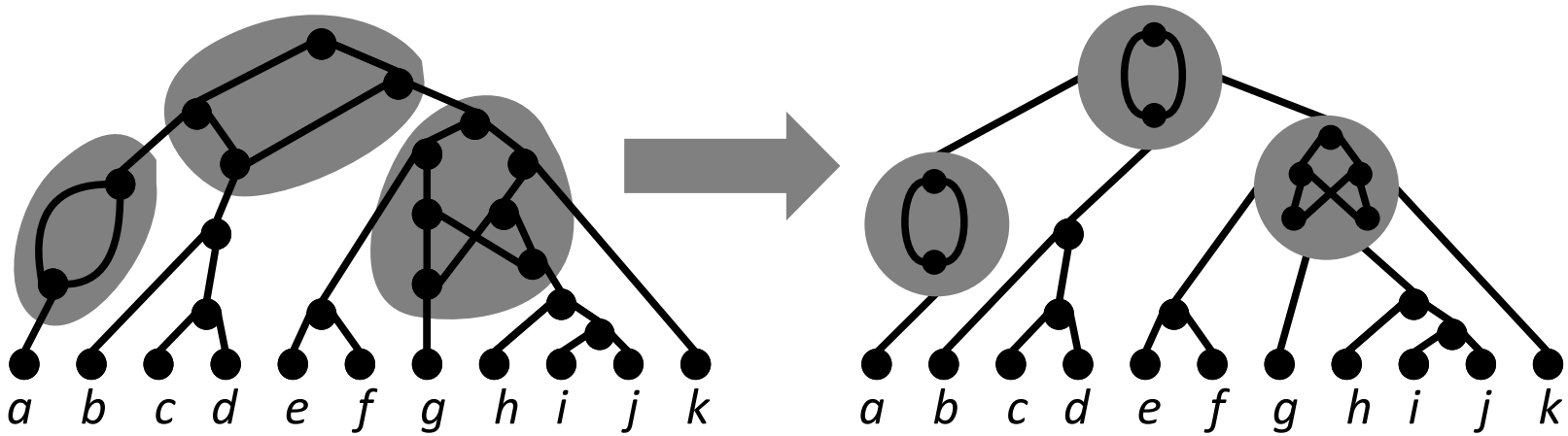
Gambette, Berry & Paul, CPM 2009

Intérêts :

- Dénombrer les réseaux de niveau *k* (taille de l'espace de recherche)
- Générer des réseaux de niveau *k* (simulation, parcours de l'espace de recherche)

Décomposition des réseaux de niveau *k*

Décomposition en blobs :



N, réseau de niveau *k*.

décomposition arborée de *N* en **générateurs**.

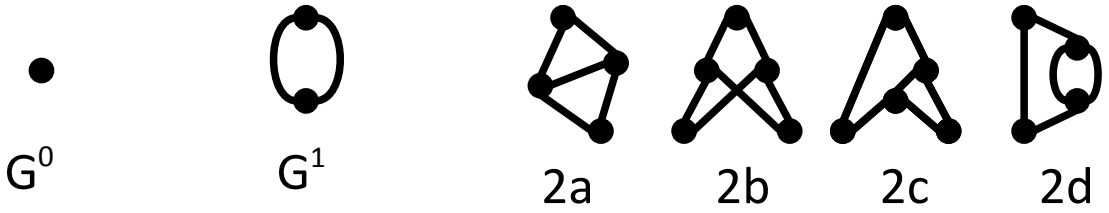
Gambette, Berry & Paul, CPM 2009

Générateurs initialement introduits pour la classe restreinte des réseaux *simples* de niveau *k*
Analyse de cas pour trouver les 4 générateurs de niveau 2
Force brute pour les 65 générateurs de niveau 3

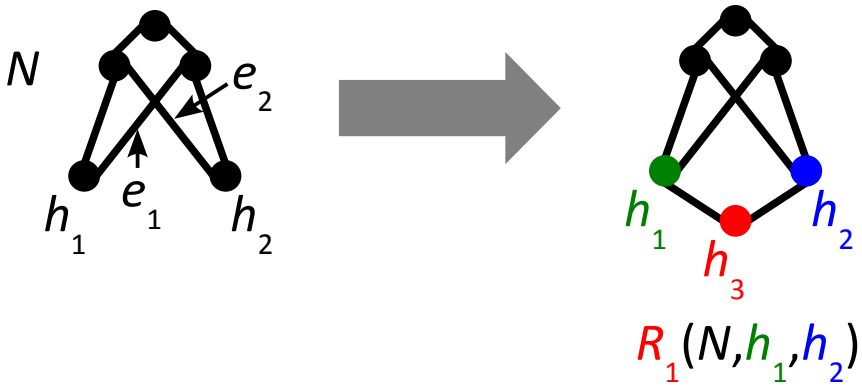
Van Iersel et al., RECOMB 2008
http://homepages.cwi.nl/~kelk/lev3gen

Décomposition des réseaux de niveau *k*

Générateur de niveau *k* :
réseau de niveau *k* sans isthme (arc dont la suppression déconnecte le réseau).



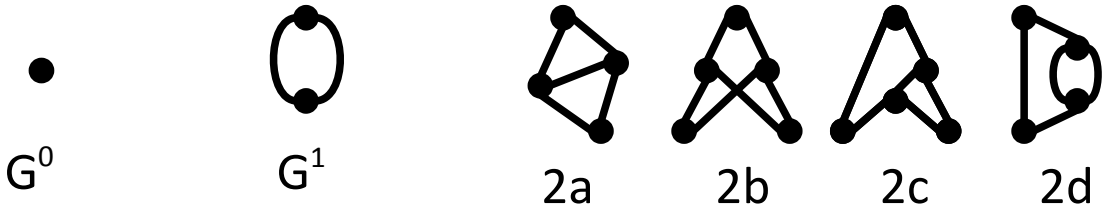
Règles de construction des générateurs de niveau *k*+1 à partir de ceux de niveau *k*



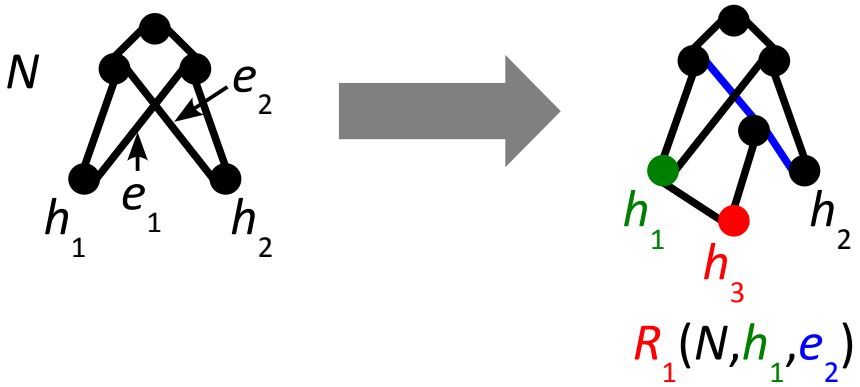
Gambette, Berry & Paul, CPM 2009

Décomposition des réseaux de niveau *k*

Générateur de niveau *k* :
réseau de niveau *k* sans isthme (arc dont la suppression déconnecte le réseau).



Règles de construction des générateurs de niveau *k*+1 à partir de ceux de niveau *k*

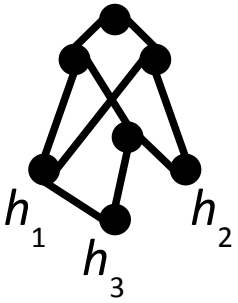


Gambette, Berry & Paul, CPM 2009

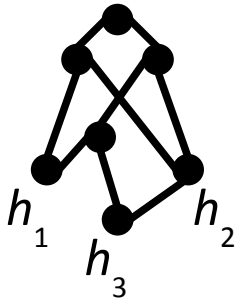
Construction des générateurs

Problème !

Certains des générateurs de niveau k+1 obtenus depuis ceux de niveau k sont isomorphes !



$R_1(N,h_1,e_2)$



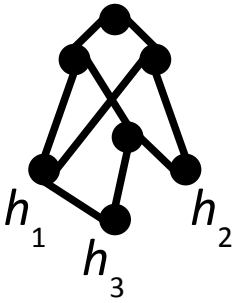
$R_1(N,h_2,e_1)$

Gambette, Berry & Paul, CPM 2009

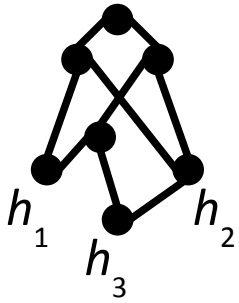
Construction des générateurs

Problème !

Certains des générateurs de niveau k+1 obtenus depuis ceux de niveau k sont isomorphes !



$R_1(N,h_1,e_2)$



$R_1(N,h_2,e_1)$

→ comptage difficile !

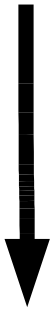
→ génération possible jusqu'à niveau 5 :
1, 4, 65, 1993, 91454

Gambette, Berry & Paul, CPM 2009

http://www.lirmm.fr/~gambette/ProgGenerators.php

Reconstruction depuis les triplets

{arbres}



{triplets}

N'
réseau
de
niveau *k*

Méthodes pour reconstruire un **réseau de niveau *k*** qui contient le **le plus de triplets** parmi ceux d'un ensemble **dense *R*** en entrée.

k = 0 :
Problème **NP-complet**

Bryant 1997, Jansson 2001, Wu 2004

Algorithme exponentiel

$O((|R|+n^2)3^n)$, Wu 2004

k = 1 :
Algorithme exponentiel

Temps $O(|R|4^n)$, espace $O(n3^n)$, Van Iersel, Kelk & Mnich, JBCB, 2009

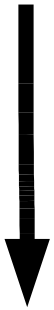
Heuristique Lev1athan

Huber, Van Iersel, Kelk & Suchecki, TCBB, 2011

http://homepages.cwi.nl/~kelk/lev1athan/

Reconstruction depuis les triplets

{arbres}



{triplets}

N' réseau de niveau *k*

Méthodes pour reconstruire un **réseau de niveau *k*** qui contient le **le plus de triplets** parmi ceux d'un ensemble **dense *R*** en entrée.

Algorithmes **FPT en** le nombre de triplets *t* de *R* à éditer

k = 0 :

R compatible avec un arbre ⟺ *R* ne contient aucune des obstructions

{a|bc,c|ab} {a|bc,c|bd,d|ab} {a|bc,c|bd,d|ac} {a|bc,a|bd,d|ac}

Arbre de recherche bornée : $O(6^t n+n^4)$

Gambette, Berry, Paul, 2008

Amélioration des obstructions + noyau : $2^{O(t^{1/3 \log t})} + O(n^4)$

Guillemot & Mnich, TAMC 2010

Noyau linéaire

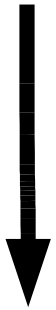
Paul, Perez & Thomassé, 2011

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Reconstruction de réseaux à partir de triplets
- Reconstruction de réseaux à partir de clades
- Sélection des données
- Visualisation de réseaux phylogénétiques
- Perspectives

Reconstruction depuis les clades souples

{arbres}



{clades}

N'

réseau à 1 couche de réticulation

Consensus de clades souples :
Dendroscope

Huson et al., BMCB, 2007

Méthode exacte rapide de reconstruction de **réseaux à 1 couche de réticulation** à partir de **clades souples**

Huson, Rupp, Berry, Gambette & Paul, ISMB 2009

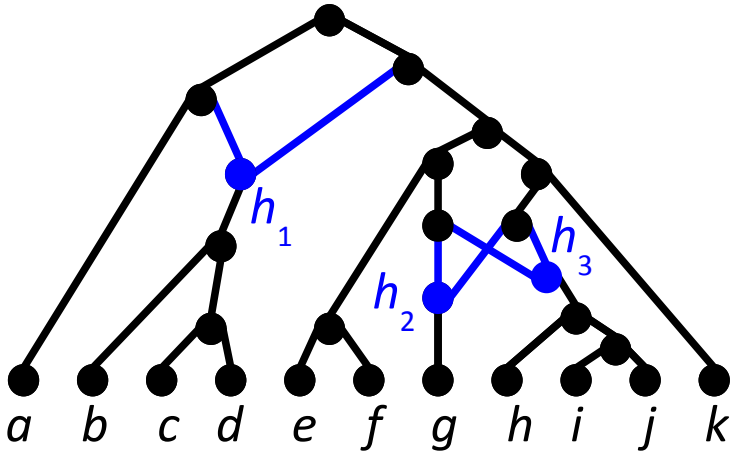
Méthode exacte de reconstruction de **réseaux de niveau *k*** à partir de **clades souples**

Iersel, Kelk, Rupp & Huson, ISMB 2010

- meilleurs résultats mais plus lente pour niveau > 2.
- pour *k* fixé, certains ensembles de clades contenus dans aucun réseau de niveau *k*.

Réseaux phylogénétiques à une couche de réticulations

Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.

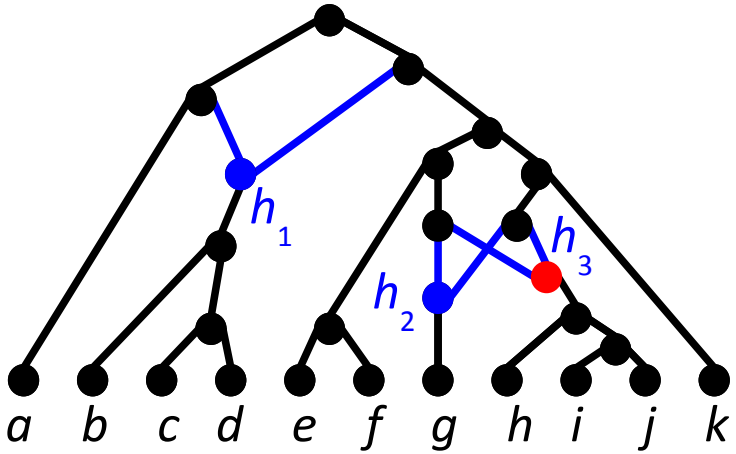


réseau à une couche de réticulation.

réseau **à une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

Réseaux phylogénétiques à une couche de réticulations

Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.

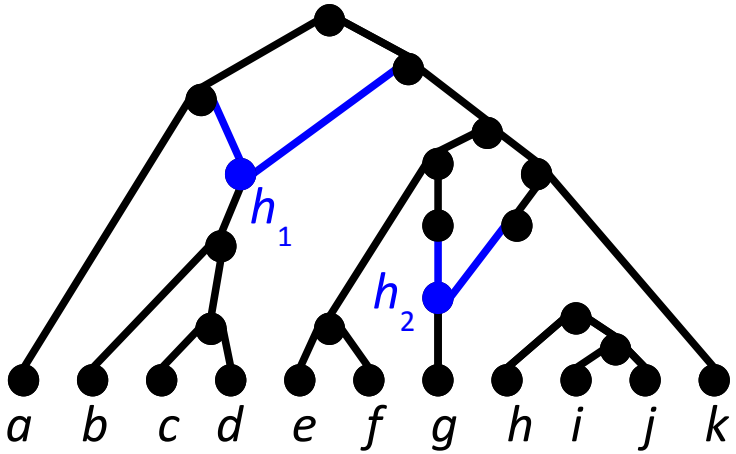


réseau à une couche de réticulation.

réseau **à une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

Réseaux phylogénétiques à une couche de réticulations

Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.

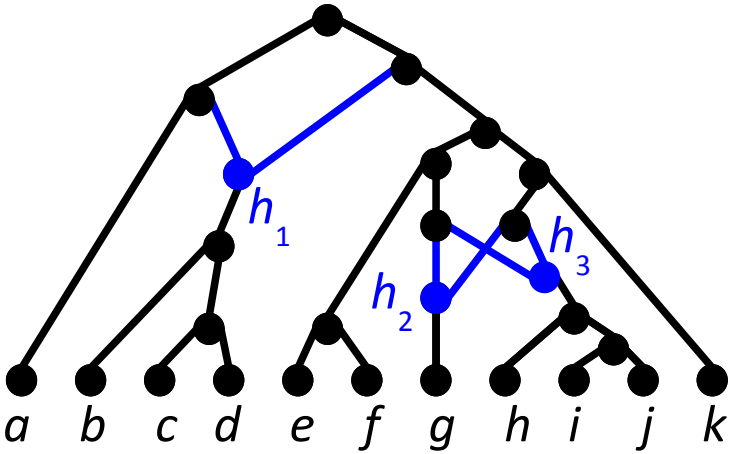


réseau à une couche de réticulation.

réseau **à une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

Réseaux phylogénétiques à une couche de réticulations

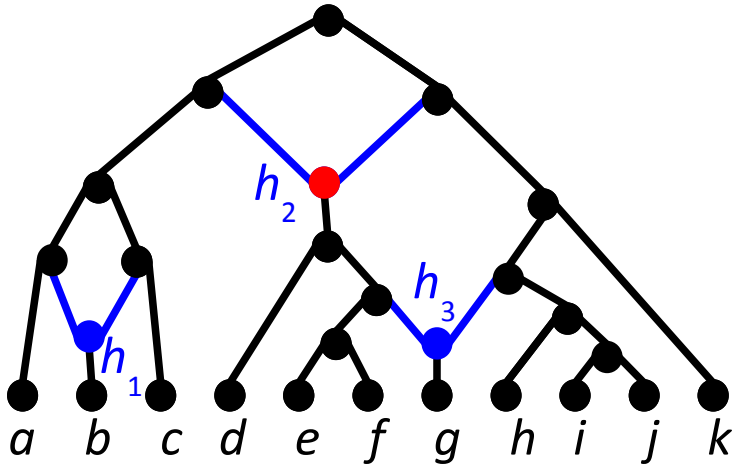
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau à une couche de réticulation.

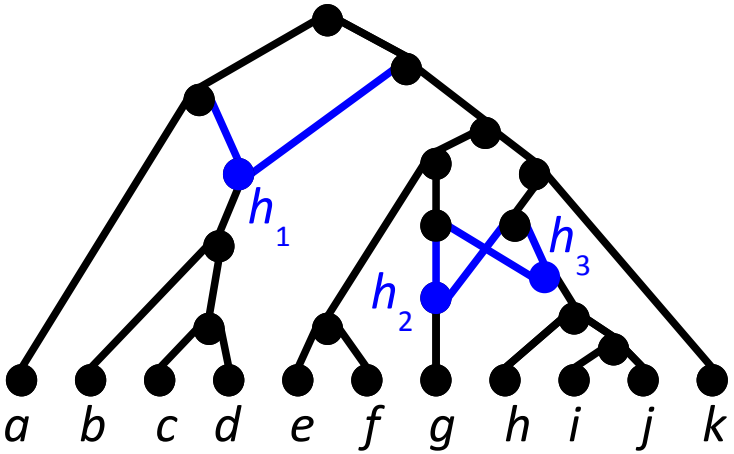
réseau **à une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à deux couches de réticulation.



Réseaux phylogénétiques à une couche de réticulations

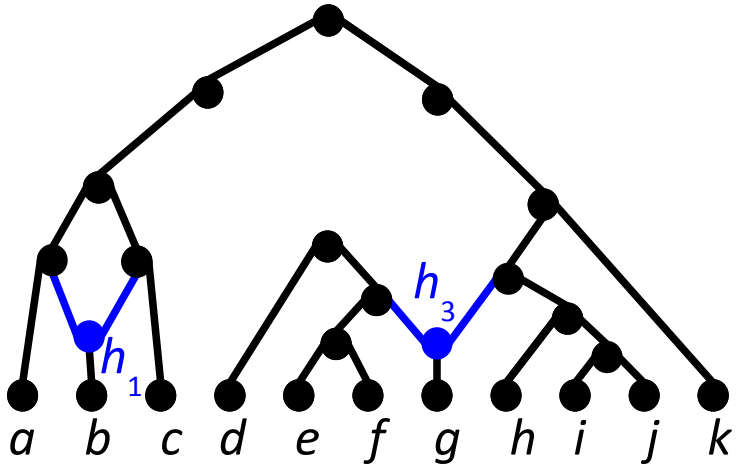
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau à une couche de réticulation.

réseau **à une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

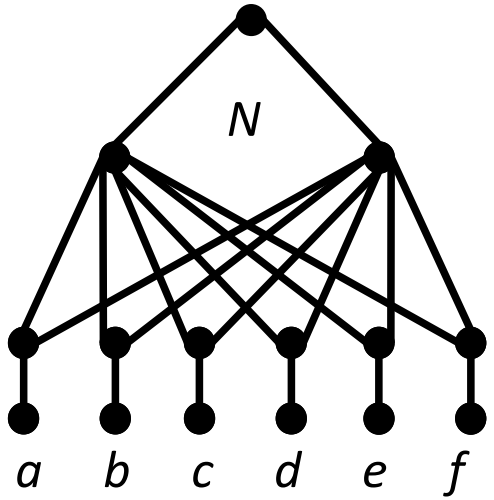
réseau à deux couches de réticulation.



Clades et réseaux à une couche de réticulation

Test de compatibilité souple **polynomial** sur les réseaux à une couche de réticulation.

Pour tout ensemble *C* de clades, il existe un **réseau à une couche de réticulation compatible** avec *C***.**



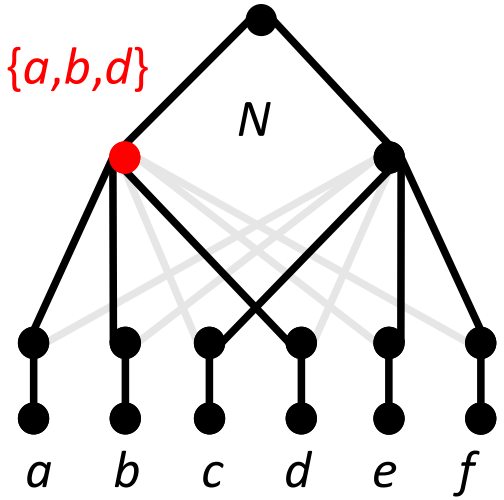
Le réseau à une couche de réticulation *N* est compatible avec
tout clade souple sur {*a*,*b*,*c*,*d*,*e*,*f*}.

Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

Clades et réseaux à une couche de réticulation

Test de compatibilité souple **polynomial** sur les réseaux à une couche de réticulation.

Pour tout ensemble *C* de clades, il existe un **réseau à une couche de réticulation compatible** avec *C***.**



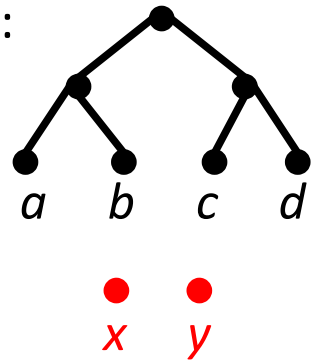
Le réseau à une couche de réticulation *N* est compatible avec
tout clade souple sur {*a*,*b*,*c*,*d*,*e*,*f*}.

Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

Une approche en deux étapes

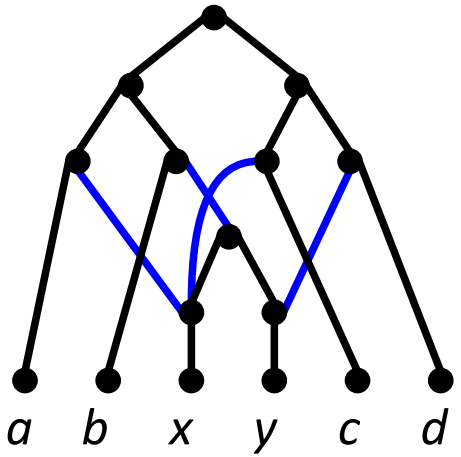
1- Trouver un **ensemble minimum de conflits** parmi les clades :
- partie sans conflits ➡ arbre,
- taxons impliqués dans des conflits ➡ sous les réticulations.

MAXIMUM COMPATIBLE SUBSET



2- Attacher à l'arbre les taxons impliqués dans des conflits avec un **nombre minimal d'arcs** :

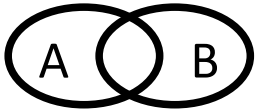
MINIMUM ATTACHMENT PROBLEM



Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'ensemble minimum de conflits

Conflit : clades ni inclus ni disjoints



Problème :
enlever un nombre minimum *t* de taxons pour supprimer tous les conflits entre les clades de *C*.

NP-complet dans le cas général

Steel & Hamel, AML, 1996

NP-complet sur un graphe connexe, sans taxons "jumeaux"

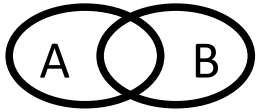
réduction depuis le cas général

Algorithme **FPT** de branchement en $O(3^t.n|C|^2)$ implémenté dans Dendroscope

Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

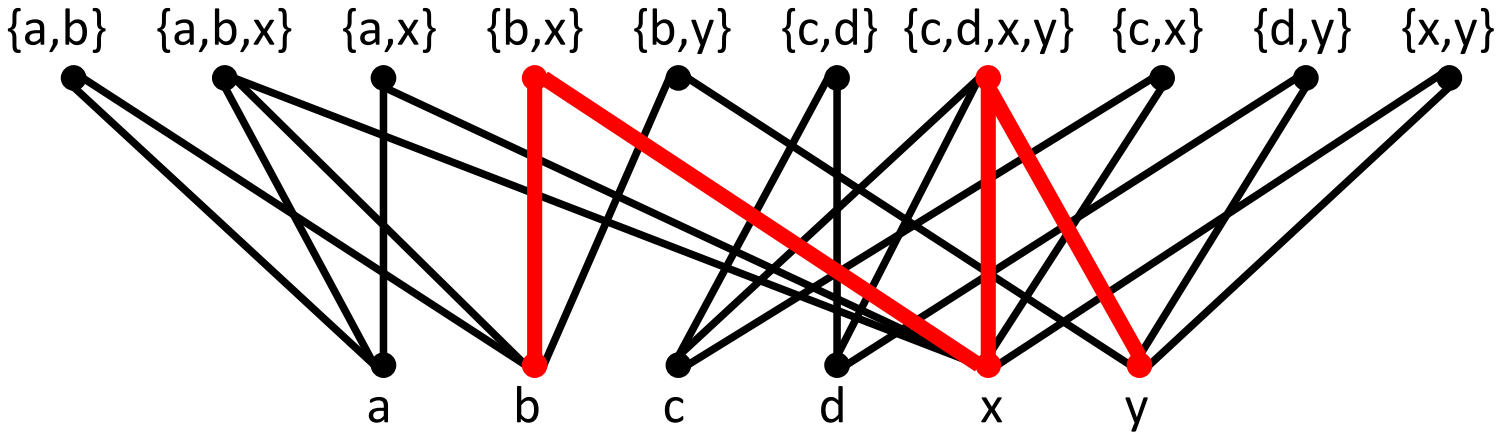
L'ensemble minimum de conflits

Conflit : clades ni inclus ni disjoints



Graphe des caractères d'un ensemble de clades,
graphe biparti avec :
- un ensemble de sommets pour les clades
- un ensemble de sommets pour les taxons
- arête quand le taxon appartient au clade

Exemple : {{a,b},{a,b,x},{a,x},{b,x},{b,y},{c,d},{c,d,x,y},{c,x},{d,y},{x,y}}

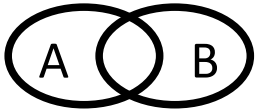


conflit = graphe "M"

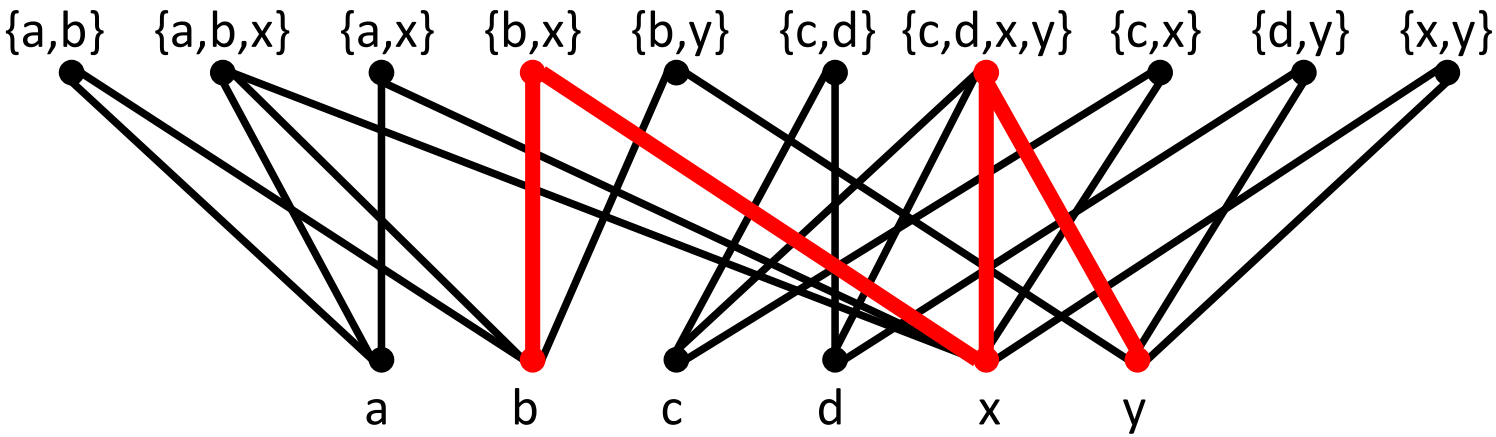
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'ensemble minimum de conflits

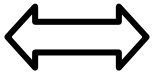
Conflit : clades ni inclus ni disjoints



Graphe des caractères :



Trouver l'ensemble minimum de conflits

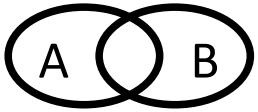


Supprimer le nombre minimum *t* de sommets-taxons tels que le graphe des caractères est un graphe “sans M”

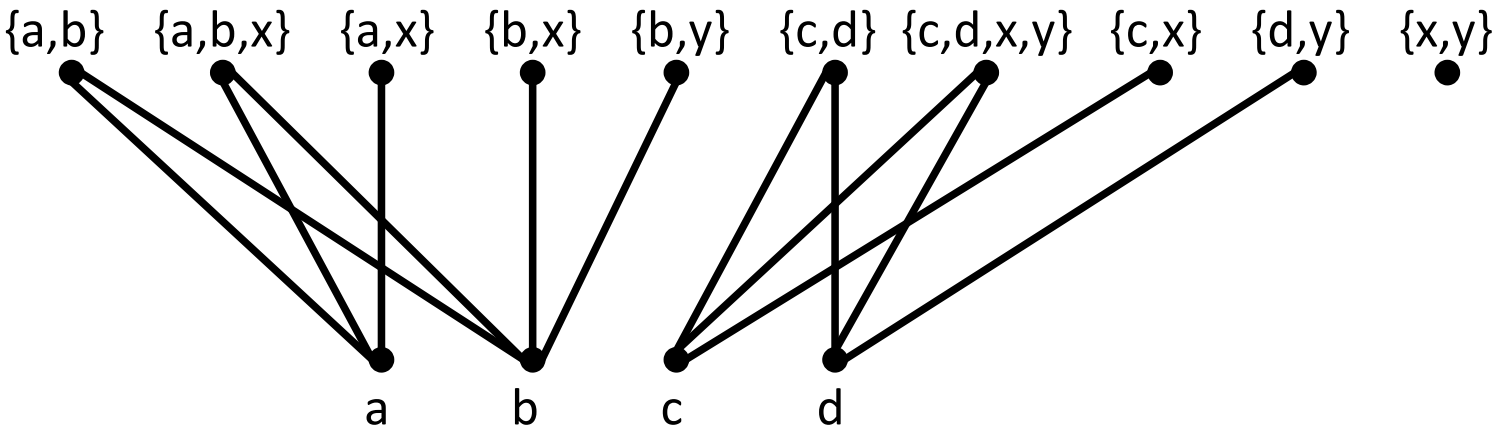
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'ensemble minimum de conflits

Conflit : clades ni inclus ni disjoints



Graphe des caractères :

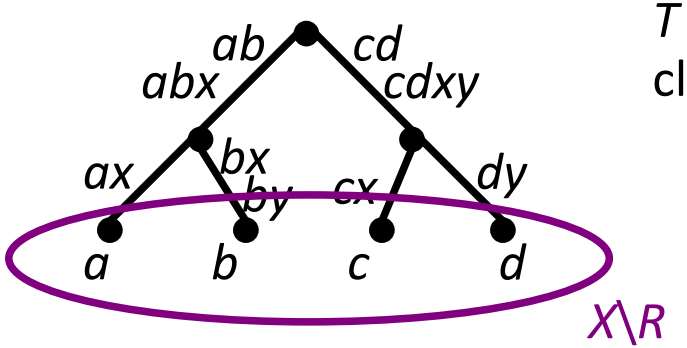


Supprimer le nombre minimum *t* de sommets-taxons tels que le graphe des caractères est un graphe “sans M” :
- algorithme FPT basique de branchement en $O^*(3^t)$
- algorithme FPT 3-Hitting-Set en $O^*(2{,}076^t)$

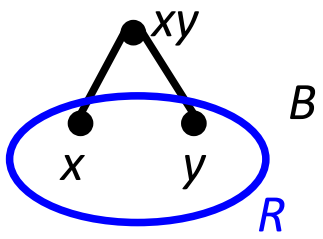
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'attachement minimum

Etape précédente :
ensemble minimum de taxons *R* tels que les clades sur *X**R* sont compatibles (avec un arbre *T*).



T : arbre représentant les clades sur *X**R*

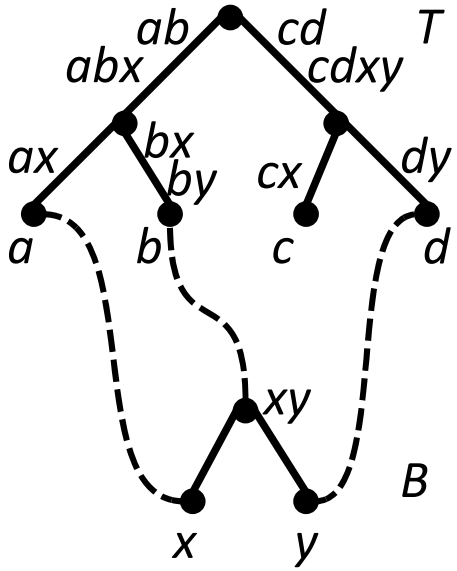


B : réseau représentant les clades maximaux sur *R* et les singletons de *R*.

Problème :
Attacher *T* à *B* avec le **minimum de liens**.

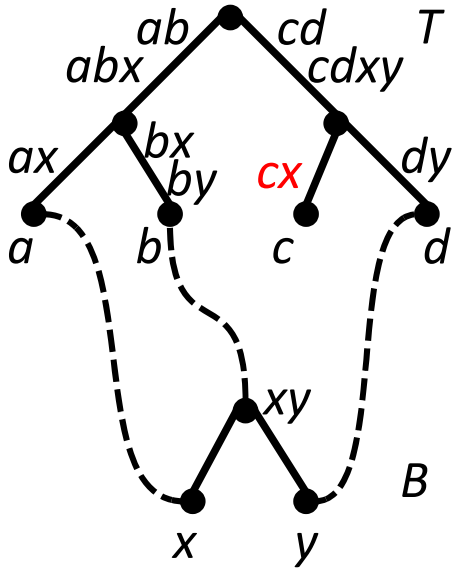
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'attachement minimum



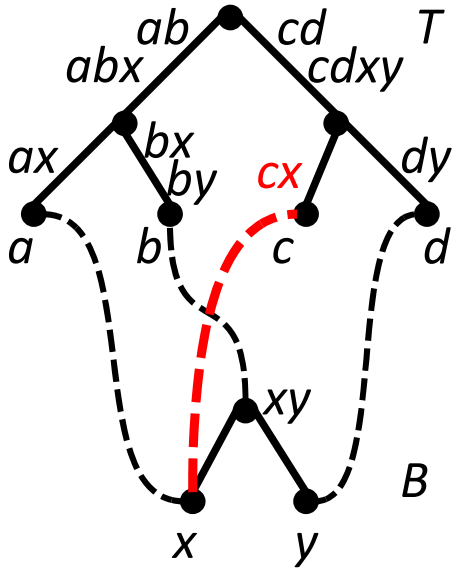
C1 - Satisfaction des clusters de *T* :
Pour tout arc *e* de l'arbre *T* et tout taxon *r* de *R* contenu dans un cluster de *e*, il existe un lien depuis un des descendants de *e* dans *T* vers le noeud correspondant à *r* dans *B*, ou un noeud de *B* correspondant à un cluster maximal qui contient *r.*

L'attachement minimum



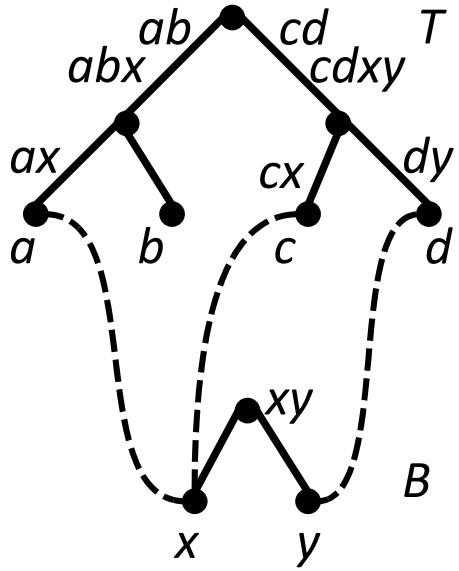
C1 - Satisfaction des clusters de *T* :
Pour tout arc *e* de l'arbre *T* et tout taxon *r* de *R* contenu dans un cluster de *e*, il existe un lien depuis un des descendants de *e* dans *T* vers le noeud correspondant à *r* dans *B*, ou un noeud de *B* correspondant à un cluster maximal qui contient *r.*

L'attachement minimum



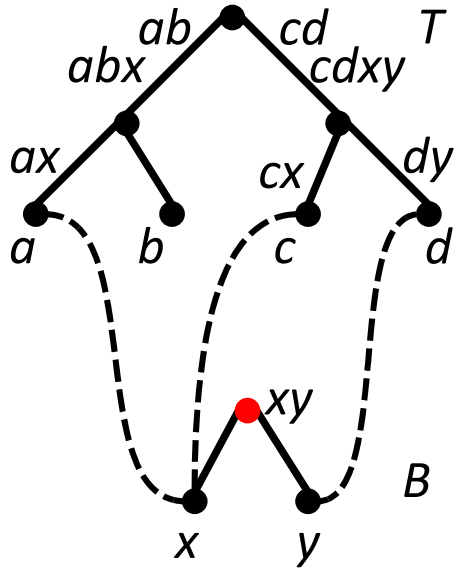
C1 - Satisfaction des clusters de *T* :
Pour tout arc *e* de l'arbre *T* et tout taxon *r* de *R* contenu dans un cluster de *e*, il existe un lien depuis un des descendants de *e* dans *T* vers le noeud correspondant à *r* dans *B*, ou un noeud de *B* correspondant à un cluster maximal qui contient *r.*

L'attachement minimum



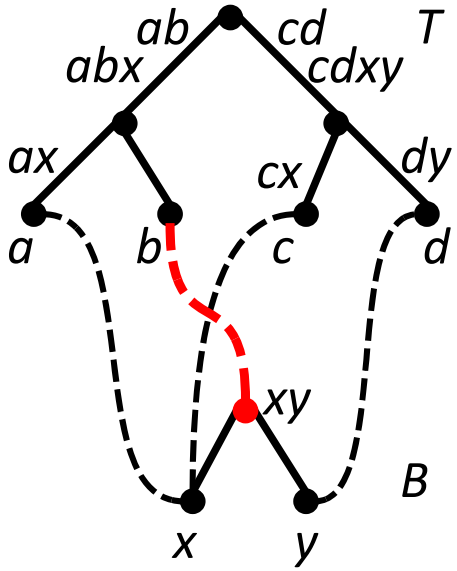
C2 - Satisfaction de la paternité des noeuds de *B* :
Tout noeud de *B* correspondant à plus d'un taxon est relié à un noeud de *T* par un lien exactement.

L'attachement minimum



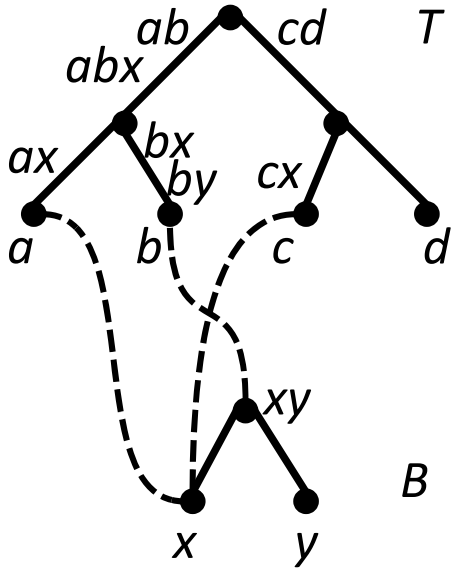
C2 - Satisfaction de la paternité des noeuds de *B* :
Tout noeud de *B* correspondant à plus d'un taxon est relié à un noeud de *T* par un lien exactement.

L'attachement minimum



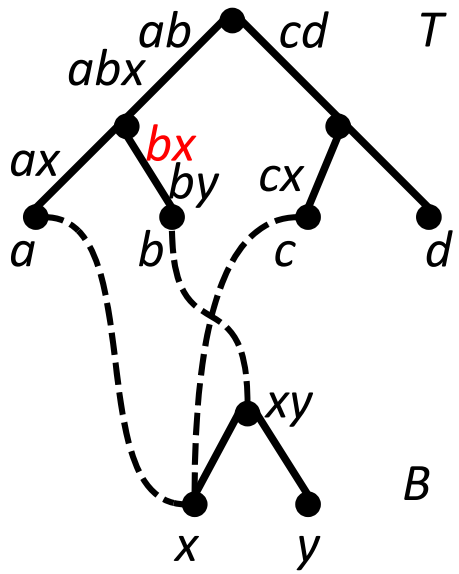
C2 - Satisfaction de la paternité des noeuds de *B* :
Tout noeud de *B* correspondant à plus d'un taxon est relié à un noeud de *T* par un lien exactement.

L'attachement minimum



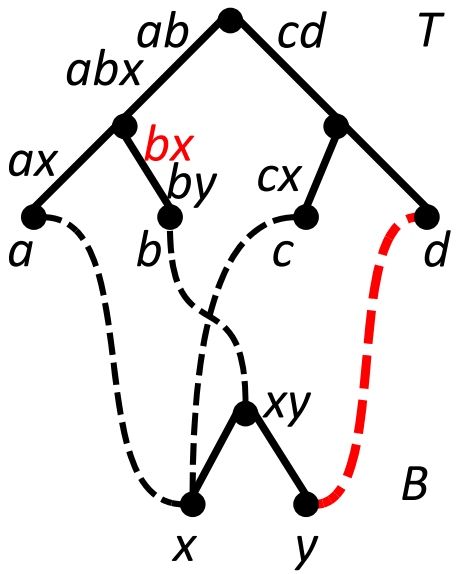
C3 – Absence de parasites des clusters de *T* dans *B* :
Pour tout arc *e* de *T*, si un cluster correspondant à *e* ne contient pas un taxon *r* de *R*, il existe un chemin d'un noeud qui ne descend pas de *e* vers le noeud associé à *r*.

L'attachement minimum



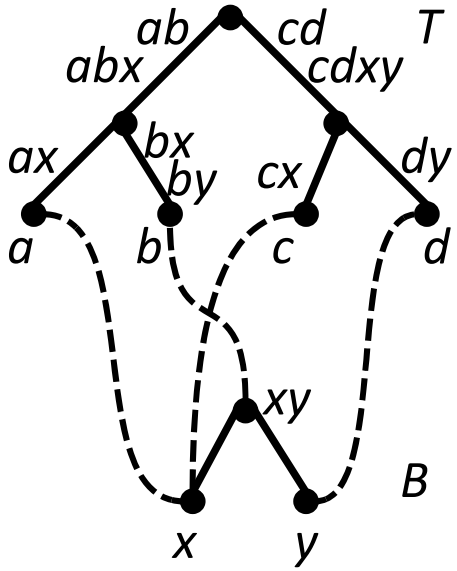
C3 – Absence de parasites des clusters de *T* dans *B* :
Pour tout arc *e* de *T*, si un cluster correspondant à *e* ne contient pas un taxon *r* de *R*, il existe un chemin d'un noeud qui ne descend pas de *e* vers le noeud associé à *r*.

L'attachement minimum



C3 – Absence de parasites des clusters de *T* dans *B* :
Pour tout arc *e* de *T*, si un cluster correspondant à *e* ne contient pas un taxon *r* de *R*, il existe un chemin d'un noeud qui ne descend pas de *e* vers le noeud associé à *r*.

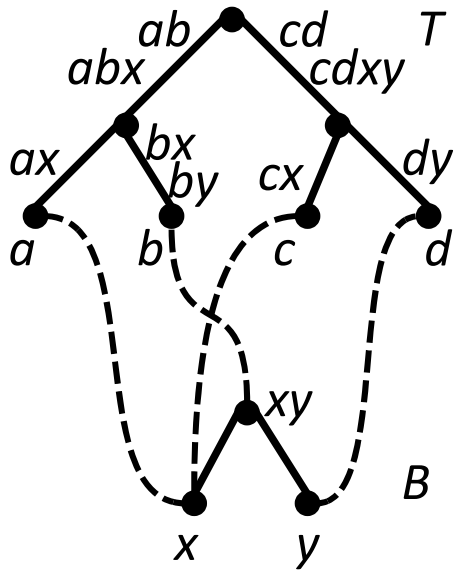
L'attachement minimum



Problème :
Trouver un attachement
respectant les contraintes C1, C2, et C3
et de taille minimale.

L'attachement minimum

Problème Minimum Attachment :
Attacher *T* à *B* avec le **minimum de liens**.



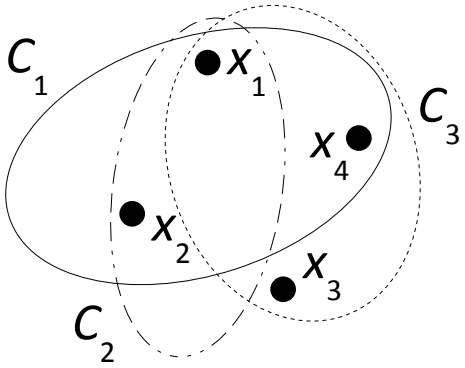
NP-complet

réduction depuis No-Inclusion SetCover
Réduction de No-Inclusion SetCover depuis SetCover

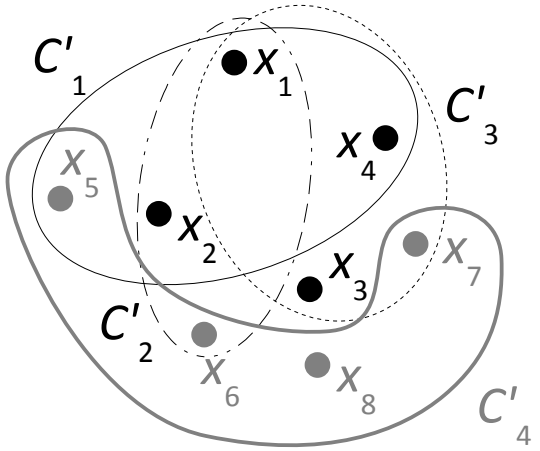
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'attachement minimum

NP-complétude de No-Inclusion Set Cover



instance de Set Cover



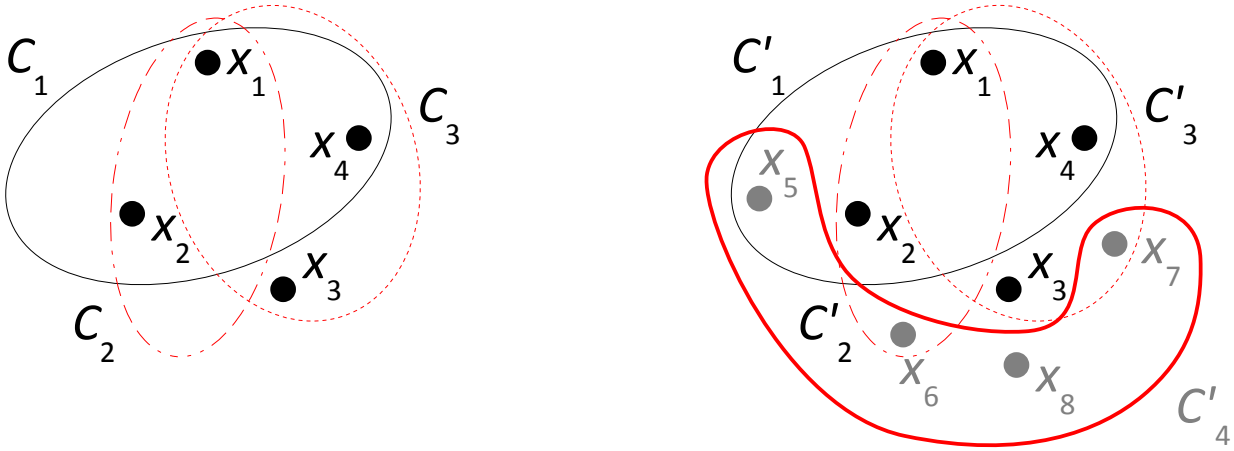
instance de No-Inclusion Set Cover

Problème Set Cover :
Nombre minimum d'ensembles dont l'union contient tous les x_i

Problème No-Inclusion Set Cover :
Problème Set Cover sur des instances sans inclusions entre les ensembles

L'attachement minimum

NP-complétude de No-Inclusion Set Cover



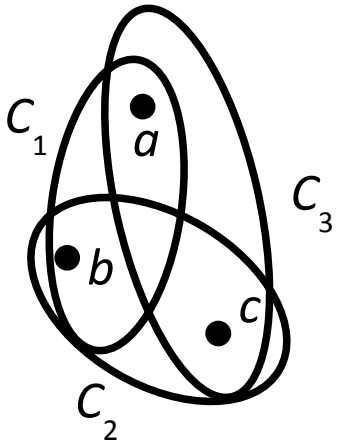
instance de Set Cover
solution de taille k

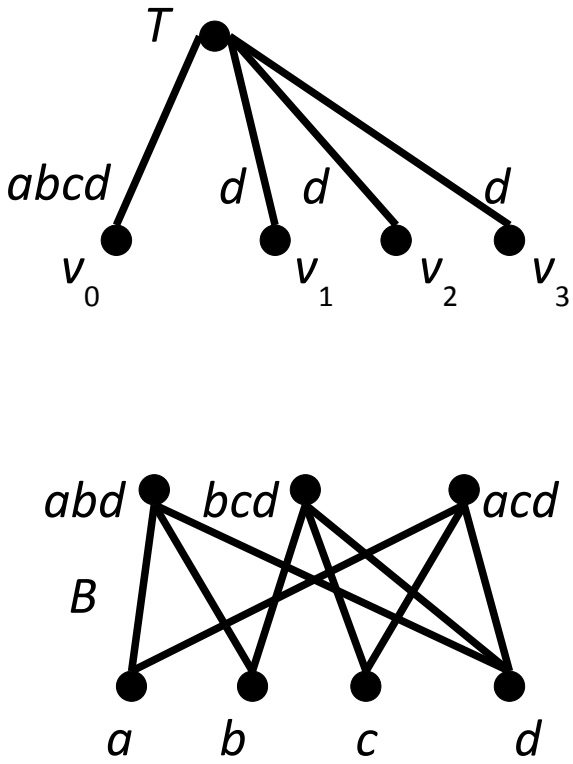
⟺

instance de No-Inclusion Set Cover
solution de taille k+1

L'attachement minimum

NP-complétude de Minimum Attachment





instance de No-Inclusion Set Cover
solution de taille k

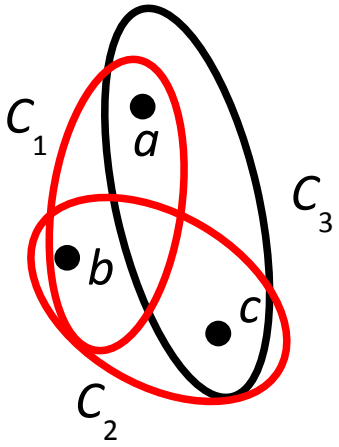
⟺

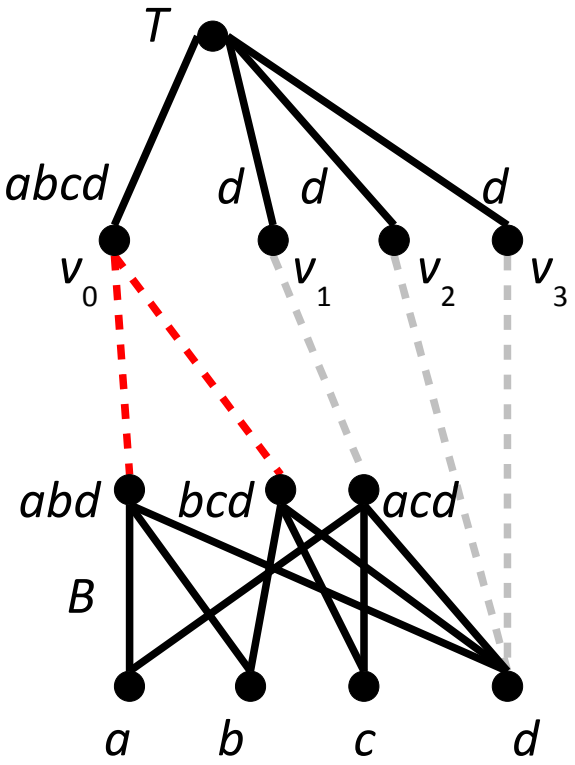
instance de Minimum Attachment
solution de taille k+m

Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'attachement minimum

NP-complétude de Minimum Attachment





instance de No-Inclusion Set Cover
solution de taille *k*

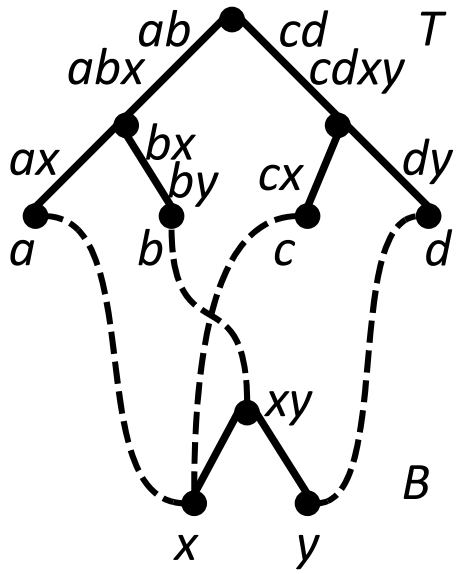
⟺

instance de Minimum Attachment
solution de taille *k*+*m*

Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'attachement minimum

Problème :
Attacher *T* à *B* avec le **minimum de liens**.



NP-complet

réduction depuis SetCover

Algorithmes :
- Séparation et évaluation

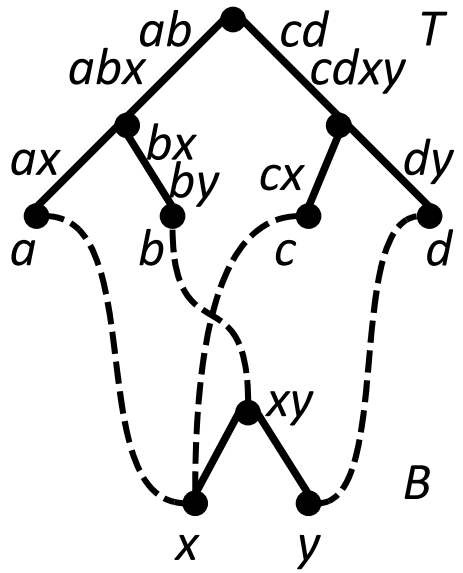
implémenté dans Dendroscope 2

- Programme linéaire en nombres entiers

Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'attachement minimum

Problème :
Attacher *T* à *B* avec le **minimum de liens**.



Programme linéaire en nombres entiers :

Minimiser $\sum_{u \in T, v \in B} a_{u,v}$

C1 - Satisfaction des clusters de *T* :

pour tout arc $e = (x, y)$ de T, pour tout taxon $r \in R(e)$,

$$\sum_{u \in T \,|\, u \preceq_T y} a_{u,v(r)} + \sum_{u \in T,\, v(C) \in B \,|\, u \preceq_T y,\, r \in C} a_{u,v(C)} \geq 1$$

C2 - Satisfaction de la paternité des noeuds de *B* : pour tout $C \in \hat{\mathcal{C}}_{|R}$, $\sum_{u \in T} a_{u,v(C)} = 1$

C3 – Absence de parasites des clusters de *T* dans *B* :

pour tout arc $e = (x, y)$ de T et pour tout $r \in R$ tel que $\mathcal{C}(e)$ contient un clade $C \in \mathcal{C}$ qui ne contient pas r,

$$\sum_{u \in T \,|\, u \not\preceq_T y} \sum_{v \in B \,|\, v(r) \preceq_B v} a_{u,v} = \sum_{u \in T \,|\, u \not\preceq_T y} a_{u,v(r)} + \sum_{u \in T,\, C' \in \hat{\mathcal{C}}_{|R} \,|\, u \not\preceq_T y,\, r \in C'} a_{u,v(C')} \geq 1.$$

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Reconstruction de réseaux à partir de triplets
- Reconstruction de réseaux à partir de clades
- Sélection des données
- Visualisation de réseaux phylogénétiques
- Perspectives

Edition des données pour arbres complets

Nécessité de données **complètes** pour les algorithmes de clades (complets) et triplets (denses)

Certains gènes ont disparu chez certaines espèces

➡ arbres de gènes incomplets

Problèmes :

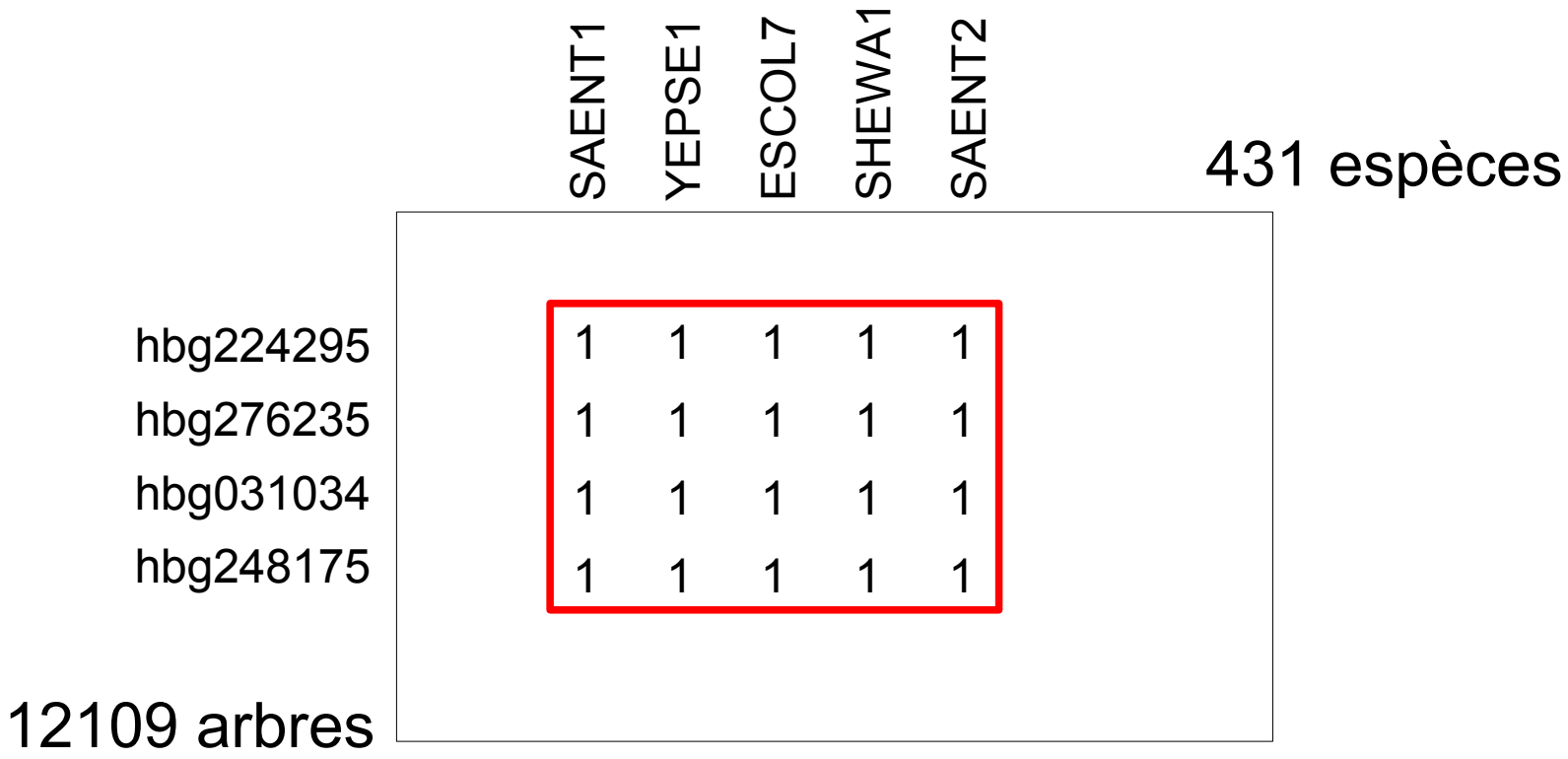
- trouver un **gros ensemble d'arbres :** ayant un **gros ensemble d'espèces** en commun

- trouver un **ensemble maximum de triplets dense**

Edition des données pour arbres complets

Problème :

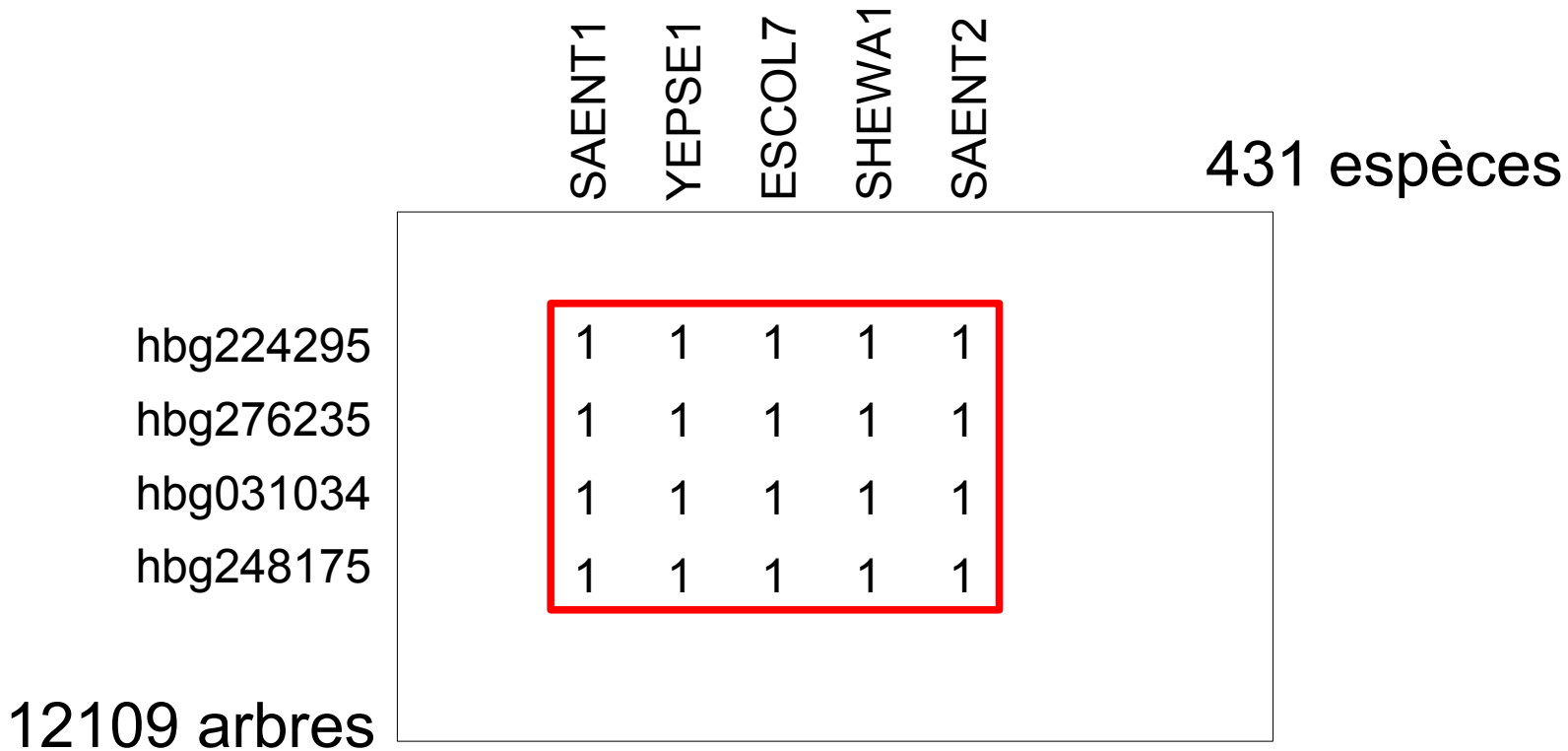
Trouver un **gros ensemble d'arbres :** ayant un **gros ensemble d'espèces** en commun



Edition des données pour arbres complets

Problème :

Trouver un **gros ensemble d'arbres :** ayant un **gros ensemble d'espèces** en commun



Rectangle de 1 de taille maximale (biclique maximum dans un biparti)

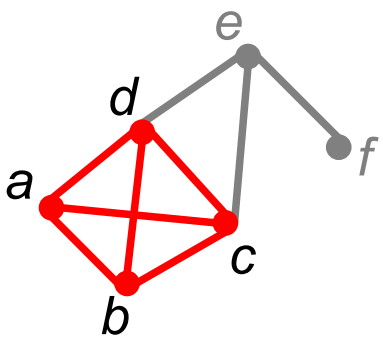
polynomial si taille = périmètre, NP-complet si taille = aire

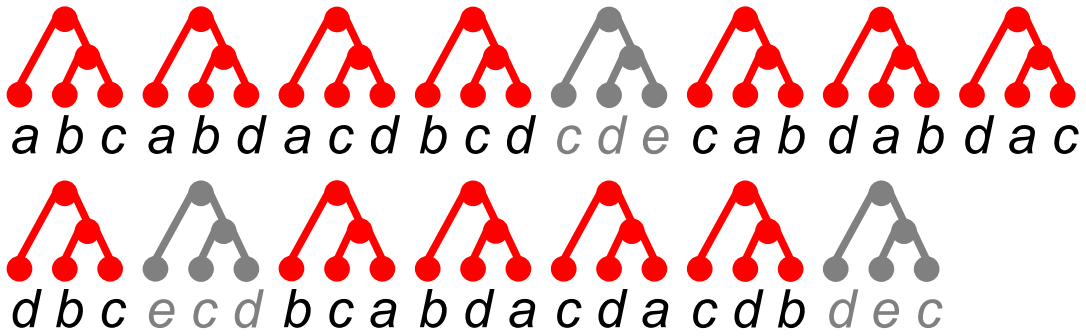
Edition des données pour triplets denses

Problème :

Trouver un **ensemble maximum dense de triplets**

NP-complet :





instance de *k*-clique
solution de taille *k*

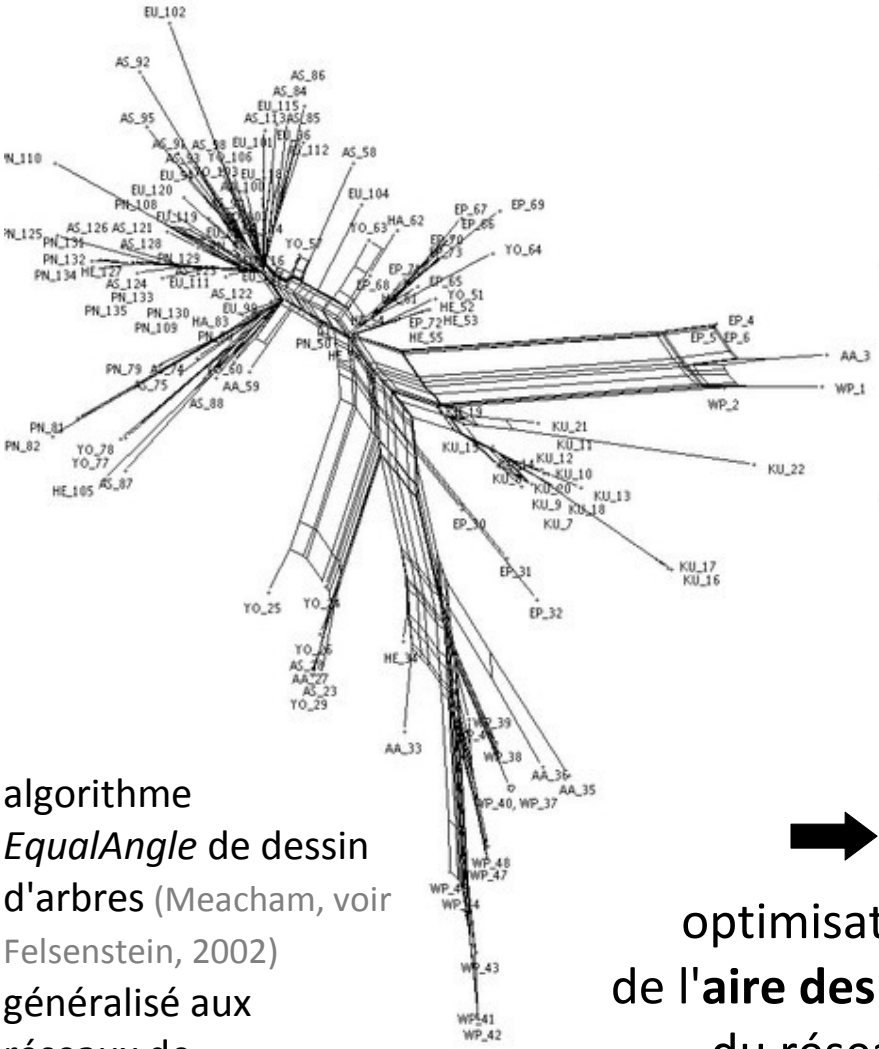
⟺

instance de *k*-Dense Triplet Set
solution de taille *k*

Plan

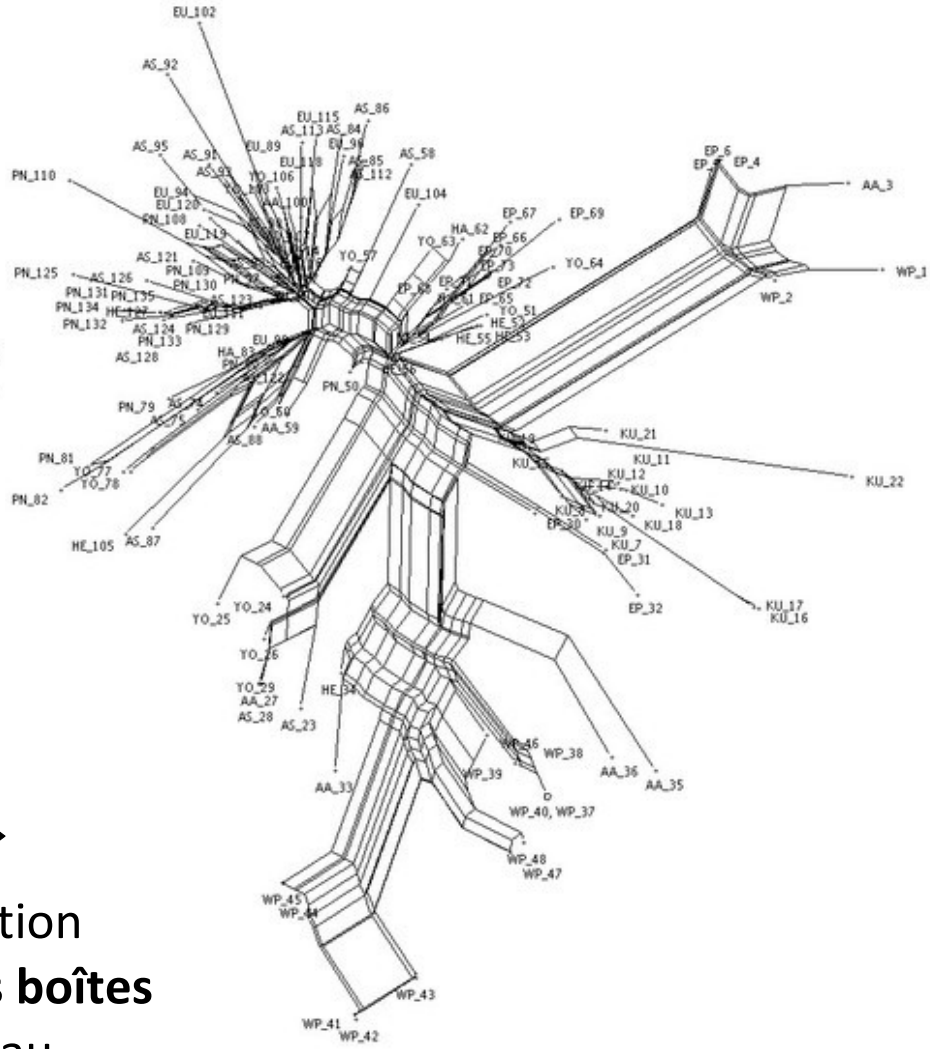
- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Reconstruction de réseaux à partir de triplets
- Reconstruction de réseaux à partir de clades
- Sélection des données
- Visualisation de réseaux phylogénétiques
- Perspectives

Visualisation des réseaux de bipartitions



algorithme *EqualAngle* de dessin d'arbres (Meacham, voir Felsenstein, 2002) généralisé aux réseaux de bipartitions.

optimisation de l'**aire des boîtes** du réseau

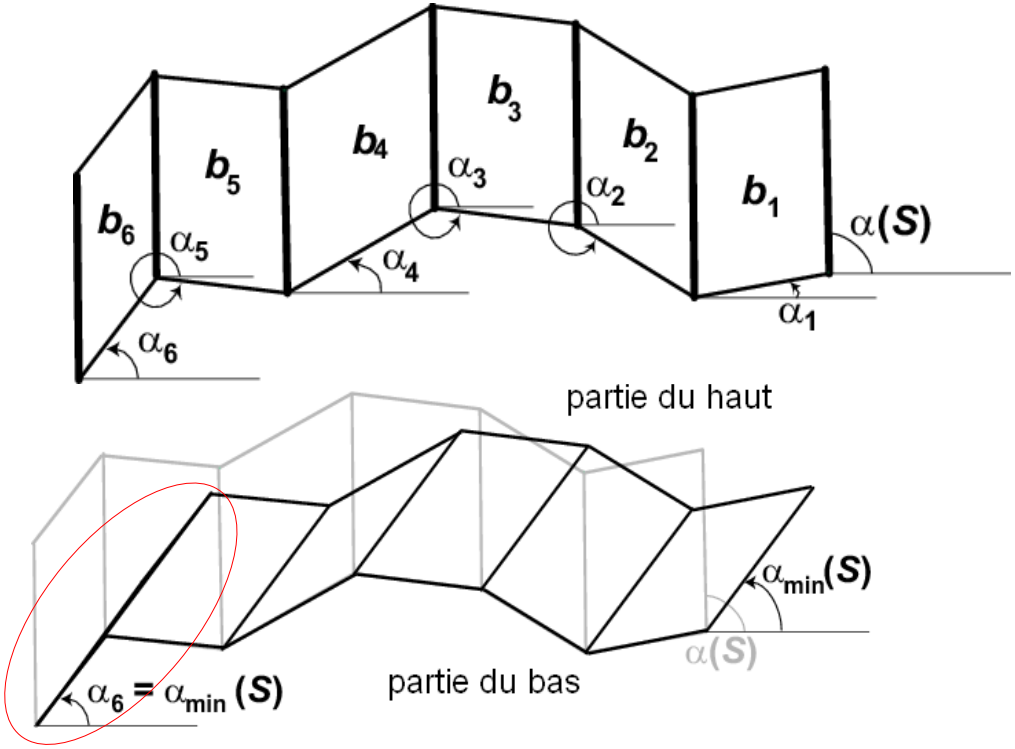


Gambette & Huson, TCBB, 2008

Algorithme Box-opening

Collisions locales :

deux **angles critiques** $\alpha_{min}(S)$ et $\alpha_{max}(S)$ pour l'ange de la bipartition $\alpha(S)$



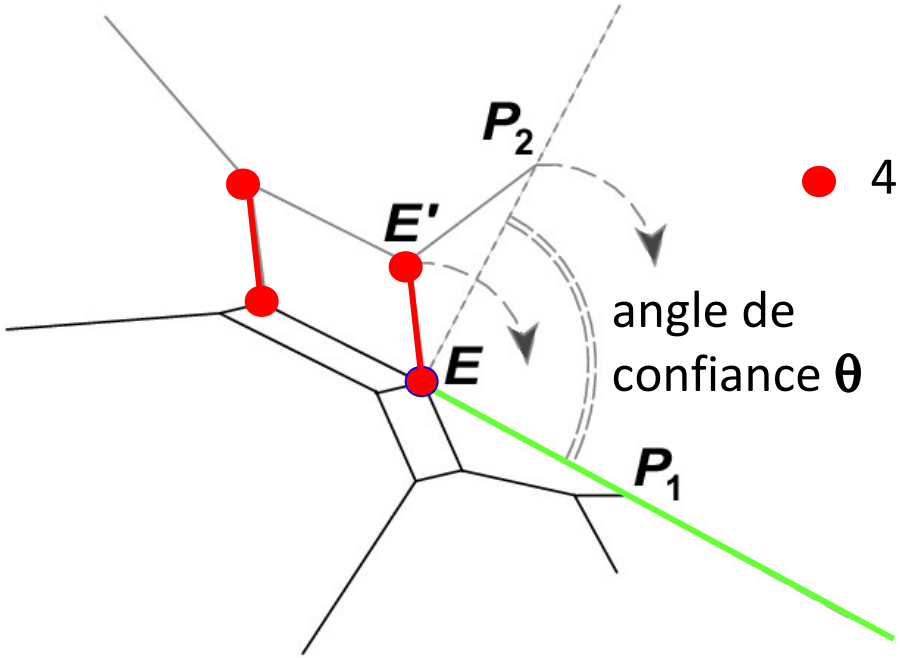
$$\alpha_{max}(S) = \alpha(S) + \min_{\text{boîte } b_i}\{(\alpha_i - \alpha(S) - \pi) \bmod 2\pi\}$$

$$\alpha_{min}(S) = \alpha(S) - \min_{\text{boîte } b_i}\{(\alpha(S) - \alpha_i) \bmod 2\pi\}$$

Gambette & Huson, TCBB, 2008

Algorithme Box-opening

Choix de l'angle $\alpha(S)$: **collisions globales**

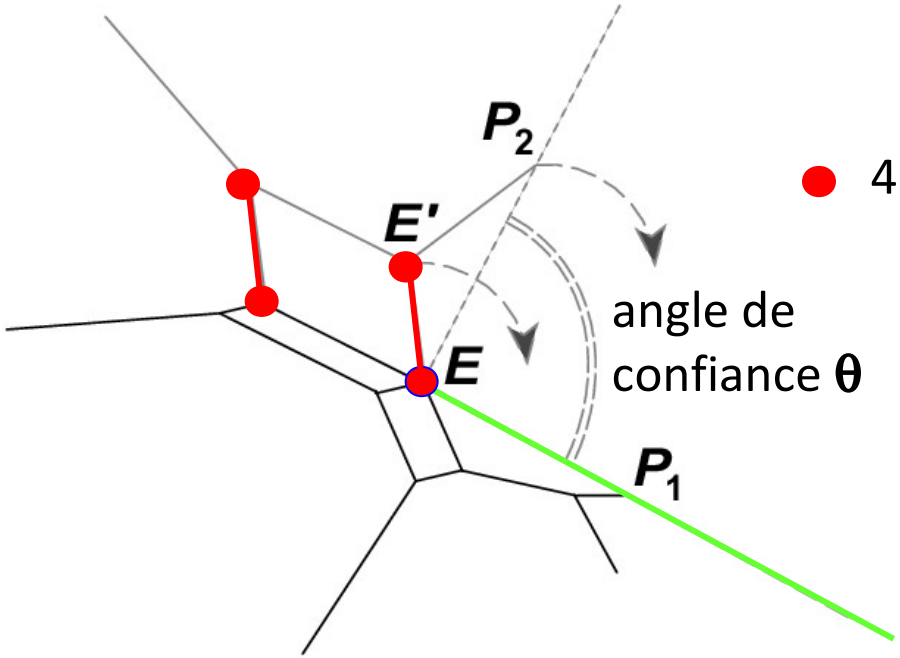


- 4 nœuds extrêmes

Gambette & Huson, TCBB, 2008

Algorithme Box-opening

Choix de l'angle $\alpha(S)$: **collisions globales**



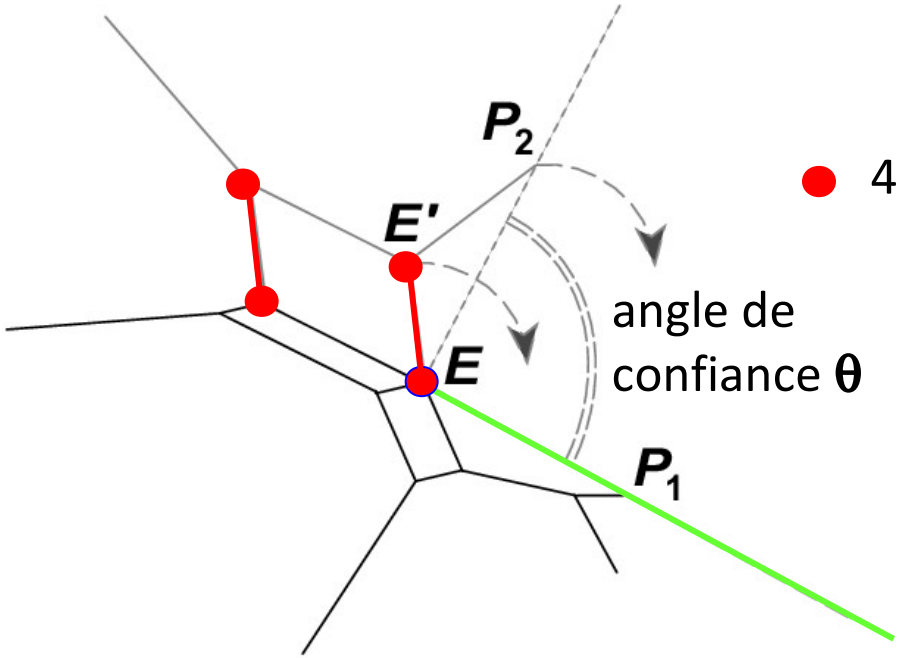
- 4 nœuds extrêmes

➡ **optimisations locales** de l'aire

Gambette & Huson, TCBB, 2008

Algorithme Box-opening

Choix de l'angle $\alpha(S)$: **collisions globales**



- 4 nœuds extrêmes

➡ **optimisations locales** de l'aire

➡ **optimisations globale** du réseau
métaheuristique (recuit simulé)

Gambette & Huson, TCBB, 2008

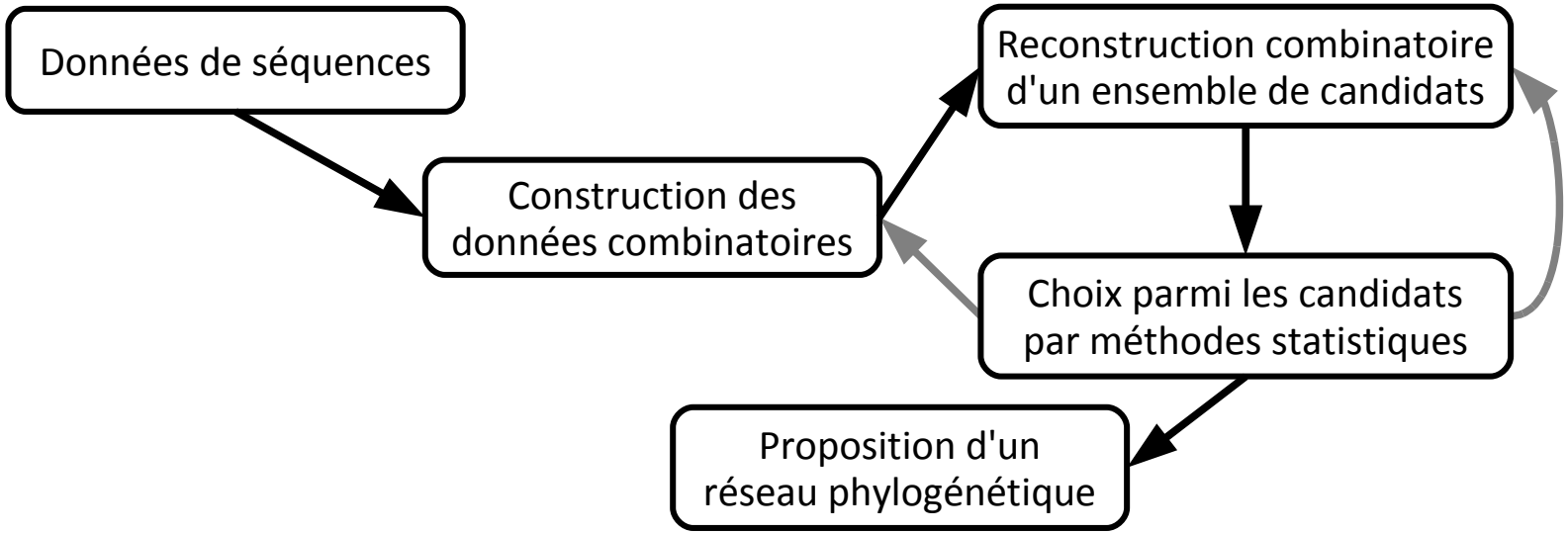
Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Reconstruction de réseaux à partir de triplets
- Reconstruction de réseaux à partir de clades
- Sélection des données
- Visualisation de réseaux phylogénétiques
- Perspectives

Perspectives de recherche

Combinatoire :
- Meilleure connaissance des réseaux de faible niveau, enracinés ou non : dénombrement, caractérisations...
- Mise à jour ou modification d'un réseau face à de nouvelles données

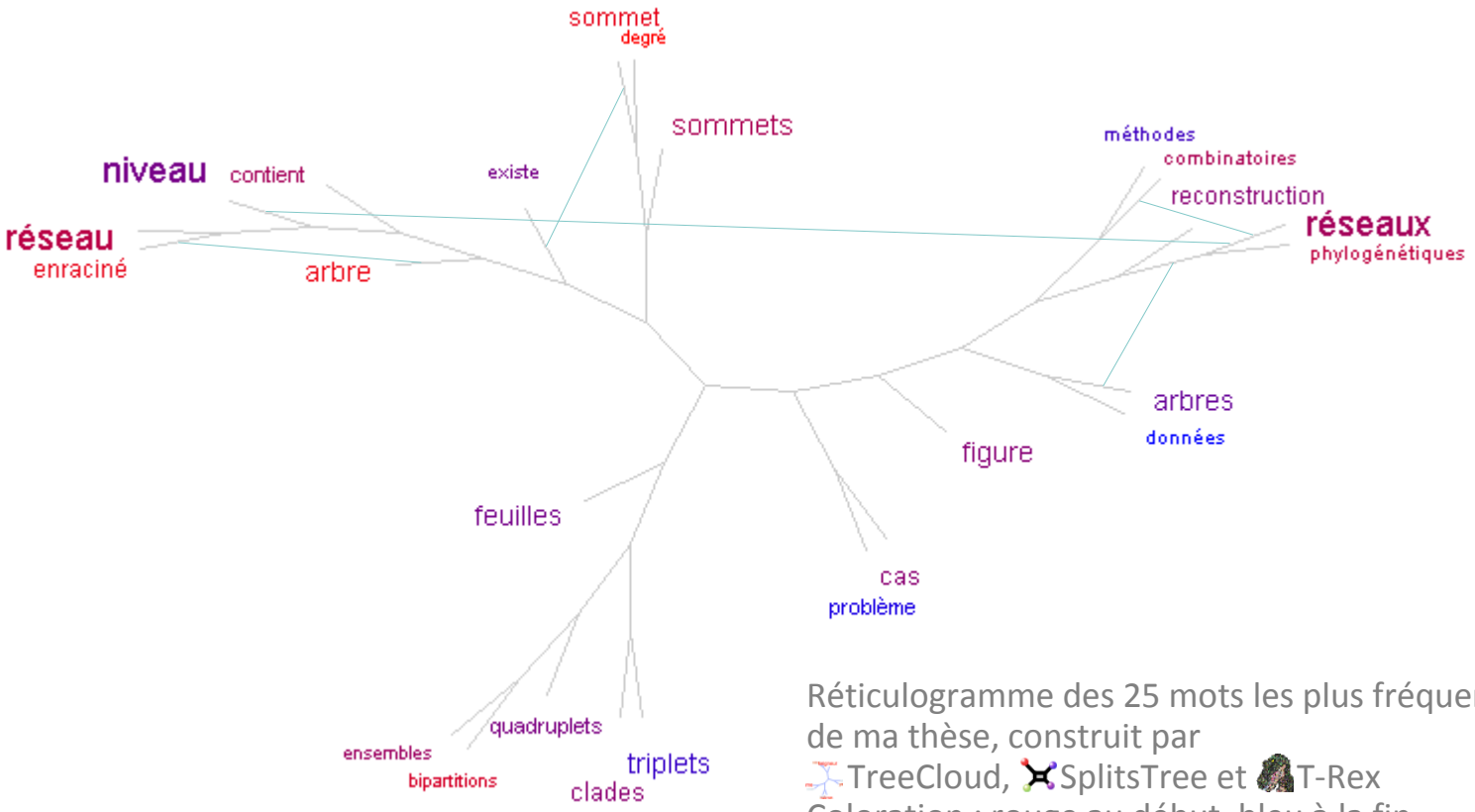
Bioinformatique :
- Fonction des gènes transférés (“autoroutes de transfert”)
- Intégration des méthodes combinatoires dans une approche statistique



Autres applications des réseaux phylogénétiques :
- visualiser la polysémie dans les nuages arborés

Merci !

Coauteurs des travaux présentés :
- Vincent Berry, Christophe Paul (LIRMM)
- Daniel Huson, Regula Rupp (Tübingen)
- Katharina Huber (East Anglia)



Réticulogramme des 25 mots les plus fréquents de ma thèse, construit par TreeCloud, SplitsTree et T-Rex
Coloration : rouge au début, bleu à la fin

http://www.treecloud.org