

GT Combinatoire Énumérative et Algébrique du LABRI
08/04/2011

***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
- Perspectives

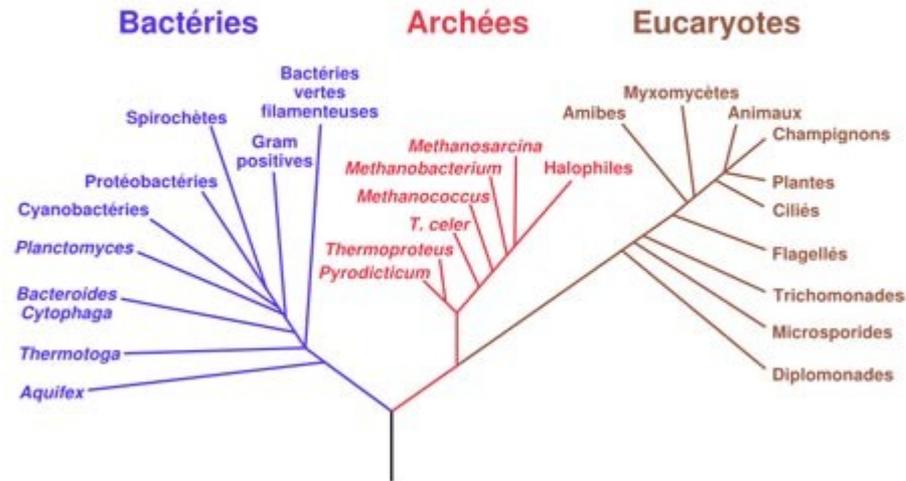
Plan

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

Les arbres phylogénétiques

Décrire l'évolution des espèces :

Arbre phylogénétique de la vie

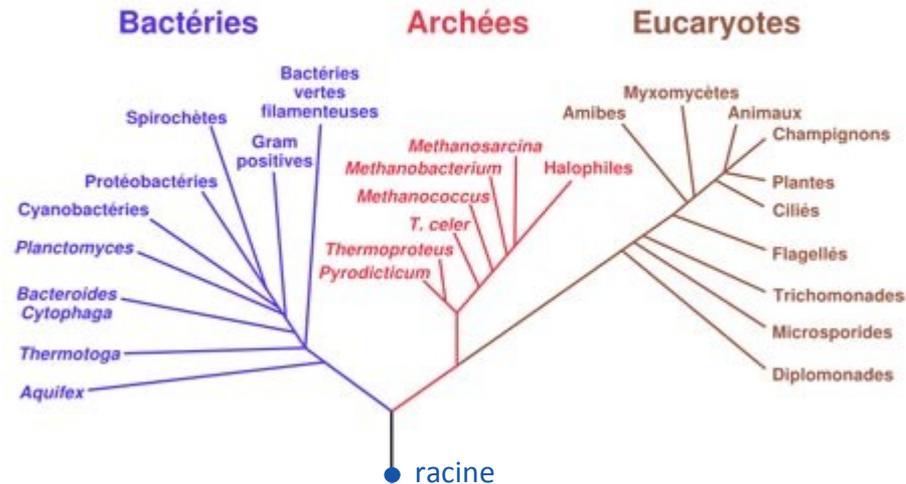


D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

Les arbres phylogénétiques

Décrire l'évolution des espèces :

Arbre phylogénétique de la vie



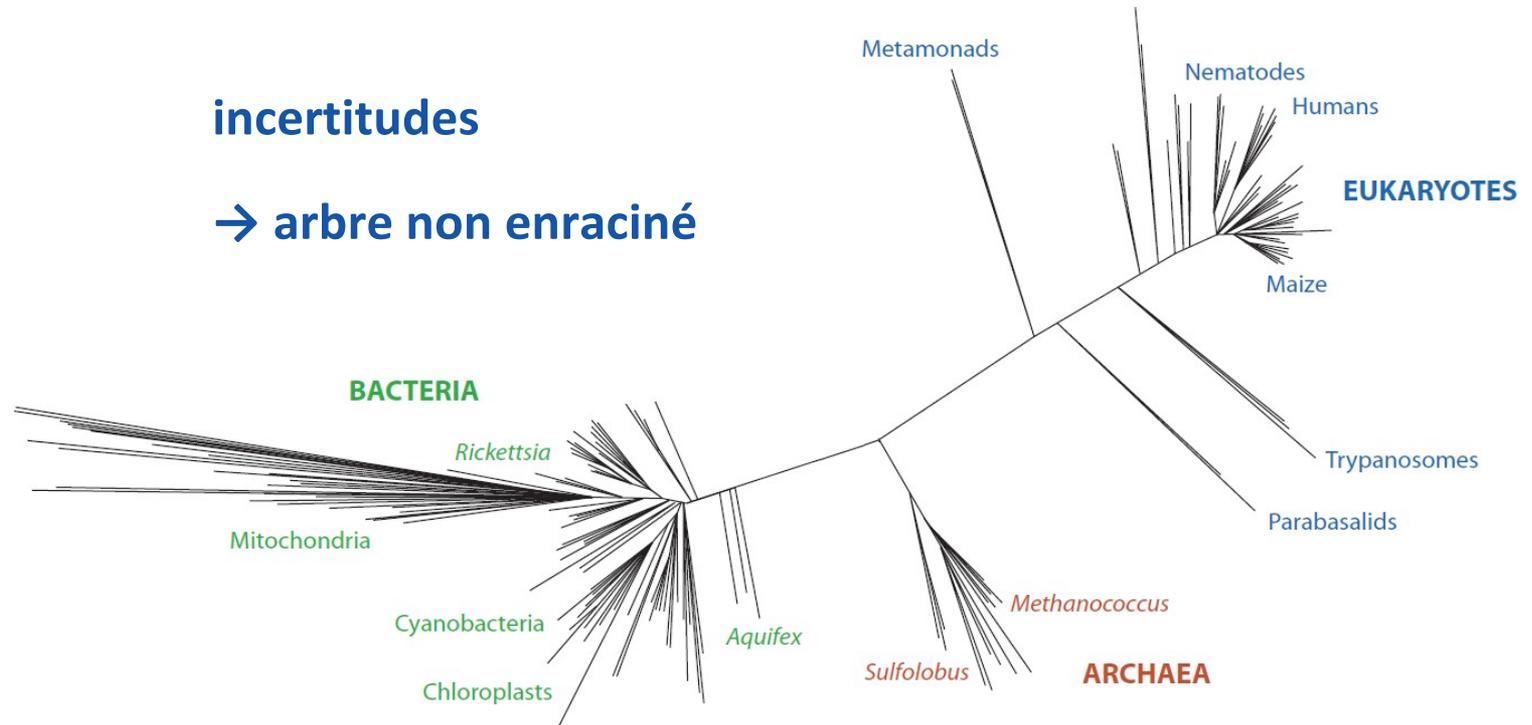
D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

Les arbres phylogénétiques

Décrire l'évolution des espèces :

incertitudes

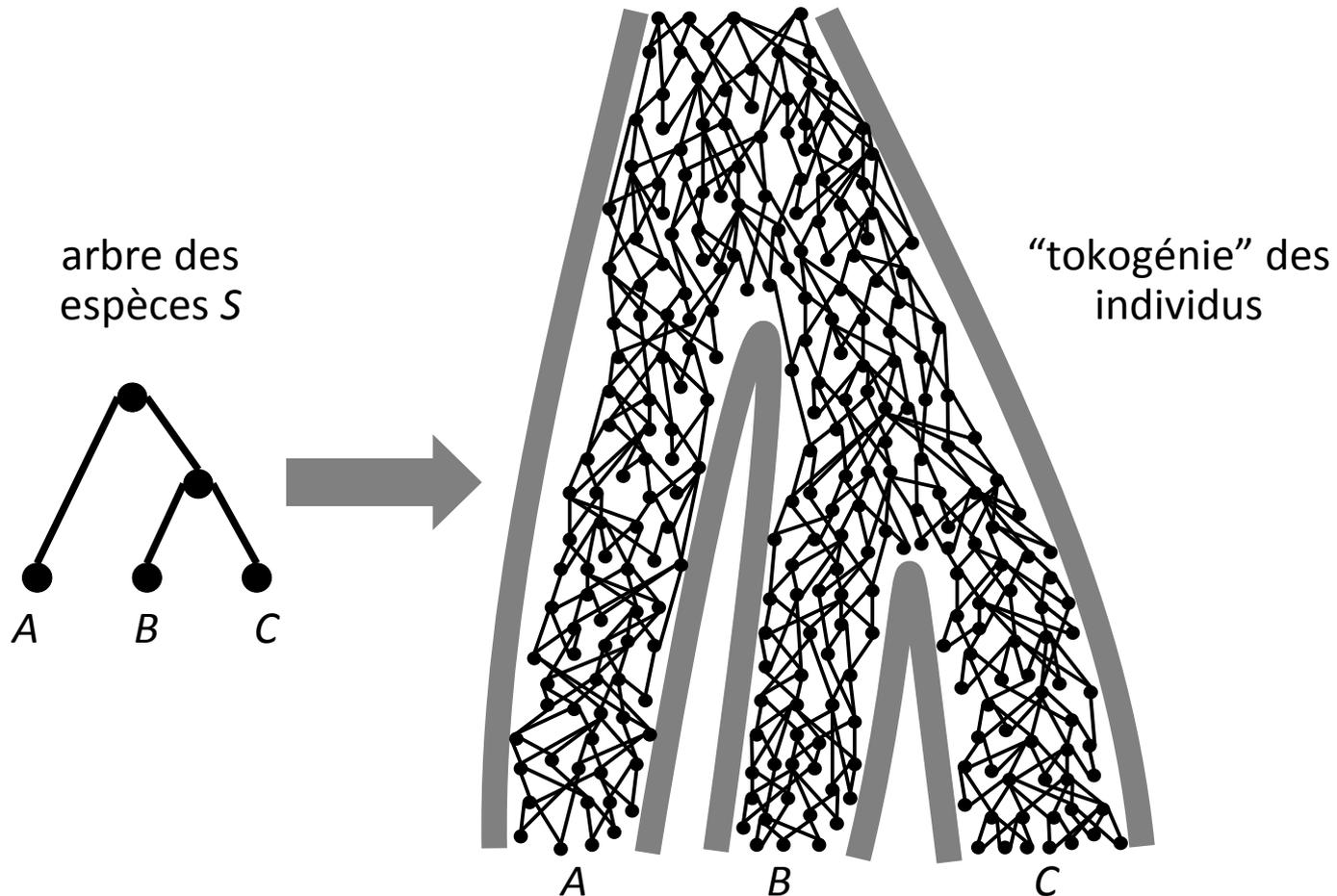
→ arbre non enraciné



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

Les arbres phylogénétiques

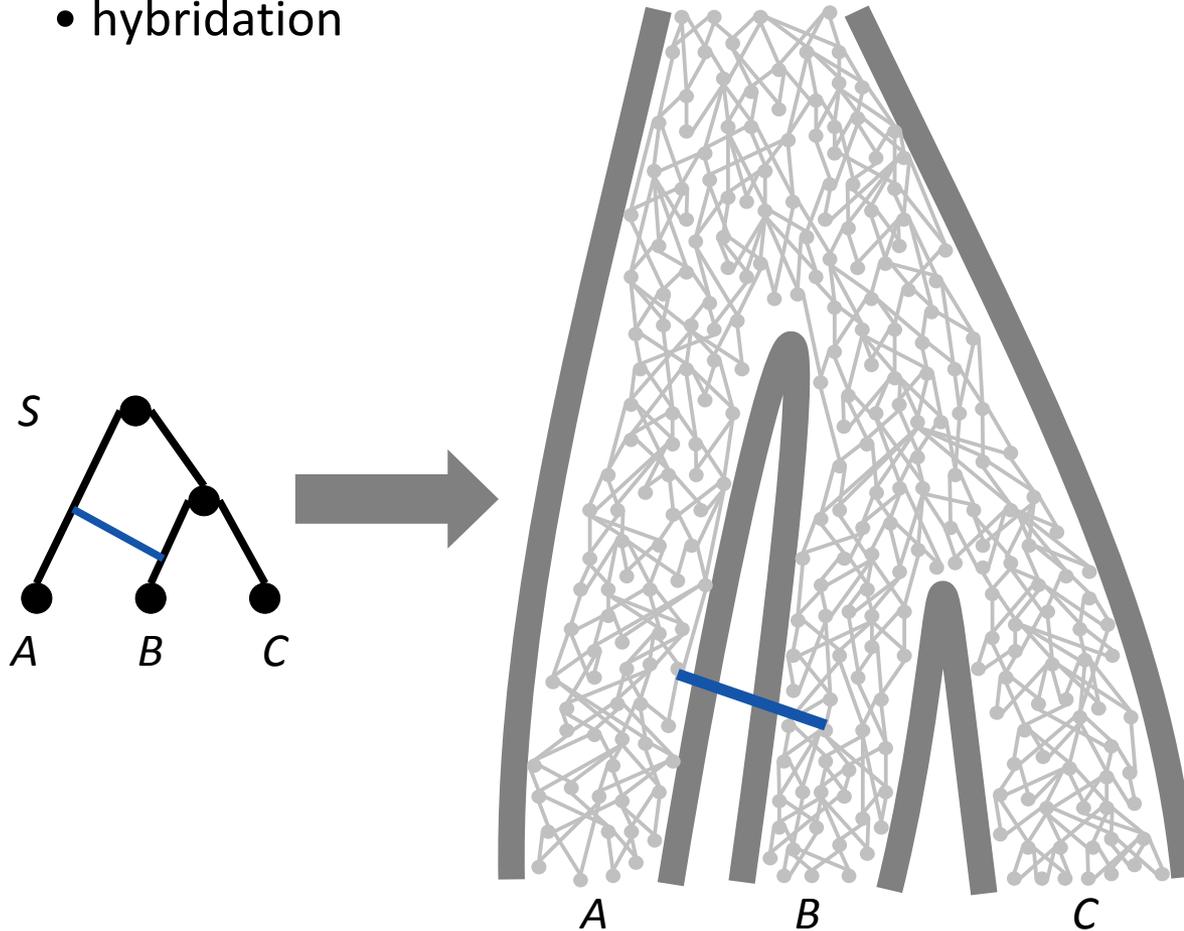
Décrire l'évolution des espèces :



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

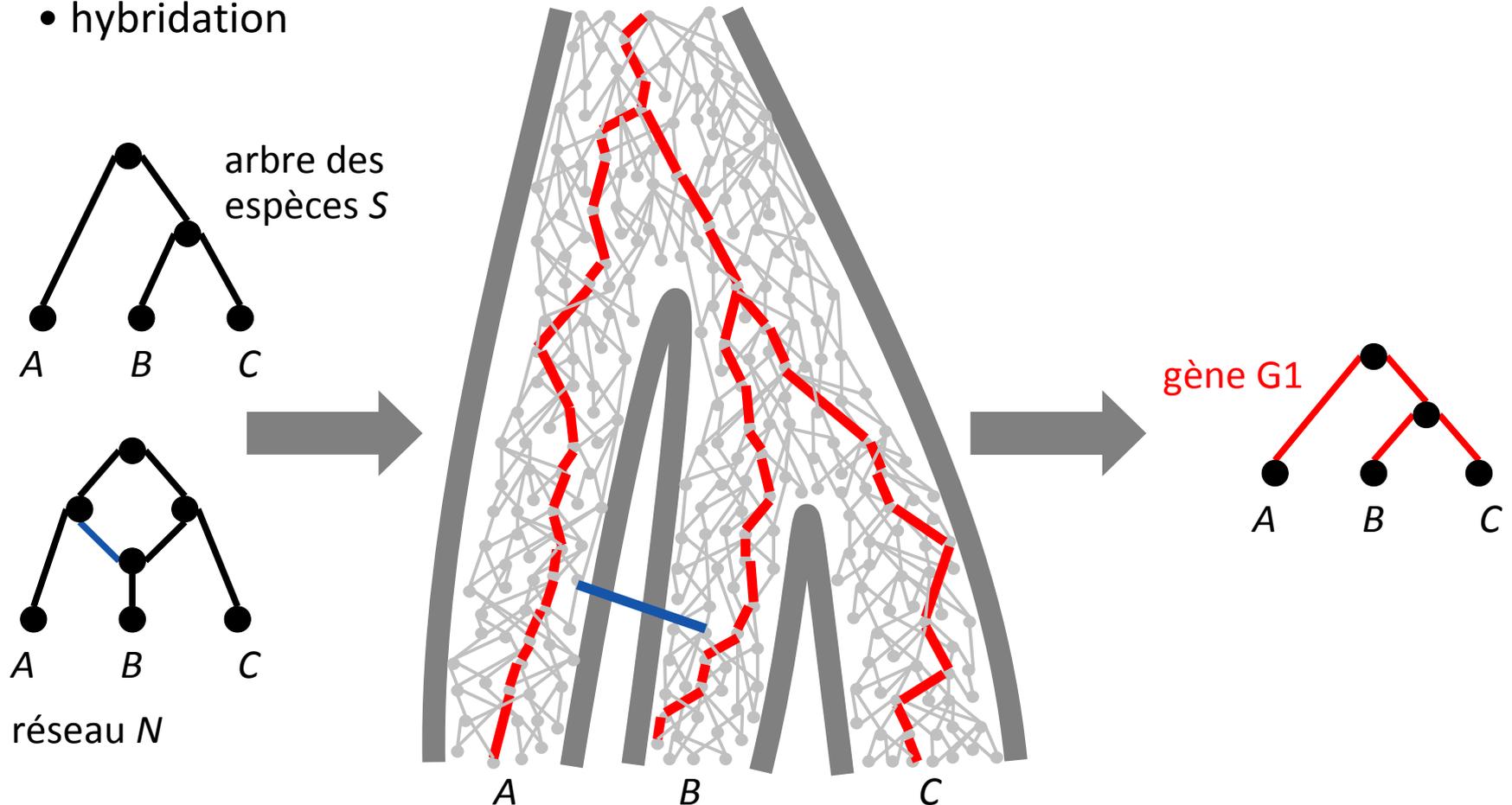
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

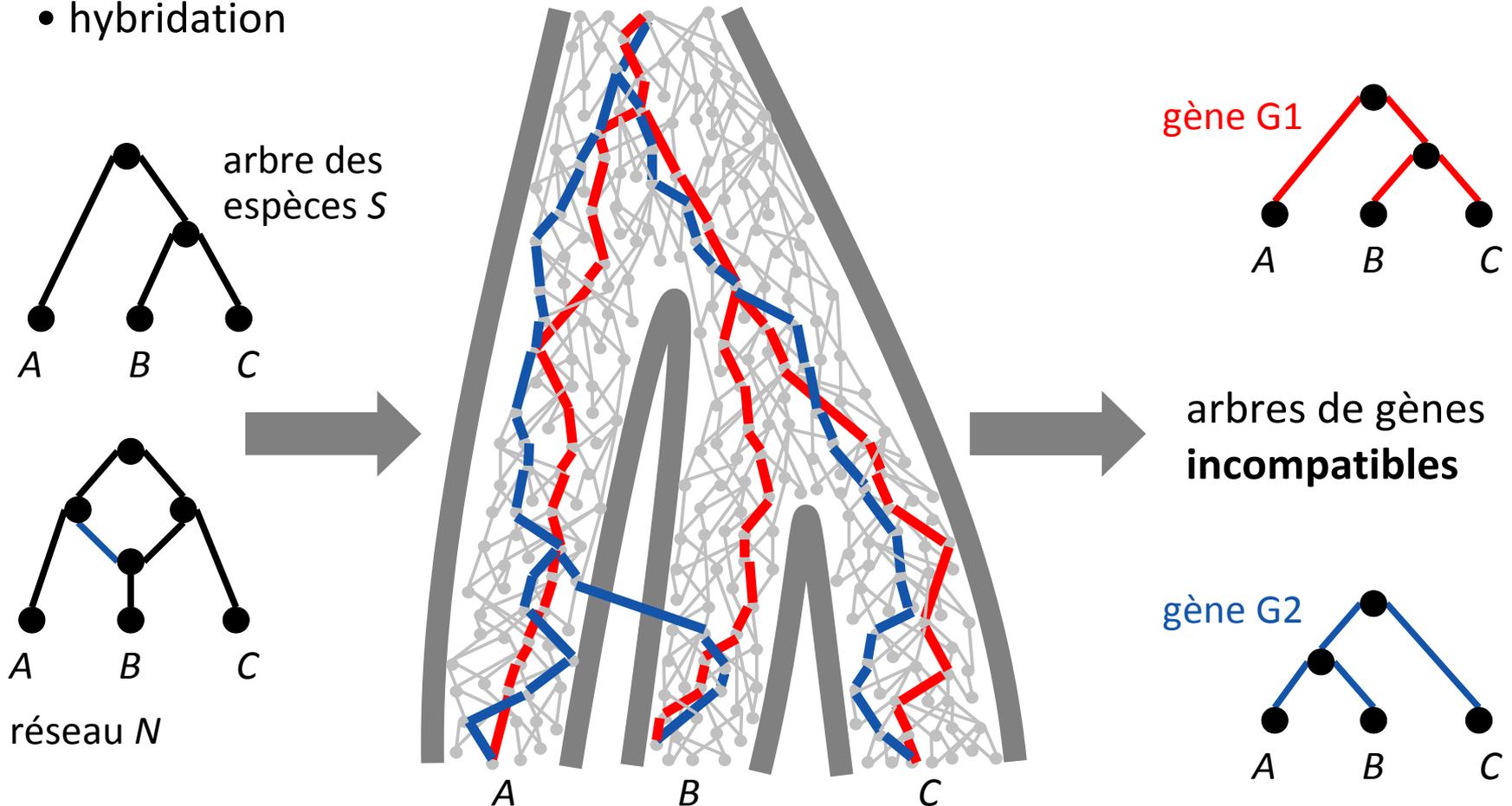
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

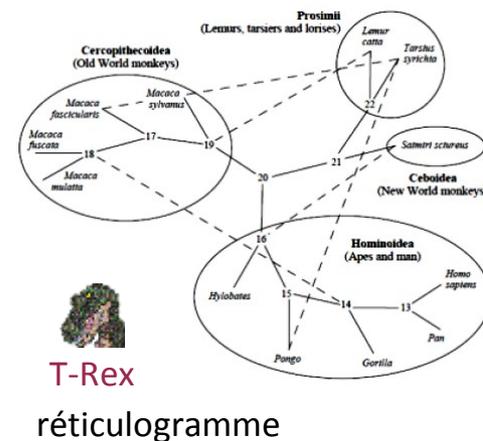
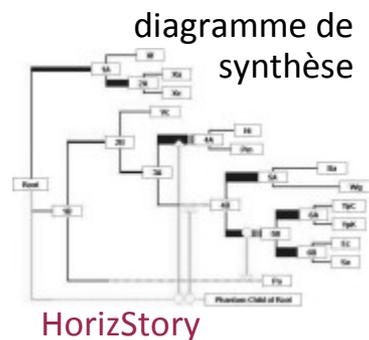
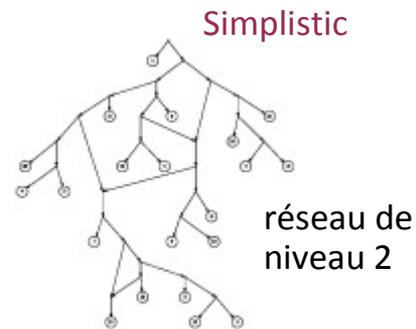
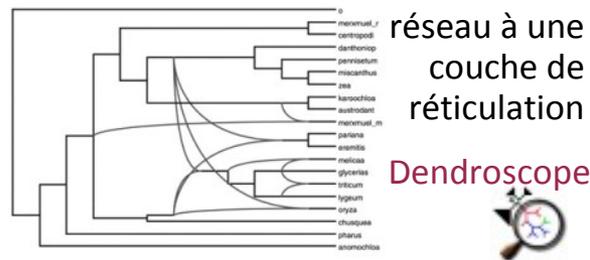
- transfert horizontal
- hybridation



Les réseaux phylogénétiques explicites

Réseaux phylogénétiques explicites :

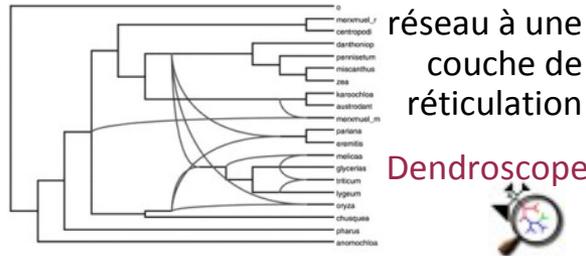
Tout sommet interprétable comme une espèce actuelle ou ancestrale



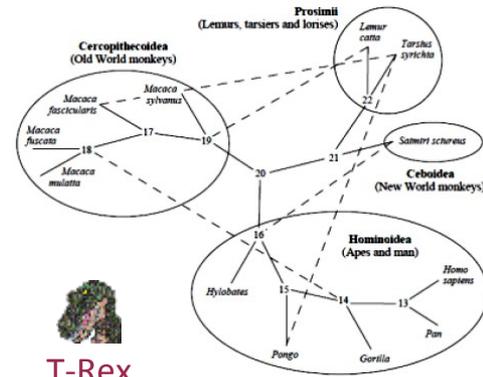
Les réseaux phylogénétiques explicites

Réseaux phylogénétiques explicites :

Tout sommet interprétable comme une espèce actuelle ou ancestrale



Dendroscope



réticulogramme

Simplistic

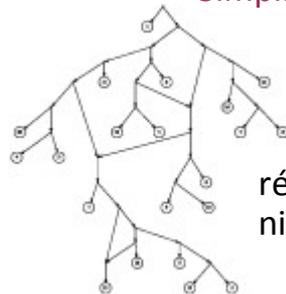
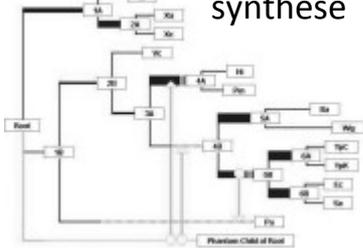


diagramme de synthèse



HorizStory

Réseau phylogénétique abstrait

réseau de bipartitions "split network"

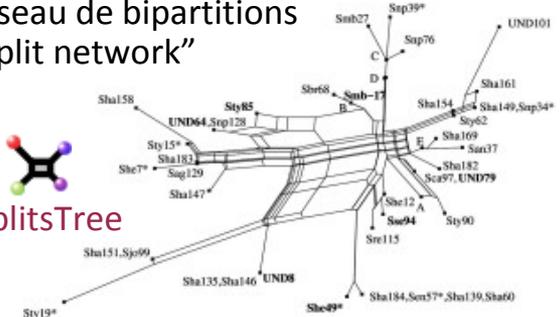


Plate-forme bibliographique

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

Who is Who in Phylogenetic Networks - Articles, Authors & Programs [RSS](#)

Index | Browse

Search: in (word length ≥ 3)

Publications - Index (All 371 publications) Selection by: Year | Category | Keyword | Author

Selection by Year

Number of publications per year on phylogenetic networks
Click on a year to display the publications

Year	Number of Publications
1975	1
1976	1
1977	1
1978	1
1979	1
1980	1
1981	1
1982	1
1983	1
1984	1
1985	1
1986	1
1987	1
1988	1
1989	1
1990	1
1991	1
1992	1
1993	1
1994	1
1995	1
1996	1
1997	1
1998	1
1999	1
2000	1
2001	1
2002	1
2003	1
2004	1
2005	1
2006	1
2007	1
2008	1
2009	17
2010	12
2011	39
2012	34
2013	44
2014	44
2015	41
2016	42
2017	1

See how the community working on phylogenetic networks evolves

Who is Who in Phylogenetic Networks - Articles, Authors & Programs [RSS](#)

Index | Browse

Search: in (word length ≥ 3)

Publications related to 'Program Dendroscope': *Dendroscope* is an interactive viewer for large phylogenetic trees and networks. Available at www.dendroscope.org.

Order by: Type | Year

Only And Or related to:

Associated keywords

abstract-network evaluation **explicit-network** FPT from-clusters from-rooted-trees galled-network level-k phylogenetic-network NP-complete **phylogenetic-network phylogeny** polynomial Program-Bio-PhyloNetwork **Program-Dendroscope** Program-HybridInterleave Program-HybridNumber Program-NetGen Program-PhyloNet Program-SplitsTree Program-TCS reconstruction software split-network survey visualization

2010

1 Steven Kelk's k...
Leo van Iersel, Steven Kelk, Regula Rupp and Daniel H. Huson. Phylogenetic Networks Do not Need to Be Complex: Using Fewer Reticulations to Represent Conflicting Clusters. In *ISMB10*, Vol. 26(12):i124-i131 of *BIO*, 2010. [Comment] [BIBTeX](#) [Google](#)

2 Robert G. Beiko. Gene sharing and genome evolution: networks in trees and trees in networks. In *Biology and Philosophy*, 2010. [Comment] [BIBTeX](#) [Google](#)

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

Who is Who in Phylogenetic Networks - Articles, Authors & Programs [RSS](#)

Index | Browse

Search: in (word length ≥ 3)

Publications related to 'Program Dendroscope': *Dendroscope* is an interactive viewer for large phylogenetic trees and networks. Available at www.dendroscope.org.

Order by: Type | Year

Only And Or related to:

Associated keywords

abstract-network evaluation **explicit-network** FPT from-clusters from-rooted-trees galled-network level-k phylogenetic-network NP-complete **phylogenetic-network phylogeny** polynomial Program-Bio-PhyloNetwork **Program-Dendroscope** Program-HybridInterleave Program-HybridNumber Program-NetGen Program-PhyloNet Program-SplitsTree Program-TCS reconstruction software split-network survey visualization

2010

1 Steven Kelk's k...
Leo van Iersel, Steven Kelk, Regula Rupp and Daniel H. Huson. Phylogenetic Networks Do not Need to Be Complex: Using Fewer Reticulations to Represent Conflicting Clusters. In *ISMB10*, Vol. 26(12):i124-i131 of *BIO*, 2010. [Comment] [BIBTeX](#) [Google](#)

Keywords: from clusters, level k phylogenetic network, Program Dendroscope, Program HybridInterleave, Program HybridNumber, reconstruction. **Note:** <http://dx.doi.org/10.1093/bioinformatics/btq202>.

2 Robert G. Beiko. Gene sharing and genome evolution: networks in trees and trees in networks. In *Biology and Philosophy*, 2010. [Comment] [BIBTeX](#) [Google](#)

Keywords: abstract network, explicit network, from rooted trees, galled network, phylogenetic network, phylogeny, Program Dendroscope, Program SplitsTree, reconstruction, split network, survey. **Note:** To appear, <http://dx.doi.org/10.1007/s10539-010-9217-3>.

Plan

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

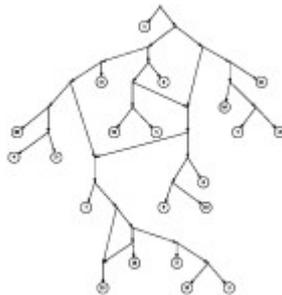
*Bandelt & Dress 1992 - Legendre & Makarenkov
2000 - Bryant & Moulton 2002*

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

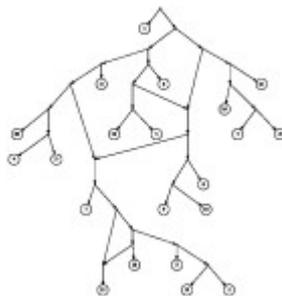
*Bandelt & Dress 1992 - Legendre & Makarenkov
2000 - Bryant & Moulton 2002*

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*

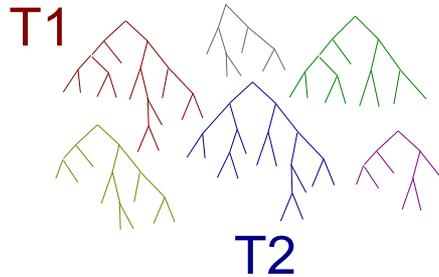


réseau *N*

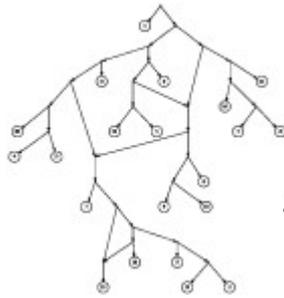
Reconstruction de réseaux phylogénétiques

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



réseau explicite



{séquences de gènes}

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

Guindon & Gascuel, SB, 2003

{arbres}

Base HOGENOM



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

Réconciliation ou consensus d'arbres

super-réseau optimal N

Reconstruction de réseaux phylogénétiques

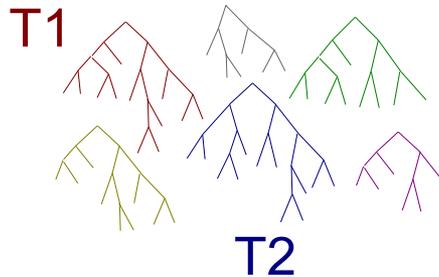
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 : TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT

G1 G2

{séquences de gènes}

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

Guindon & Gascuel, SB, 2003



{arbres}

Base HOGENOM

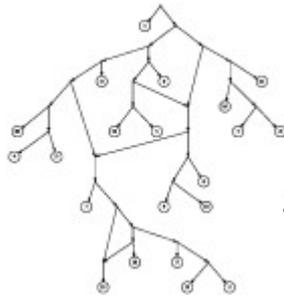


Dufayard, Duret, Penel, Gouy,
Rechenmann & Perrière, BioInf, 2005

> 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 :

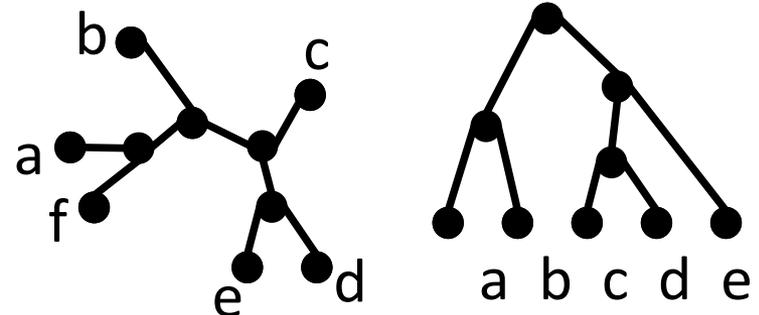
Reconstruire le **super-réseau** d'un ensemble d'arbres est
difficile.

Idée :

reconstituer un réseau contenant tous les :

triplets
quadruplets
clades
bipartitions

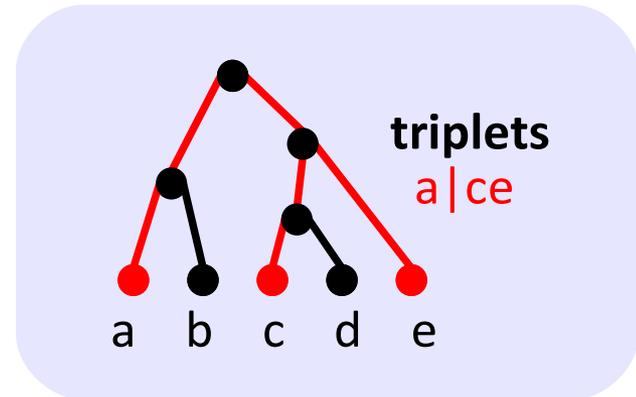
des arbres en entrée ?



Triplets et quadruplets, clades et bipartitions

Idée :

reconstituer un réseau contenant tous les :



des arbres en entrée ?

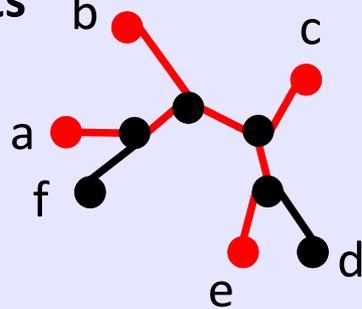
Triplets et quadruplets, clades et bipartitions

Idée :

reconstituer un réseau contenant tous les :

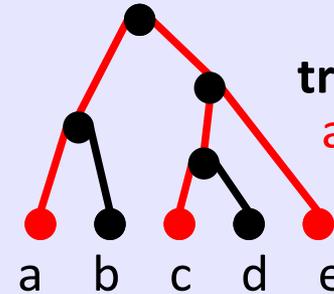
quadruplets

$ab|ce$



triplets

$a|ce$



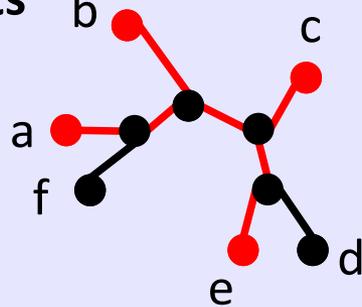
des arbres en entrée ?

Triplets et quadruplets, clades et bipartitions

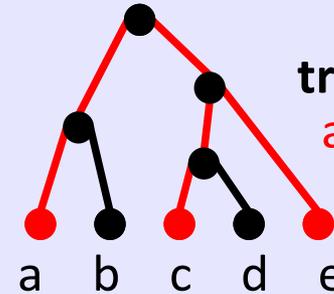
Idée :

reconstituer un réseau contenant tous les :

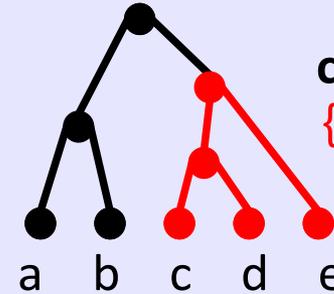
quadruplets
 $ab|ce$



triplets
 $a|ce$



clades
 $\{c,d,e\}$



des arbres en entrée ?

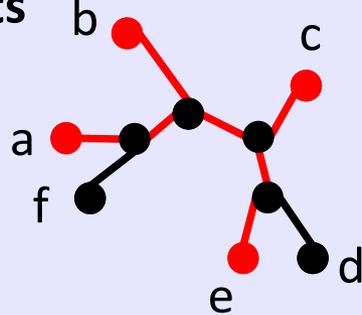
Triplets et quadruplets, clades et bipartitions

Idée :

reconstituer un réseau contenant tous les :

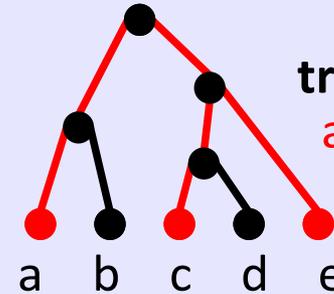
quadruplets

$ab|ce$



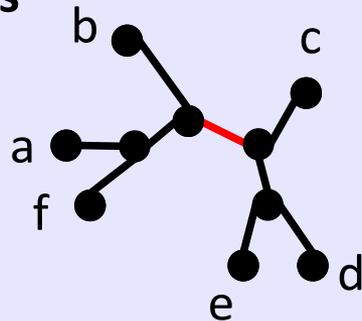
triplets

$a|ce$



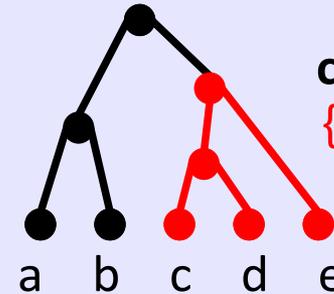
bipartitions

$\{a,b,f\}$
 $\{c,d,e\}$



clades

$\{c,d,e\}$

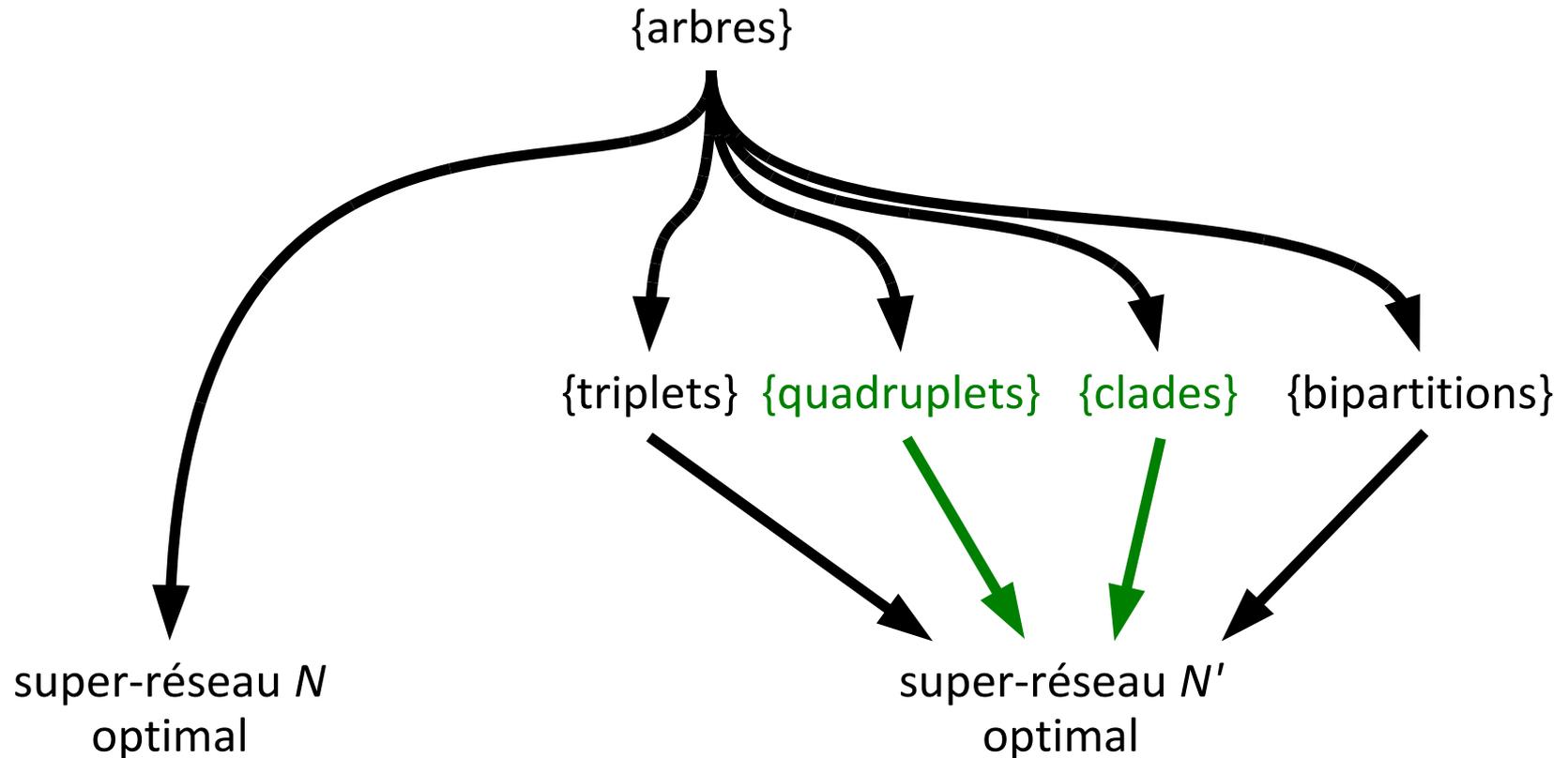


des arbres en entrée ?

Reconstruction combinatoire de réseaux phylogénétiques

Idée :

modifier le type de données à traiter

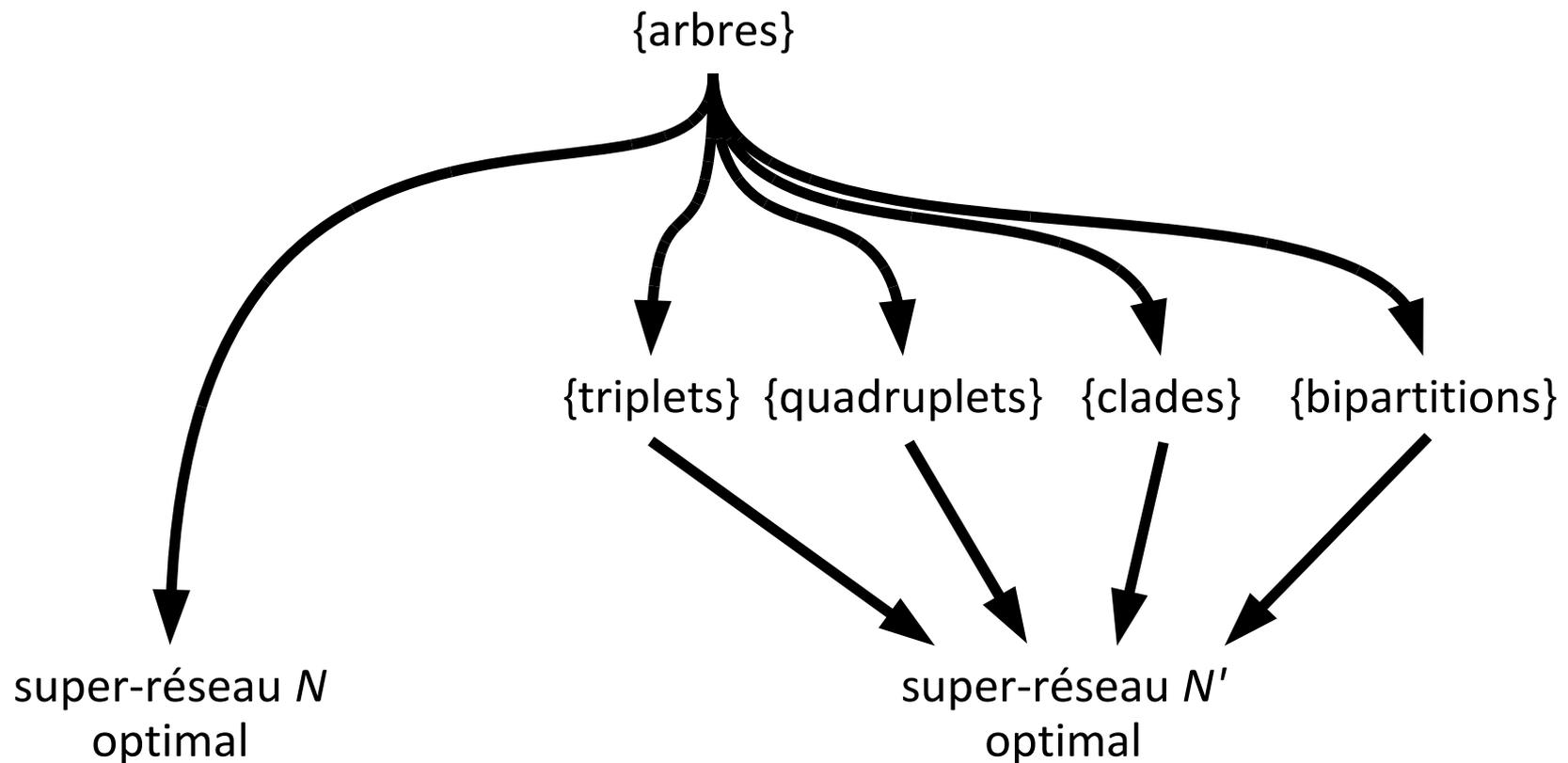


Huson, Rupp, Berry, Gambette & Paul, IMSB 2009
Gambette, Berry & Paul, 2011

Reconstruction combinatoire de réseaux phylogénétiques

Idée :

modifier le type de données à traiter

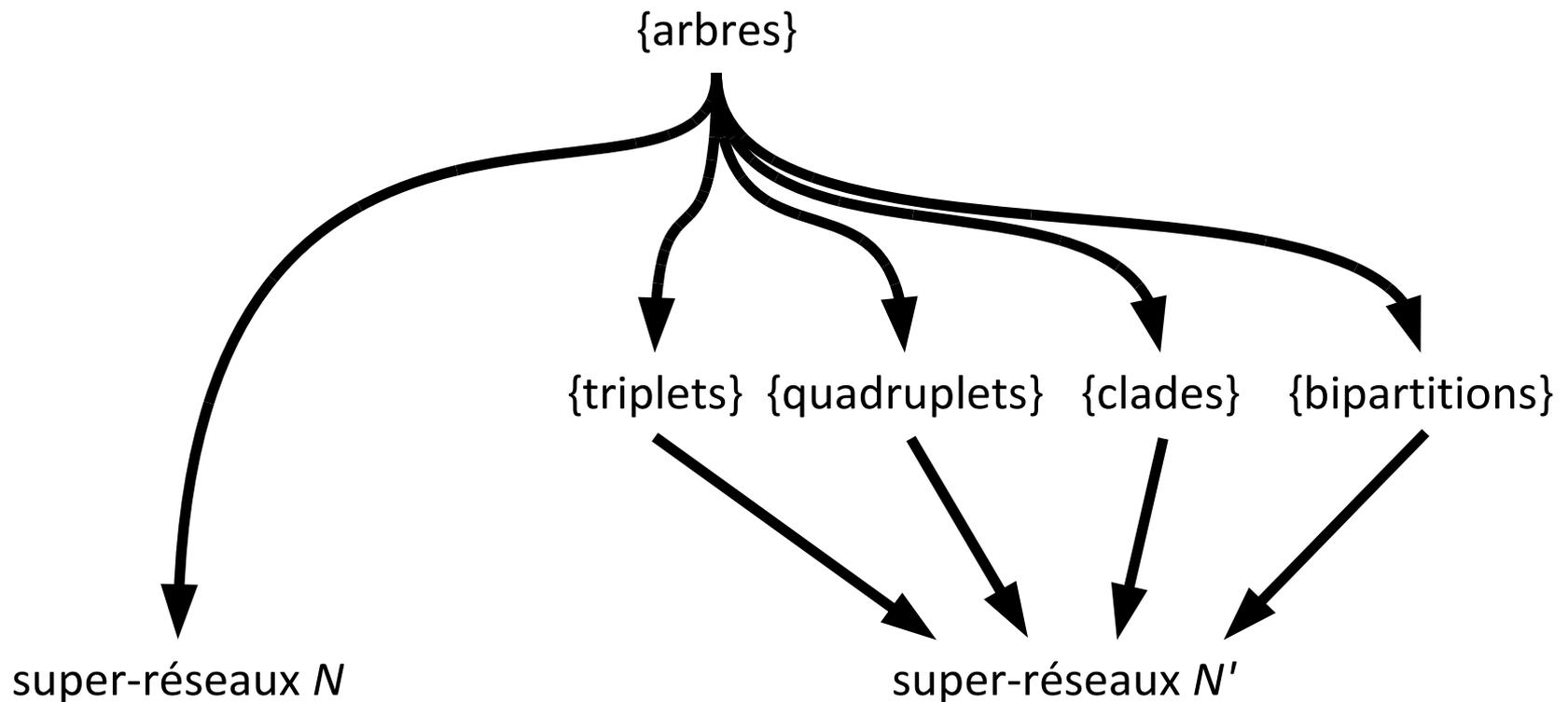


$N=N'$?

Reconstruction combinatoire de réseaux phylogénétiques

Idée :

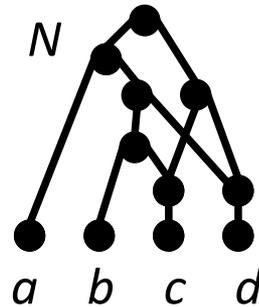
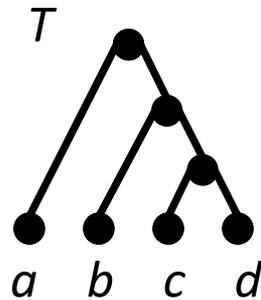
modifier le type de données à traiter



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

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** .

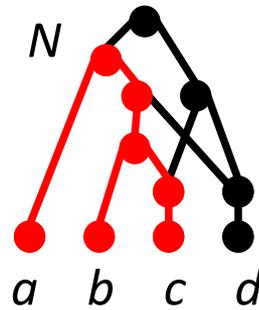
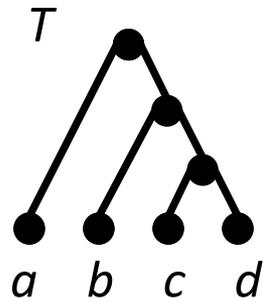


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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** .

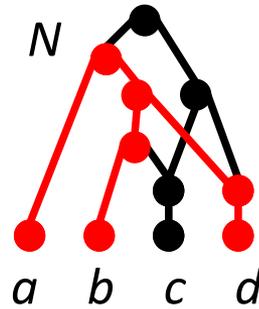
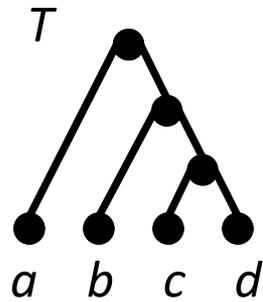


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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** .

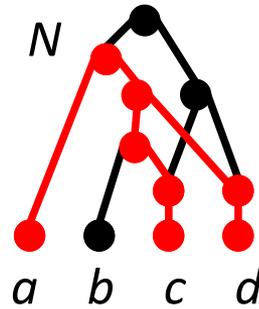
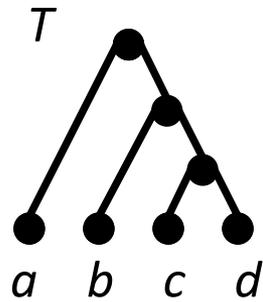


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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 .

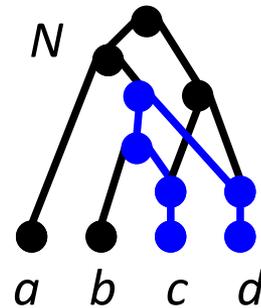
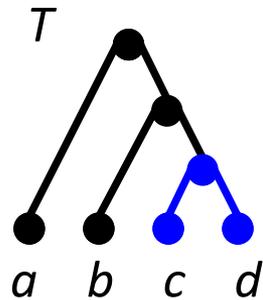


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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 .

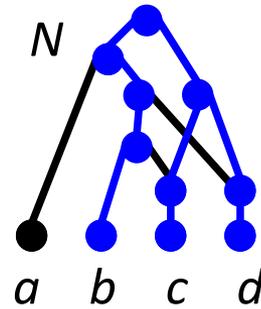
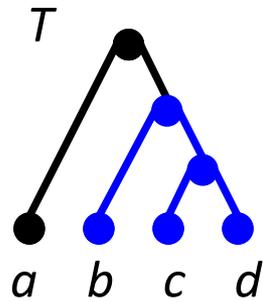


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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 .

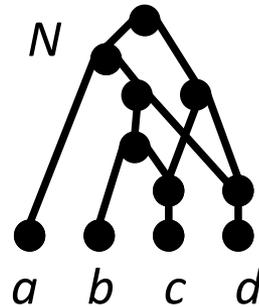
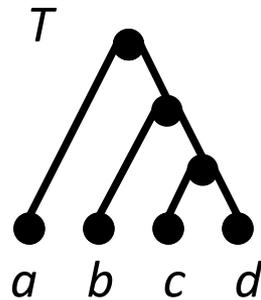


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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

Reconstruction combinatoire de réseaux phylogénétiques

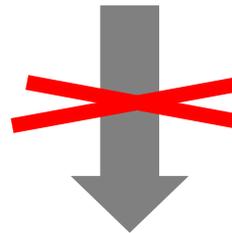
Un réseau qui contient l'ensemble de **tous les triplets ou clades d'un arbre T** ne contient **pas forcément T** .



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

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

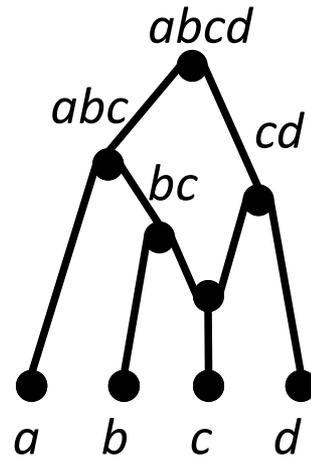
contient les clades / triplets d'un arbre T



contient T .

Clades stricts et souples

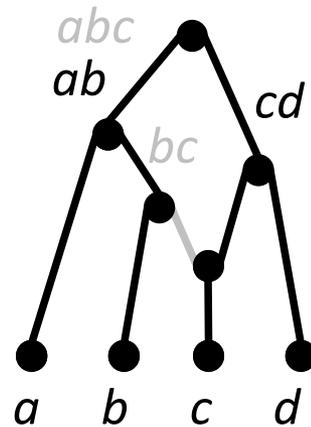
Clade "strict" : ensemble des feuilles sous un noeud du réseau



Clades stricts et souples

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

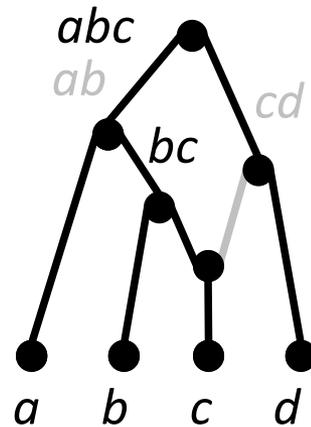
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



Clades stricts et souples

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

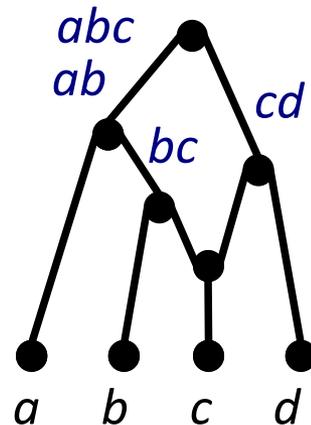
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



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 seulement compatibles** avec N peut être de taille **exponentielle**.

Tester si un **clade souple** appartient à un réseau : **NP-complet**.

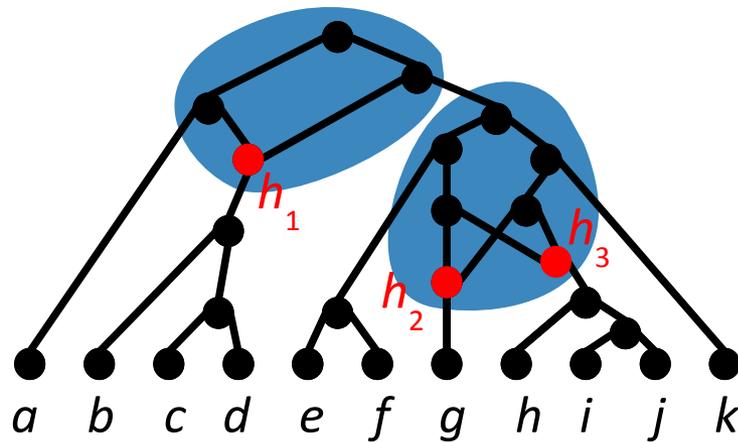
Kanj, Nakhleh, Than & Xia, TCS, 2008

Plan

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

Réseaux phylogénétiques restreints

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

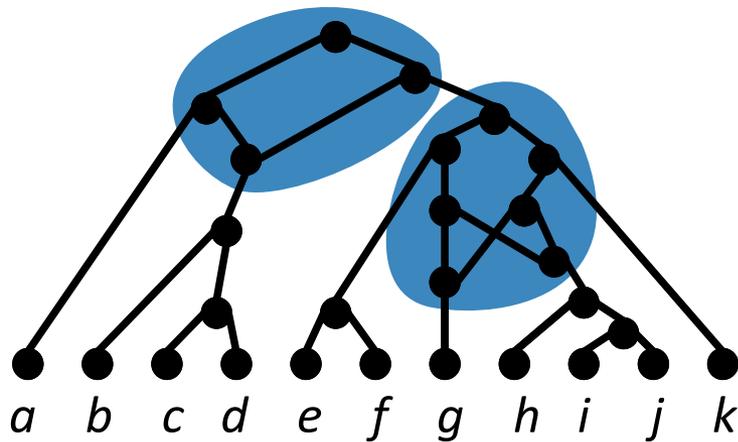


réseau de niveau 2

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

Réseaux phylogénétiques restreints

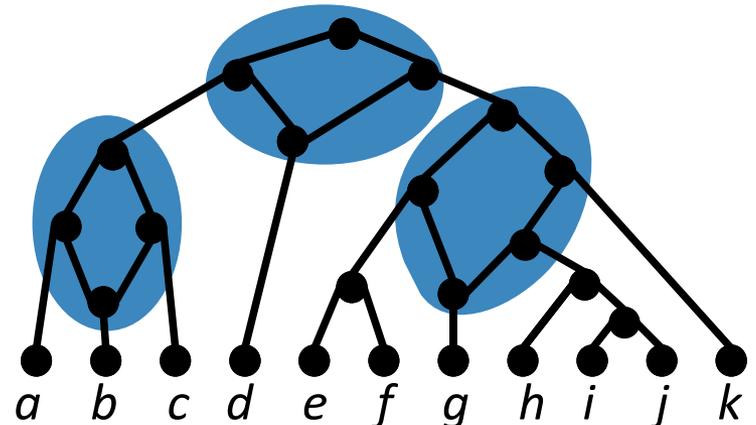
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau de niveau 2

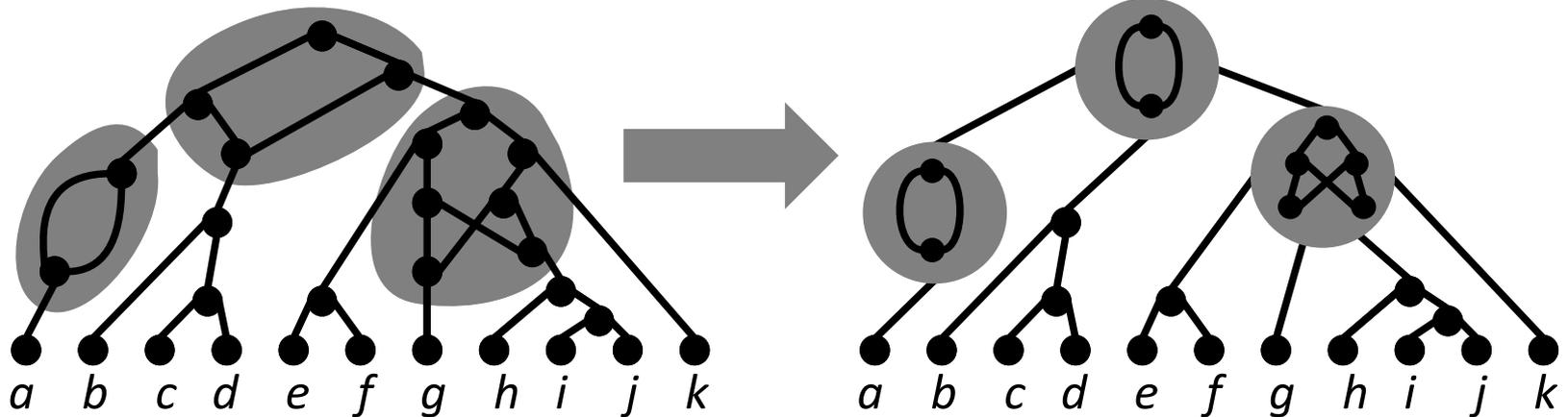
niveau =
nombre maximum d'hybridations
par partie non arborée (*blob*).

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** de niveau au plus k .

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

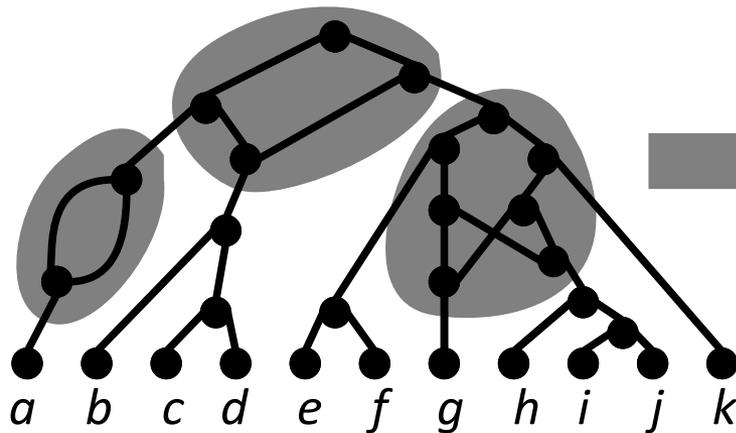
Force brute pour les 65 générateurs de niveau 3

Van Iersel et al., RECOMB 2008

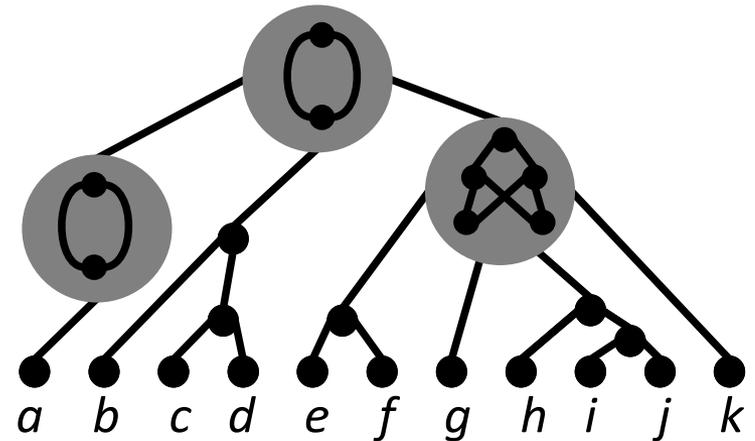
<http://homepages.cwi.nl/~kelk/lev3gen>

Décomposition des réseaux de niveau k

Décomposition en blobs :



N , réseau de niveau k .



décomposition arborée de N en
générateurs de niveau au plus k .

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**

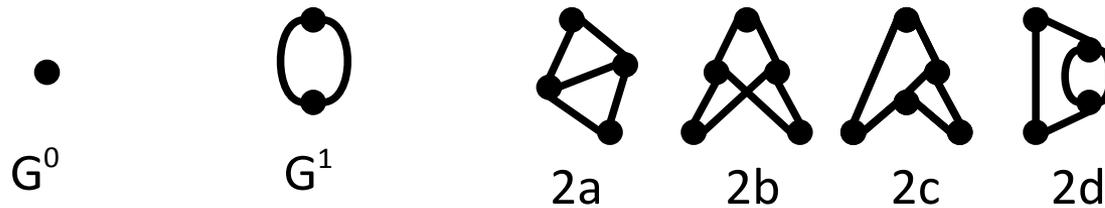
Van Iersel & Kelk, arXiv, 2009

Kelk, Scornavacca & Van Iersel, soumis, 2011

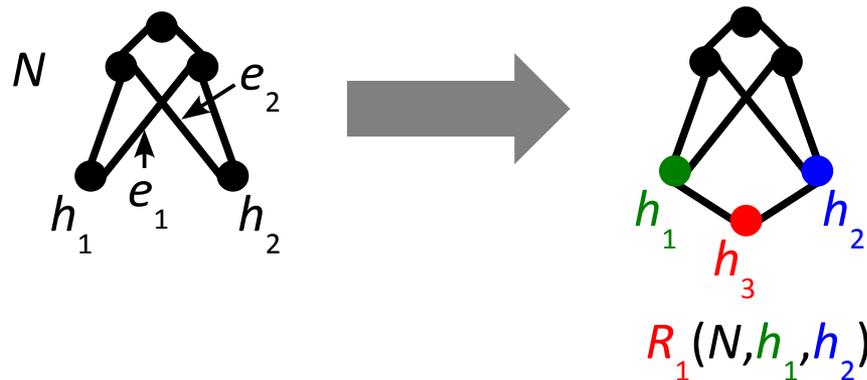
Décomposition des réseaux de niveau k

Générateur de niveau k :

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



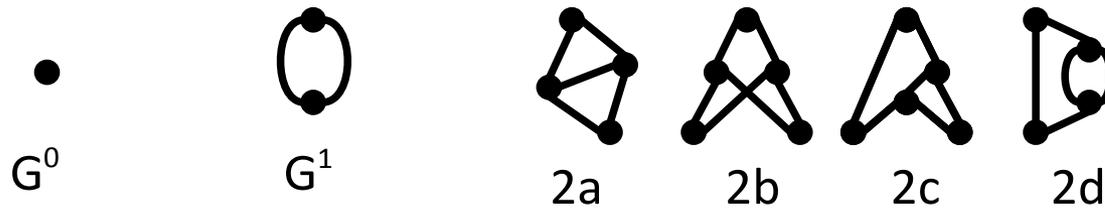
Règles de construction des générateurs de niveau $k+1$ à partir de ceux de niveau k



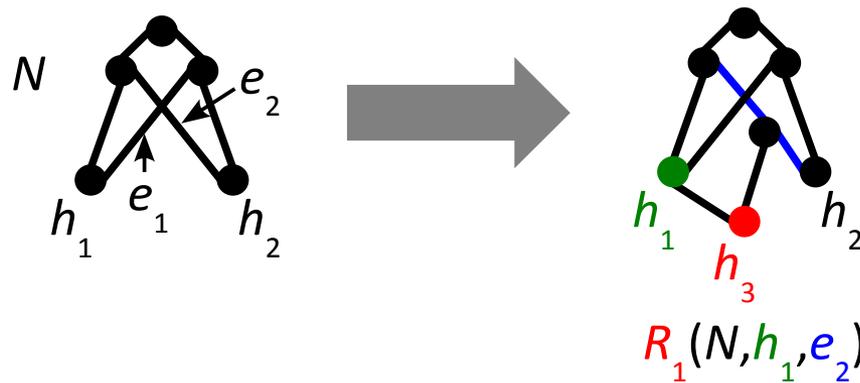
Décomposition des réseaux de niveau k

Générateur de niveau k :

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



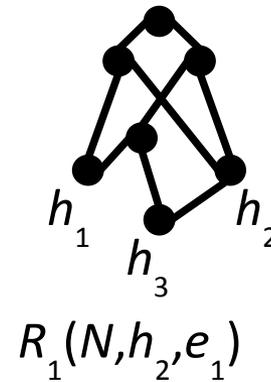
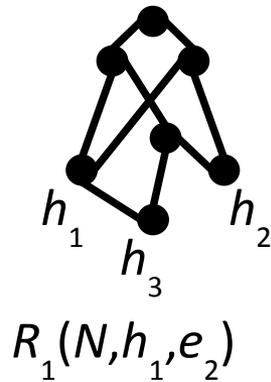
Règles de construction des générateurs de niveau $k+1$ à partir de ceux de niveau k



Construction des générateurs

Problème !

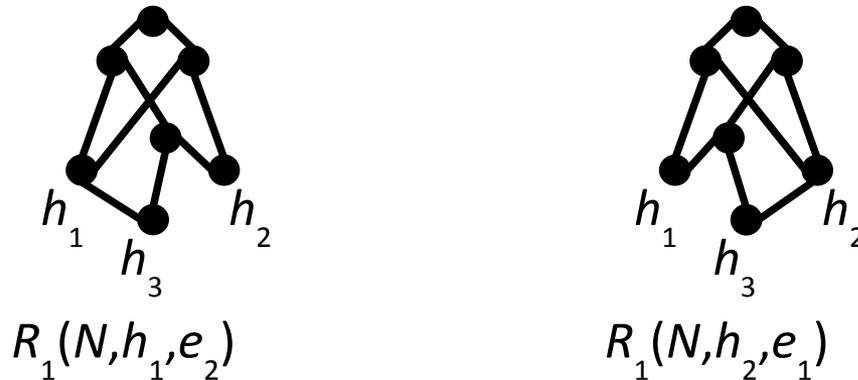
Certains des générateurs de niveau $k+1$ obtenus depuis ceux de niveau k sont isomorphes !



Construction des générateurs

Problème !

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



- comptage difficile !
- génération possible jusqu'à niveau 5 :
1, 4, 65, 1993, 91454
- borne inférieure

Borne inférieure

Borne inférieure :

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

Il y a un **nombre exponentiel** de générateurs !

Idée :

Coder tout nombre entre 0 et $2^{k-1}-1$ par un générateur de niveau k .

Borne inférieure

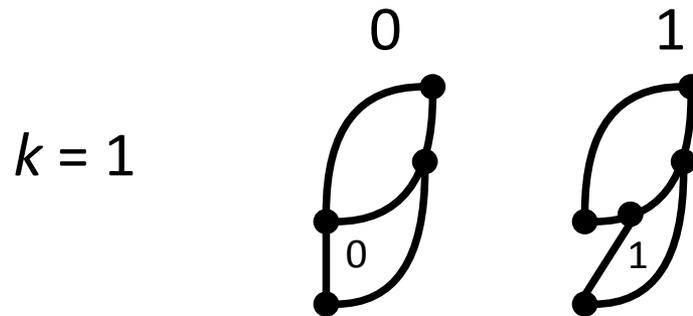
Borne inférieure :

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

Il y a un **nombre exponentiel** de générateurs !

Idée :

Coder tout nombre entre 0 et $2^{k-1}-1$ par un générateur de niveau k .



Borne inférieure

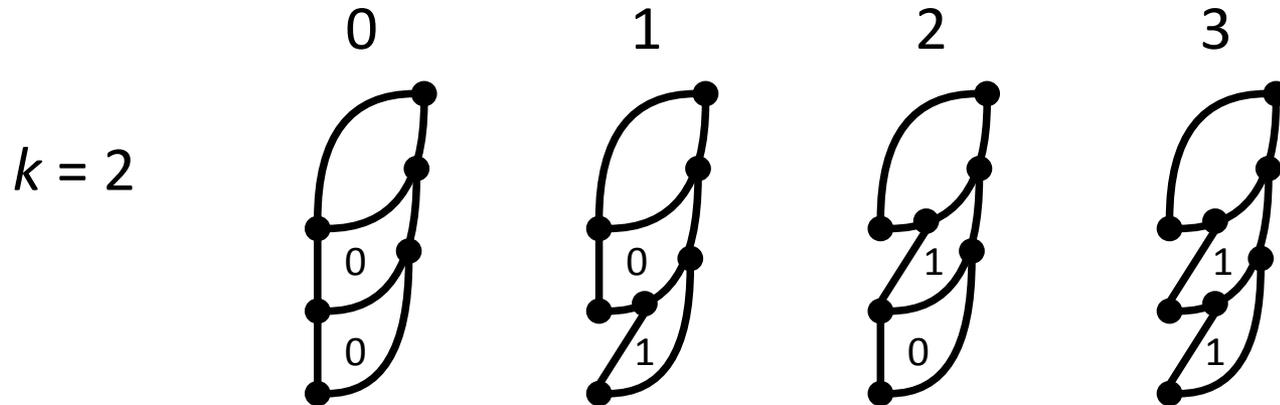
Borne inférieure :

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

Il y a un **nombre exponentiel** de générateurs !

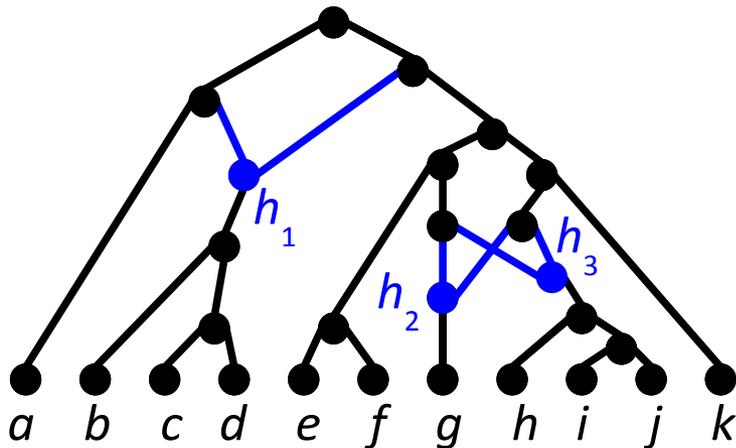
Idée :

Coder tout nombre entre 0 et $2^{k-1}-1$ par un générateur de niveau k .



Réseaux phylogénétiques restreints

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

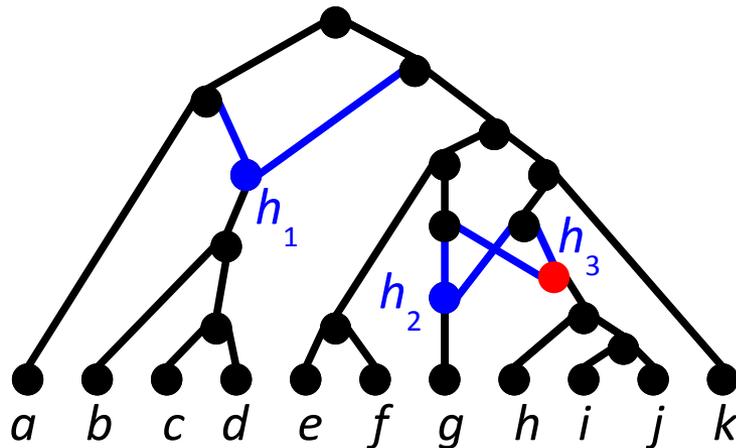


réseau à **une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à une couche de réticulation.

Réseaux phylogénétiques restreints

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

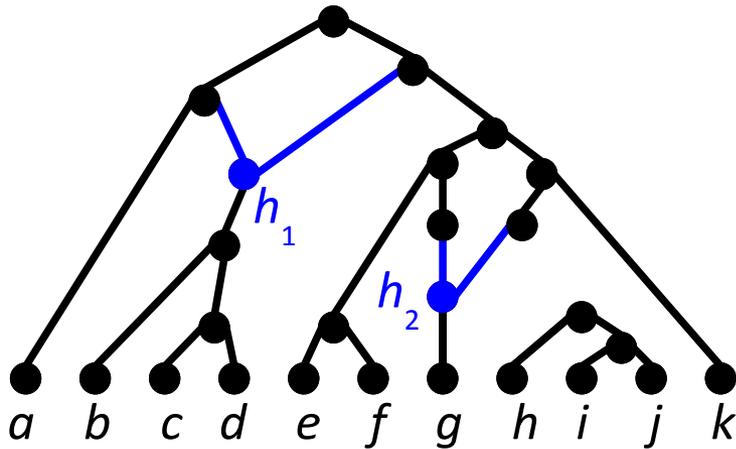


réseau à **une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à une couche de réticulation.

Réseaux phylogénétiques restreints

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

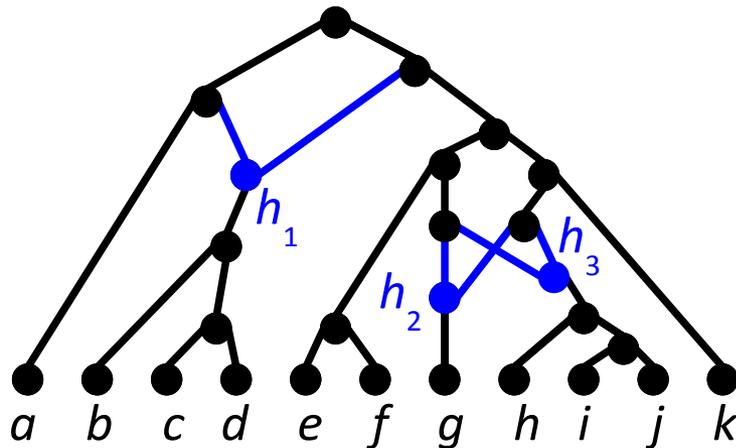


réseau à **une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à une couche de réticulation.

Réseaux phylogénétiques restreints

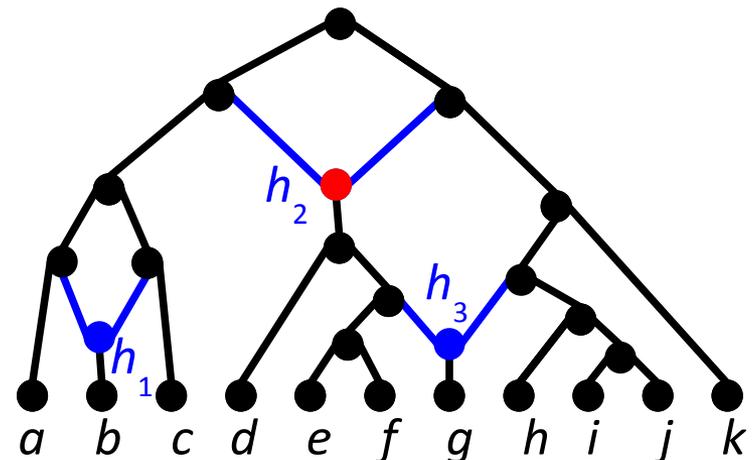
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau à une couche de réticulation.

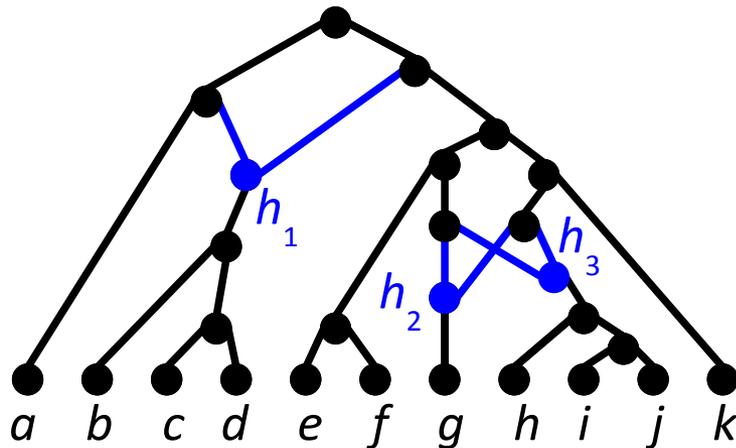
réseau à **une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à deux couches de réticulation.



Réseaux phylogénétiques restreints

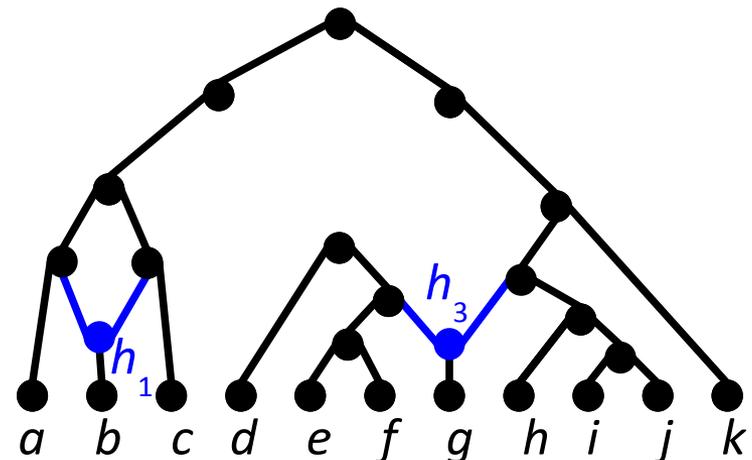
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



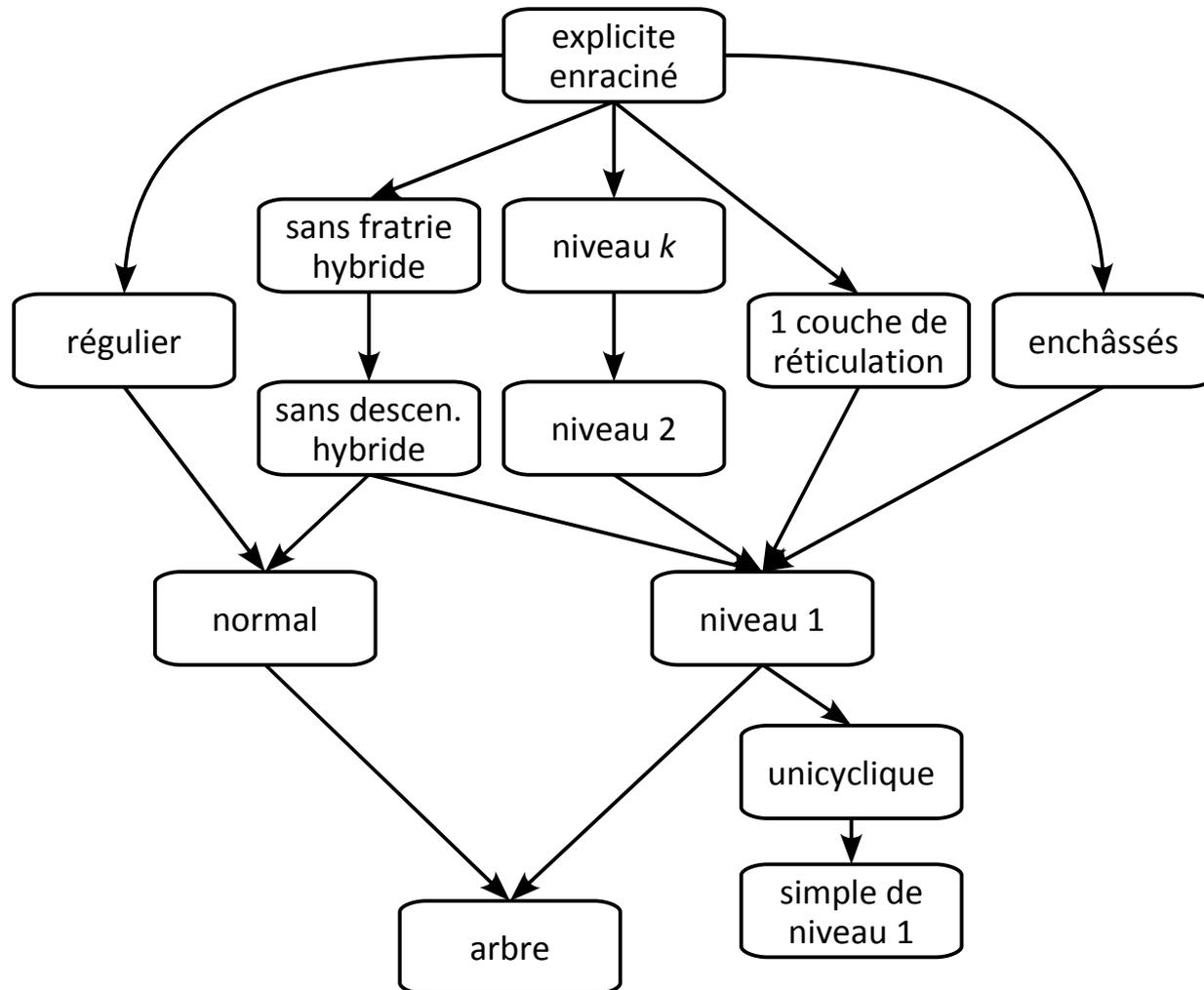
réseau à une couche de réticulation.

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

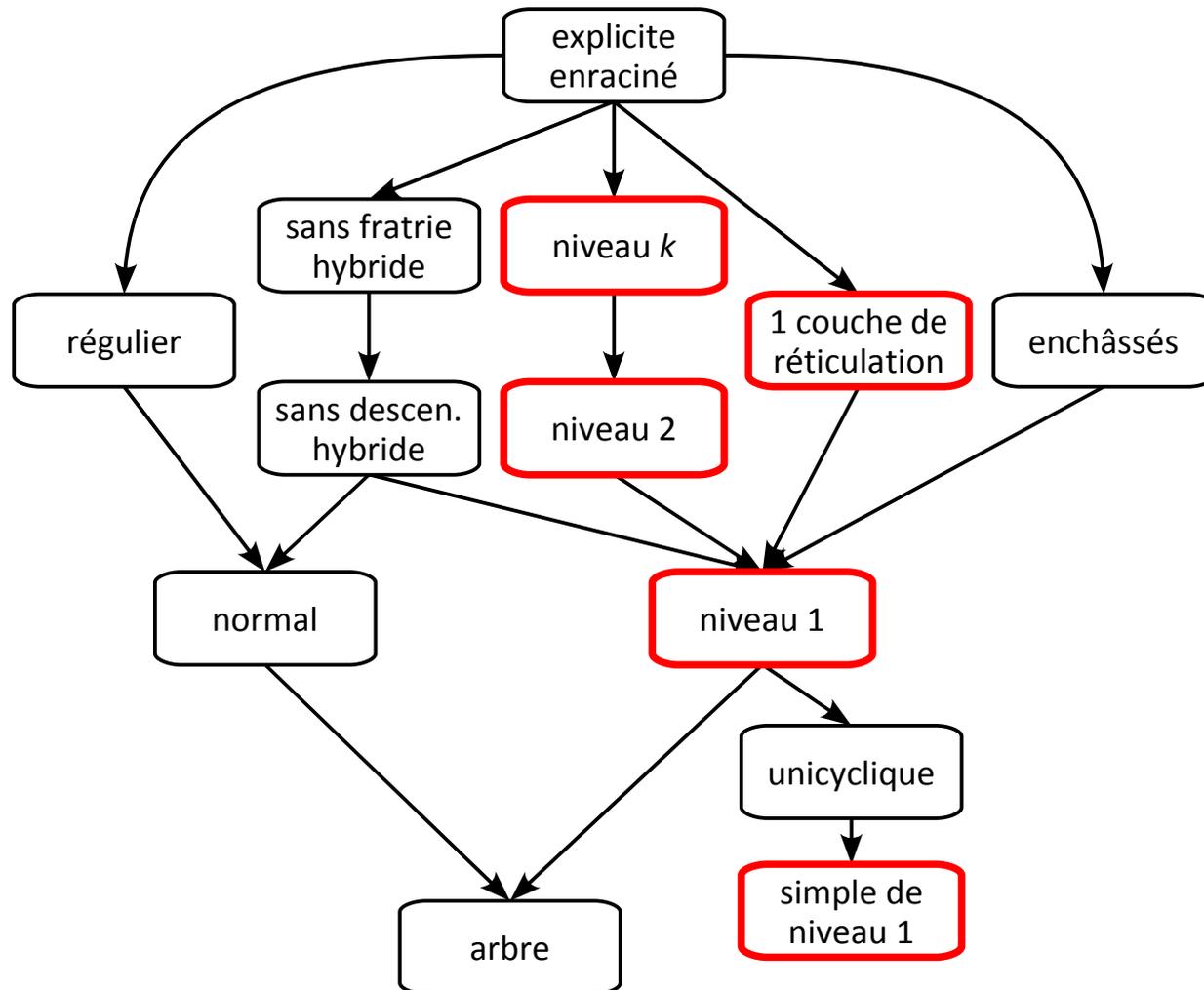
réseau à deux couches de réticulation.



Hiérarchie de sous-classes de réseaux



Hiérarchie de sous-classes de réseaux



Plan

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

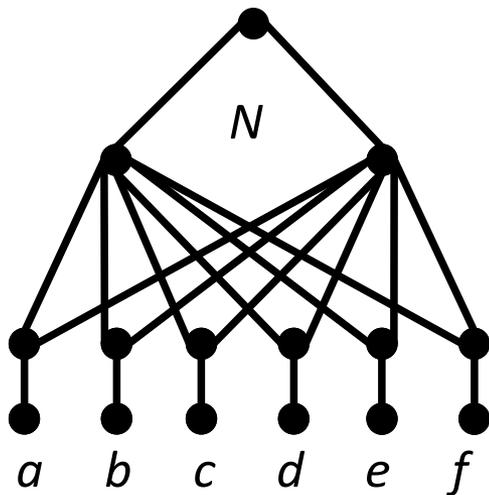


meilleurs résultats mais plus lente pour niveau > 2 .
pour k fixé, certains ensembles de clades contenus
dans aucun réseau de niveau k .

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 .

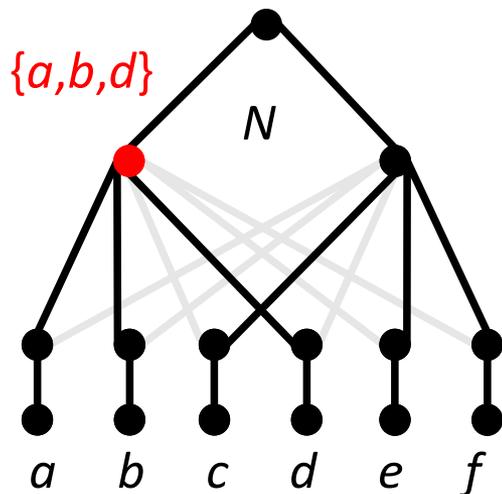


Le réseau à une couche de réticulation N est compatible avec tout clade souple sur $\{a, b, c, d, e, f\}$.

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 .

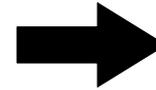


Le réseau à une couche de réticulation N est compatible avec tout clade souple sur $\{a,b,c,d,e,f\}$.

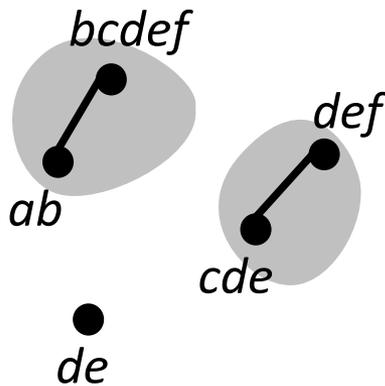
Clades et réseaux à une couche de réticulation

Décomposition

décomposition de l'ensemble de clades

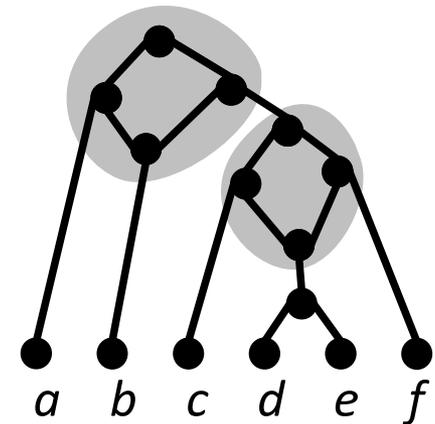


décomposition du réseau



graphe des conflits

Gusfield, Eddhu, Langley,
JBCB, 2004

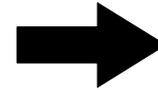


réseau reconstruit

Clades et réseaux à une couche de réticulation

Décomposition

décomposition de l'ensemble de clades

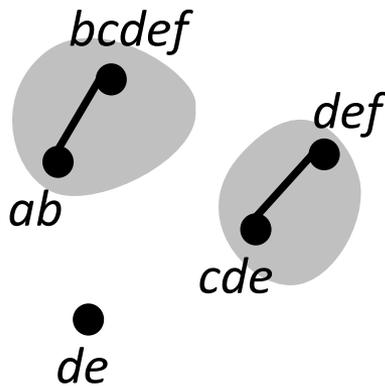


décomposition du réseau

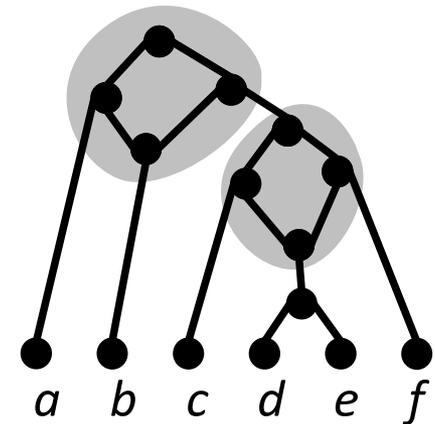
composante connexe
du graphe des conflits



blob



graphe des conflits

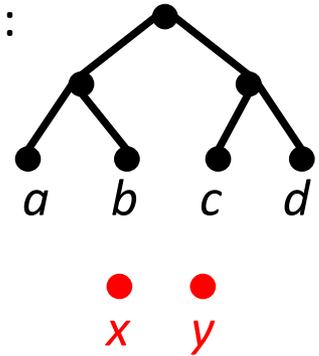


réseau reconstruit

Une approche en deux étapes

1- Trouver un **ensemble minimum de conflits** parmi les clades :

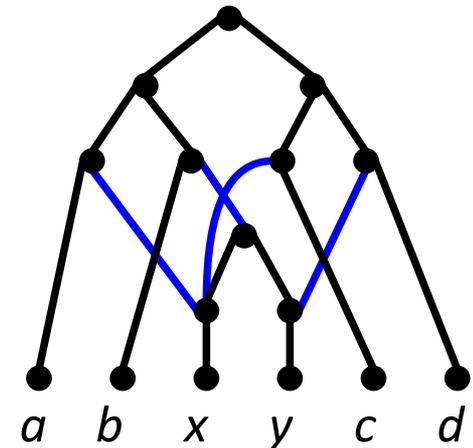
- partie sans conflits ➡ arbre,
- taxons impliqués dans des **conflits** ➡ sous les réticulations.



MAXIMUM COMPATIBLE SUBSET

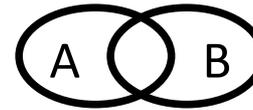
2- Attacher à l'arbre les taxons impliqués dans des conflits avec un **nombre minimal d'arcs** :

MINIMUM ATTACHMENT PROBLEM



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

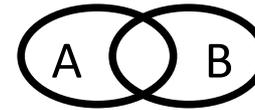
NP-complet sur un graphe des conflits connexe, sans taxons “jumeaux”

réduction depuis le cas général

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

L'ensemble minimum de conflits

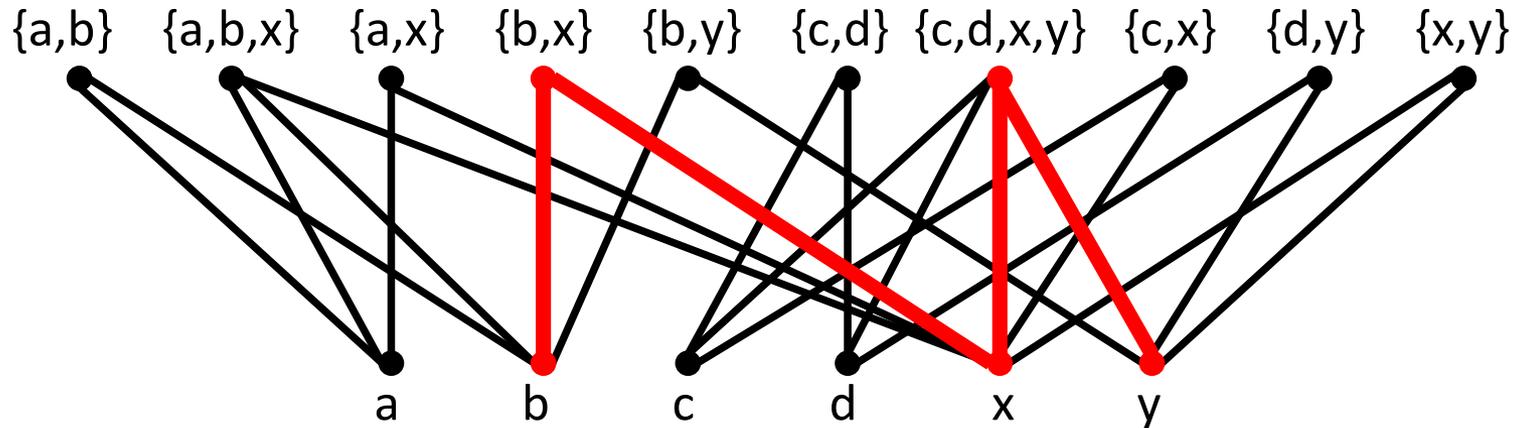
Conflit : clades ni inclus ni disjoints



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

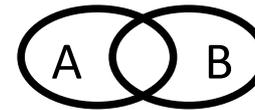
Exemple : $\{\{a,b\},\{a,b,x\},\{a,x\},\{b,x\},\{b,y\},\{c,d\},\{c,d,x,y\},\{c,x\},\{d,y\},\{x,y\}\}$



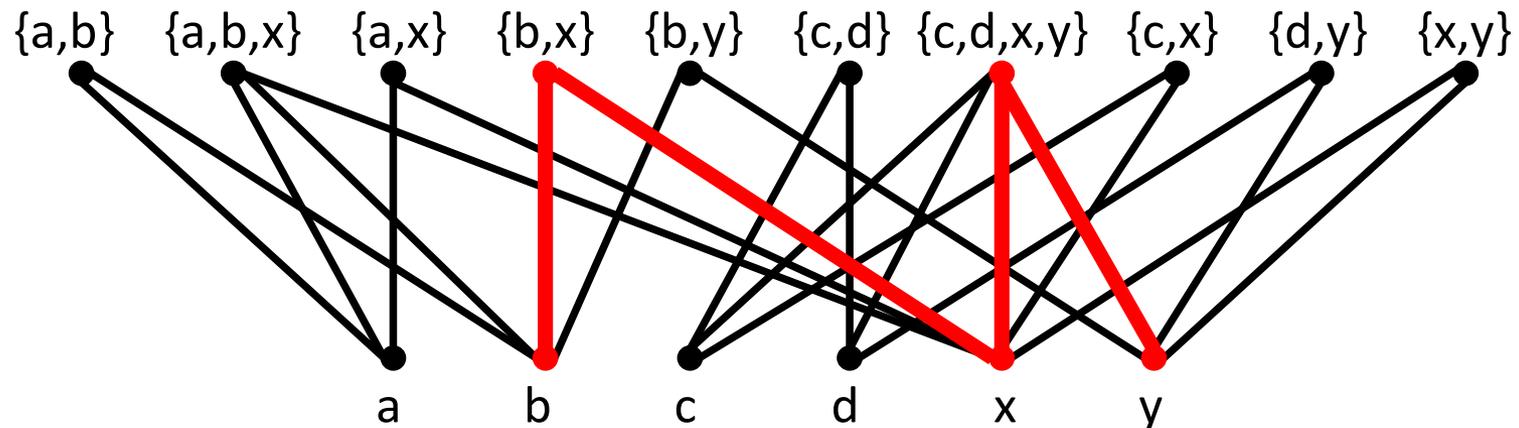
conflit = graphe "M"

L'ensemble minimum de conflits

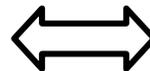
Conflit : clades ni inclus ni disjoints



Graphe des caractères :



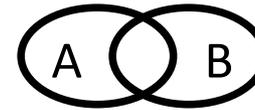
Trouver l'ensemble minimum de conflits



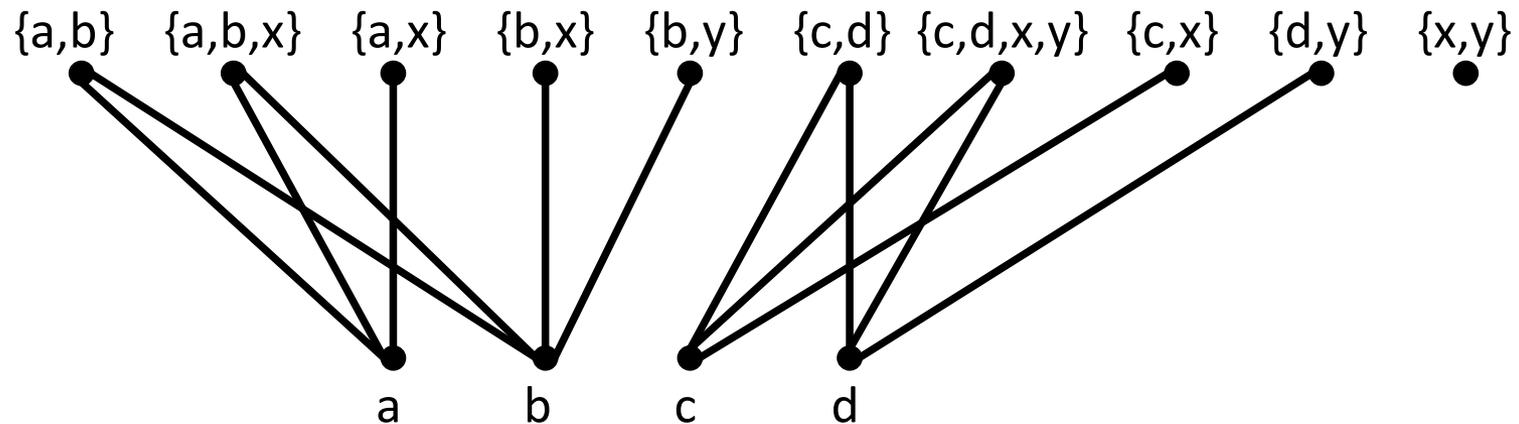
Supprimer le nombre minimum t de sommets-taxons tels que le graphe des caractères est un graphe "sans M"

L'ensemble minimum de conflits

Conflit : clades ni inclus ni disjoints



Graphe des caractères :



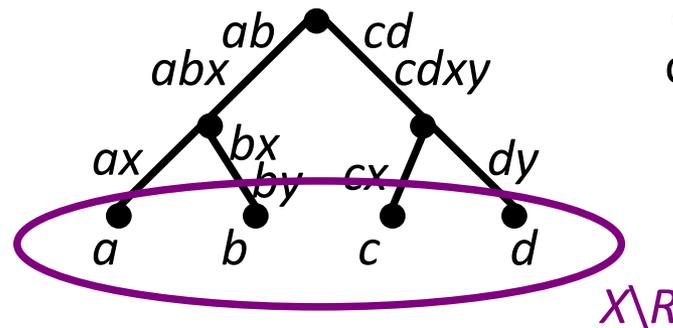
Supprimer le nombre minimum t de sommets-taxons tels que le graphe des caractères est un graphe “sans M” :

- algorithme FPT basique de branchement en $O^*(3^t)$
- algorithme FPT 3-Hitting-Set en $O^*(2,076^t)$

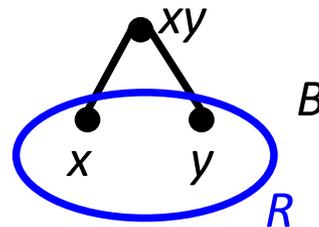
L'attachement minimum

Etape précédente :

ensemble minimum de taxons R tels que les clades sur $X \setminus R$ sont compatibles (avec un arbre T).



T : arbre représentant les clades sur $X \setminus R$



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

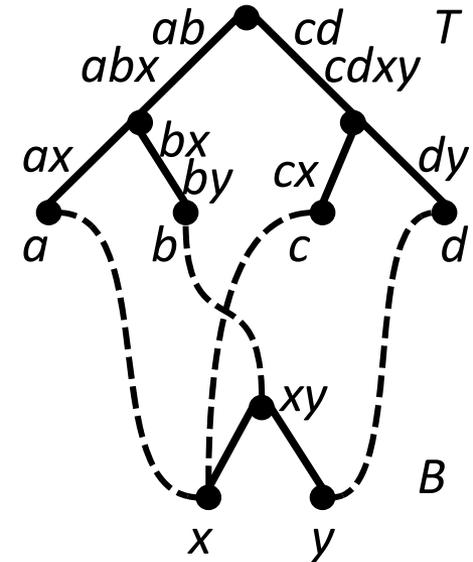
Problème :

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

L'attachement minimum

Problème :

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



NP-complet

Algorithmes :

- Séparation et évaluation
- Programme linéaire en nombres entiers

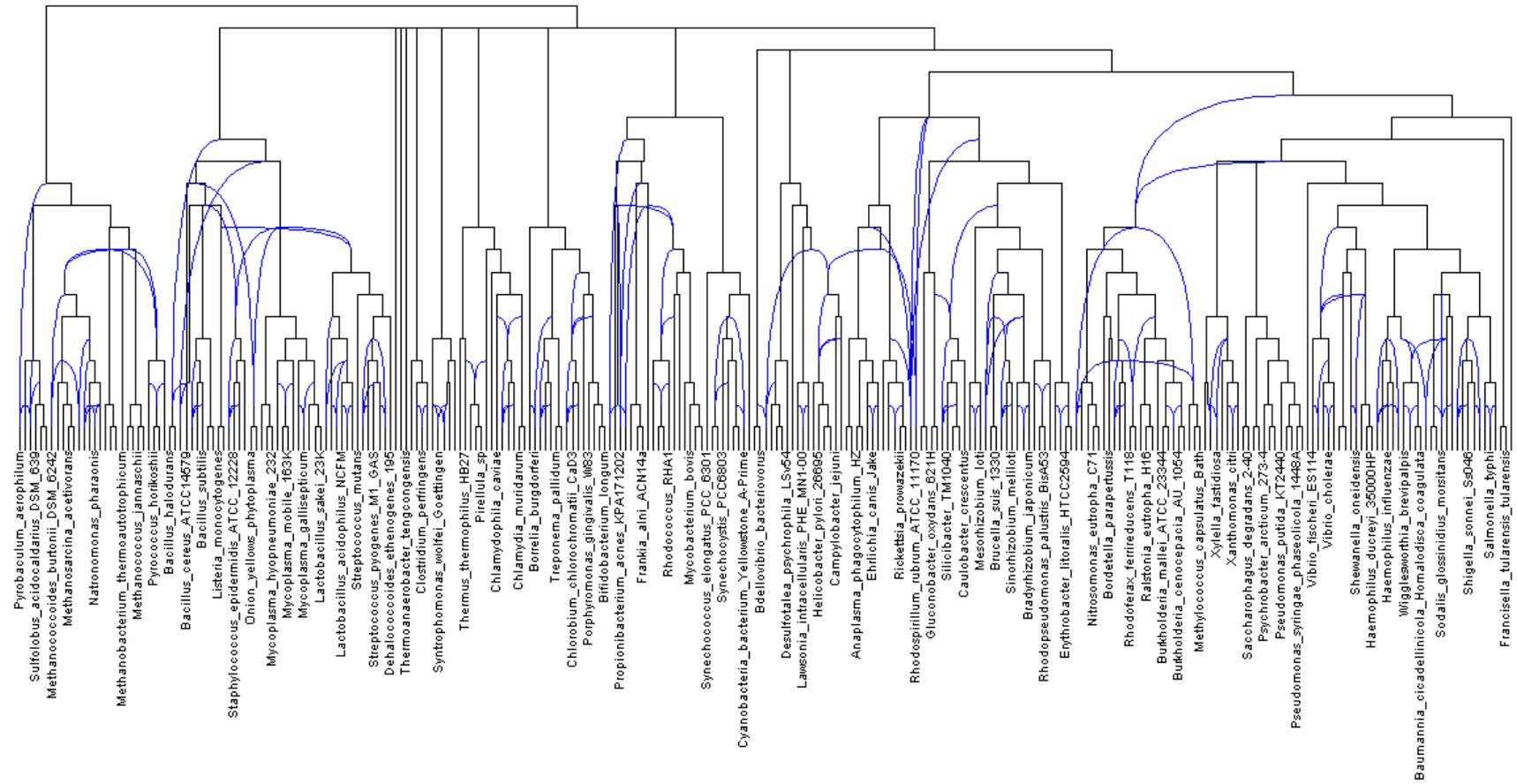
réduction depuis SetCover

implémenté dans Dendroscope 2

Exemples de résultats

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

Auch, Steigle, Huson & Henz, 2009



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

Exemples de résultats

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

24 Enterobacterales

2 Pasteurellales

1 Aeromonadales

9 Alteromonadales

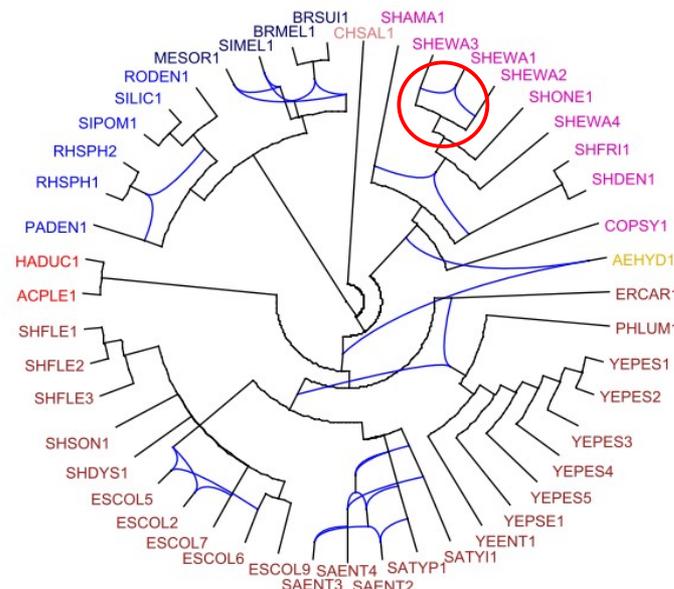
1 Oceanospirillales

6 Rhodobacterales

4 Rhizobiales

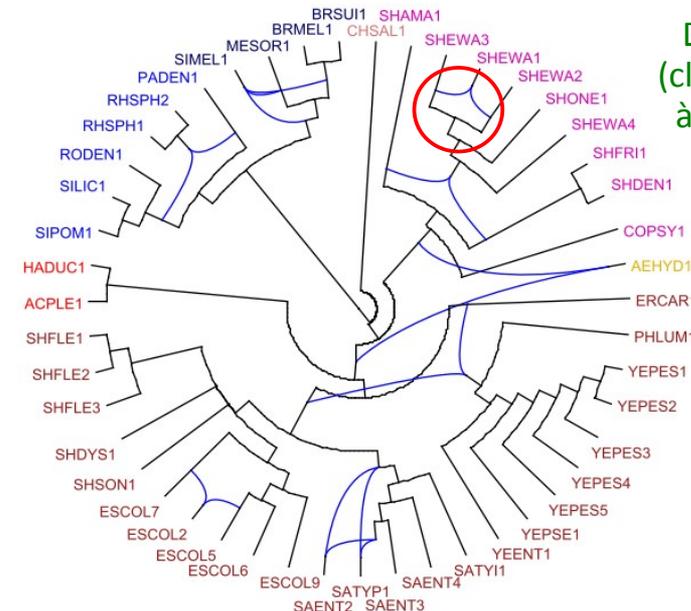
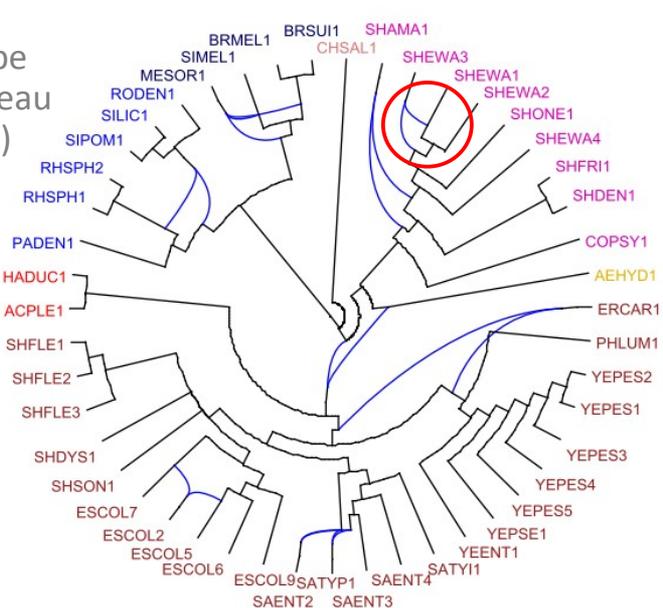


Réseaux contenant les clades souples
présents dans au moins 20% des arbres



Dendroscope
(réseau de
clades)
<1 sec.

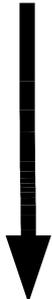
Dendroscope
(clades, réseau
de niveau 2)
2 sec.



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

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,

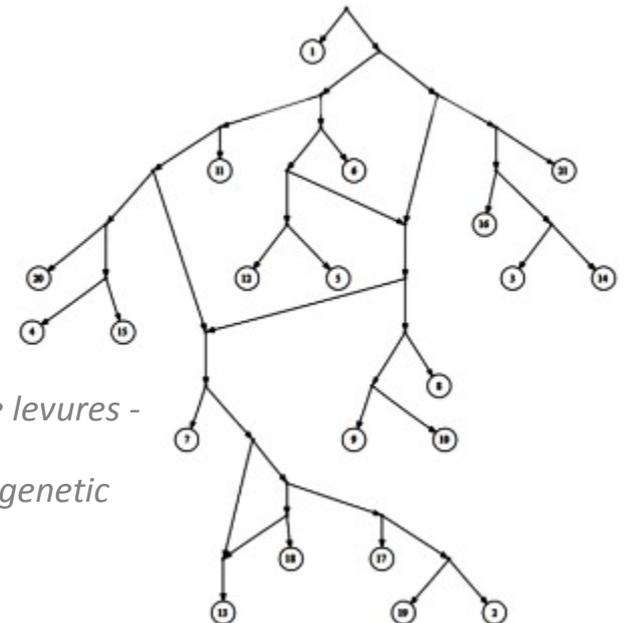
To & Habib, CPM'09 : $O(n^{5k+4})$ pour niveau k

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



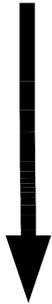
*Réseau phylogénétique de levures -
Van Iersel et al. :*

*Constructing level-2 phylogenetic
networks from triplets.*

RECOMB 2008

Reconstruction depuis les quadruplets

{arbres}



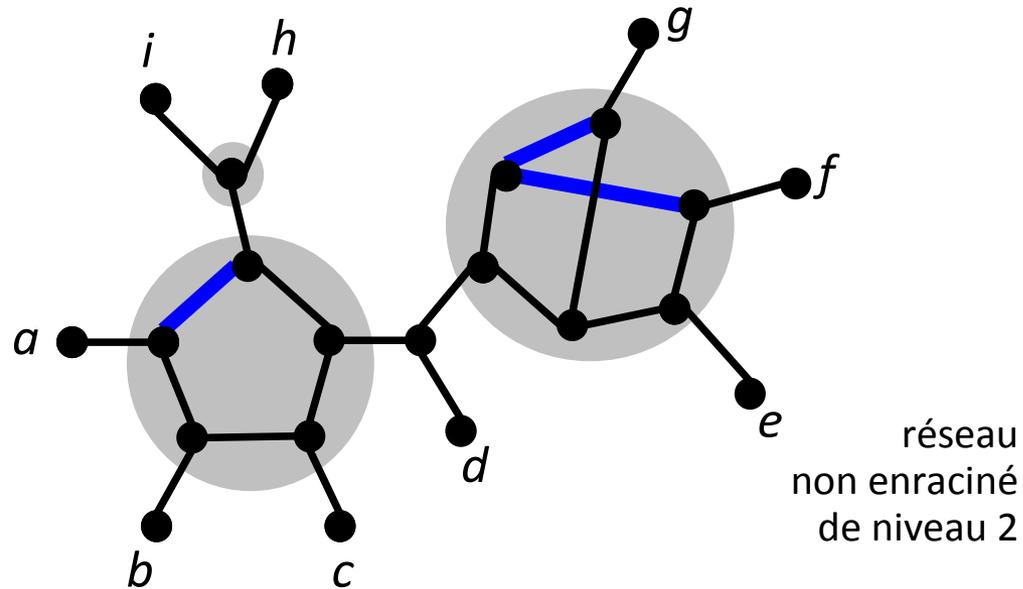
{quadruplets}



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

Reconstruction depuis les quadruplets

{arbres}

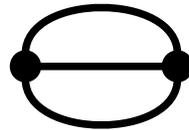


{quadruplets}

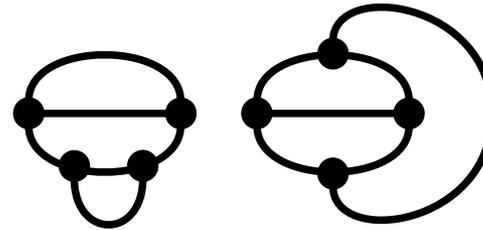


N' réseau
non enraciné
de niveau k

Générateurs de niveau k :
multigraphes biconnexes 3-réguliers sans boucle
à $2k-2$ sommets



niveau 2



niveau 3

Reconstruction depuis les quadruplets

{arbres}

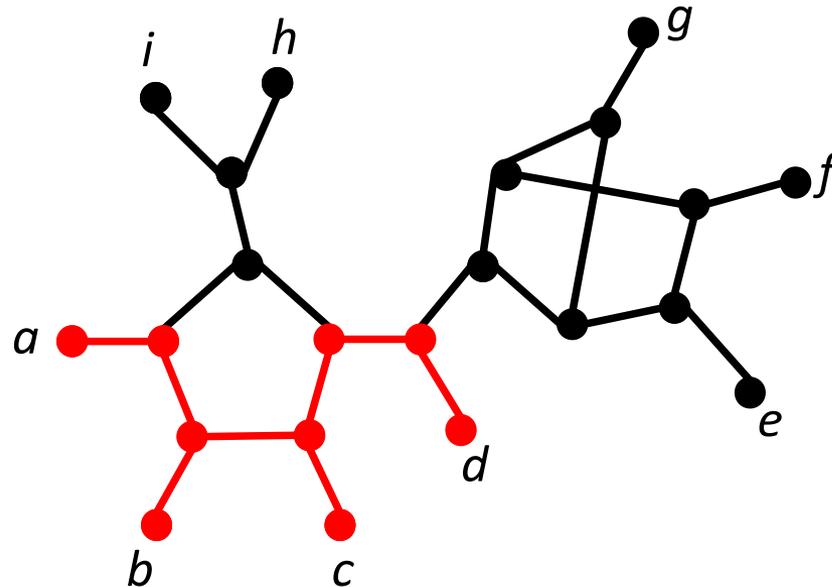


{quadruplets}



N' réseau
non enraciné
de niveau k

Quadruplet $ab|cd$



Calcul des quadruplets d'un réseau : $O(n^6)$
"2-disjoint paths" dans un graphe de degré max 3

Reconstruction depuis les quadruplets

{arbres}

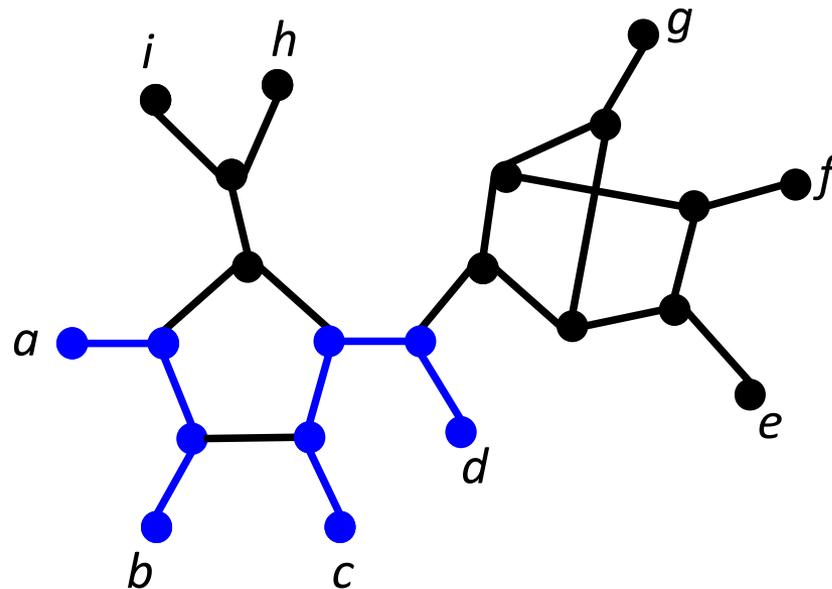


{quadruplets}



N' réseau
non enraciné
de niveau k

Quadruplet $ab|cd$



Calcul des quadruplets d'un réseau : $O(n^6)$
"2-disjoint paths" dans un graphe de degré max 3

Reconstruction d'arbres

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

Reconstruction de réseaux de niveau k

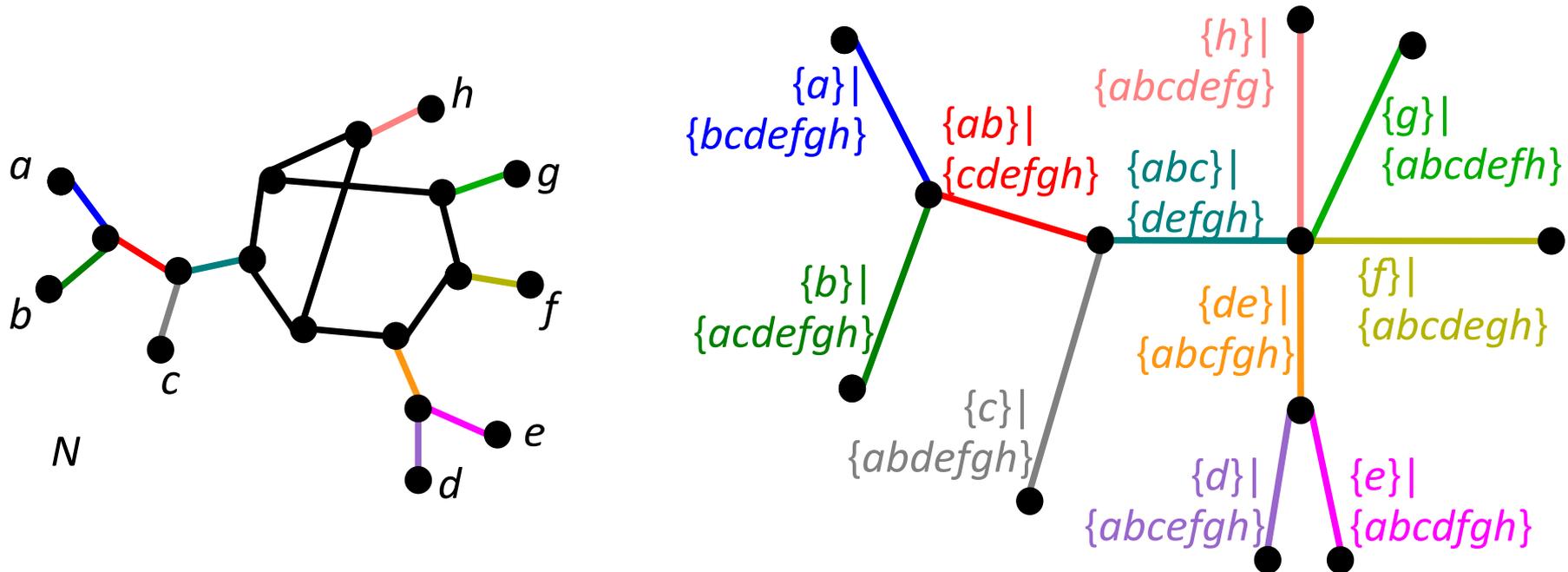
	non enraciné		enraciné	
	niveau 1	niveau $k > 1$	niveau 1	niveau $k > 1$
général	NP-complet <i>Grünwald, Moulton & Spillner, DAM, 2009</i>	?	NP-complet <i>Jansson, Nguyen & Sung, SJOC, 2006</i>	NP-complet <i>Van Iersel, Kelk & Mnich, JBCB, 2009</i>
dense <i>au moins un quadruplet pour tout ensemble de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

Décomposition des ensembles de quadruplets

$Q(N)$, ensemble de **tous** les quadruplets
contenus dans un réseau N de niveau k



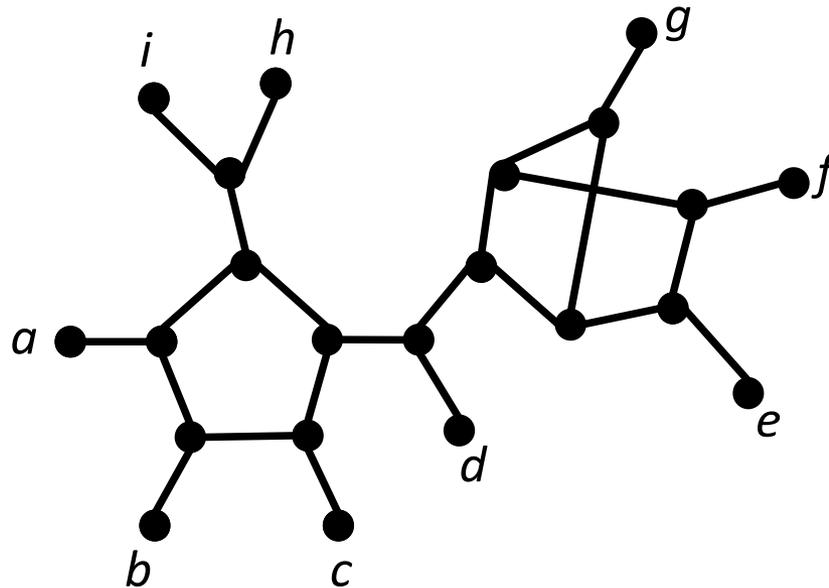
Les **SN-bipartitions** de $Q(N)$
sont en bijection avec les **isthmes** de N .



Bipartitions et réseaux non enracinés

Bipartition :

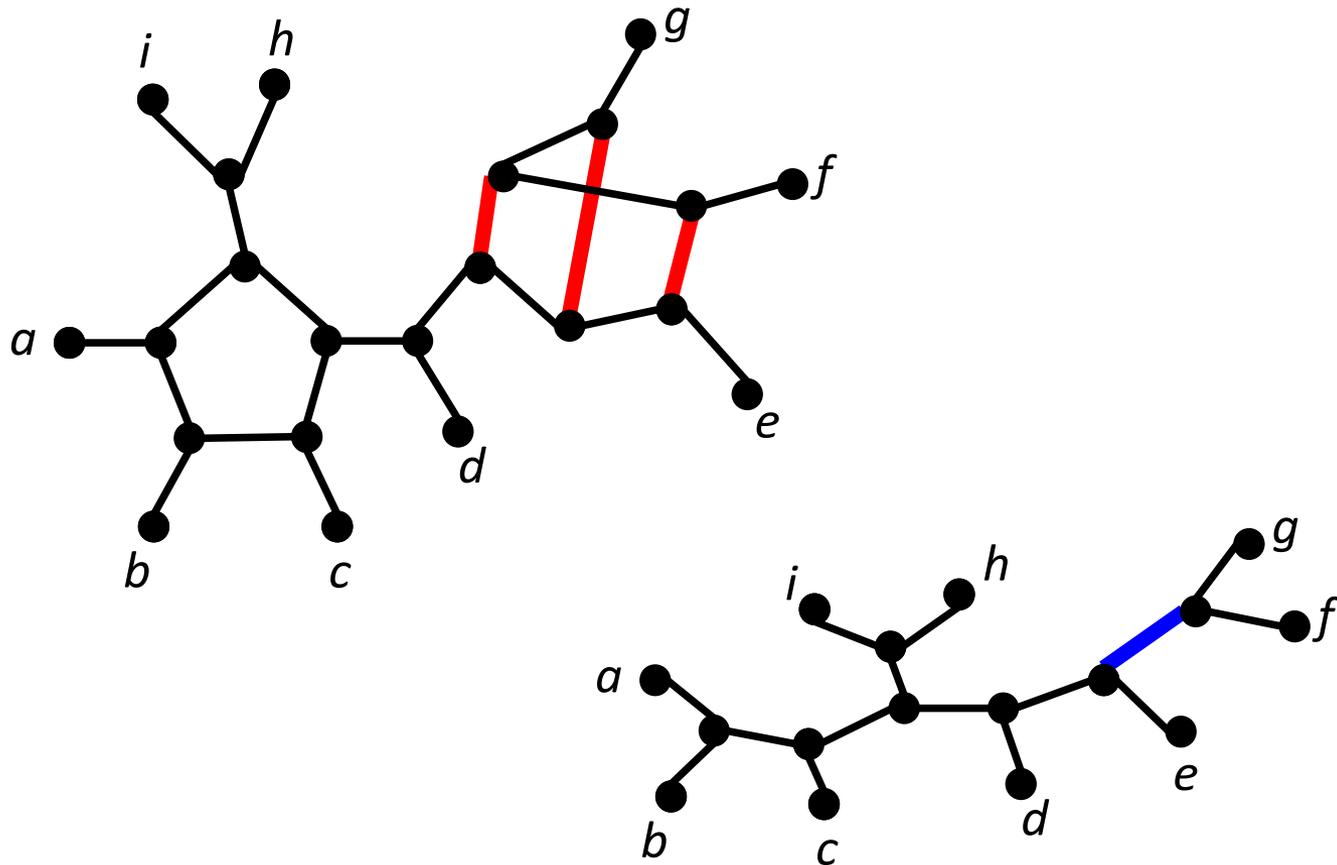
- ➡ Bipartition d'un arbre contenu dans le réseau
- ➡ Feuilles séparées par une coupe minimale du réseau



Bipartitions et réseaux non enracinés

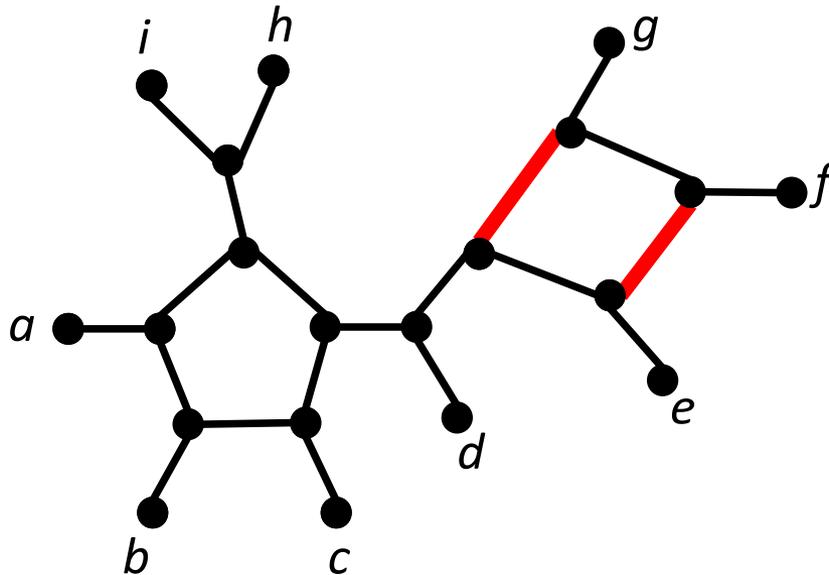
Bipartition :

- ➡ **Bipartition** d'un arbre contenu dans le réseau
- ➡ Feuilles séparées par une **coupe minimale** du réseau

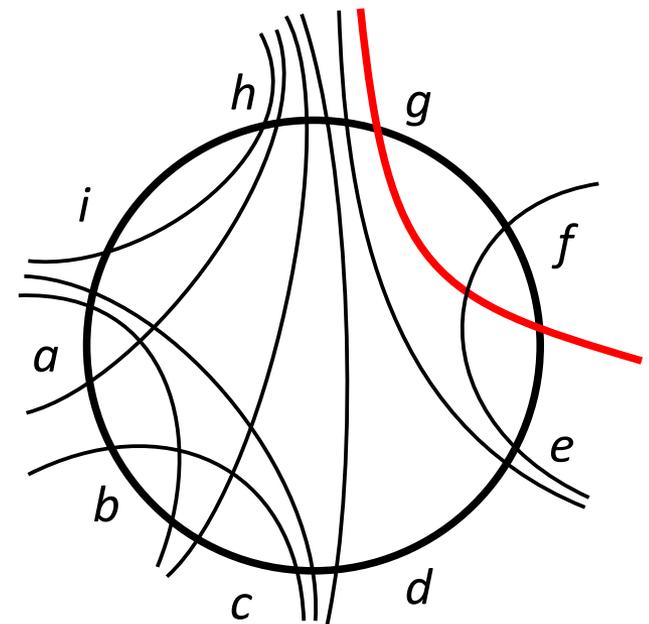


Bipartitions et réseaux non enracinés

Bipartitions d'un réseau de niveau 1 :

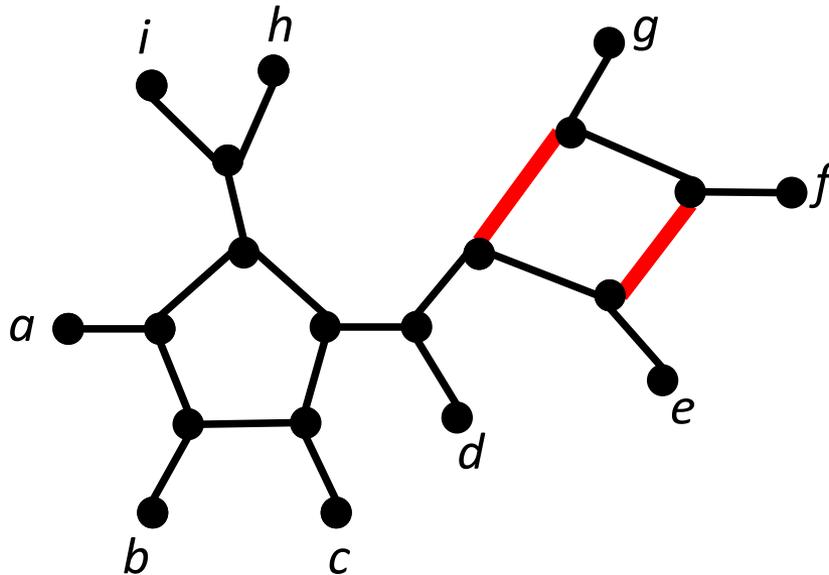


Ensemble **circulaire** de bipartitions
"circular split system"

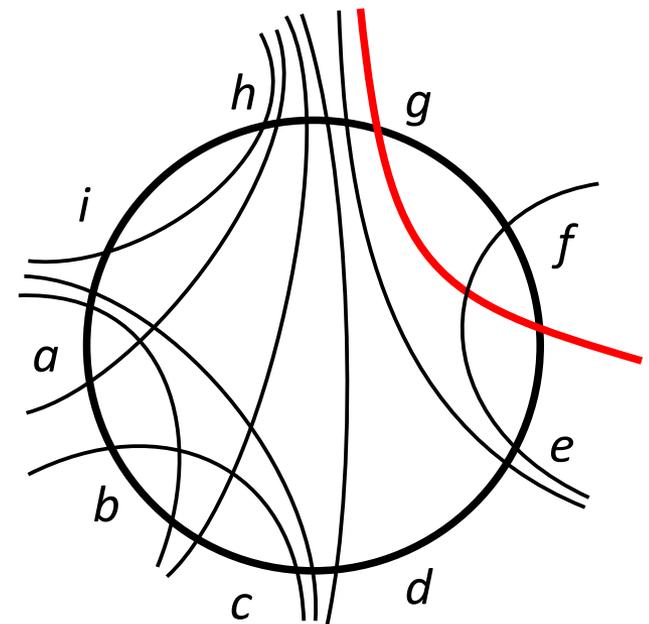


Bipartitions et réseaux non enracinés

Bipartitions d'un réseau de niveau 1 :



Ensemble **circulaire** de bipartitions
“*circular split system*”



Liens avec les réseaux phylogénétiques abstraits,
T-theory

Gambette, Berry & Paul, 2011

(Dress, Bandelt, Moulton...)

Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Restrictions sur les réseaux phylogénétiques
- Méthodes combinatoires de reconstruction
- **Limites**
- 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 :

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

- arbres à partir de triplets :

$O^*(3^t)$, Guillemot & Berry, TCBB, 2007,
 $O(n^4 + 2^{O(t^{1/3} \log t)})$, 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 :

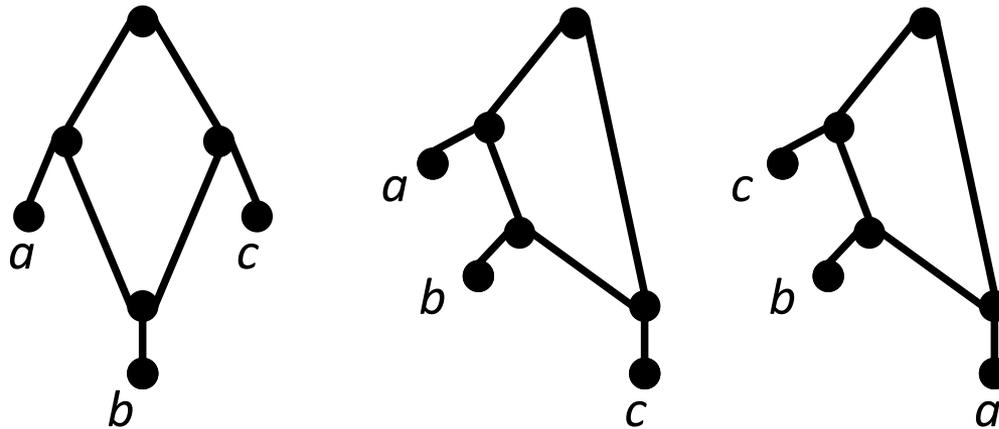
Problème NP-complet

Réduction de Clique

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.

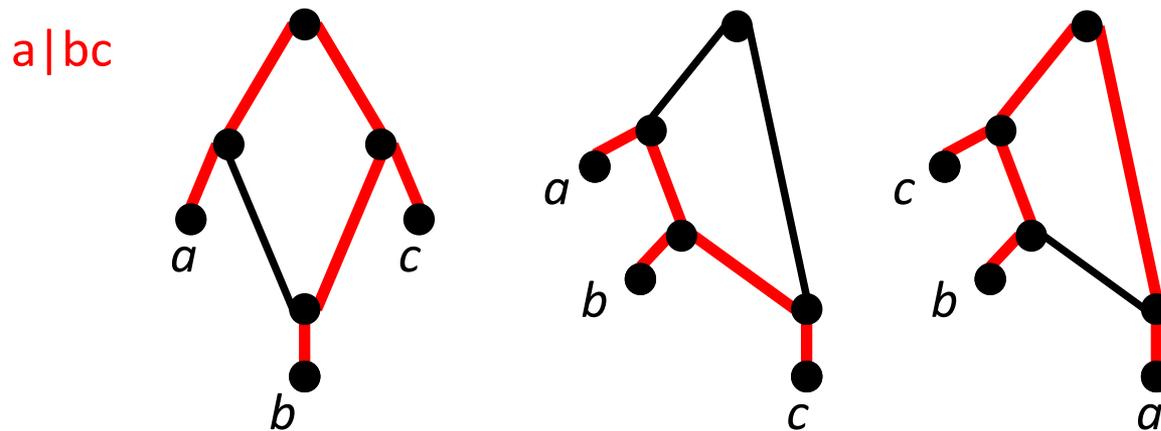


Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 sommets)

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.

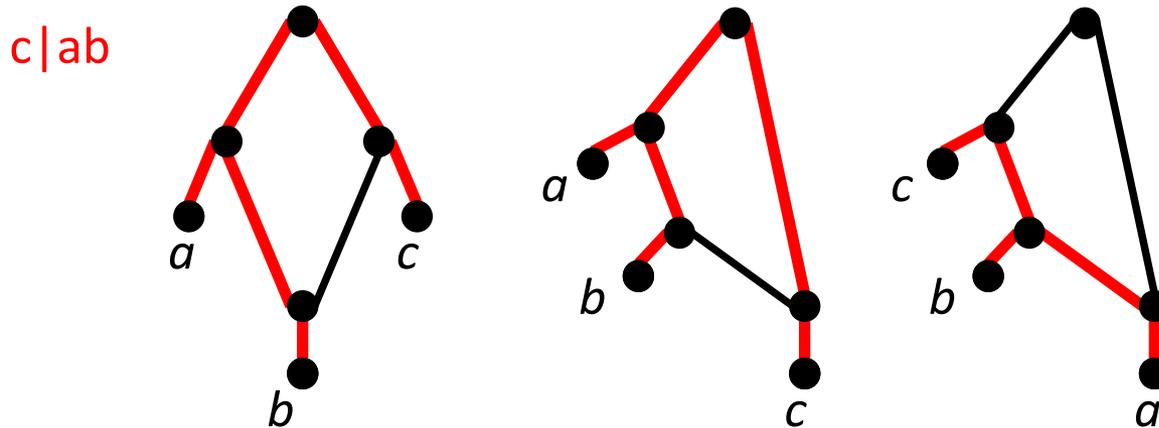


Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 sommets)

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.

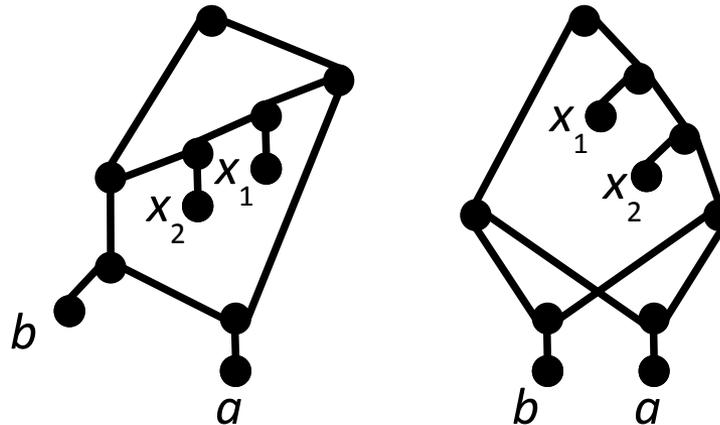


Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 sommets)

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.

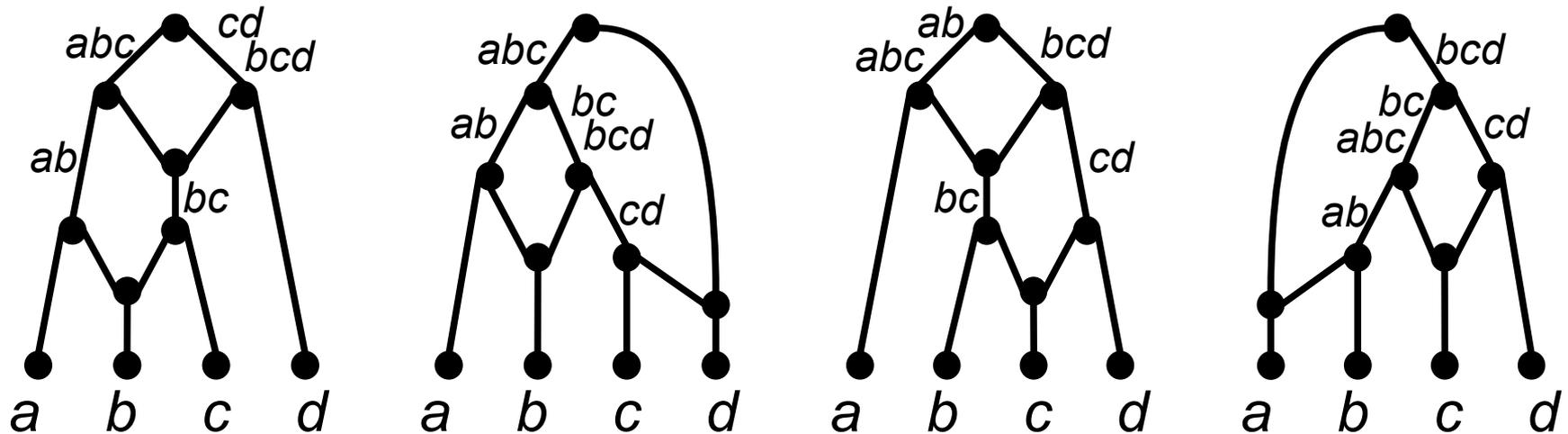


2 réseaux de niveau 2 avec le même ensemble de triplets

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.



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

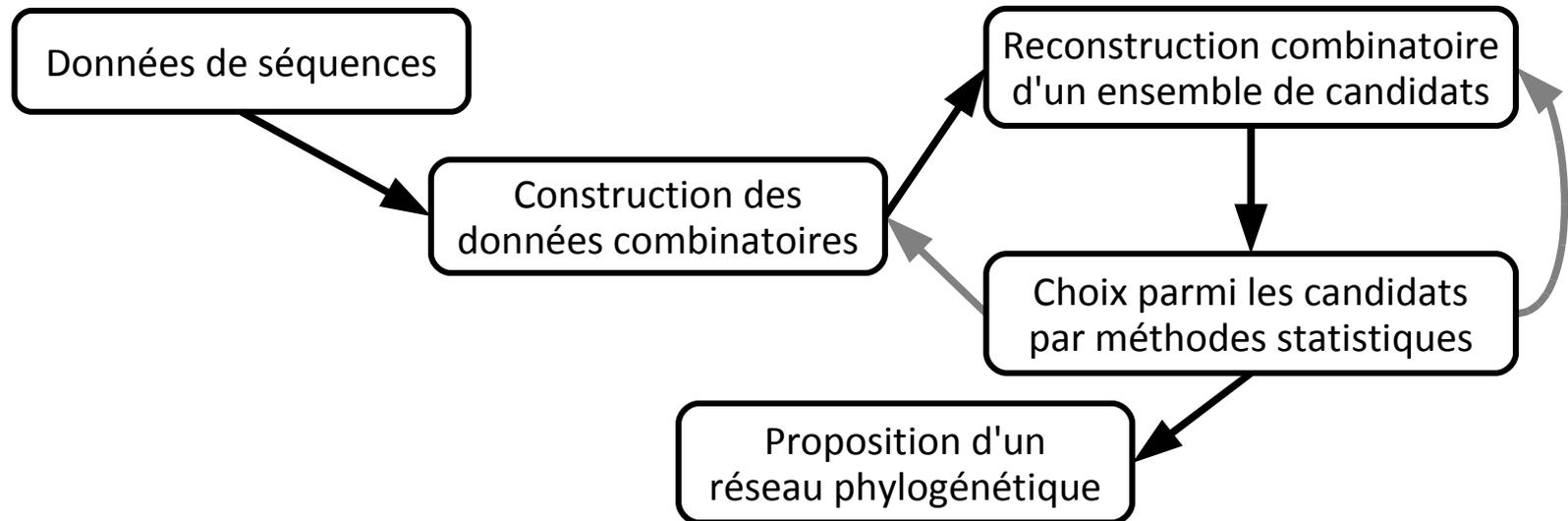
Plan

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

Perspectives de recherche

Bioinformatique :

- Intégration des méthodes combinatoires dans une approche statistique



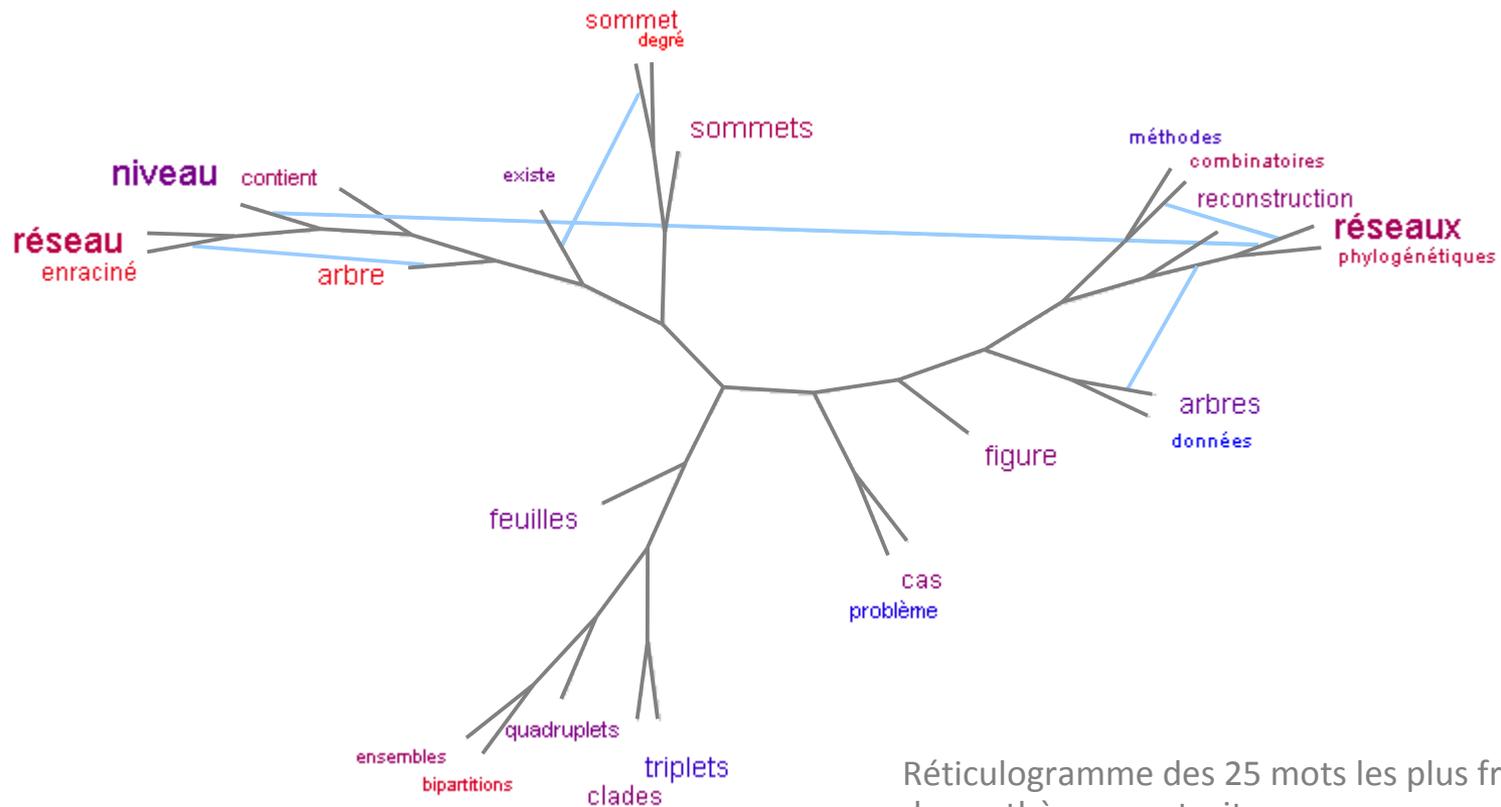
Combinatoire :

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

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>