Séminaire du LIGM - 05/04/2011

Méthodes combinatoires de reconstruction de réseaux phylogénétiques

Philippe Gambette

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- Limites
- Application pratique
- Bilan
- Perspectives

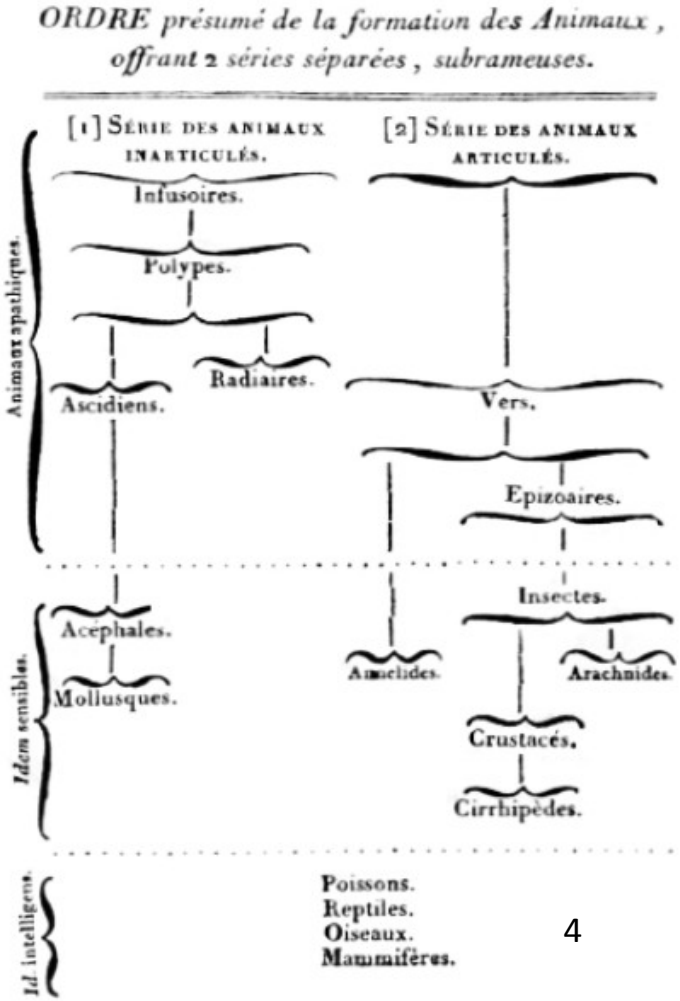
Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- Limites
- Application pratique
- Bilan
- 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)

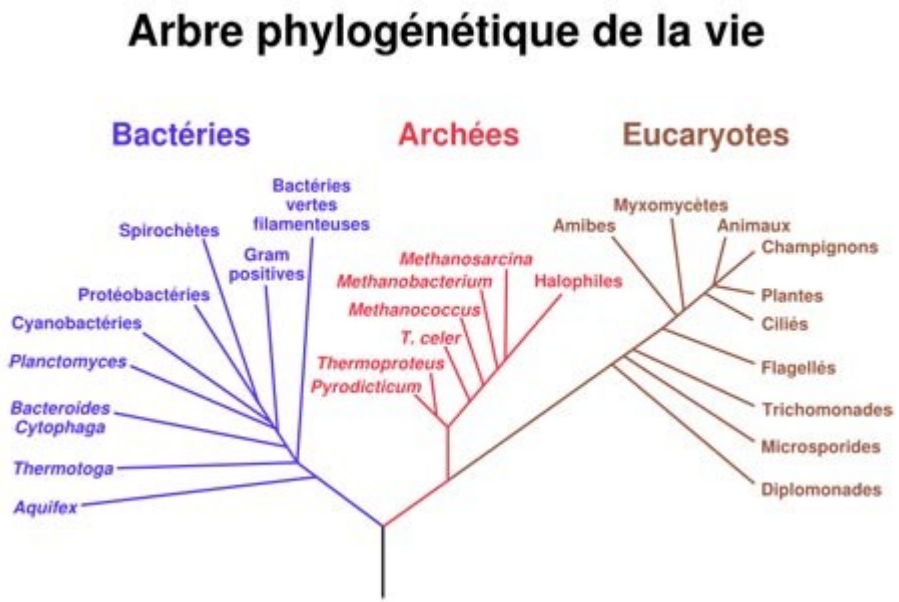
4

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

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

modélisation



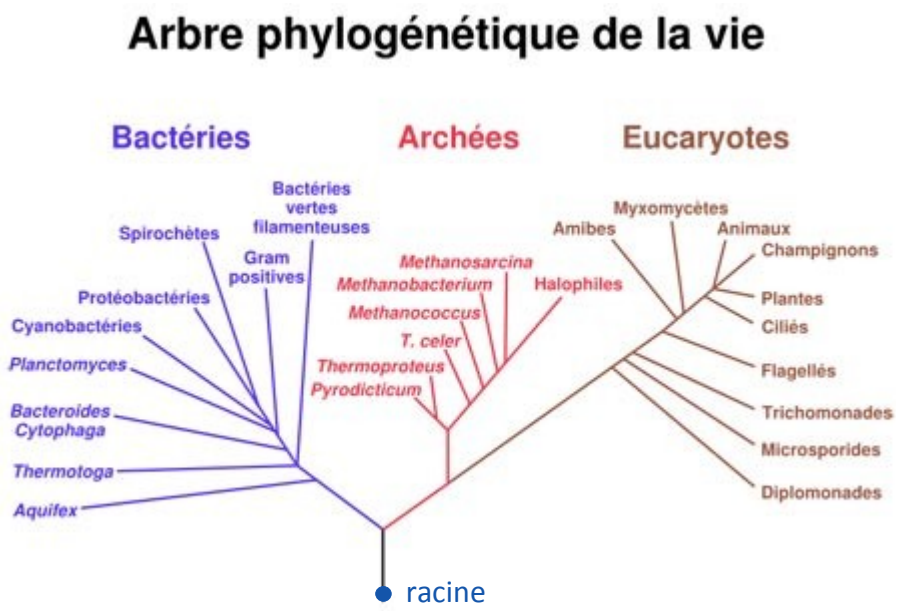
D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

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

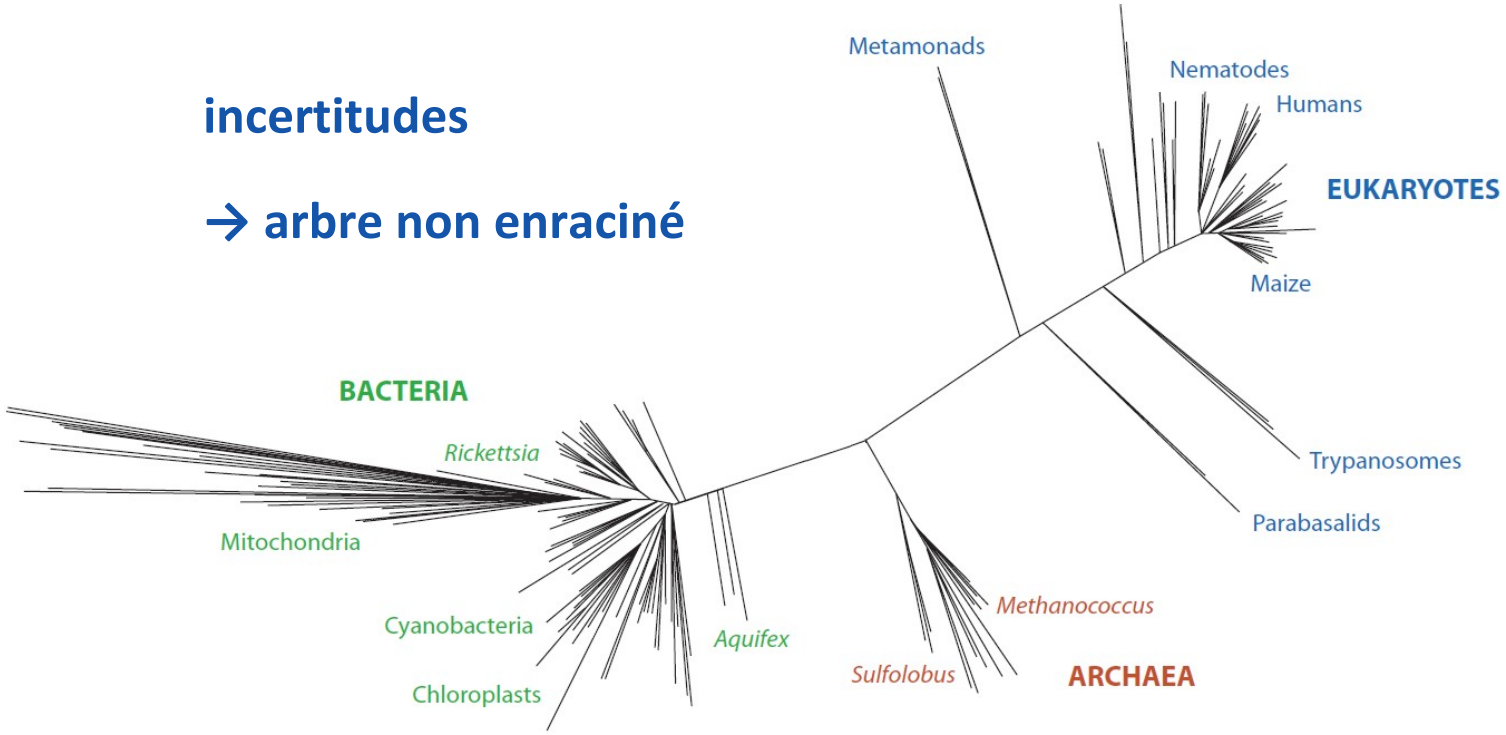
Arbre phylogénétique d'un ensemble d'espèces :

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

modélisation

incertitudes

→ arbre non enraciné



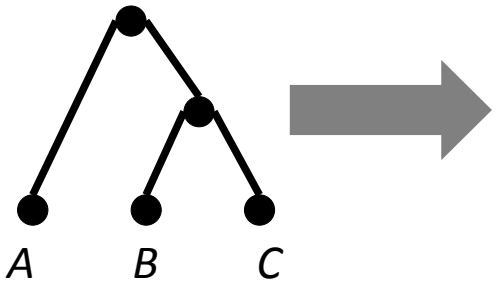
D'après Christophe Blumrich, David S. Spencer,
cité dans Doolittle : Uprooting the Tree of Life, Scientific American (Fév. 2000)

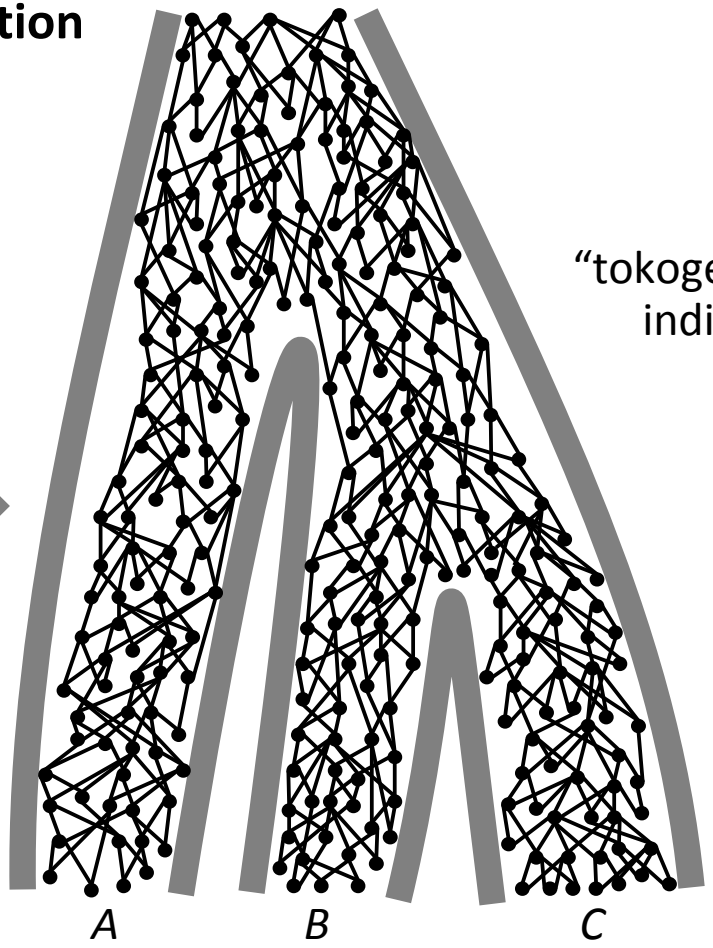
Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

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

arbre des espèces *S*



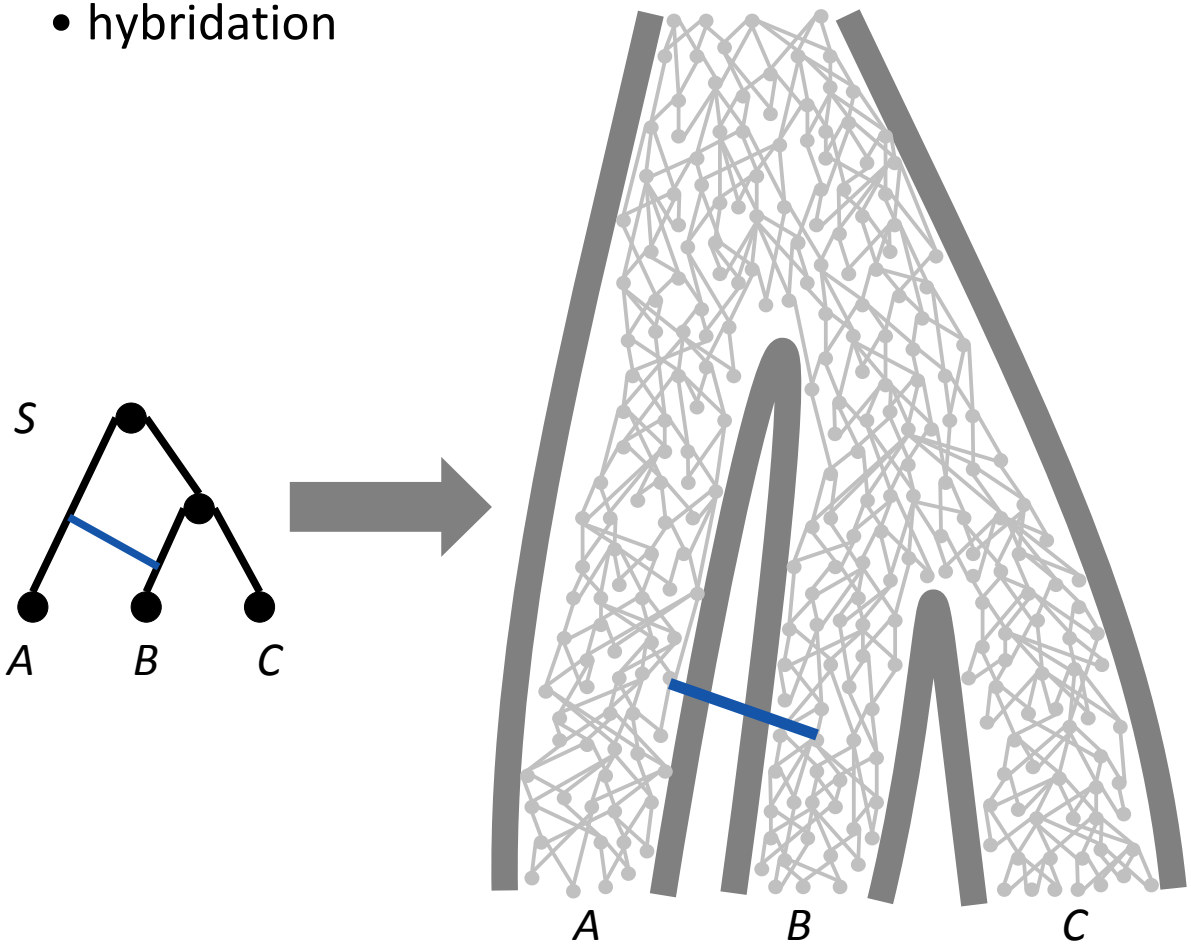


"tokogénie" des individus

Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

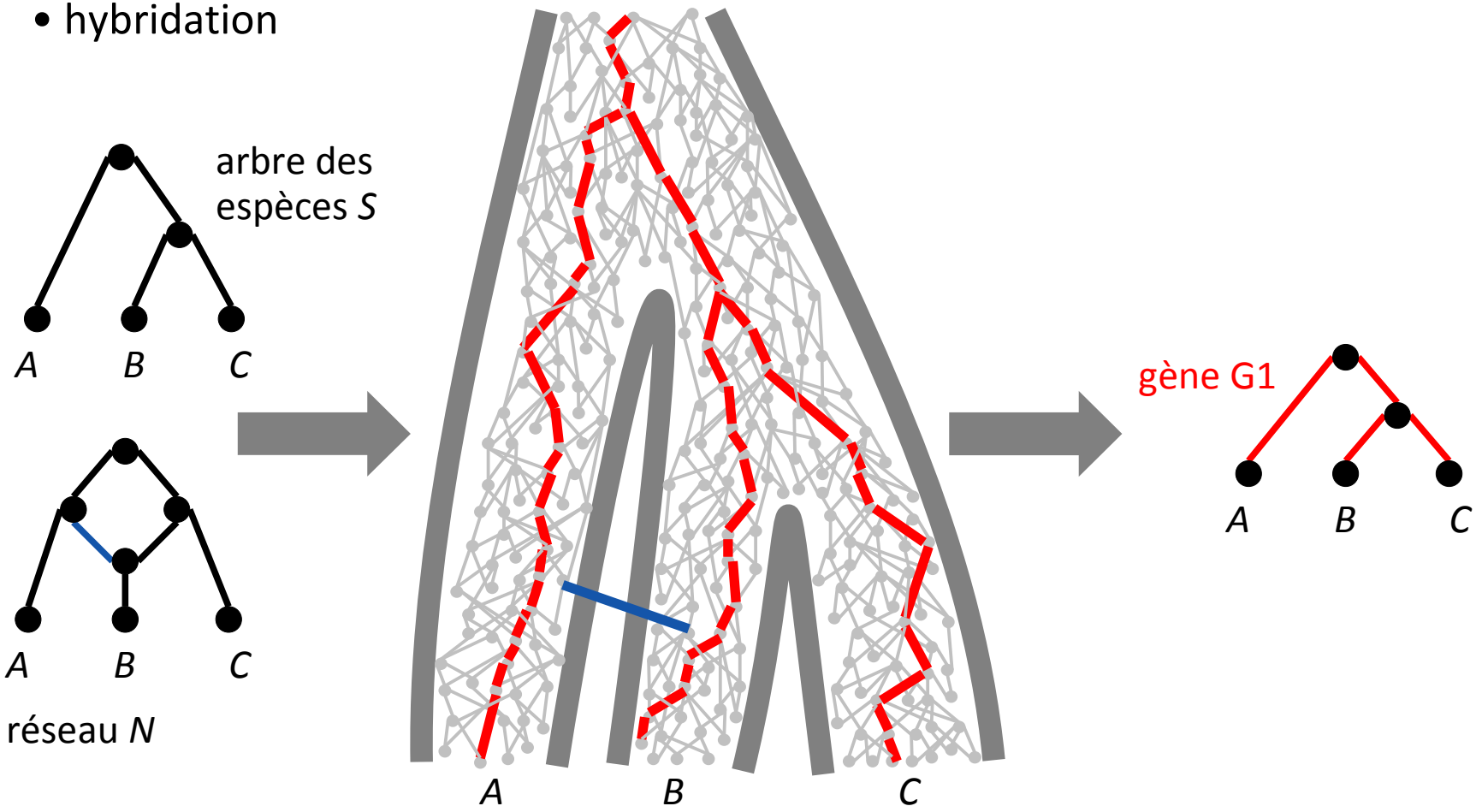
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

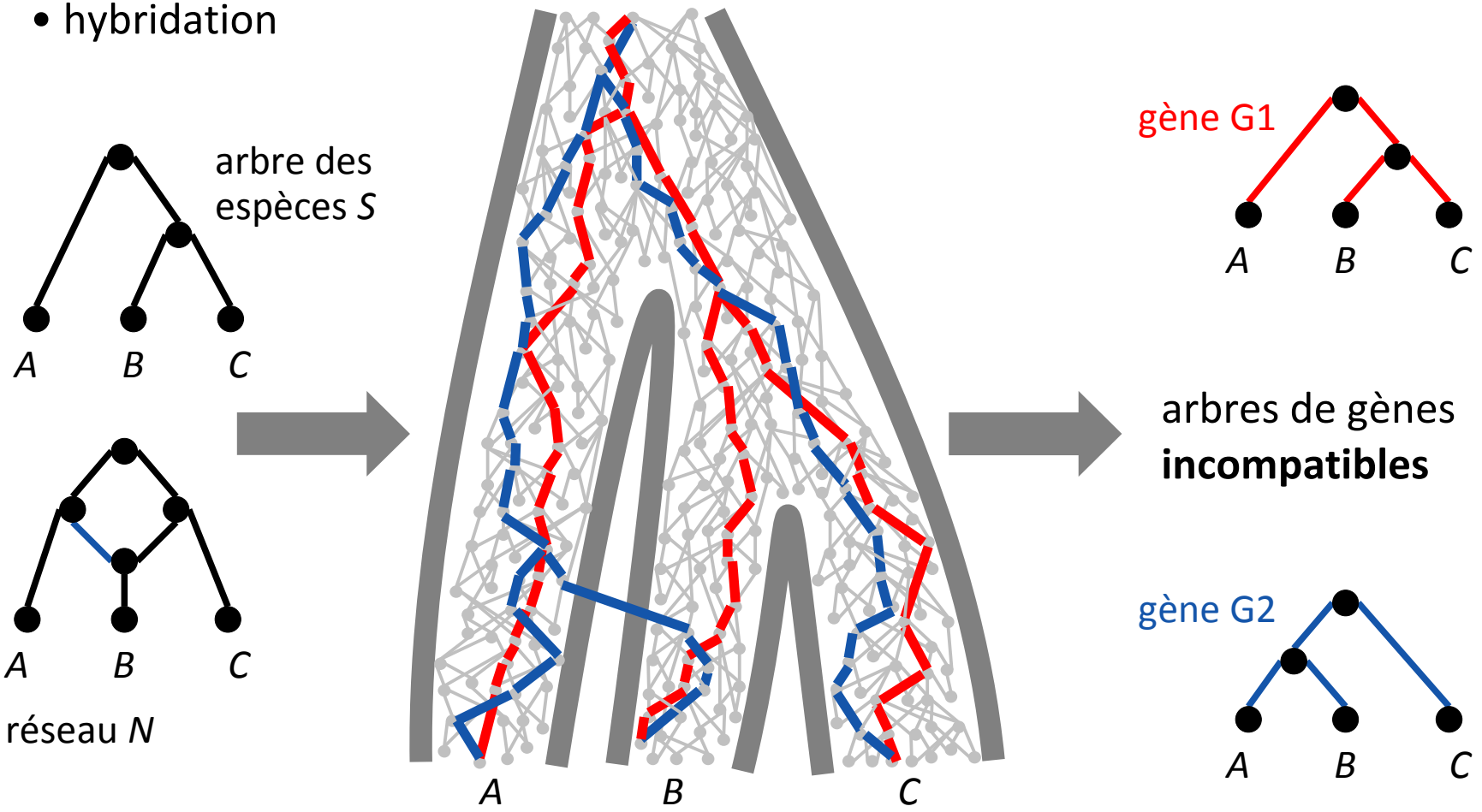
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

- transfert horizontal
- hybridation

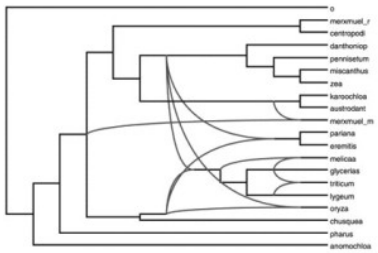


Les réseaux phylogénétiques

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **explicites**

modélisation de l'évolution



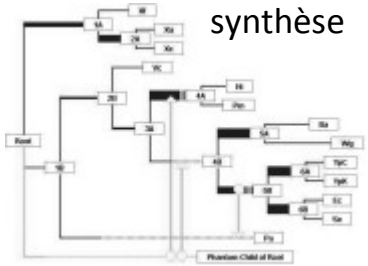
réseau à une couche de réticulation

Dendroscope

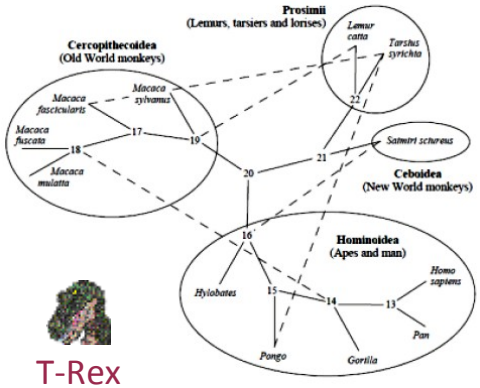
Simplistic

réseau de niveau 2

diagramme de synthèse



HorizStory



T-Rex

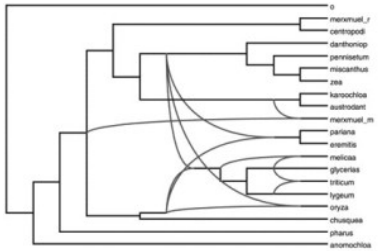
réticulogramme

Les réseaux phylogénétiques

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **explicites**

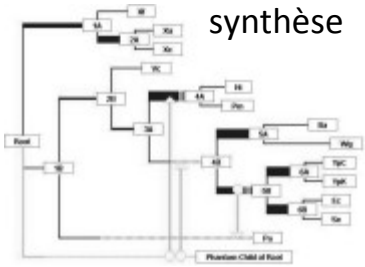
modélisation de l'évolution



réseau à une couche de réticulation

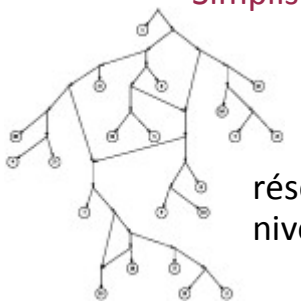
Dendroscope

diagramme de synthèse



HorizStory

Simplistic

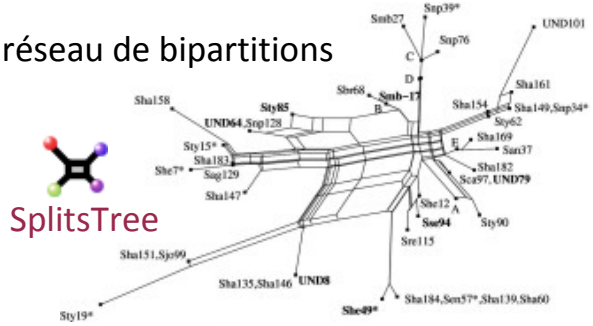


réseau de niveau 2

- réseaux phylogénétiques **abstraits**

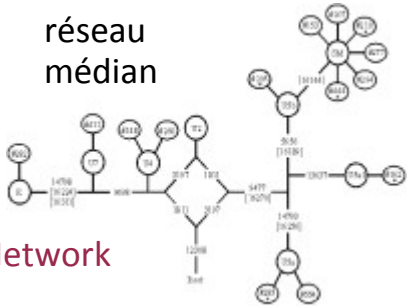
classification, visualisation de données

réseau de bipartitions



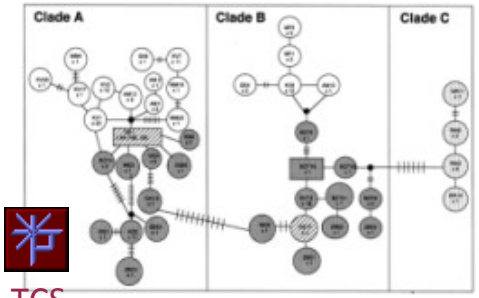
SplitsTree

réseau médian



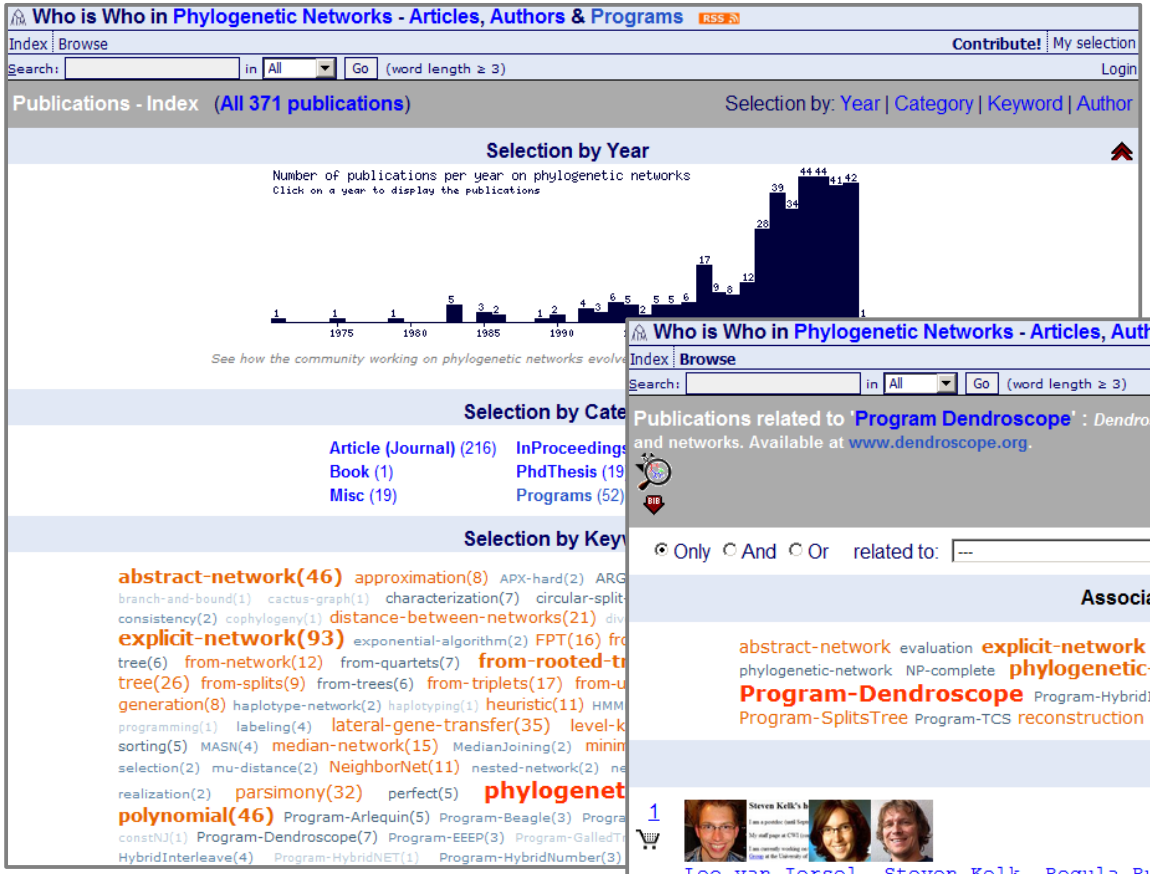
Network

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

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- Limites
- Application pratique
- Bilan
- 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

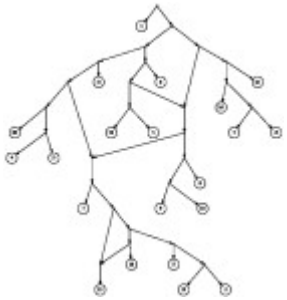
Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002

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

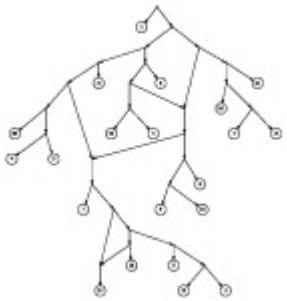
Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002

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

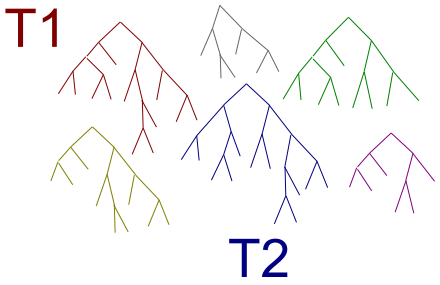
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

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

Guindon & Gascuel, SB, 2003



{arbres}

Base HOGENOM

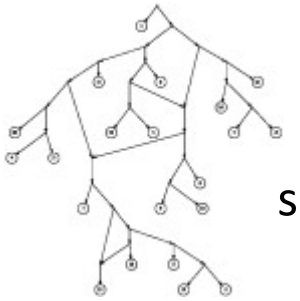
Phyl-ARIANE

ANR

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

Réconciliation ou consensus d'arbres

réseau explicite



super-réseau optimal *N*

Reconstruction de réseaux phylogénétiques

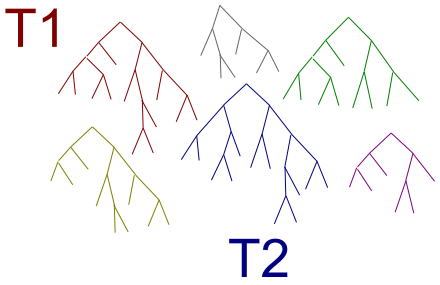
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

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

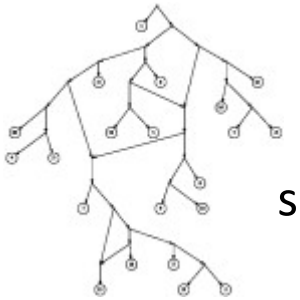
ANR

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

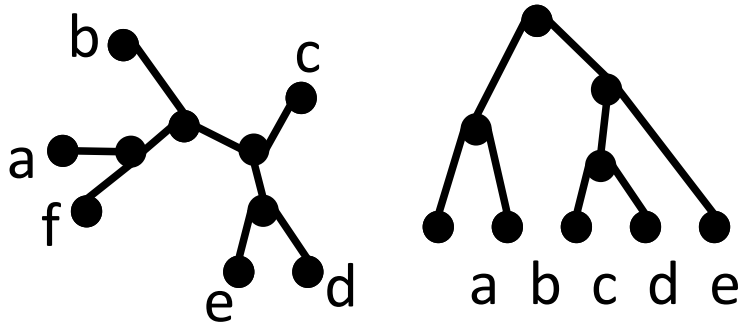
Problème :
Reconstruire le **super-réseau** d'un ensemble d'arbres est

difficile.

Idée :
reconstuire un réseau contenant tous les :

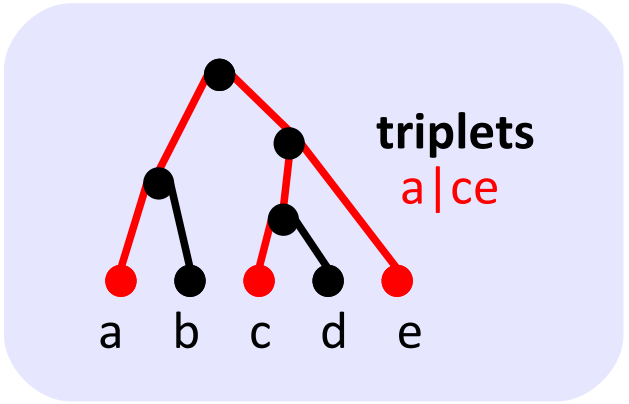
triplets
quadruplets
clades
bipartitions

des arbres en entrée ?



Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :

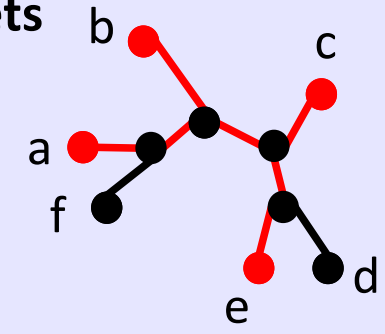


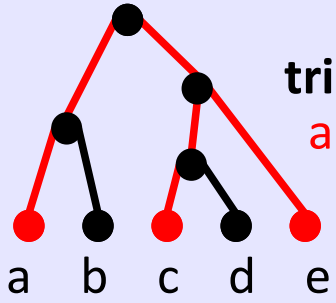
des arbres en entrée ?

Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :

quadruplets
ab|ce





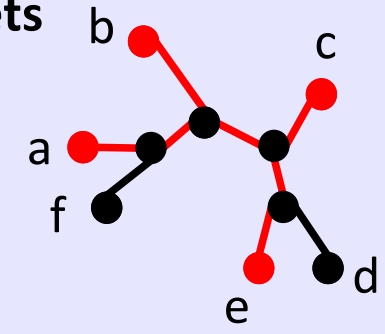
triplets
a|ce

des arbres en entrée ?

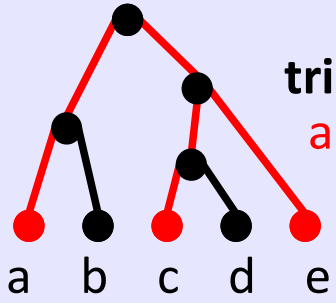
Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :

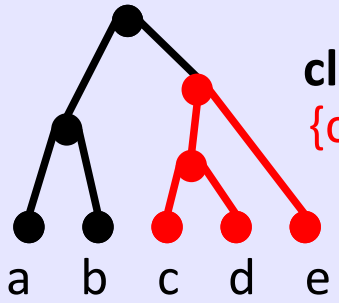
quadruplets
ab|ce



triplets
a|ce



clades
{c,d,e}

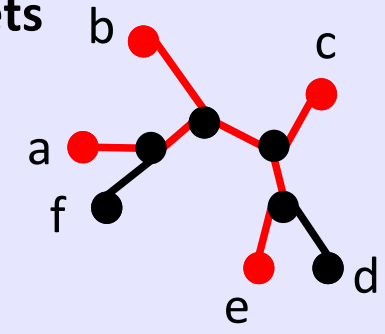


des arbres en entrée ?

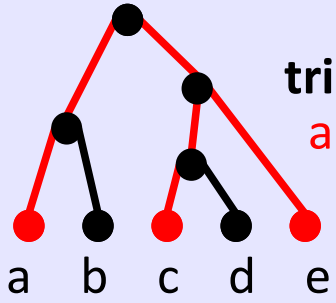
Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :

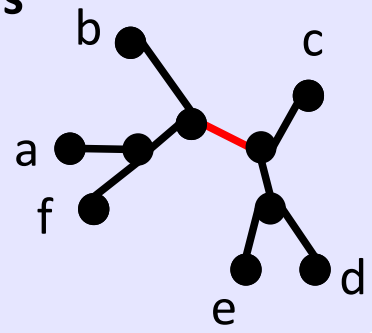
quadruplets
ab|ce



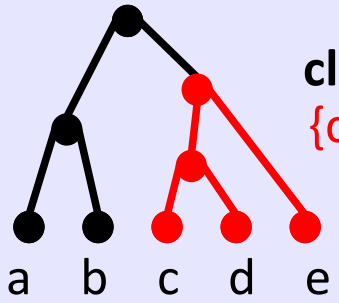
triplets
a|ce



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



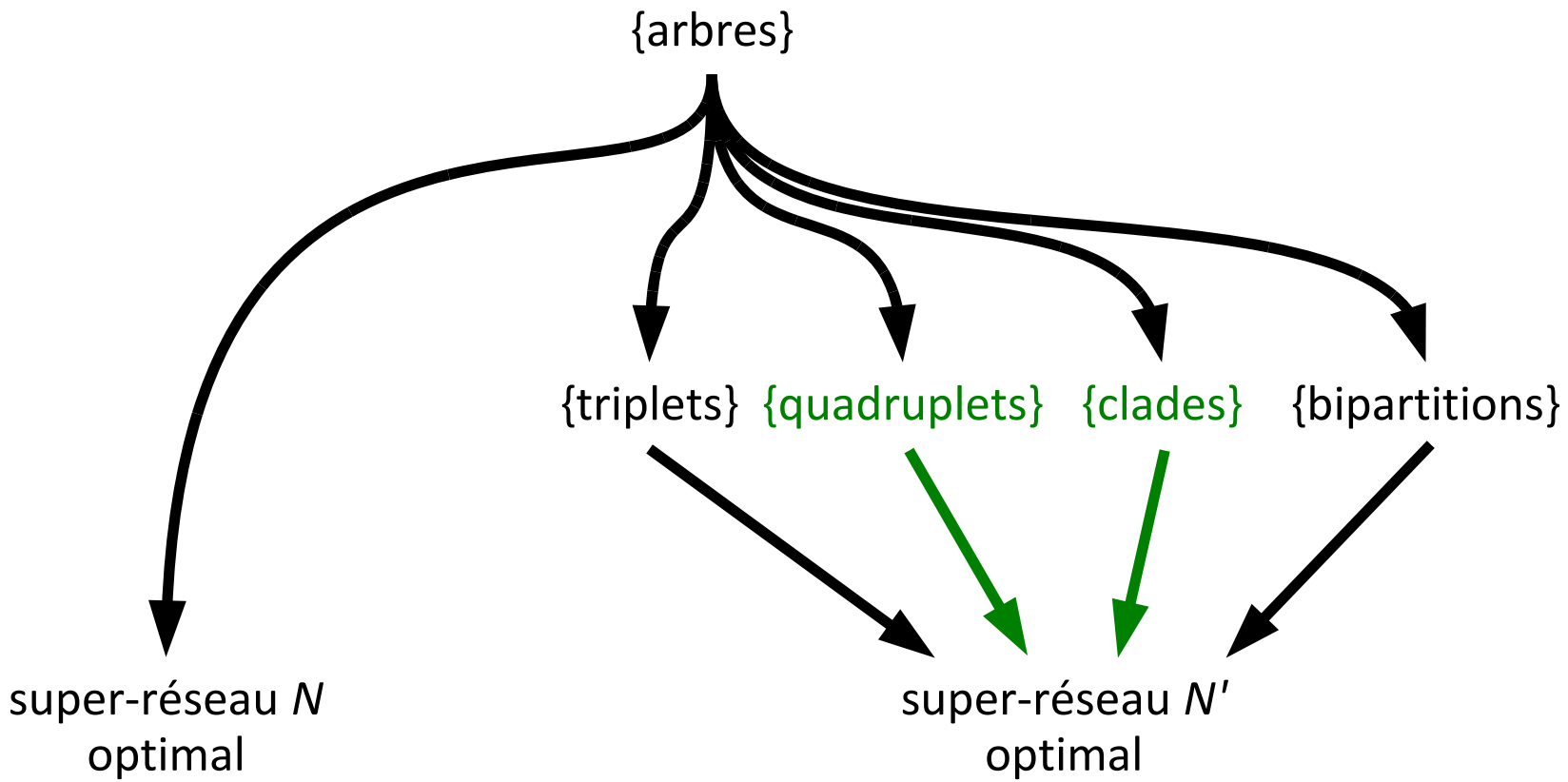
clades
{c,d,e}



des arbres en entrée ?

Reconstruction combinatoire de réseaux phylogénétiques

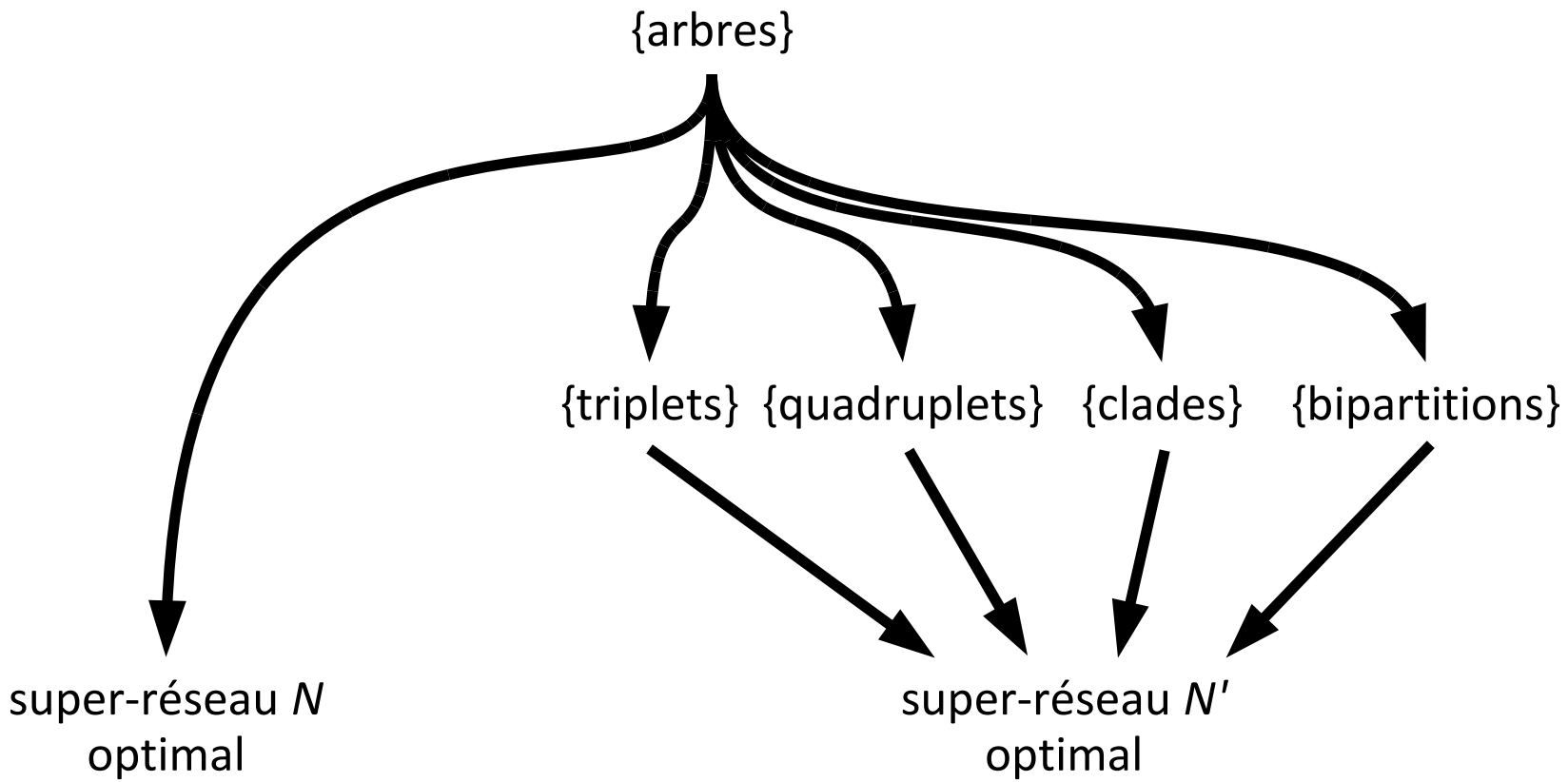
Idée :
modifier le type de données à traiter



Huson, Rupp, Berry, Gambette & Paul, IMSB 2009
Gambette, Berry & Paul, 2010

Reconstruction combinatoire de réseaux phylogénétiques

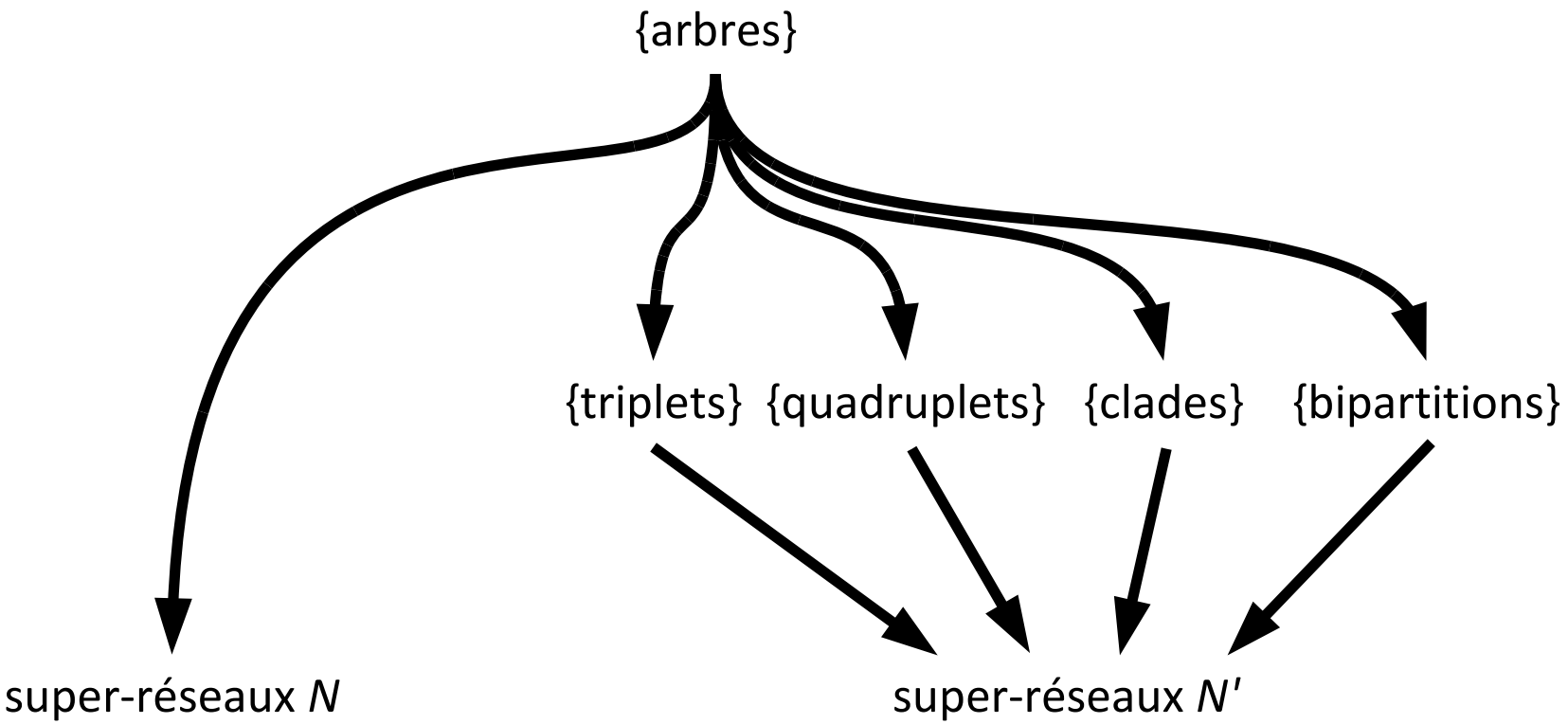
Idée :
modifier le type de données à traiter



N=N' ?

Reconstruction combinatoire de réseaux phylogénétiques

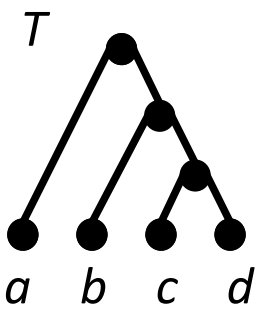
Idée :
modifier le type de données à traiter

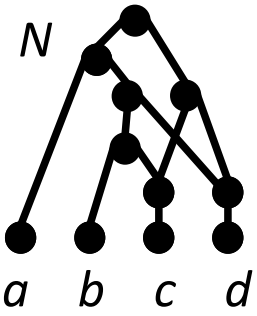


$$\{ N \} \subseteq \{ N' \}$$

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***.



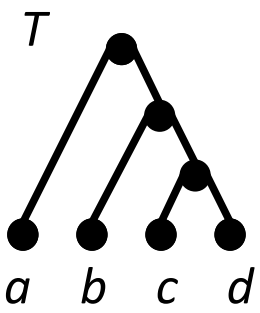


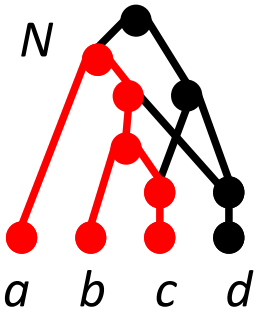
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



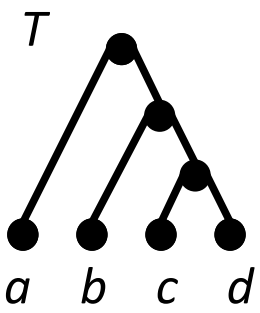


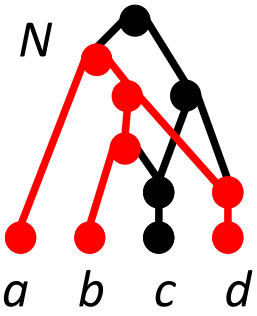
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



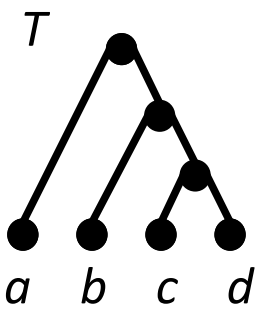


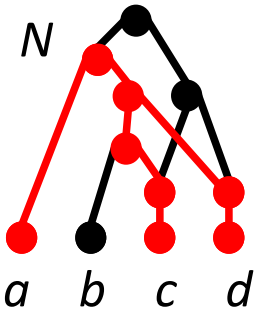
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



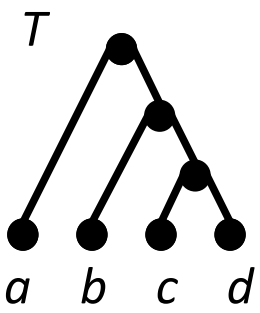


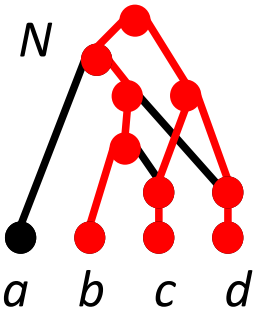
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



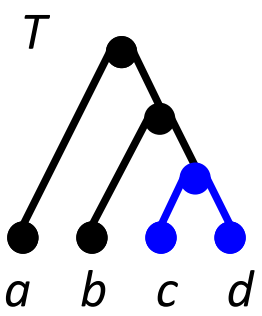


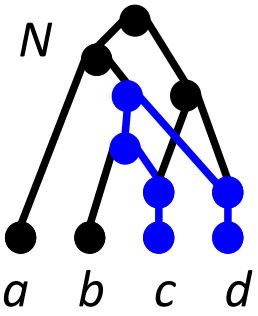
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



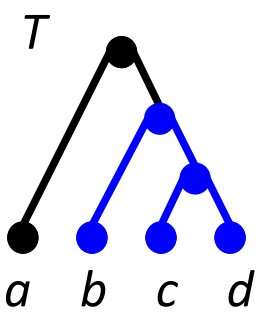


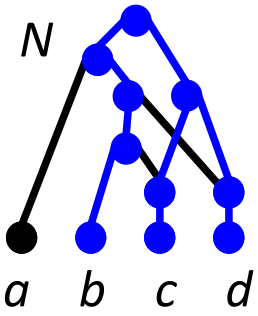
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



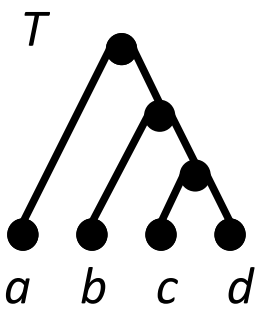


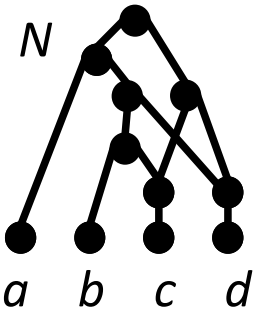
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.

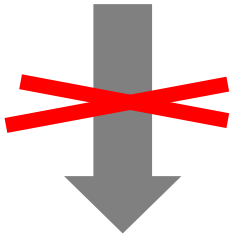




contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

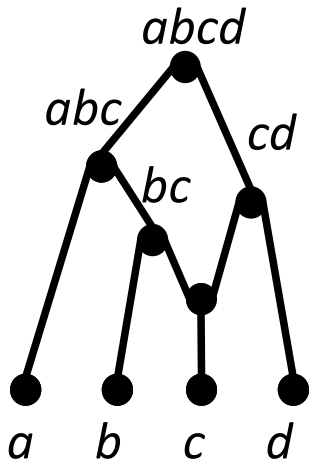
contient les clades / triplets d'un arbre *T*



contient *T*.

Clades stricts et souples

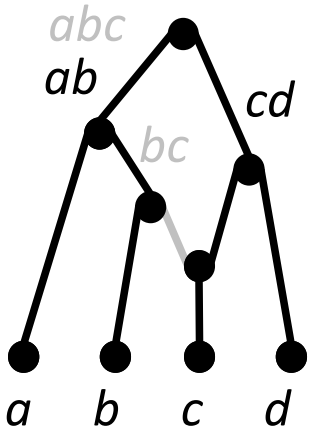
Clade “strict” : ensemble des feuilles sous un noeud du réseau



Clades stricts et souples

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

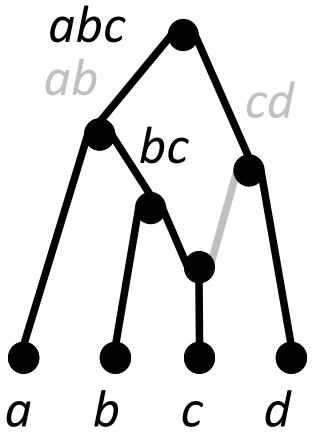
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



Clades stricts et souples

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

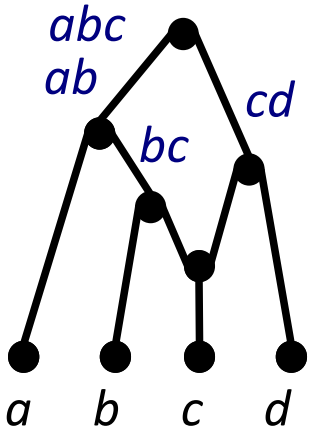
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



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.**

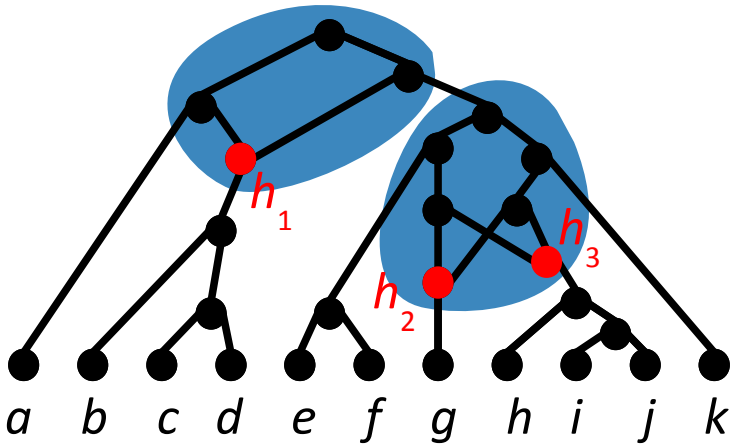
Kanj, Nakhleh, Than & Xia, TCS, 2008

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- Limites
- Application pratique
- Perspectives

Réseaux phylogénétiques restreints

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



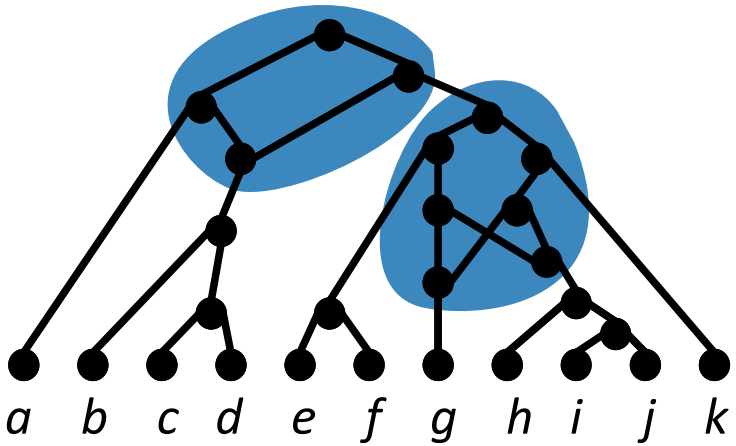
réseau de niveau 2

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

Choy, Jansson, Sadakane & Sung, TCS, 2005

Réseaux phylogénétiques restreints

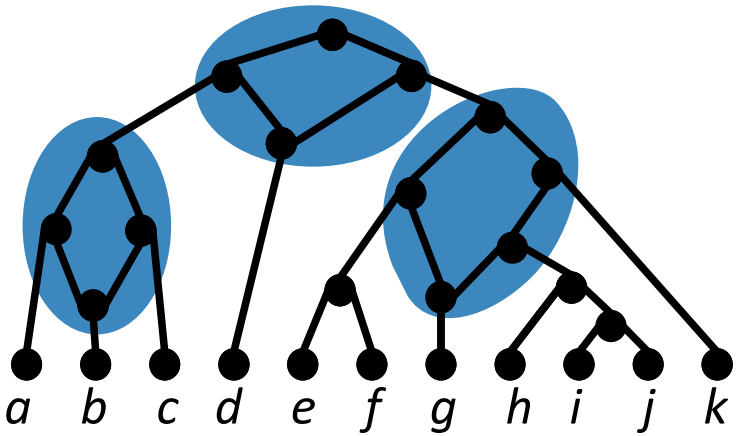
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau de niveau 2

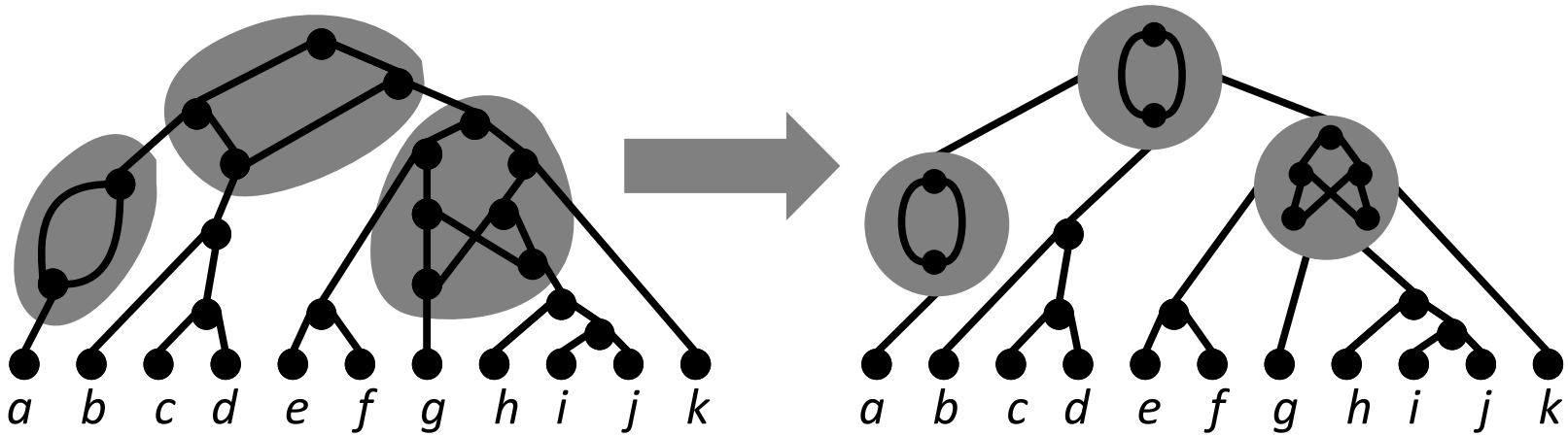
niveau =
nombre maximum d'hybridations
par partie non arborée (*blob*).

réseau de niveau 1
(“*galled tree*”)



Décomposition des réseaux de niveau *k*

Décomposition en blobs :



N, réseau de niveau *k*.

décomposition arborée de *N* en générateurs.

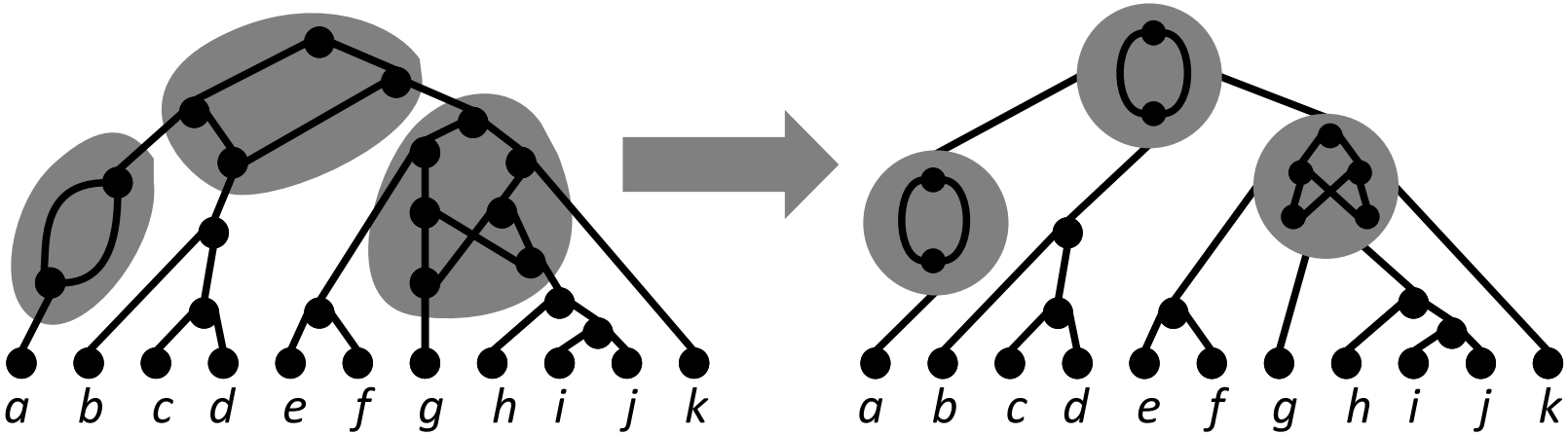
Gambette, Berry & Paul, CPM 2009

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*

Décomposition en blobs :



N, réseau de niveau *k*.

décomposition arborée de *N* en générateurs.

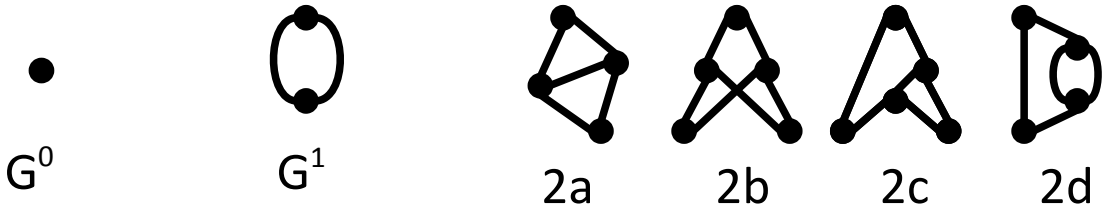
Gambette, Berry & Paul, CPM 2009

Conséquence algorithmique :
Pour *k* fixé, **algorithme polynomial** de décision et de reconstruction des réseaux de niveau *k* à partir de **clades souples** ou **triplets**

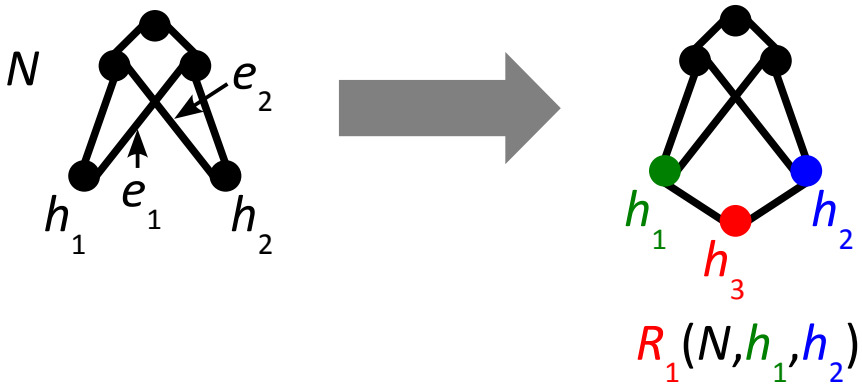
Van Iersel & Kelk, arXiv, 2009
Kelk, Scornavacca & Van Iersel, soumis, 2011

Décomposition des réseaux de niveau *k*

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



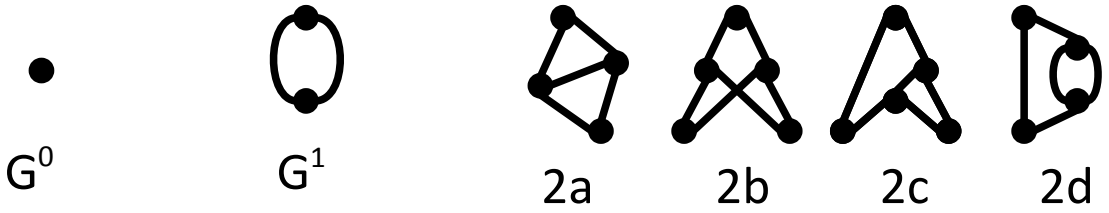
Règles de construction des générateurs de niveau *k*+1 à partir de ceux de niveau *k*



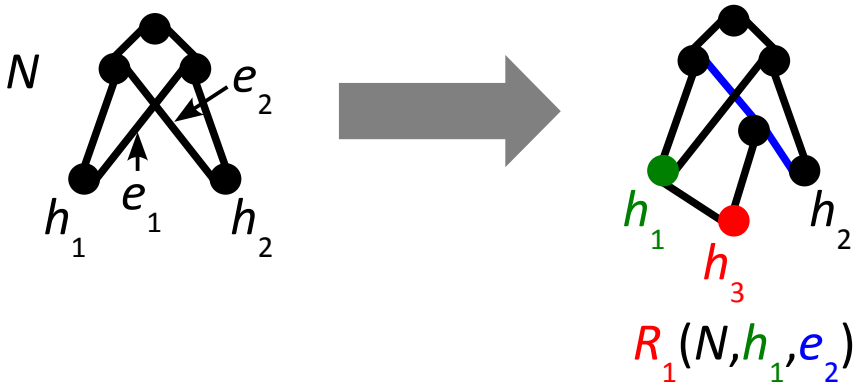
Gambette, Berry & Paul, CPM 2009

Décomposition des réseaux de niveau *k*

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



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

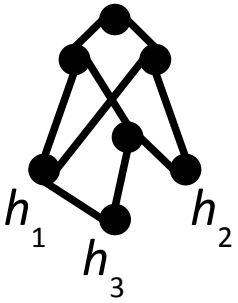


Gambette, Berry & Paul, CPM 2009

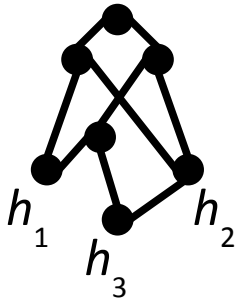
Construction des générateurs

Problème !

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



$R_1(N,h_1,e_2)$



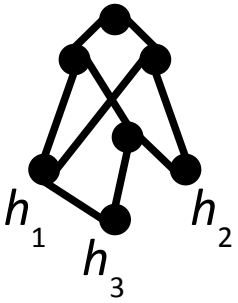
$R_1(N,h_2,e_1)$

Gambette, Berry & Paul, CPM 2009

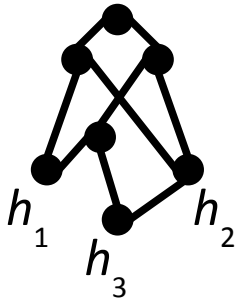
Construction des générateurs

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 !

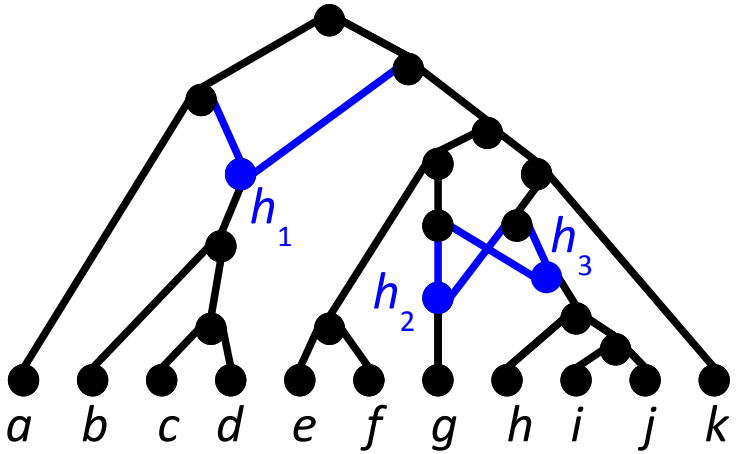
→ génération possible jusqu'à niveau 5 :
1, 4, 65, 1993, 91454

Gambette, Berry & Paul, CPM 2009

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

Réseaux phylogénétiques restreints

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

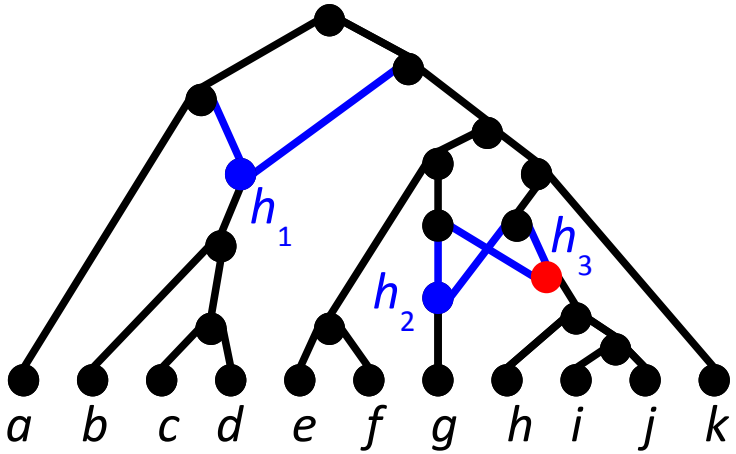


réseau à une couche de réticulation.

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

Réseaux phylogénétiques restreints

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

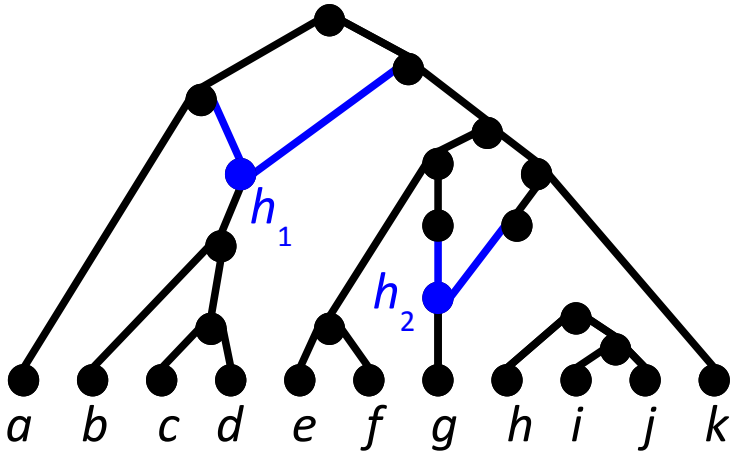


réseau à une couche de réticulation.

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

Réseaux phylogénétiques restreints

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

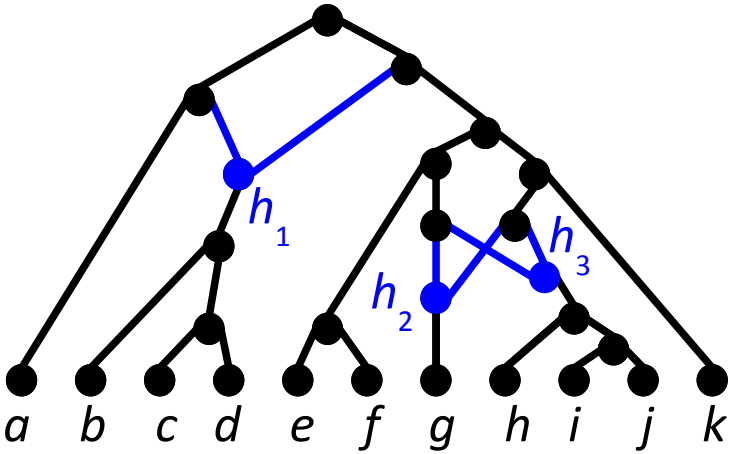


réseau à une couche de réticulation.

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

Réseaux phylogénétiques restreints

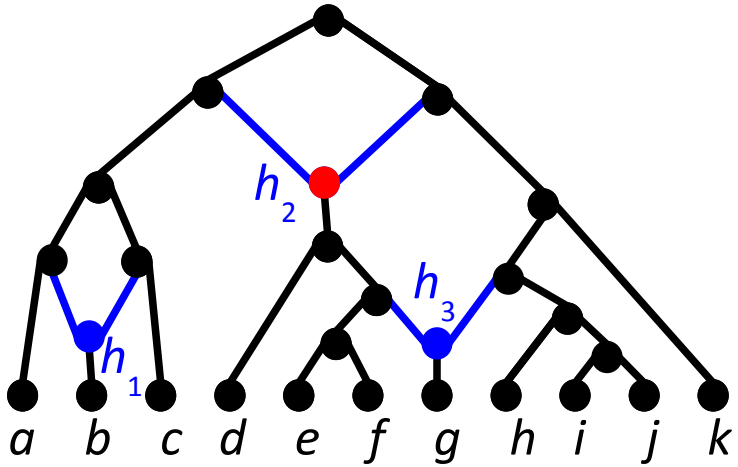
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau à une couche de réticulation.

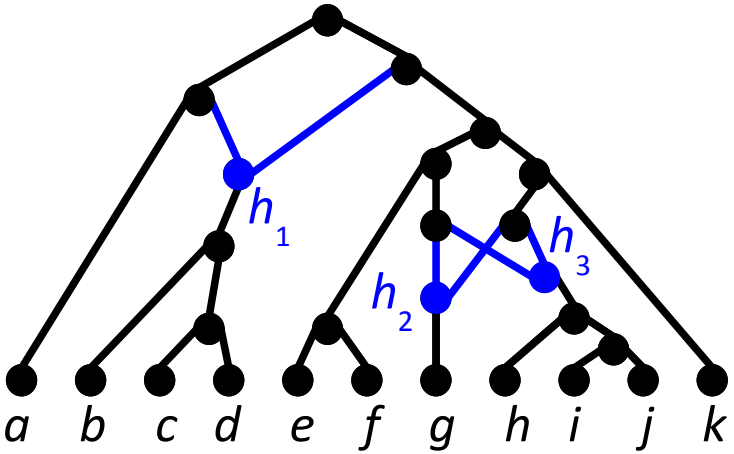
réseau **à une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à deux couches de réticulation.



Réseaux phylogénétiques restreints

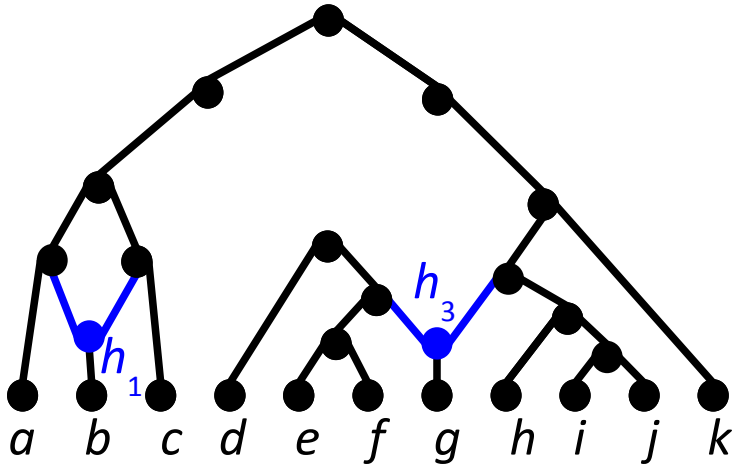
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



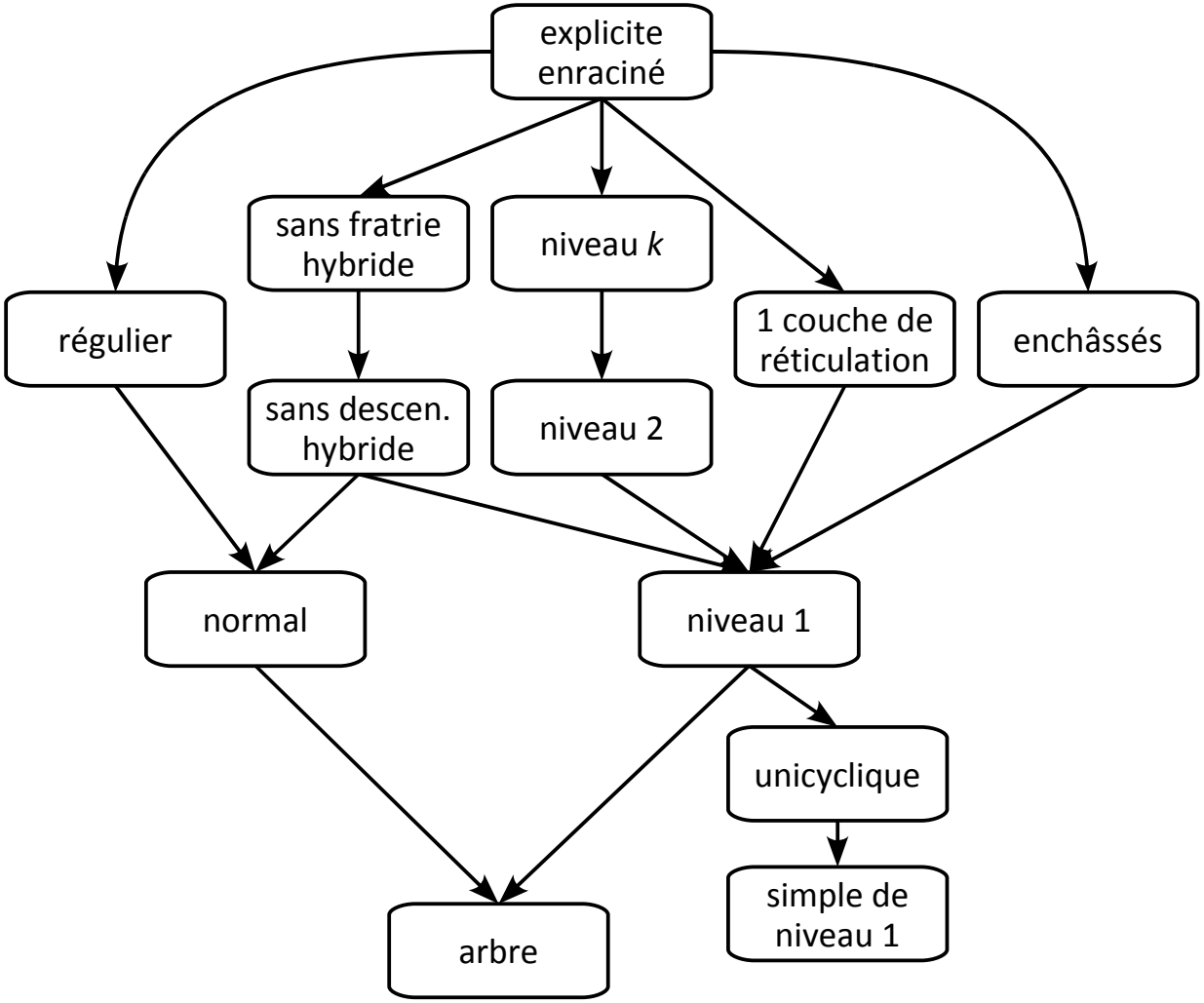
réseau à une couche de réticulation.

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

réseau à deux couches de réticulation.

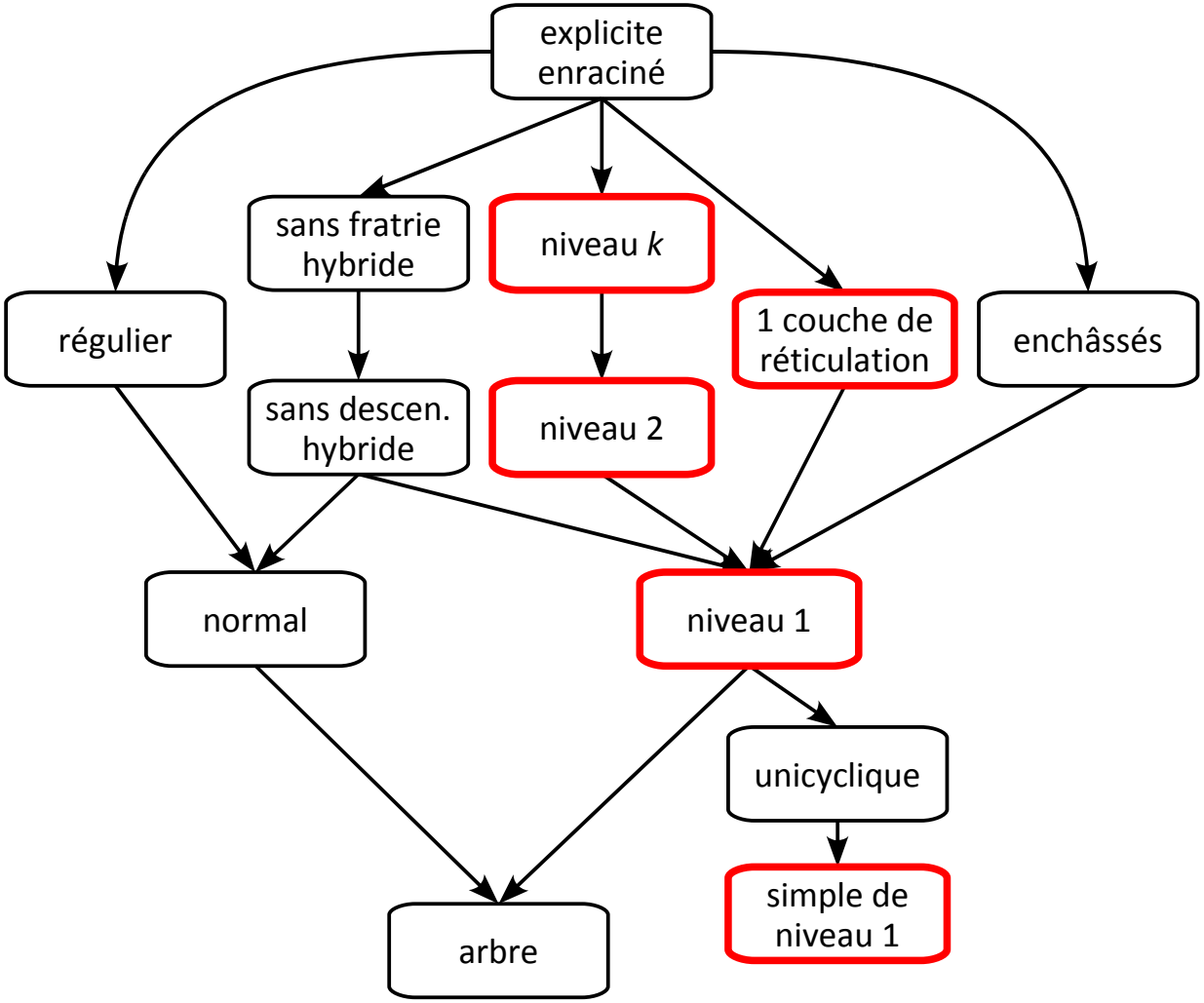


Hiérarchie de sous-classes de réseaux



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

Hiérarchie de sous-classes de réseaux



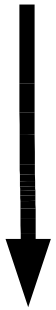
http://www.lirmm.fr/~gambette/RePhylogeneticNetworks.php

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- Limites
- Application pratique
- Bilan
- 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

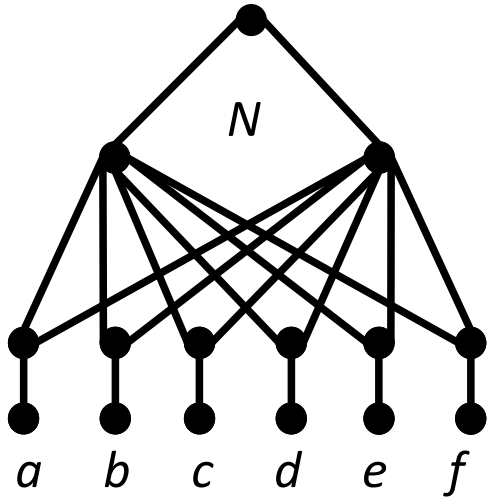
- meilleurs résultats mais plus lente pour niveau > 2.
- pour *k* fixé, certains ensembles de clades contenus dans aucun réseau de niveau *k*.

http://www.dendroscope.org

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***.**



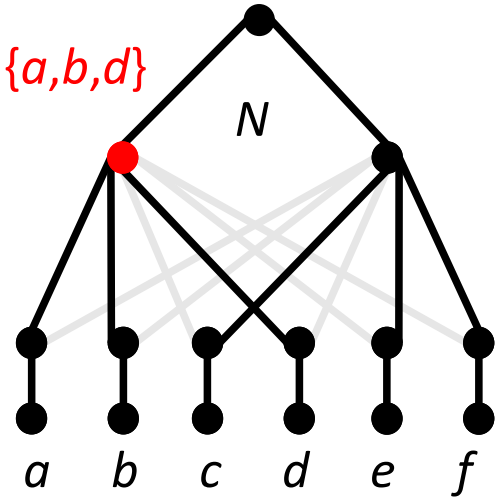
Le réseau à une couche de réticulation *N* est compatible avec
tout clade souple sur {*a*,*b*,*c*,*d*,*e*,*f*}.

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***.**



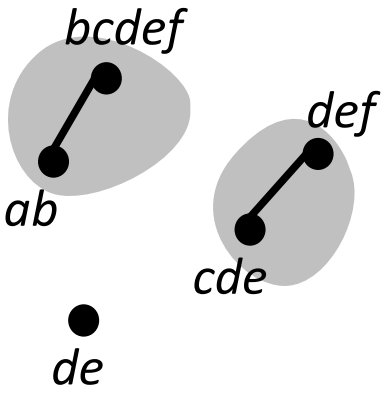
Le réseau à une couche de réticulation *N* est compatible avec
tout clade souple sur {*a*,*b*,*c*,*d*,*e*,*f*}.

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

Clades et réseaux à une couche de réticulation

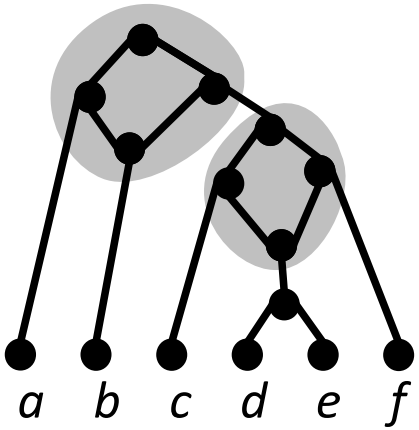
Décomposition

décomposition de l'ensemble de clades → décomposition du réseau



graphe des conflits

Gusfield, Eddhu, Langley, *JBCB*, 2004



réseau reconstruit

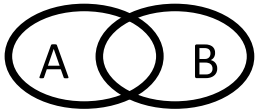
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

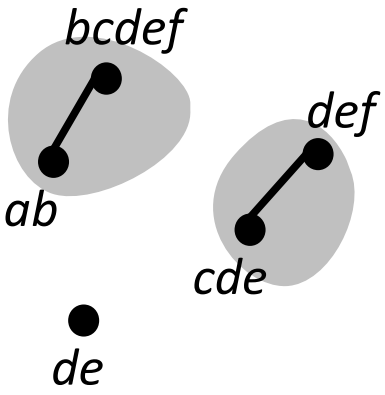
Clades et réseaux à une couche de réticulation

Décomposition

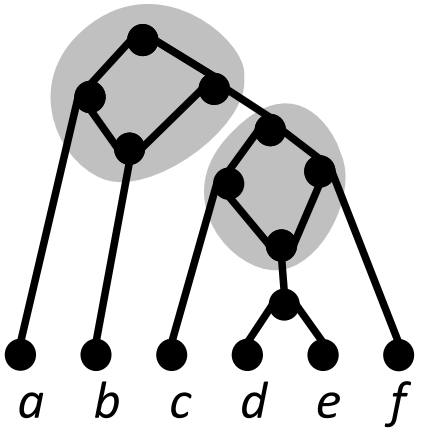
décomposition de l'ensemble de clades → décomposition du réseau

conflit :
clades ni inclus ni disjoints





graphe des conflits



réseau reconstruit

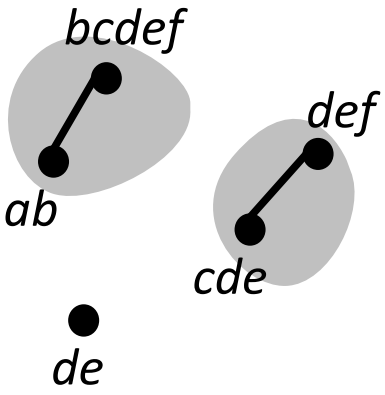
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

Clades et réseaux à une couche de réticulation

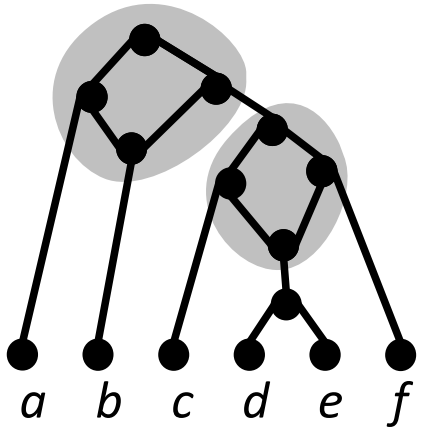
Décomposition

décomposition de l'ensemble de clades → décomposition du réseau

composante connexe du graphe des conflits ↔ **blob**



graphe des conflits



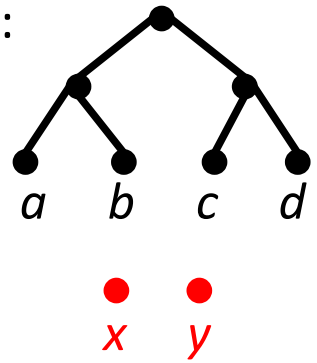
réseau reconstruit

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

Une approche en deux étapes

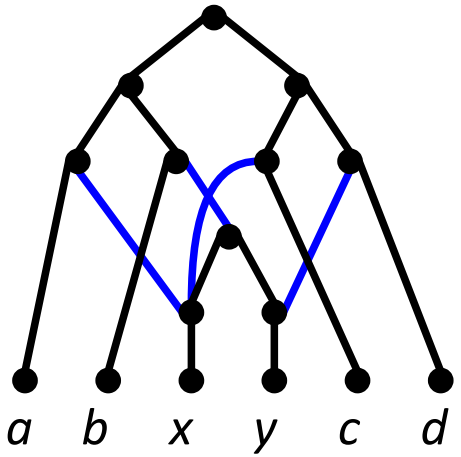
1- Trouver un **ensemble minimum de conflits** parmi les clades :
- partie sans conflits ➡ arbre,
- taxons impliqués dans des conflits ➡ sous les réticulations.

MAXIMUM COMPATIBLE SUBSET



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

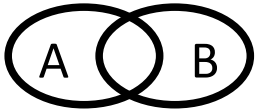
MINIMUM ATTACHMENT PROBLEM



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 des conflits connexe, sans taxons "jumeaux"

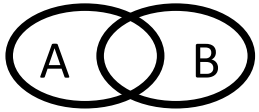
réduction depuis le cas général

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

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

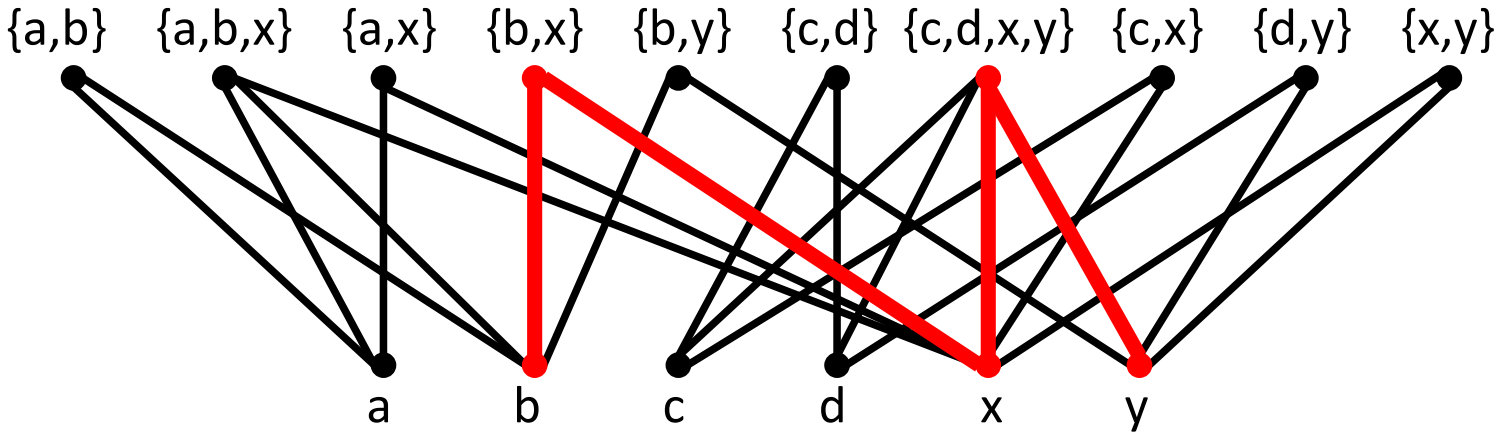
L'ensemble minimum de conflits

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}}

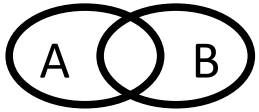


conflit = graphe "M"

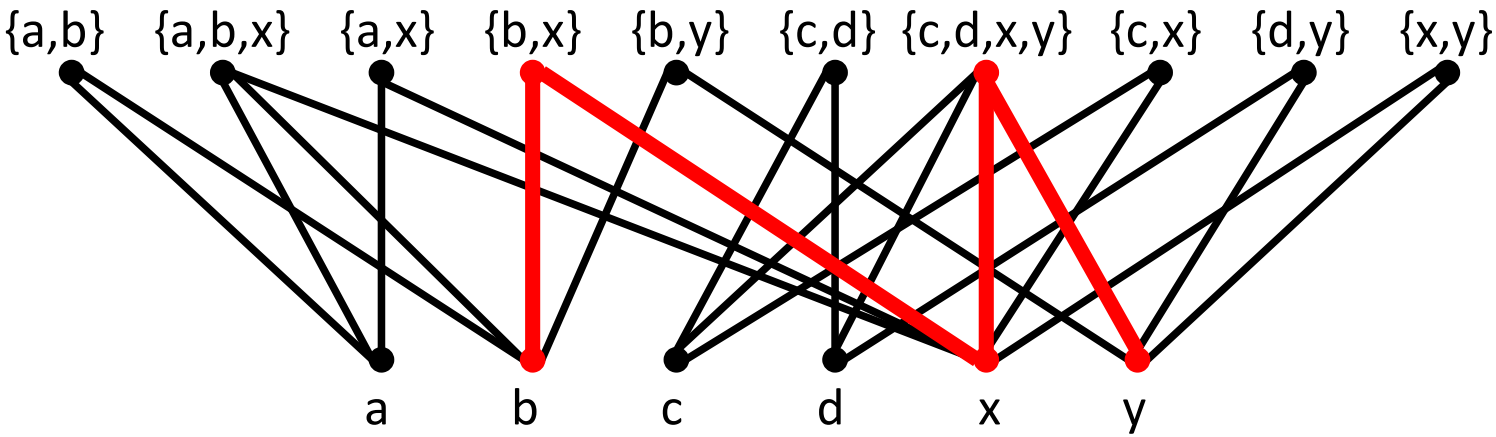
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'ensemble minimum de conflits

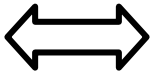
Conflit : clades ni inclus ni disjoints



Graphe des caractères :



Trouver l'ensemble minimum de conflits

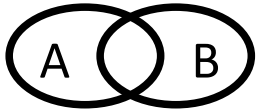


Supprimer le nombre minimum *t* de sommets-taxons tels que le graphe des caractères est un graphe “sans M”

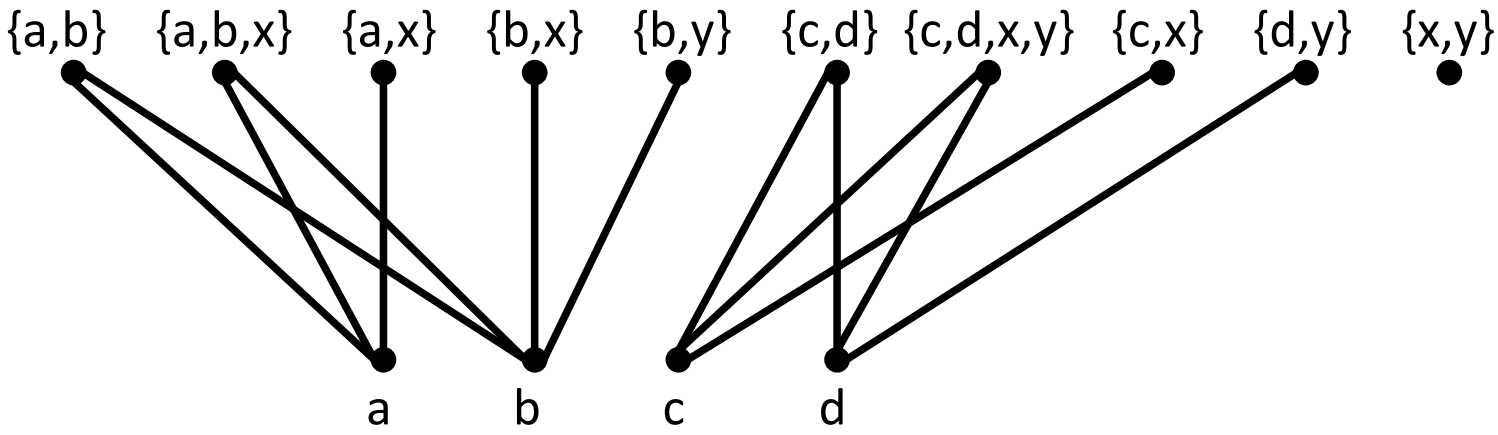
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

L'ensemble minimum de conflits

Conflit : clades ni inclus ni disjoints



Graphe des caractères :

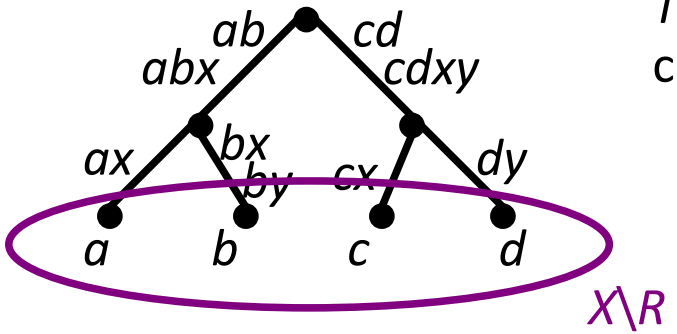


Supprimer le nombre minimum *t* de sommets-taxons tels que le graphe des caractères est un graphe “sans M” :
- algorithme FPT basique de branchement en $O^*(3^t)$
- algorithme FPT 3-Hitting-Set en $O^*(2{,}076^t)$

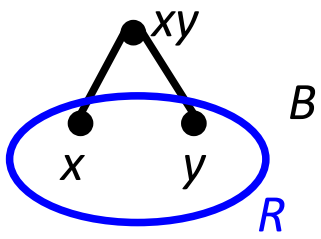
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

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*



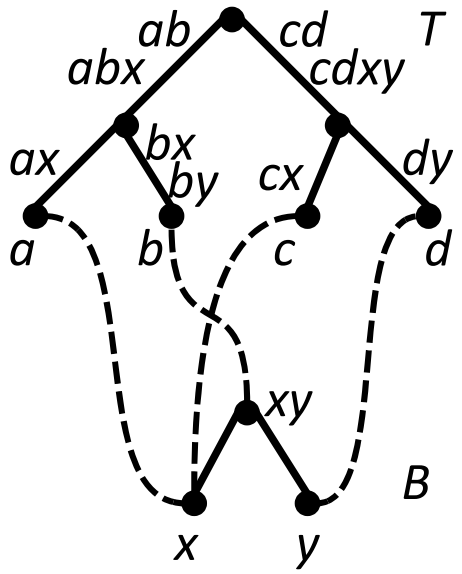
B : réseau représentant les clades maximaux sur *R* et les singletons de *R*.

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

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

L'attachement minimum

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



NP-complet

réduction depuis SetCover

Algorithmes :
- Séparation et évaluation

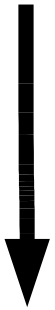
implémenté dans Dendroscope 2

- Programme linéaire en nombres entiers

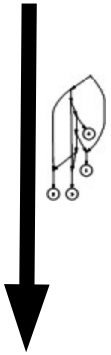
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

Reconstruction depuis les triplets

{arbres}



{triplets}



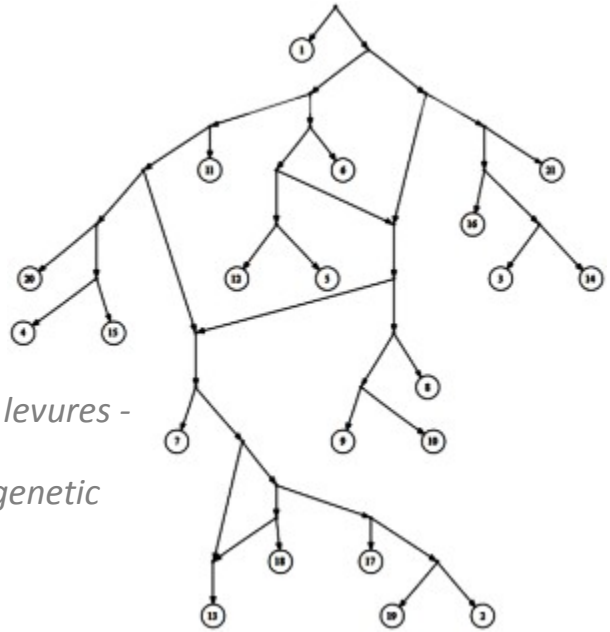
N'
réseau
de niveau *k*

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

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

Reconstruction depuis les quadruplets

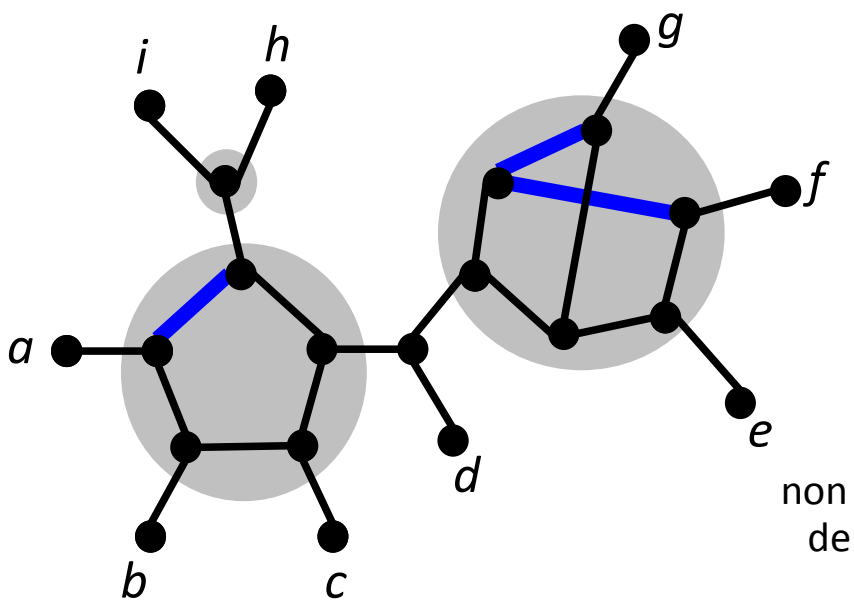
{arbres}

{quadruplets}

N' réseau
non enraciné
de niveau *k*

Réseau **non enraciné de niveau *k***

➡ nombre maximum d'arêtes à supprimer,
par blob, pour obtenir un arbre



réseau
non enraciné
de niveau 2

Orientation enracinée : réseau de niveau *k*
Niveau **invariant** selon l'orientation

Gambette, Berry & Paul, 2011

Reconstruction d'arbres

	non enraciné / quadruplets	enraciné / triplets
général	**NP-complet** *Steel, JOC, 1992*	**polynomial** *Aho, Sagiv, Szymanski & Ullman, SJOC, 1981* *Henzinger, King & Warnow, ALG, 1999* *Jansson, Ng, Sadakane & Sung, ALG, 2005*
dense *au moins un quadruplet pour tout ensemble de 4 feuilles*	$\boldsymbol{O(n^4)}$ *Berry & Gascuel, TCS, 2000*	$\boldsymbol{O(n^3)}$ *Aho et al., SJOC, 1981*

Reconstruction de réseaux de niveau *k*

	non enraciné		enraciné	
	niveau 1	niveau k>1	niveau 1	niveau k>1
général	NP-complet *Grünewald, Moulton & Spillner, DAM, 2009*	?	NP-complet *Jansson, Nguyen & Sung, SJOC, 2006*	NP-complet *Van Iersel, Kelk & Mnich, JBCB, 2009*
dense *au moins un quadruplet pour tout ensemble de 4 feuilles*	**?** **(décomposition en temps polynomial)**	?	$O(n^3)$ *Jansson, Nguyen & Sung, SJOC, 2006*	$O(n^{5k+4})$ *To & Habib, CPM 2009*
complet *tous les quadruplets du réseau*	$O(n^4)$	? **(décomposition en temps polynomial)**	$O(n^3)$ *Jansson, Nguyen & Sung, SJOC, 2006*	$O(n^{3k+3})$ *Van Iersel & Kelk, ALG, 2010*

Gambette, Berry & Paul, 2011

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- Limites
- Application pratique
- Bilan
- Perspectives

Explosion combinatoire

• Les restrictions sur les réseaux reconstruits n'empêchent pas l'**explosion combinatoire**

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

Rappel :
Un réseau de niveau *k* se décompose en
un arbre de générateurs choisis parmi un ensemble fini.

Nombre exponentiel de générateurs de niveau *k* :

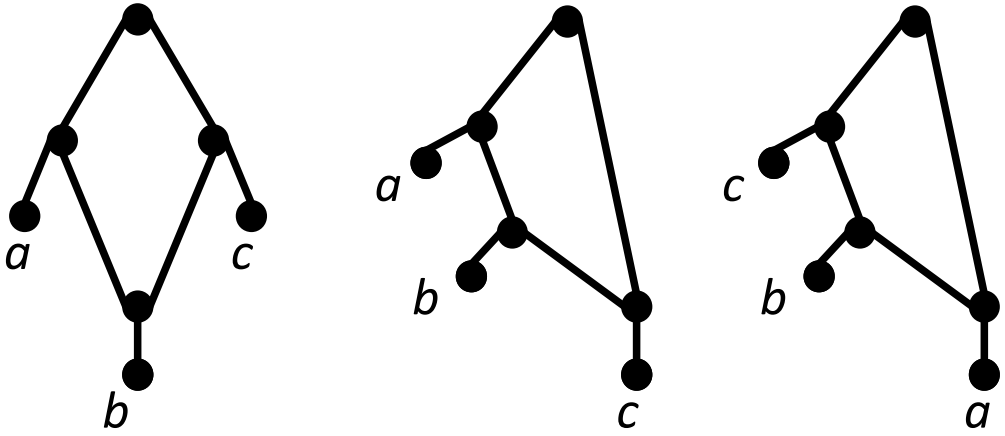
$$g_k \geq 2^{k-1}$$

Gambette, Berry & Paul, CPM, 2009

Ambiguïté des solutions

- Les restrictions sur les réseaux reconstruits n'empêchent pas l'**explosion combinatoire**
- **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)

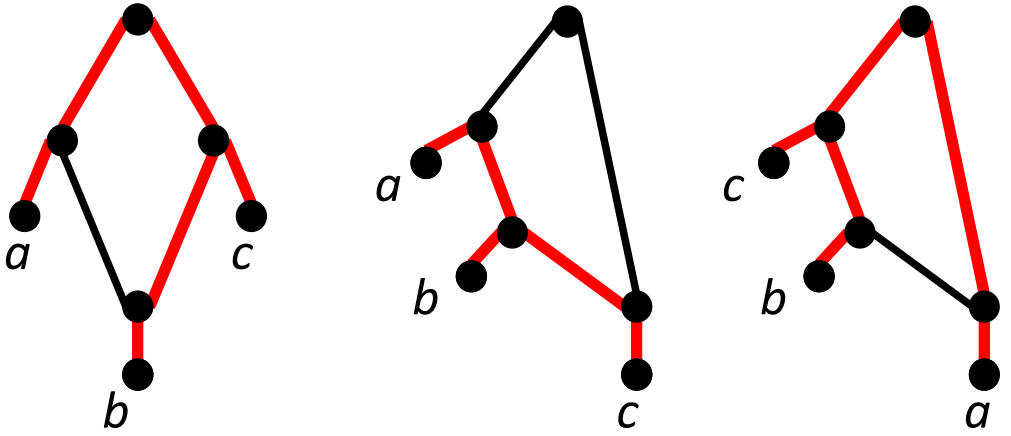
Gambette & Huber, 2011

Ambiguïté des solutions

- Les restrictions sur les réseaux reconstruits n'empêchent pas l'**explosion combinatoire**
- **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



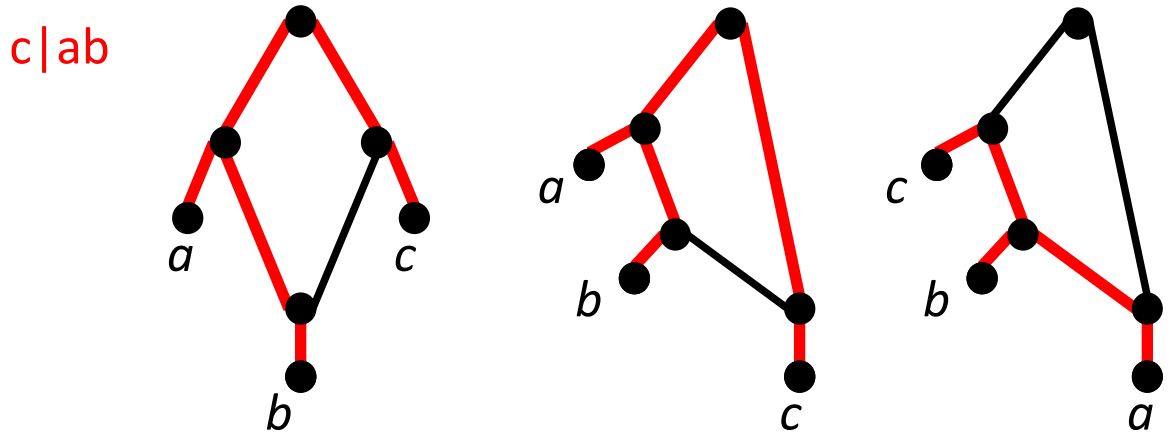
Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 somments)

Gambette & Huber, 2011

Ambiguïté des solutions

- Les restrictions sur les réseaux reconstruits n'empêchent pas l'**explosion combinatoire**
- **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.



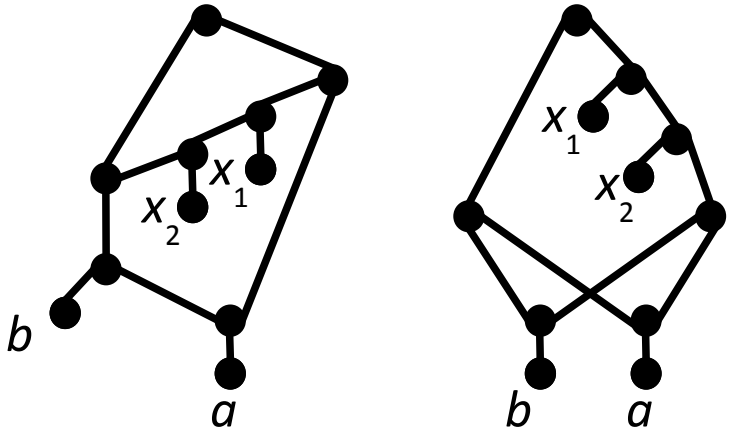
Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 somments)

Gambette & Huber, 2011

Ambiguïté des solutions

- Les restrictions sur les réseaux reconstruits n'empêchent pas l'**explosion combinatoire**
- **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.



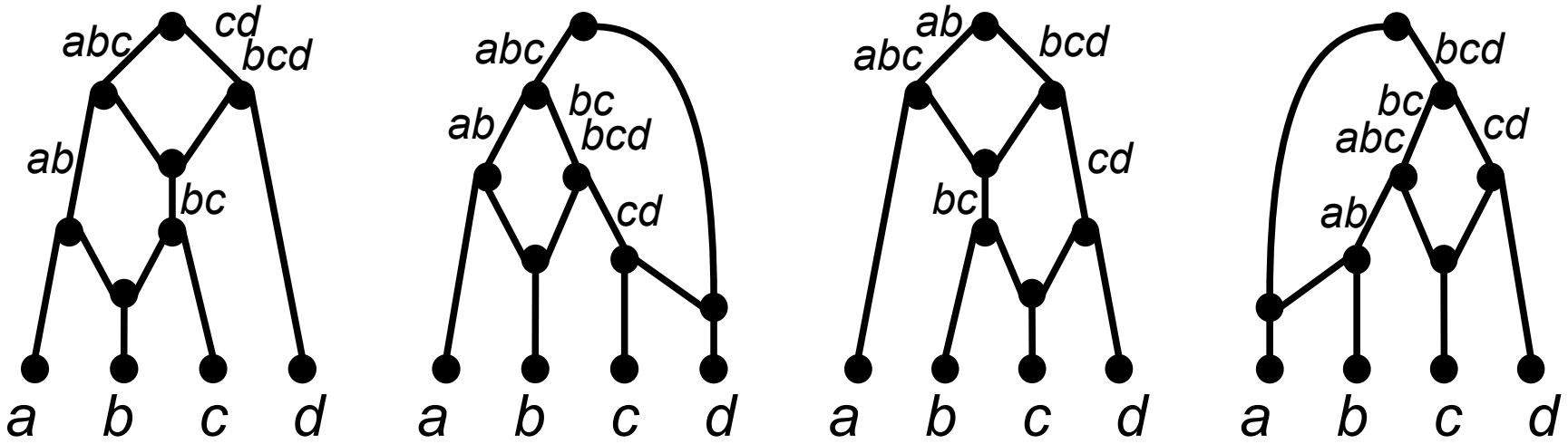
2 réseaux de niveau 2 avec le même ensemble de triplets

Gambette & Huber, 2011

Ambiguïté des solutions

- Les restrictions sur les réseaux reconstruits n'empêchent pas l'**explosion combinatoire**
- **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.



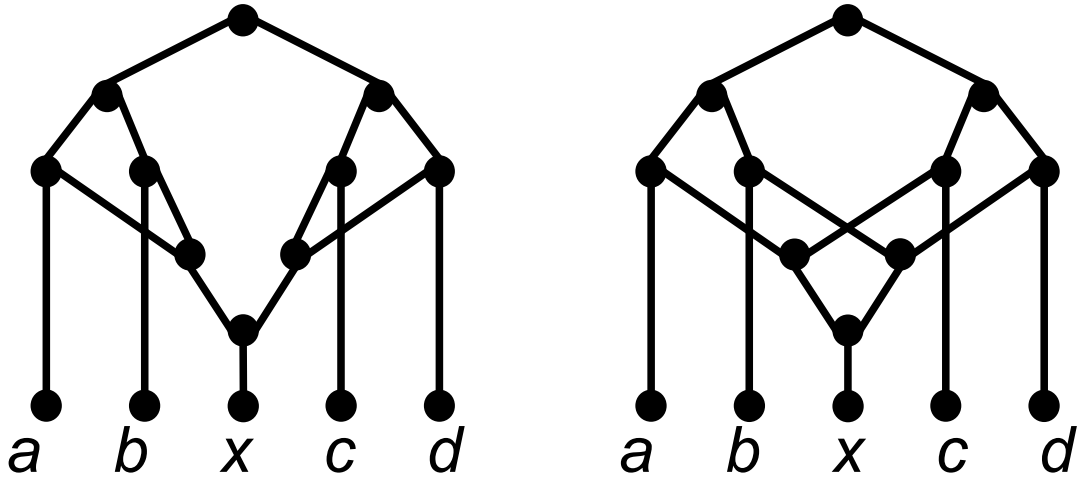
4 réseaux de niveau 2 avec le même ensemble de triplets et de clades

Gambette & Huber, 2011

Ambiguïté des solutions

- Les restrictions sur les réseaux reconstruits n'empêchent pas l'**explosion combinatoire**
- **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 3 avec le même ensemble d'arbres, de triplets, de clades

Huson, Rupp & Scornavacca, 2011

Ambiguïté des solutions

- Les restrictions sur les réseaux reconstruits n'empêchent pas l'**explosion combinatoire**
- **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.

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

Gambette & Huber, 2011

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- Limites
- Application pratique
- Perspectives

Bruit dans les données

• Approches de **filtres** :
Ne considérer que les clades, triplets, avec un bon support.
Ex : clades présents dans 20% des arbres en entrée

• Approches d'**édition des données** :
Corriger les données au minimum pour obtenir un réseau restreint :
- arbres à partir de clades :

$O^*(3^l)$, Huson, Rupp, Berry, Gambette & Paul, ISMB 2009

- arbres à partir de triplets :

$O^*(3^l)$, Guillemot & Berry, TCBB, 2007,
$O(n^4 + 2^{O(t\text{^}(1/3)\log t)})$, Guillemot & Mnich, TAMC 2009.

Problème ouvert :
réseaux de niveau 1 à partir de clades, triplets, quadruplets, bipartitions

Silence dans les données

Nécessité d'avoir des clades **complets**, des **ensembles denses** de **triplets** ou **quadruplets** :

arbres en entrée sur le **même ensemble de taxons**

• **Compléter les données :**
Inférence de triplets

Bryant & Steel, AAM, 1995

Inférence de bipartitions

Huson, Dezulian, Klöpper & Steel, TCBB, 2004
Grünewald, Huber & Wu, BMB, 2008

• **Sélectionner les données :**

Arbres de gènes :
Rectangles maximaux / bicliques maximales

Ensemble dense maximum de triplets :
Problème NP-complet

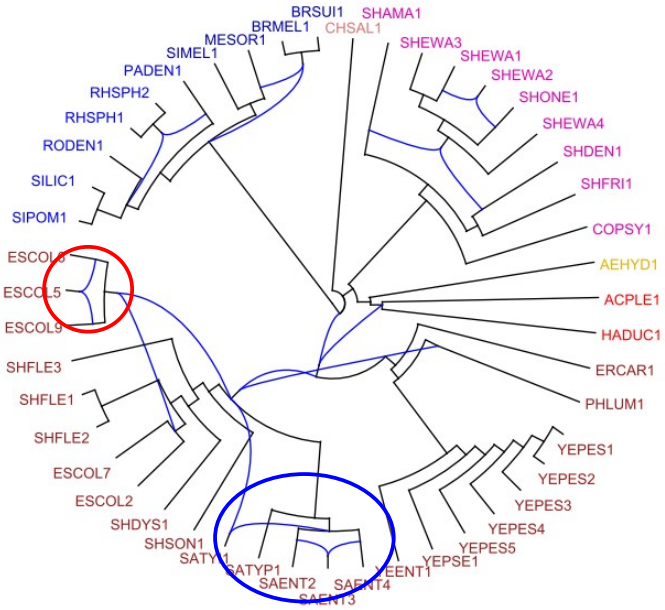
Réduction de Clique

Exemples de résultats

16 arbres de la base HOGENOM sur **47 taxons** (protéobactéries)

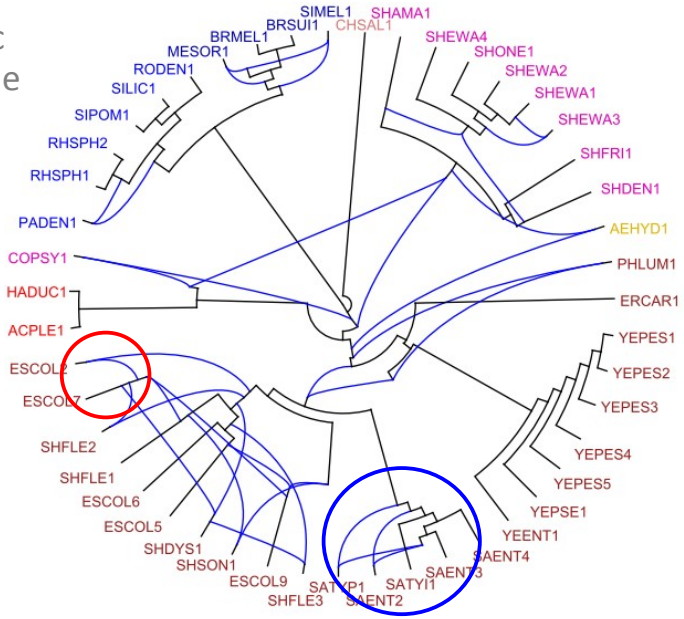
24 Enterobacteriales
2 Pasteurellales
1 Aeromonadales
9 Alteromonadales
1 Oceanospirillales
6 Rhodobacterales
4 Rhizobiales

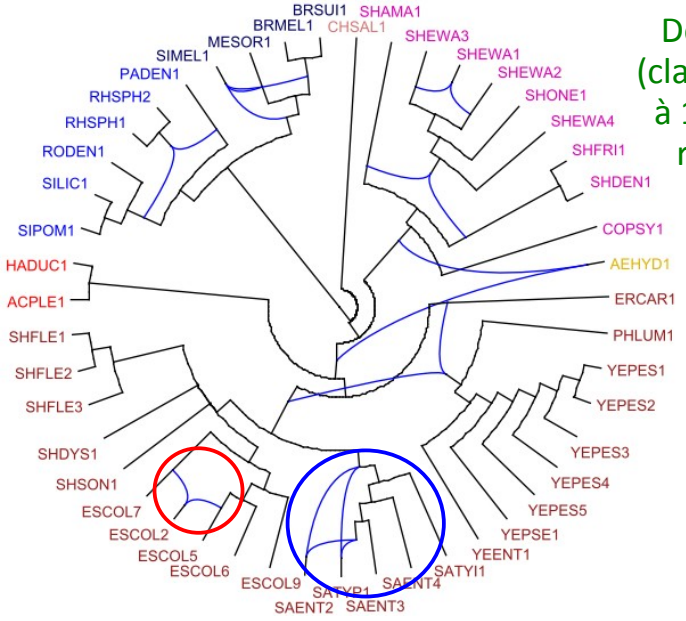
Réseaux contenant les triplets, clades souples, présents dans au moins 20% des arbres



Lev1athan (heuristique triplets, niveau 1) 24 sec.

Simplistic (réseau de niveau 7 à partir de triplets) 63 sec.



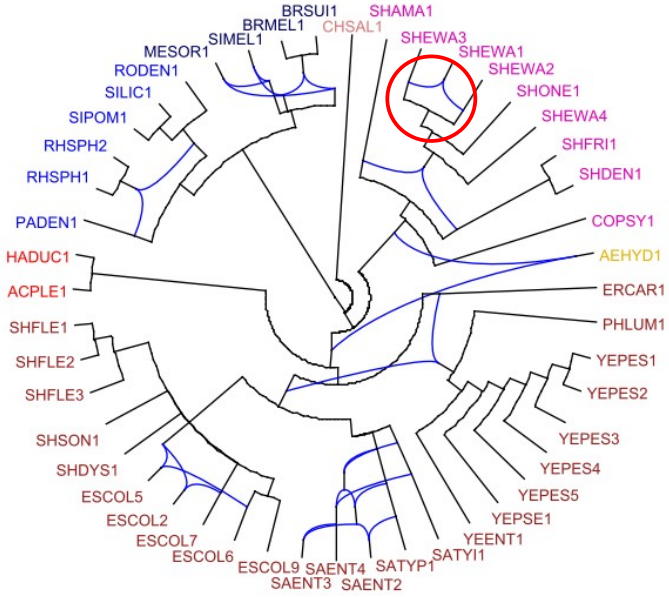


Dendroscope (clades, réseau à 1 couche de réticulation) <1 sec.

Exemples de résultats

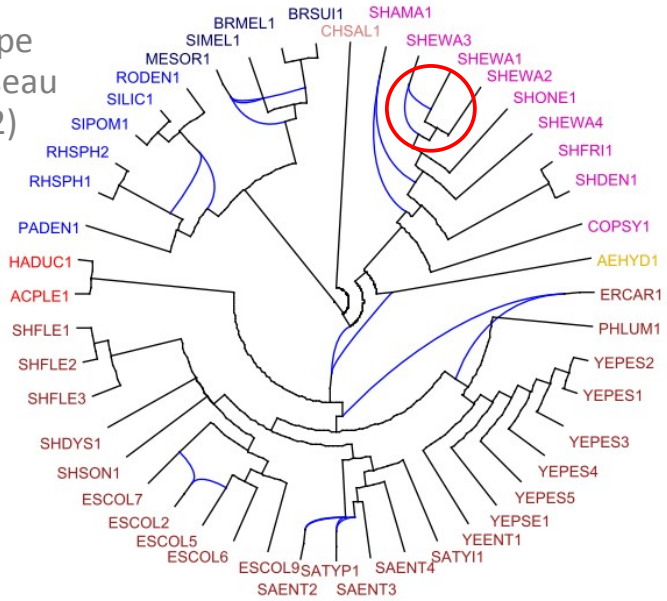
16 arbres de la base HOGENOM sur **47 taxons** (protéobactéries)
24 Enterobacteriales
2 Pasteurellales
1 Aeromonadales
9 Alteromonadales
1 Oceanospirillales
6 Rhodobacterales
4 Rhizobiales

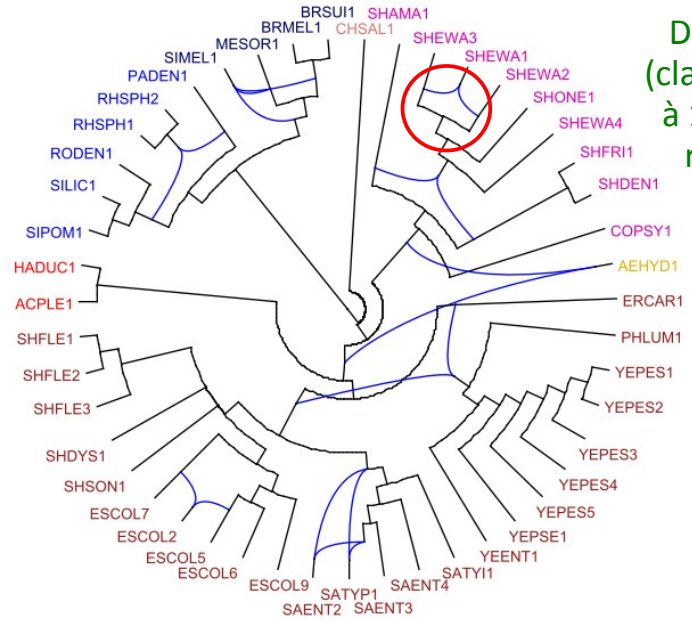
Réseaux contenant les clades souples présents dans au moins 20% des arbres



Dendroscope (réseau de clades) <1 sec.

Dendroscope (clades, réseau de niveau 2) 2 sec.



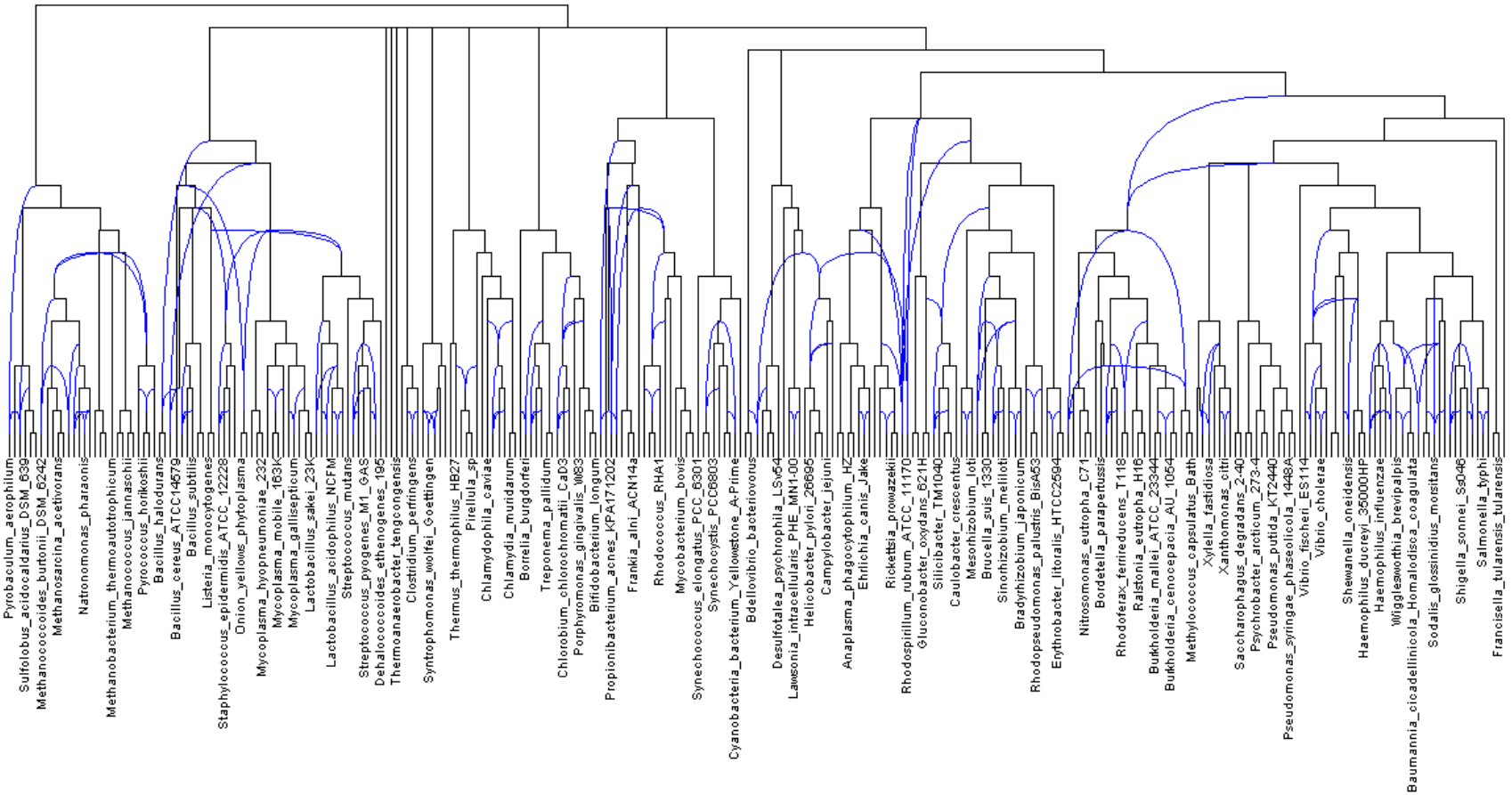


Dendroscope (clades, réseau à 1 couche de réticulation) <1 sec.

Exemples de résultats

9 arbres sur **279 espèces** de procaryotes
Clades dans au moins 2 arbres

Auch, Steigele, Huson & Henz, 2009



Dendroscope
(clades, réseau à 1 couche de réticulation)
2 sec.

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- Limites
- Application pratique
- Bilan
- Perspectives

Bilan

Implémentations

Implémentation Java, dans Dendroscope, de l'algorithme FPT de suppression des conflits

Dendroscope http://www.dendroscope.org

Implémentation Java de l'algorithme de construction des générateurs, de calcul du niveau

Generators http://generators.gambette.com

Implémentations PHP de fonctions supplémentaires à BibAdmin pour la bibliographie interactive sur les réseaux phylogénétiques

Who is Who in Phylogenetic Networks http://www.atgc-montpellier.fr/phylnet

Références

Quartets and Unrooted Phylogenetic Networks

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

The Structure of Level-*k* Phylogenetic Networks

Gambette, Berry & Paul, ***CPM 2009***

On Encodings of Phylogenetic Networks of Bounded Level

Gambette & Huber, soumis à *JMB*

Quartets and Unrooted Phylogenetic Networks

Gambette, Berry & Paul, soumis à *JBCB*

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- Limites
- Application pratique
- Bilan
- Perspectives

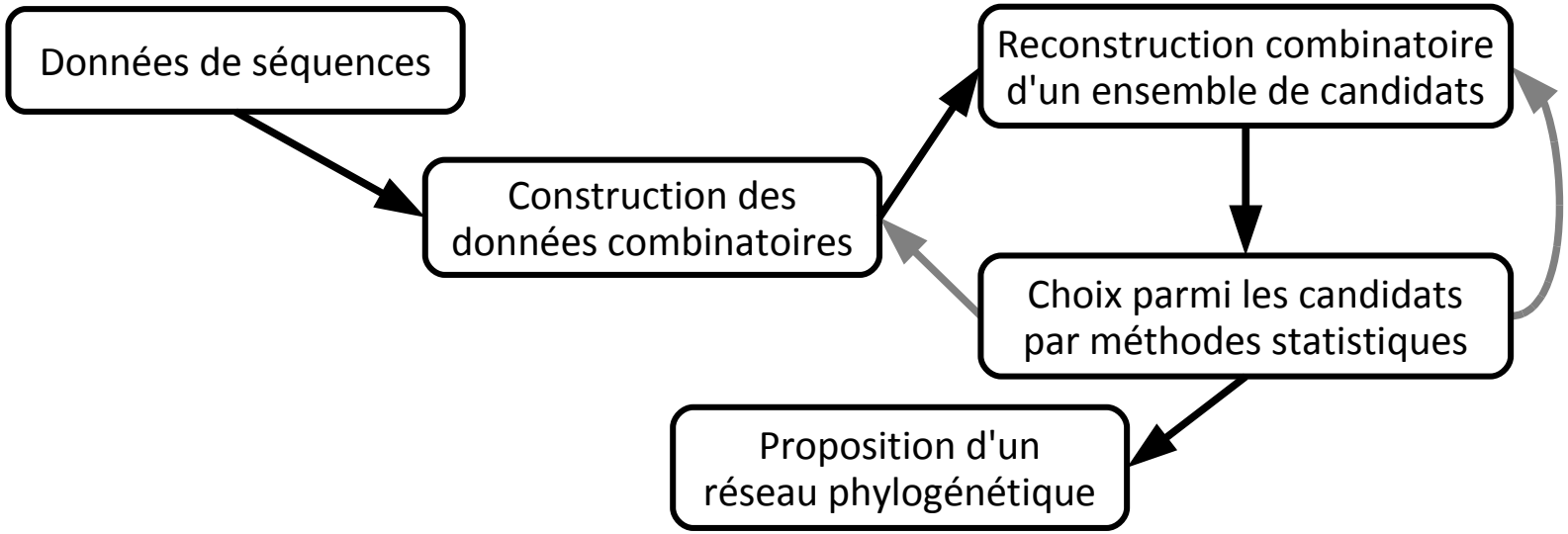
Perspectives de recherche

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

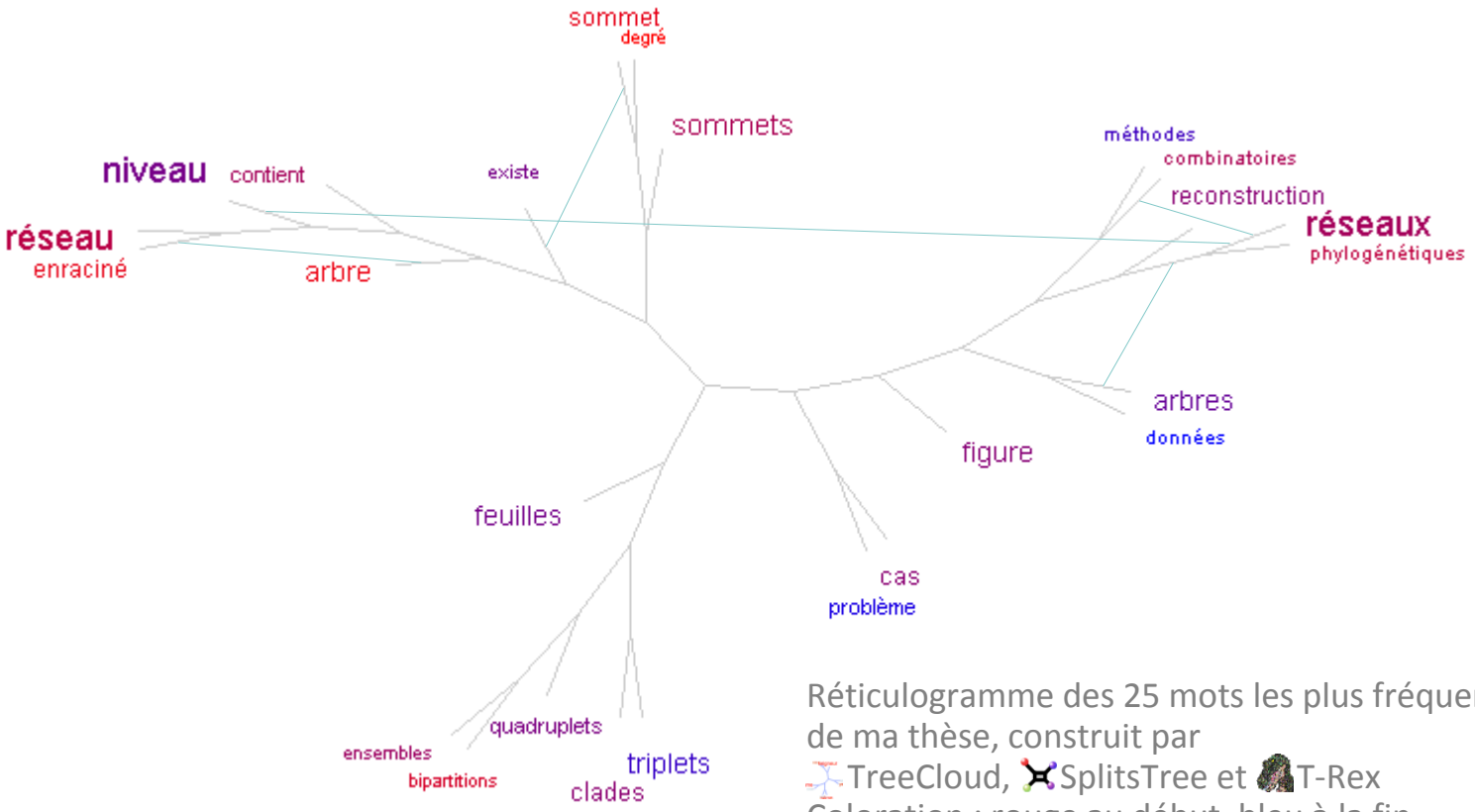
Bioinformatique :

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



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