

30/11/2010

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

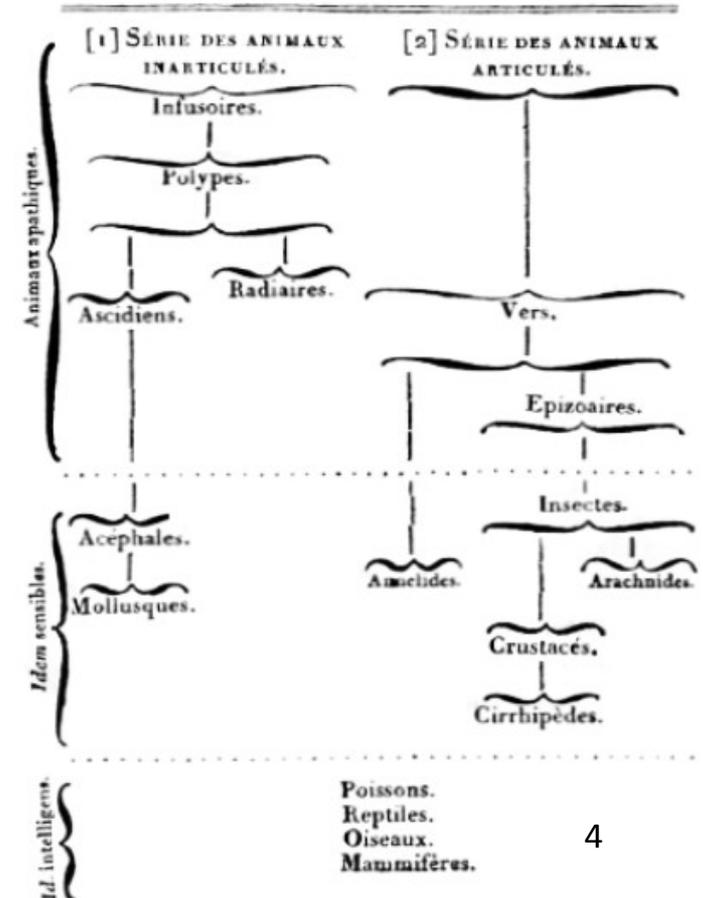
Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

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

classification

*ORDRE présumé de la formation des Animaux ,
offrant 2 séries séparées , subrameuses.*



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

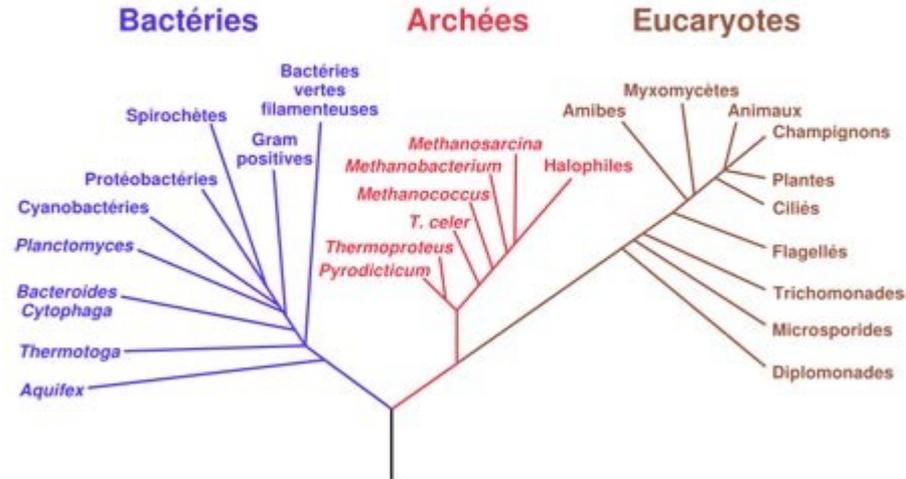
Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

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

modélisation

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)

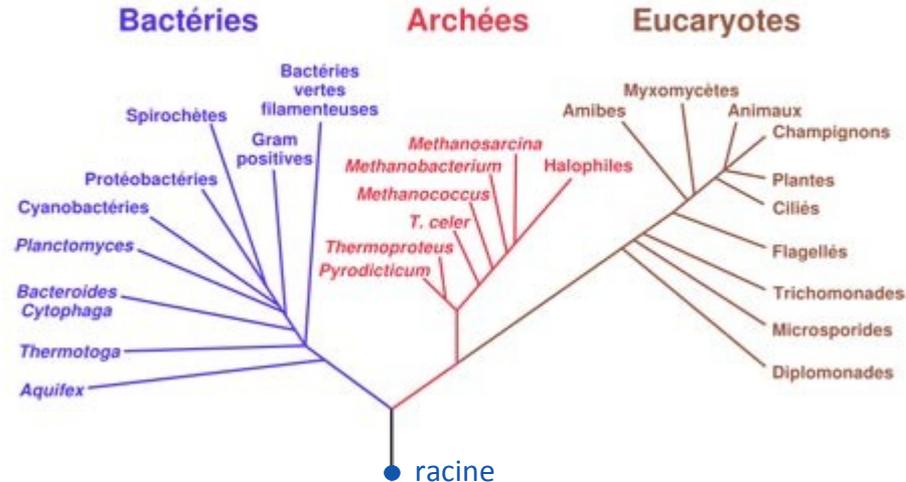
Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

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

modélisation

Arbre phylogénétique de la vie

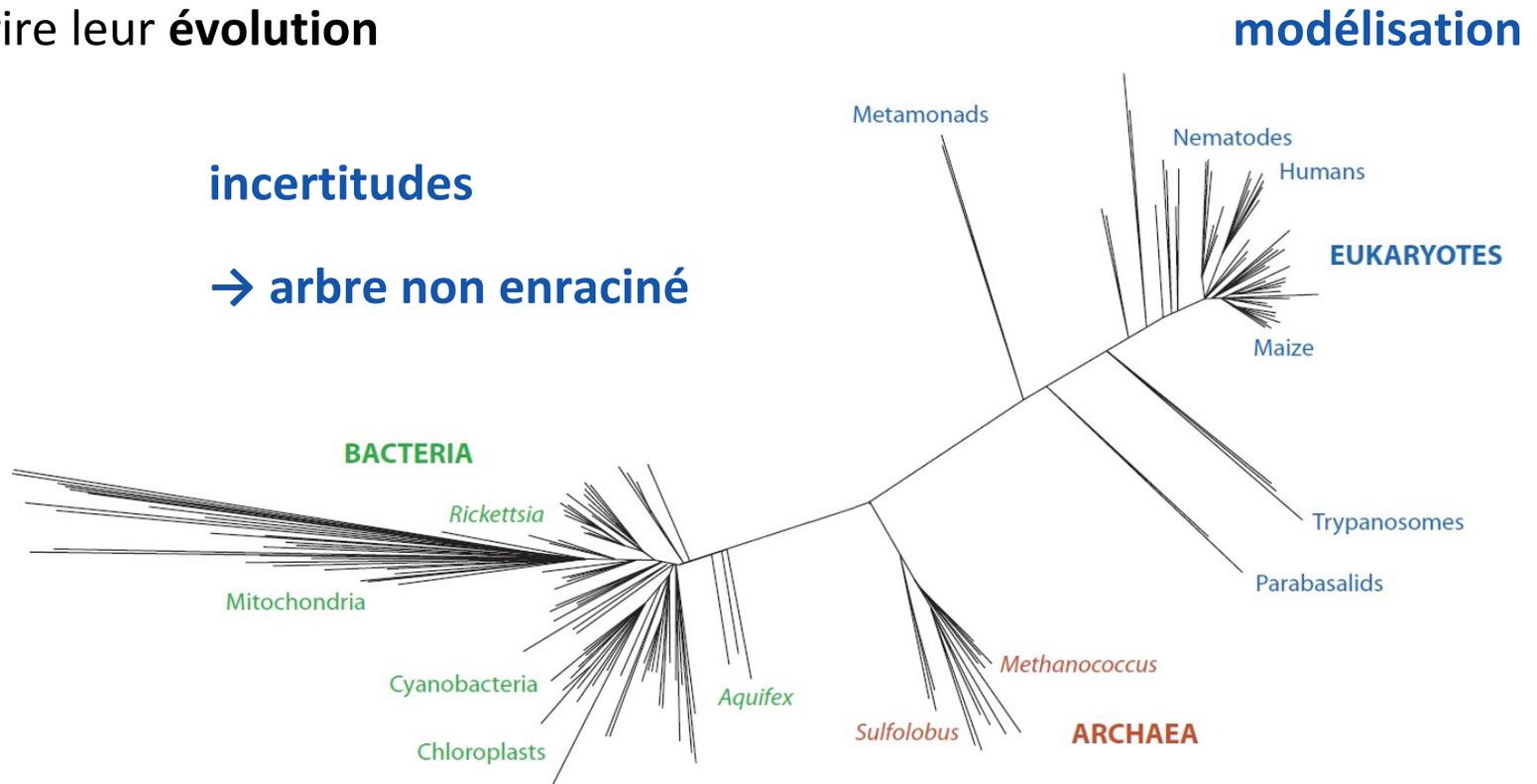


D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

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

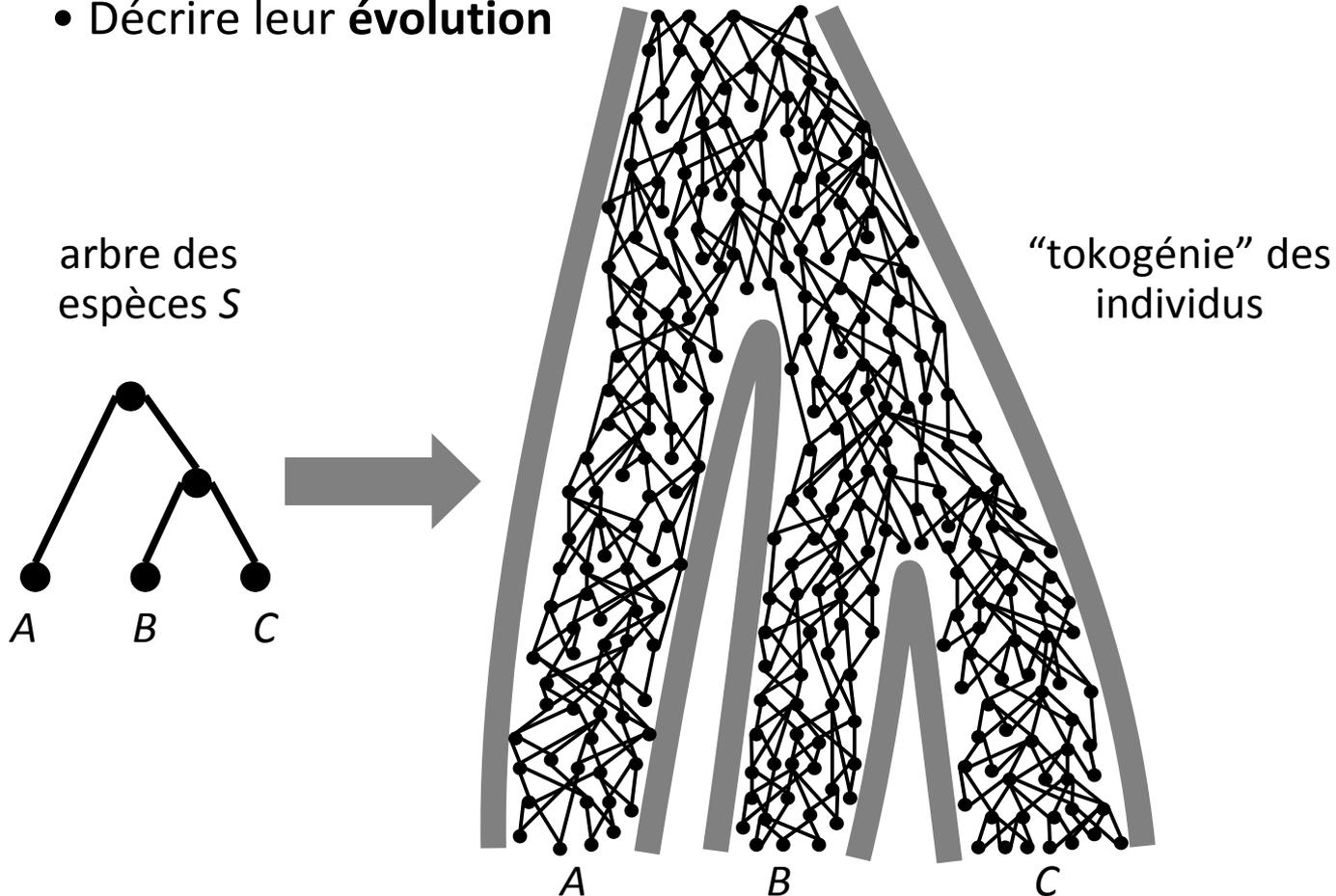


D'après Christophe Blumrich, David S. Spencer,
cité dans Doolittle : Uprooting the Tree of Life, Scientific American (Fév. 2000)

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

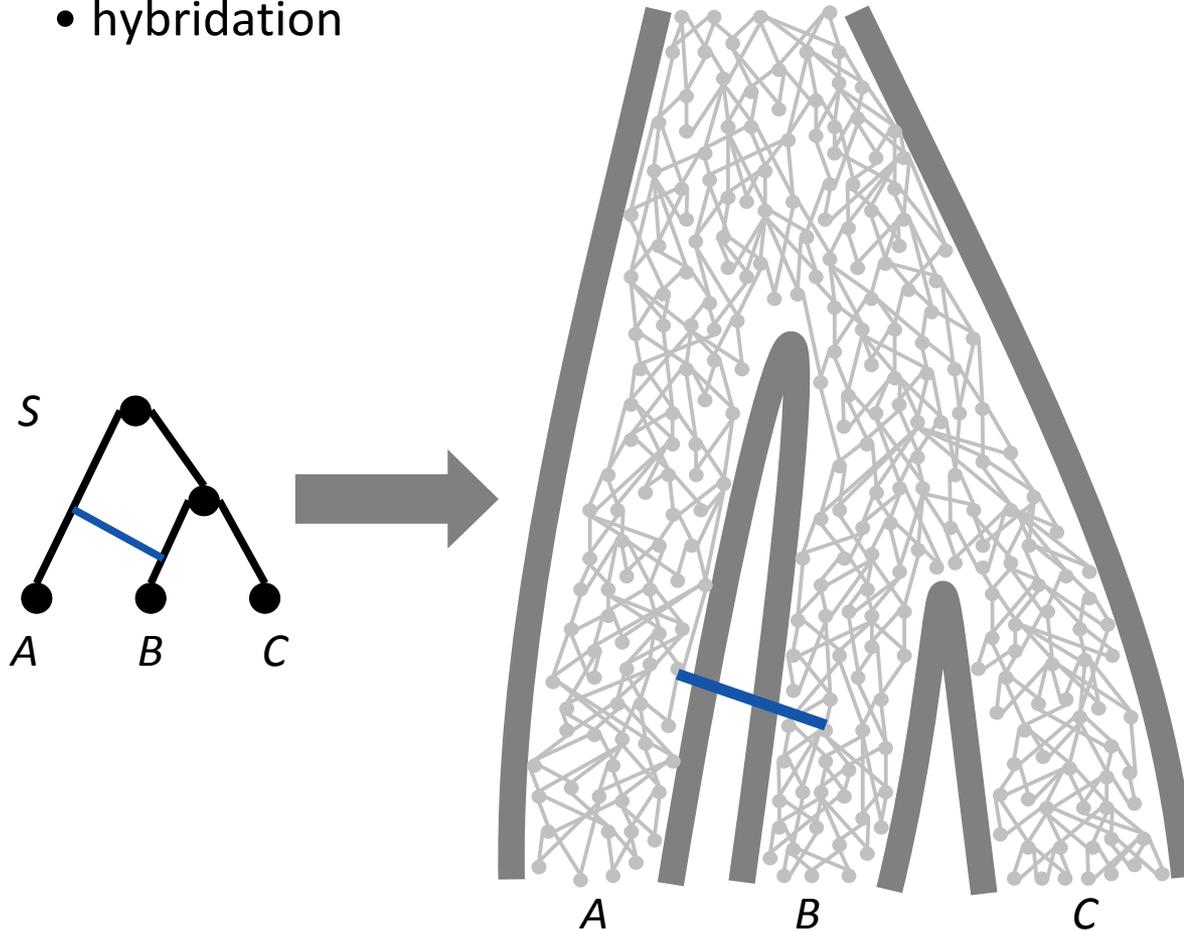
- Les organiser en fonction de caractères communs
- Décrire leur **évolution**



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

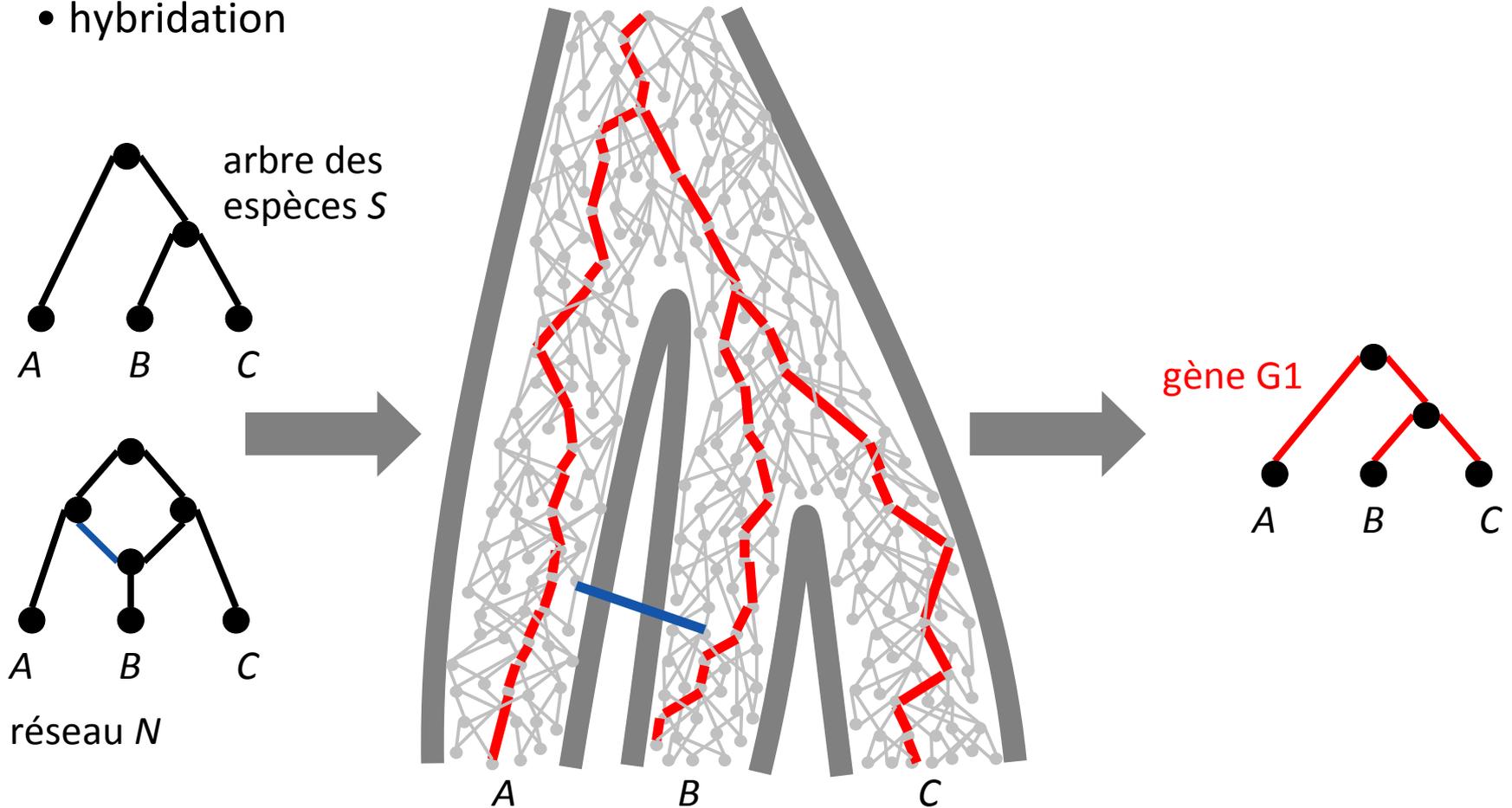
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

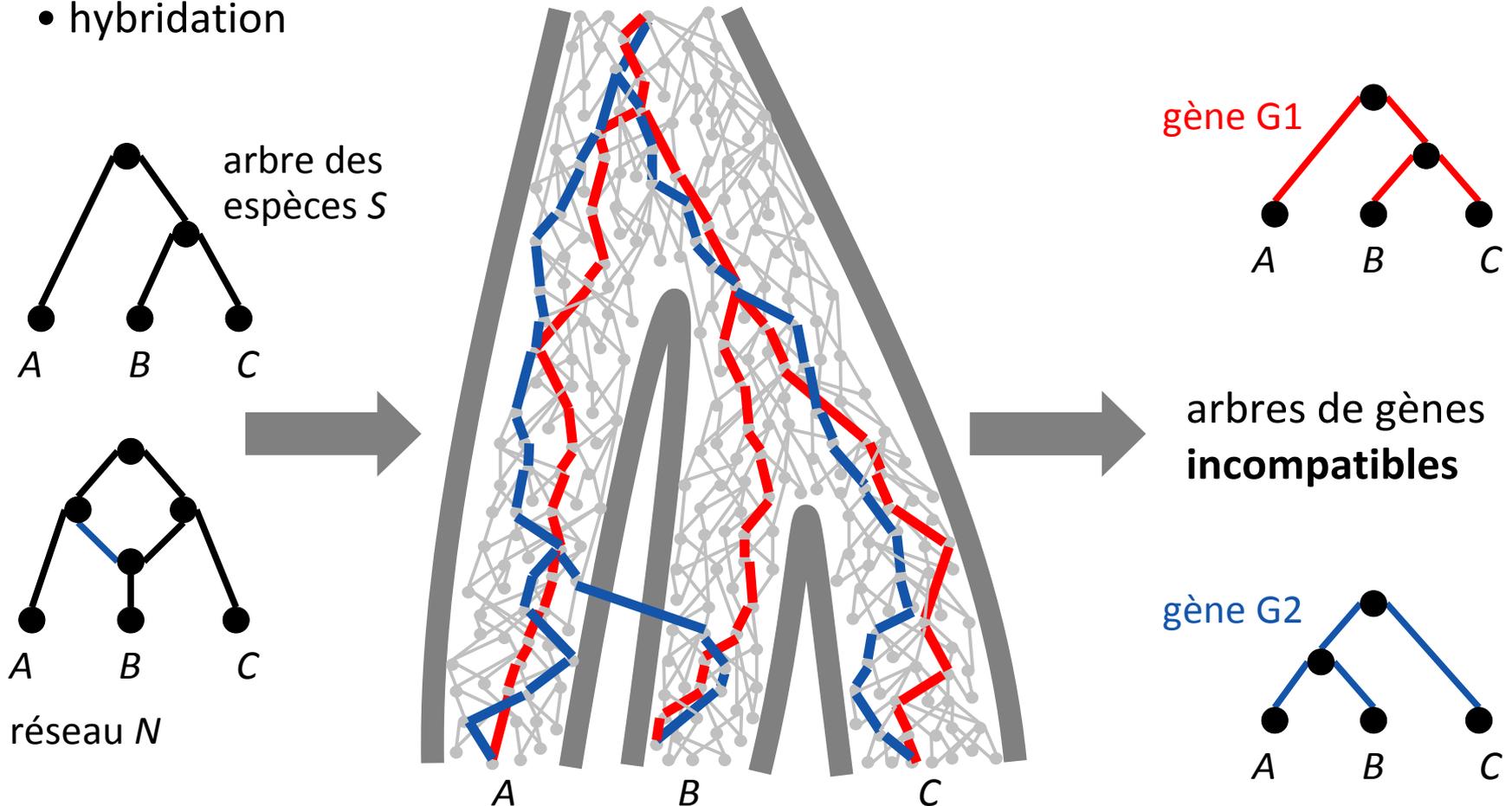
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

- transfert horizontal
- hybridation

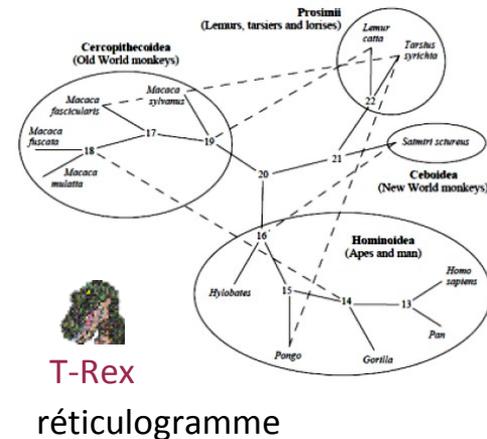
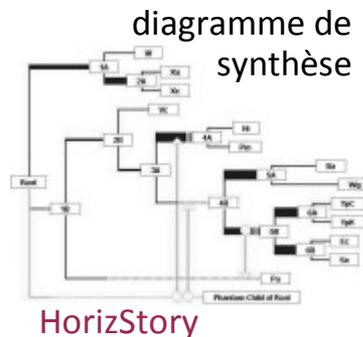
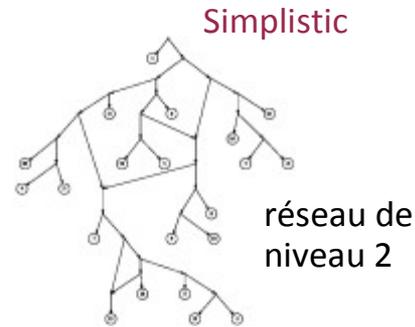
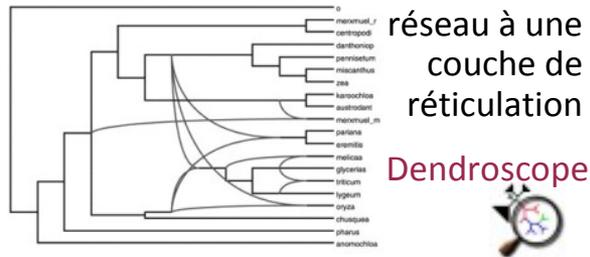


Les réseaux phylogénétiques

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **explicités**

modélisation de l'évolution

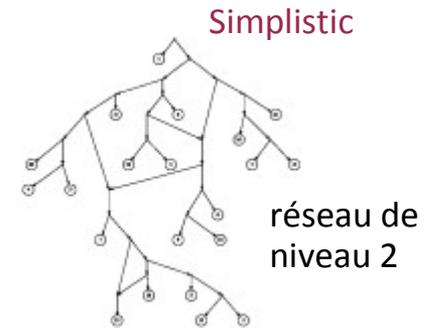
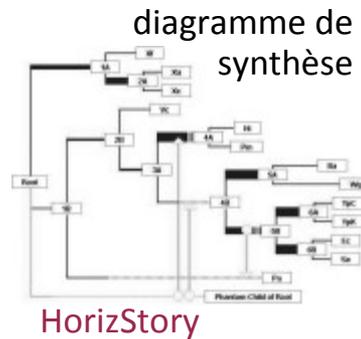
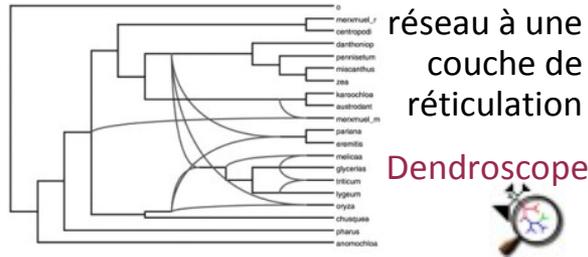


Les réseaux phylogénétiques

Réseau phylogénétique : réseau représentant des données d'évolution

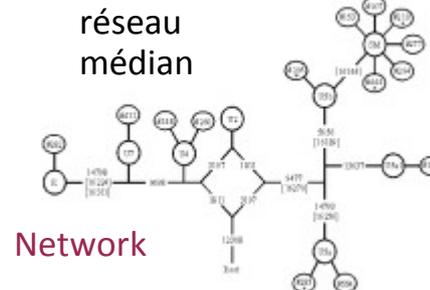
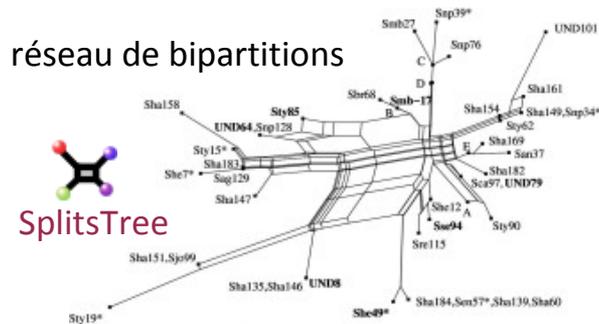
- réseaux phylogénétiques **explicités**

modélisation de l'évolution



- réseaux phylogénétiques **abstraites**

classification, visualisation de données



réseau couvrant minimum

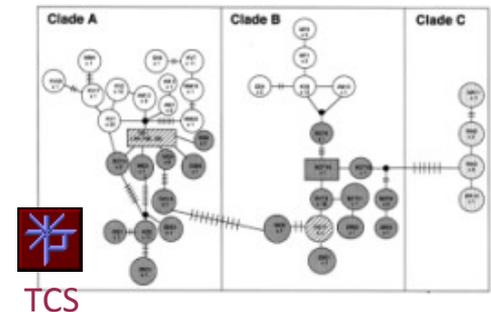


Plate-forme bibliographique

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

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

Index | Browse

Search: in All Go (word length ≥ 3) [Login](#)

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	1
2010	44

See how the community working on phylogenetic networks evolves

Selection by Category

- Article (Journal) (216)
- Book (1)
- Misc (19)
- InProceedings (19)
- PhdThesis (19)
- Programs (52)

Selection by Keyword

abstract-network(46) approximation(8) APX-hard(2) ARG branch-and-bound(1) cactus-graph(1) characterization(7) circular-split-consistency(2) cophylogeny(1) distance-between-networks(21) divergence(1) **explicit-network(93)** exponential-algorithm(2) FPT(16) from-tree(6) from-network(12) from-quartets(7) **from-rooted-tree(26)** from-splits(9) from-trees(6) from-triplets(17) from-generation(8) haplotype-network(2) haplotyping(1) heuristic(11) HMM programming(1) labeling(4) lateral-gene-transfer(35) level-k-sorting(5) MASN(4) **median-network(15)** MedianJoining(2) minin-selection(2) mu-distance(2) NeighborNet(11) nested-network(2) realization(2) parsimony(32) perfect(5) **phylogenetic-network(46)** Program-Arlequin(5) Program-Beagle(3) Program-constNJ(1) Program-Dendroscope(7) Program-EEEEP(3) Program-GalledTree(4) Program-HybridInterleave(4) Program-HybridNET(1) Program-HybridNumber(3)

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

Index | Browse

Search: in All Go (word length ≥ 3) [Login](#)

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: Show

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>.

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

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

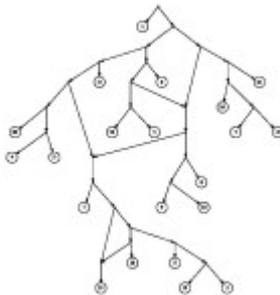
*Bandelt & Dress 1992 - Legendre & Makarenkov
2000 - Bryant & Moulton 2002*

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

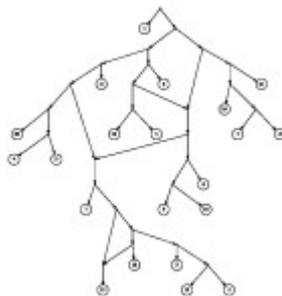
*Bandelt & Dress 1992 - Legendre & Makarenkov
2000 - Bryant & Moulton 2002*

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*

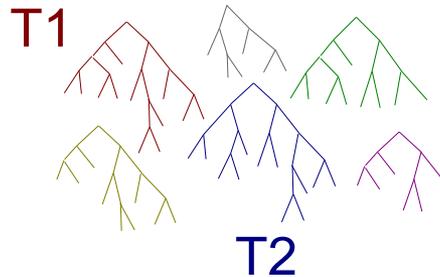


réseau *N*

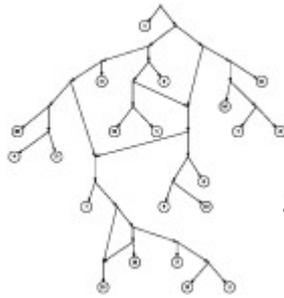
Reconstruction de réseaux phylogénétiques

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



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

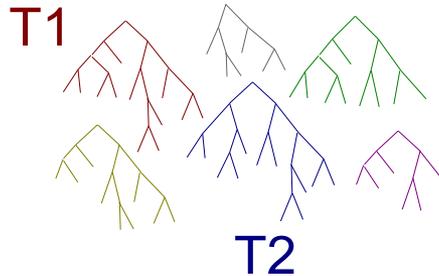
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAAT
espèce 5 : TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAAT

G1 G2

{séquences de gènes}

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

Guindon & Gascuel, SB, 2003



{arbres}

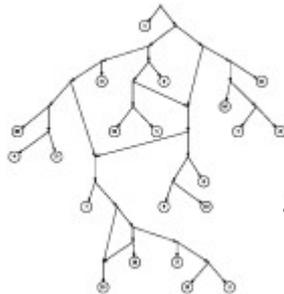
Base HOGENOM



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

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 :

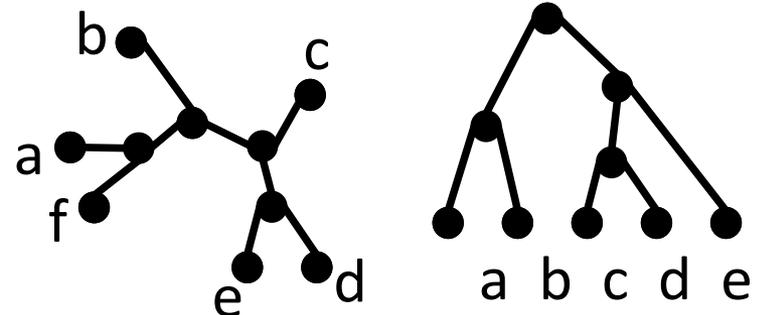
Reconstruire le **super-réseau** d'un ensemble d'arbres est
difficile.

Idée :

reconstituer un réseau contenant tous les :

triplets
quadruplets
clades
bipartitions

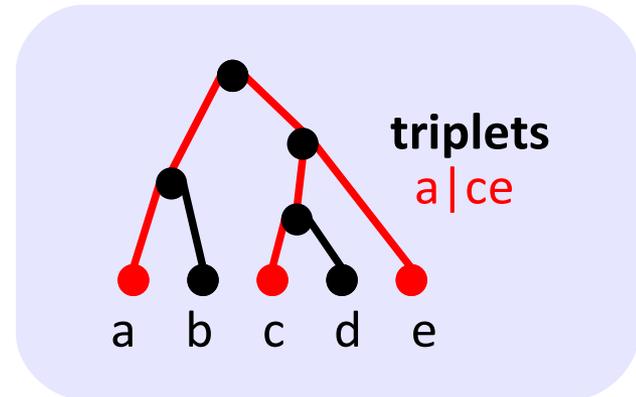
des arbres en entrée ?



Triplets et quadruplets, clades et bipartitions

Idée :

reconstituer un réseau contenant tous les :



des arbres en entrée ?

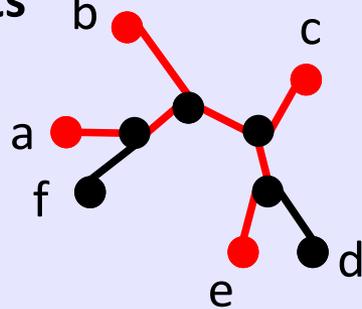
Triplets et quadruplets, clades et bipartitions

Idée :

reconstituer un réseau contenant tous les :

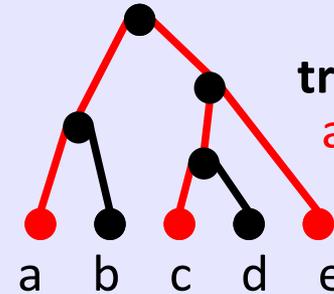
quadruplets

ab|ce



triplets

a|ce



des arbres en entrée ?

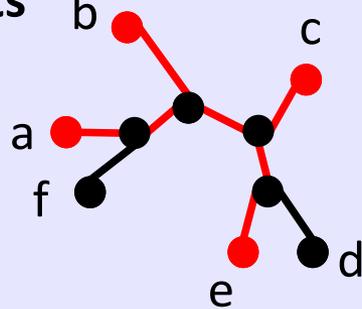
Triplets et quadruplets, clades et bipartitions

Idée :

reconstituer un réseau contenant tous les :

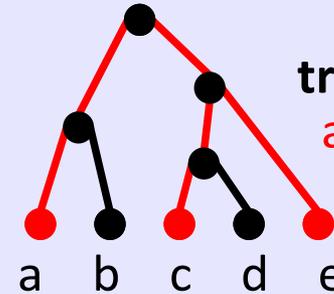
quadruplets

$ab|ce$



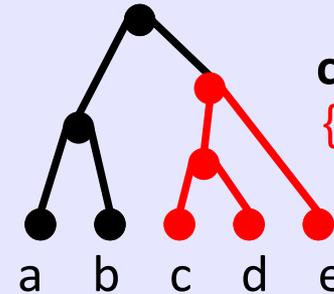
triplets

$a|ce$



clades

$\{c,d,e\}$



des arbres en entrée ?

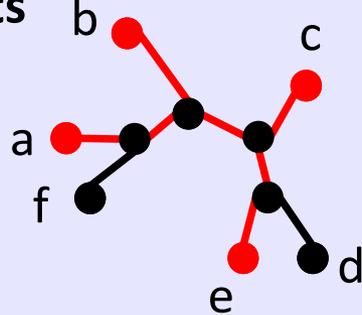
Triplets et quadruplets, clades et bipartitions

Idée :

reconstituer un réseau contenant tous les :

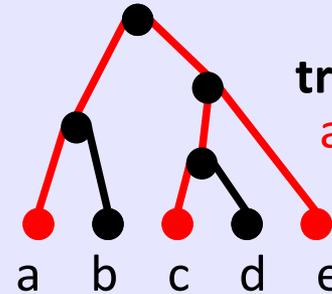
quadruplets

$ab|ce$



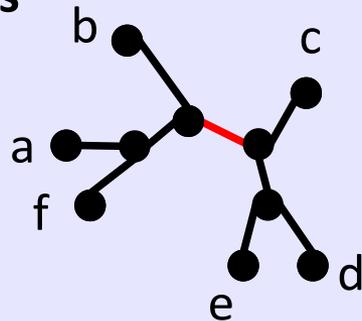
triplets

$a|ce$



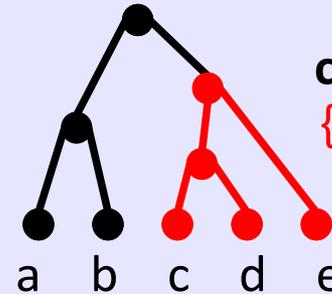
bipartitions

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



clades

$\{c,d,e\}$

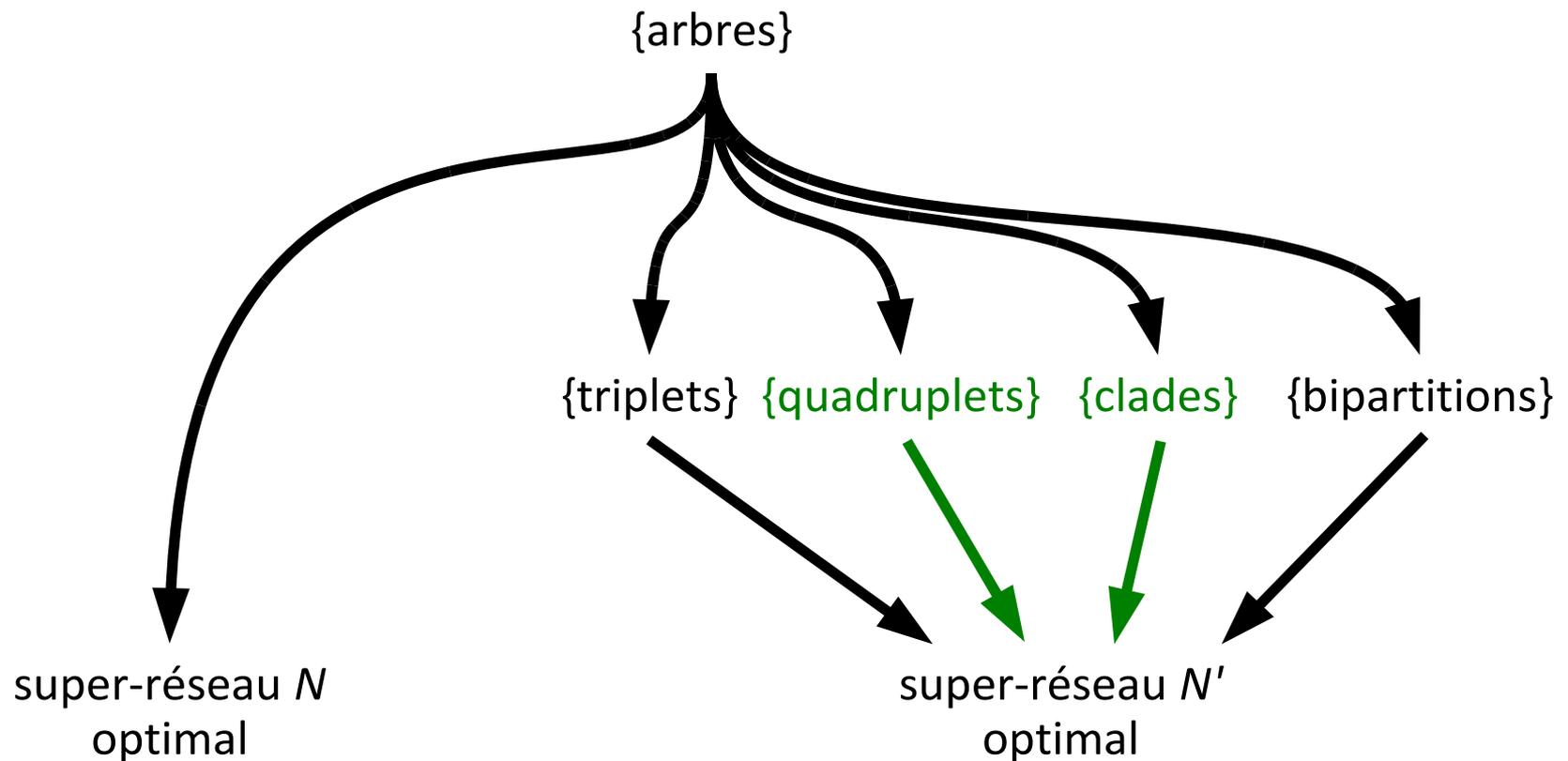


des arbres en entrée ?

Reconstruction combinatoire de réseaux phylogénétiques

Idée :

modifier le type de données à traiter

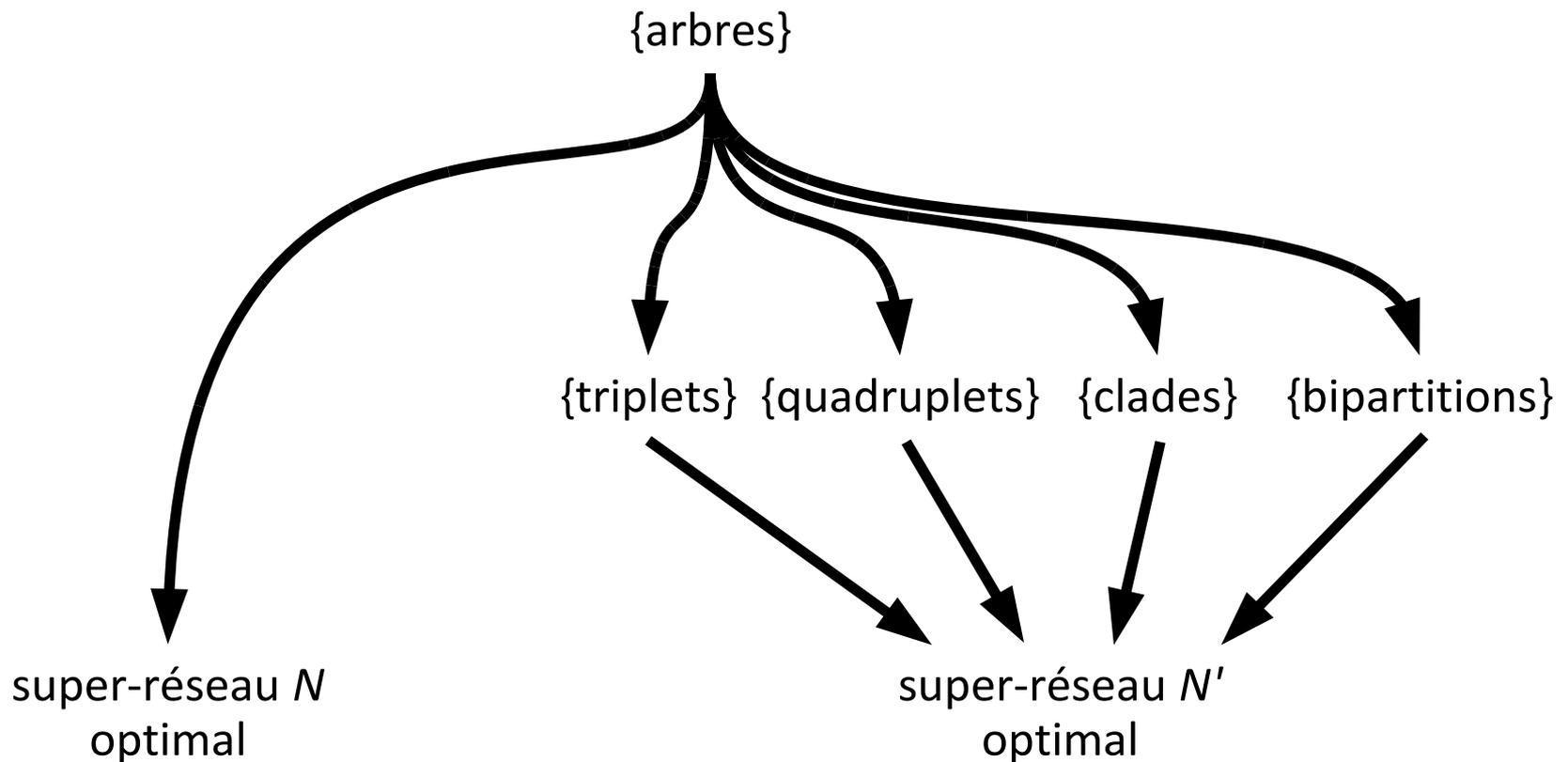


Huson, Rupp, Berry, Gambette & Paul, IMSB 2009
Gambette, Berry & Paul, 2010

Reconstruction combinatoire de réseaux phylogénétiques

Idée :

modifier le type de données à traiter

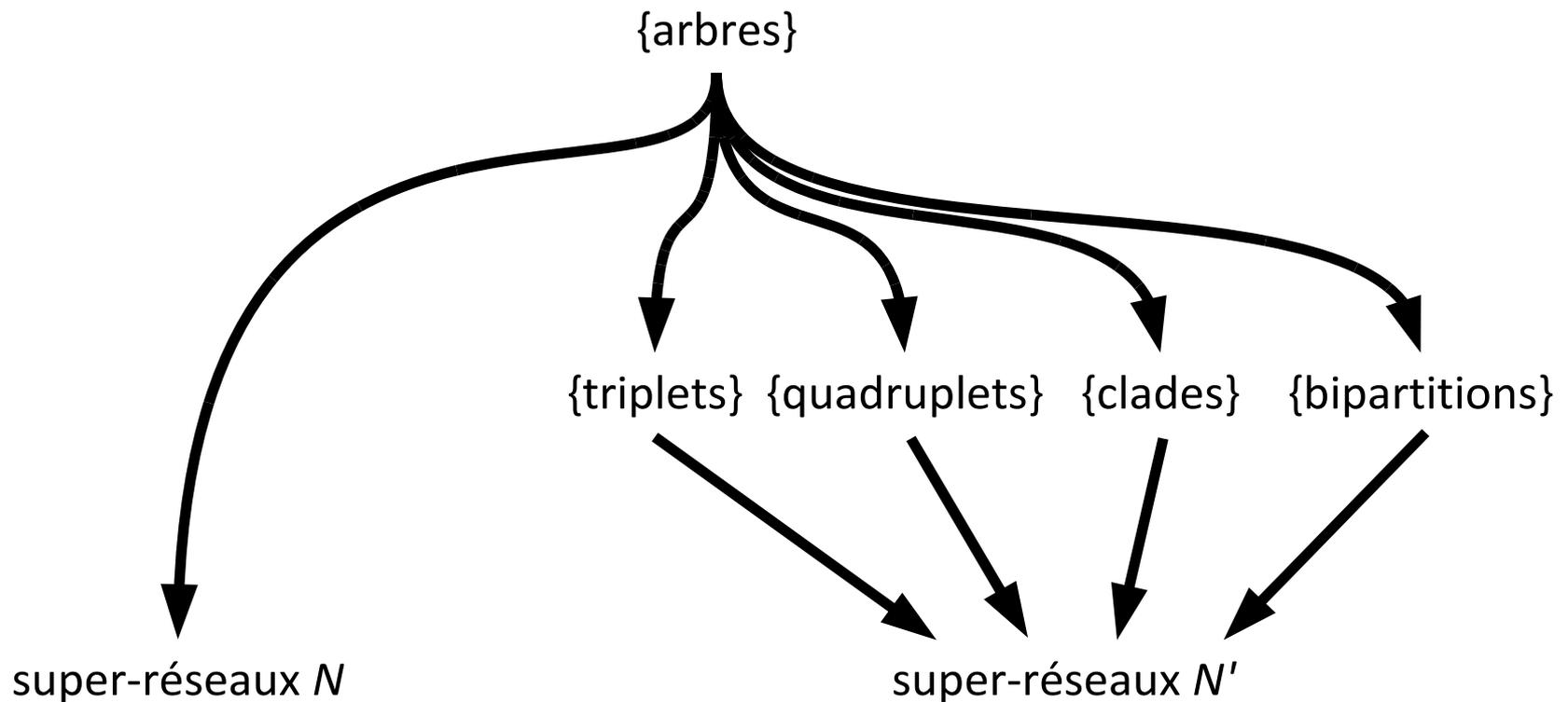


$N=N'$?

Reconstruction combinatoire de réseaux phylogénétiques

Idée :

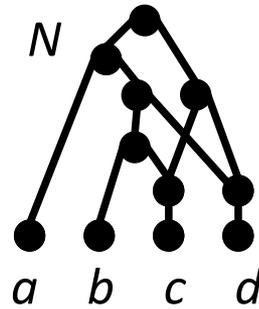
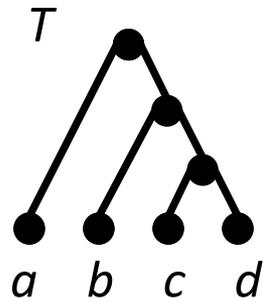
modifier le type de données à traiter



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

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** .

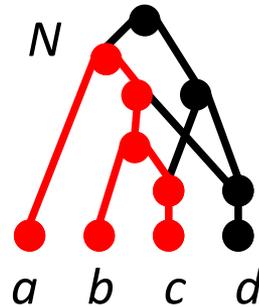
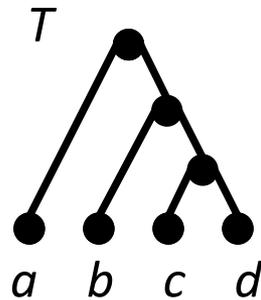


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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 .

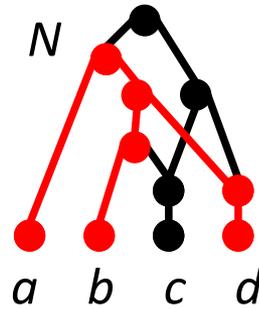
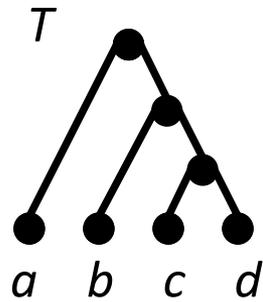


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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** .

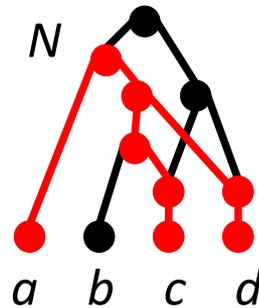
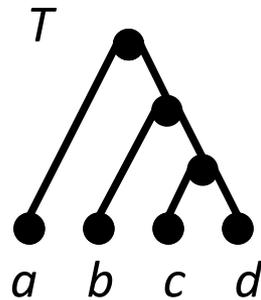


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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** .

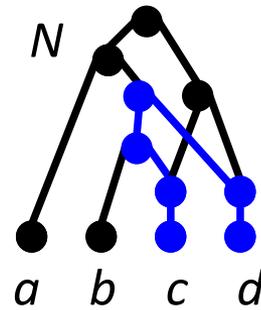
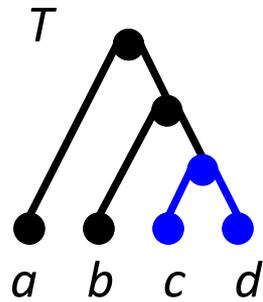


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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 .

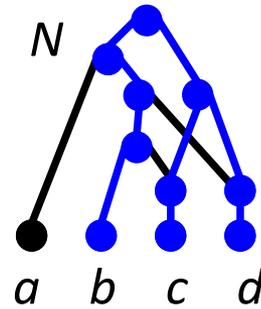
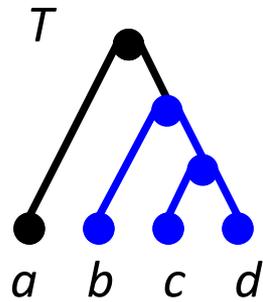


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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 .

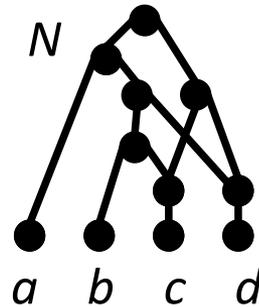
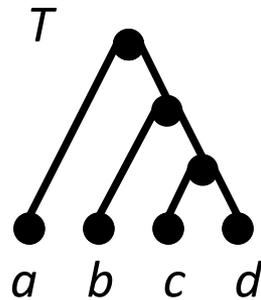


contient $\{a|bc, a|bd, a|cd, b|cd\}$
mais pas T

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

Reconstruction combinatoire de réseaux phylogénétiques

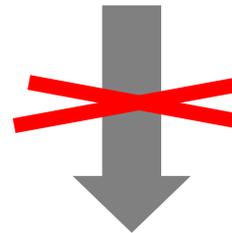
Un réseau qui contient l'ensemble de **tous les triplets ou clades** d'un arbre T ne contient **pas forcément** T .



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

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

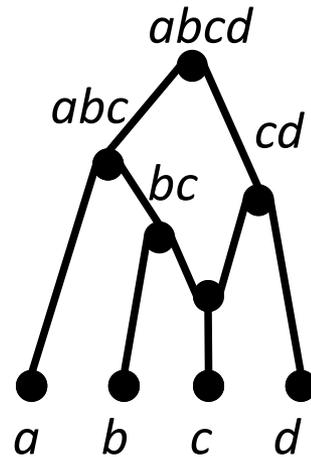
contient les clades / triplets d'un arbre T



contient T .

Clades stricts et souples

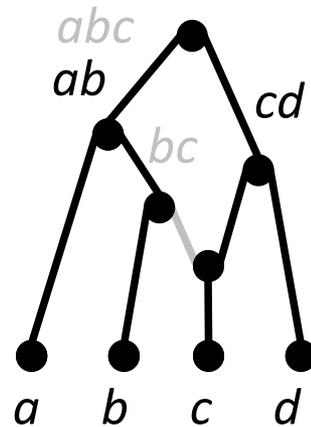
Clade “strict” : ensemble des feuilles sous un noeud du réseau



Clades stricts et souples

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

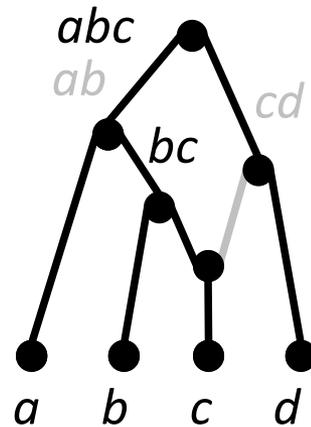
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



Clades stricts et souples

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

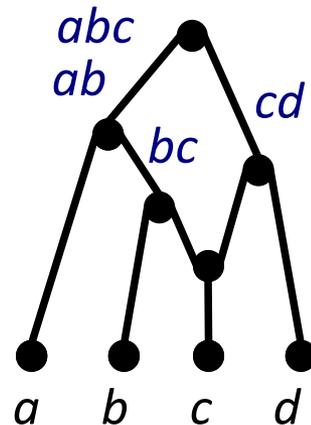
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



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**.

Test de compatibilité souple **NP-complet**.

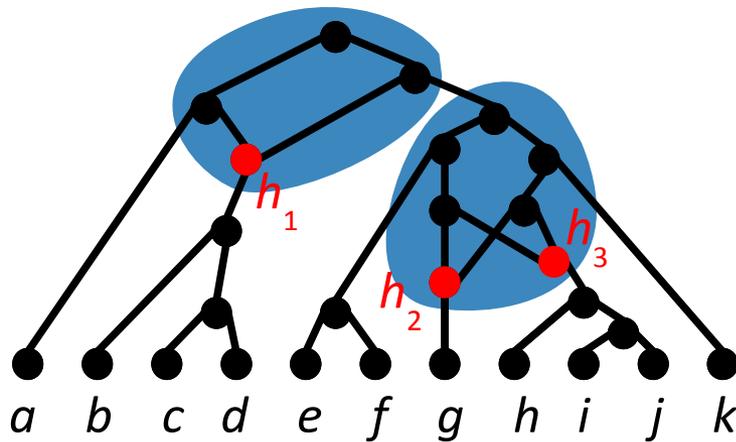
Kanj, Nakhleh, Than & Xia, TCS, 2008

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**.

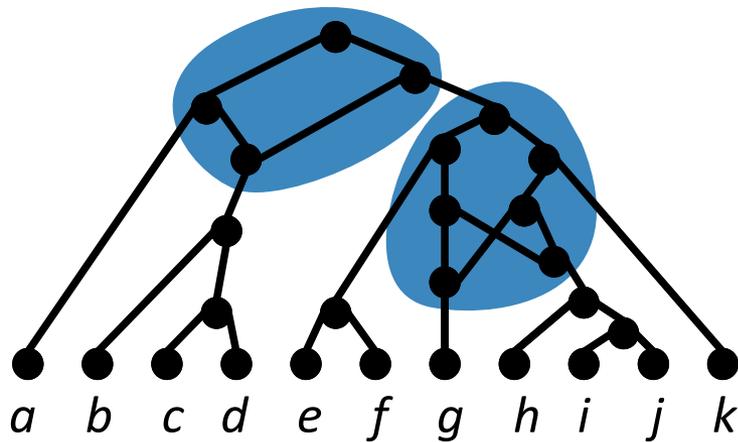


réseau de niveau 2

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

Réseaux phylogénétiques restreints

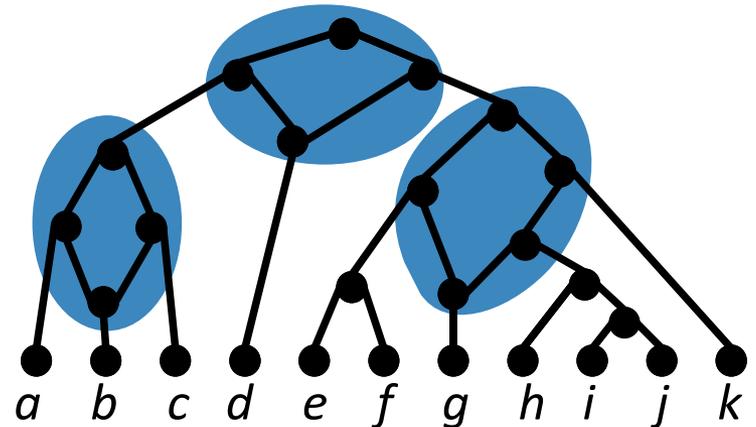
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau de niveau 2

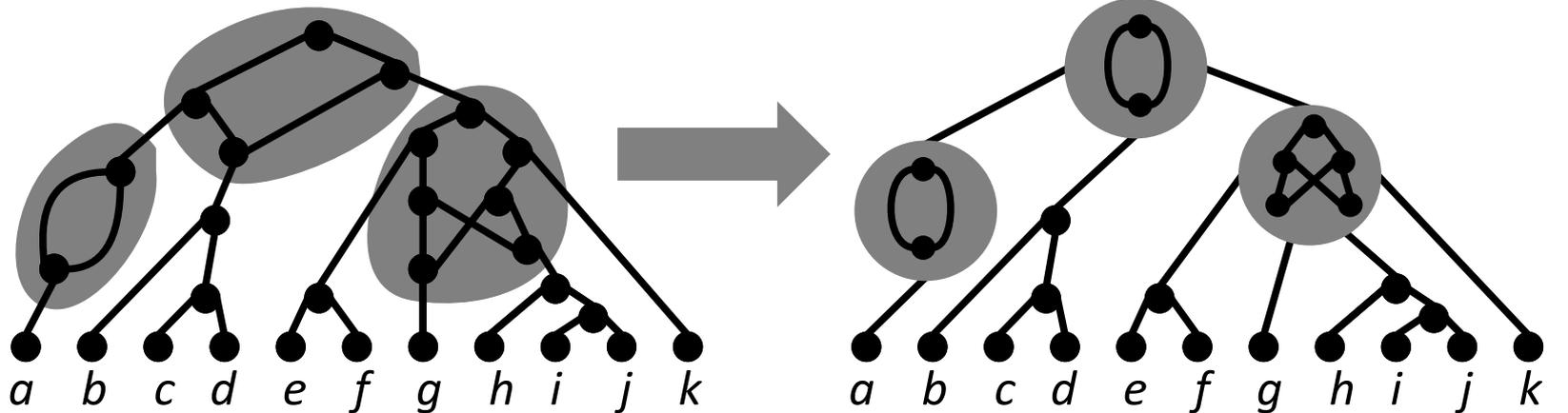
niveau =
nombre maximum d'hybridations
par partie non arborée (*blob*).

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

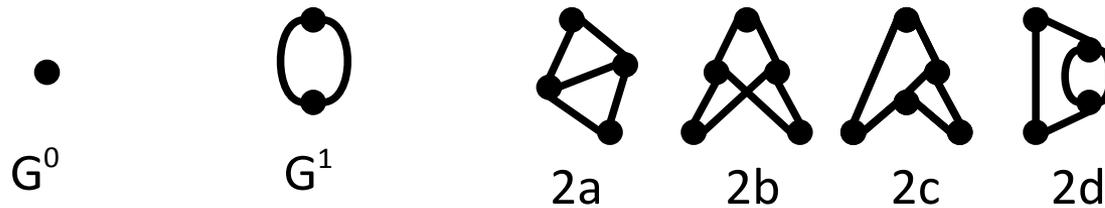
Van Iersel et al., RECOMB 2008

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

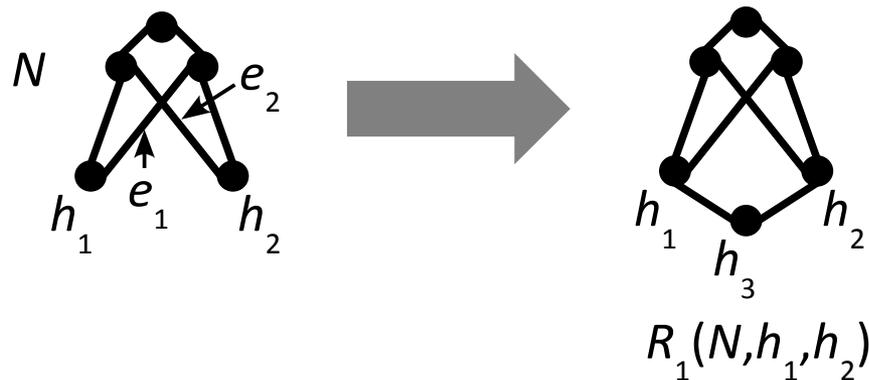
Décomposition des réseaux de niveau k

Générateur de niveau k :

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



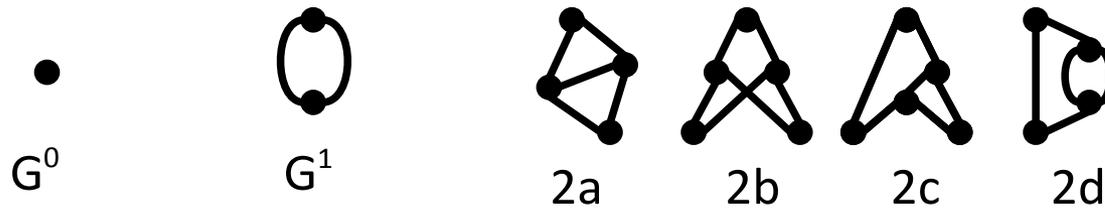
Règles de construction des générateurs de niveau $k+1$ à partir de ceux de niveau k



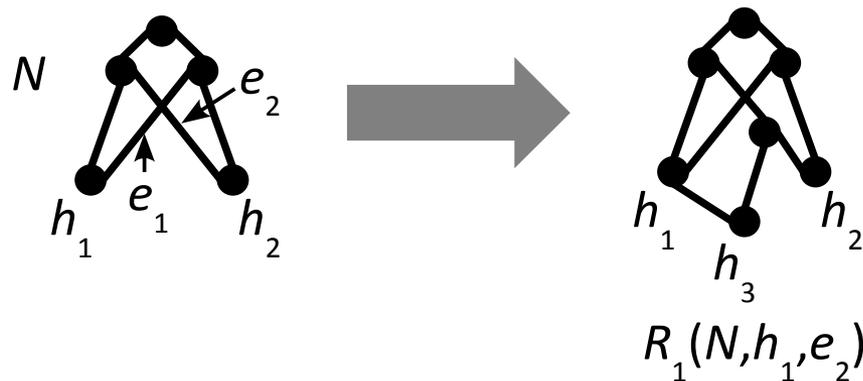
Décomposition des réseaux de niveau k

Générateur de niveau k :

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



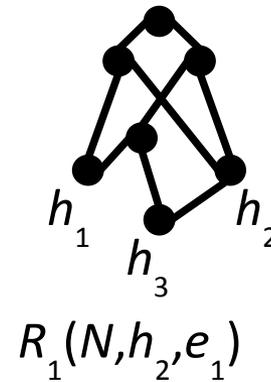
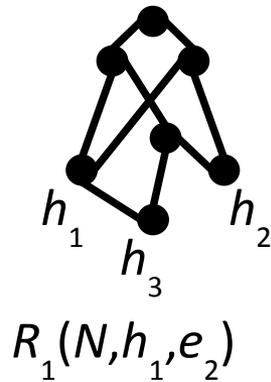
Règles de construction des générateurs de niveau $k+1$ à partir de ceux de niveau k



Construction des générateurs

Problème !

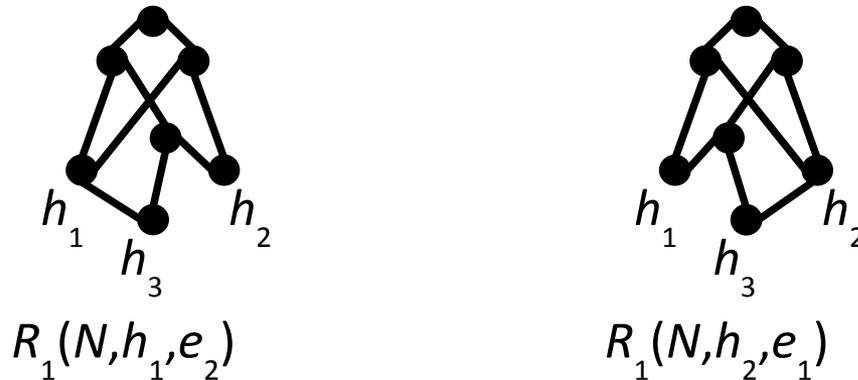
Certains des générateurs de niveau $k+1$ obtenus depuis ceux de niveau k sont isomorphes !



Construction des générateurs

Problème !

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

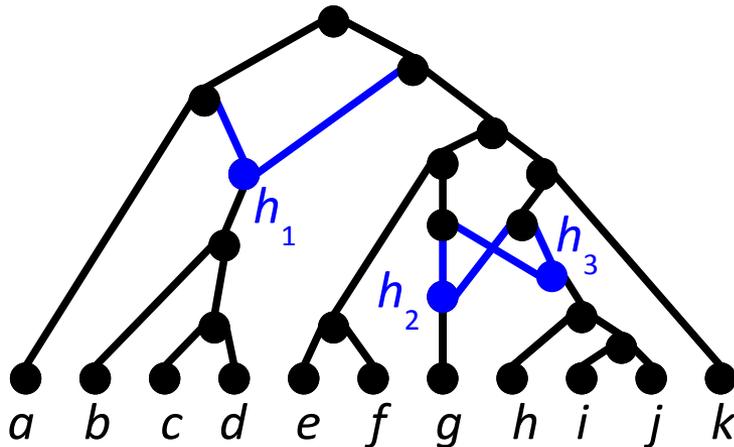


→ comptage difficile !

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

Réseaux phylogénétiques restreints

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

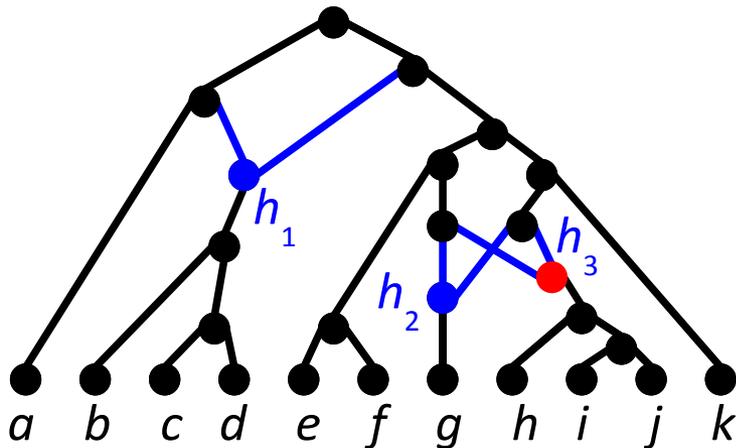


réseau à **une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à une couche de réticulation.

Réseaux phylogénétiques restreints

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

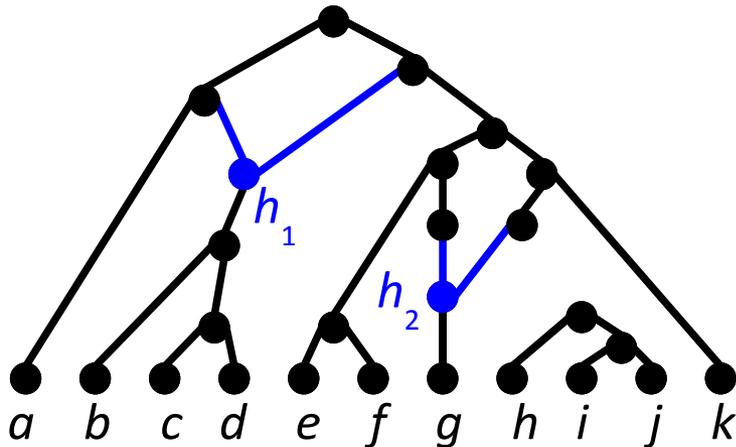


réseau à **une couche de réticulation**
("galled network") : la suppression
d'un noeud de réticulation
déconnecte le réseau.

réseau à une couche de
réticulation.

Réseaux phylogénétiques restreints

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

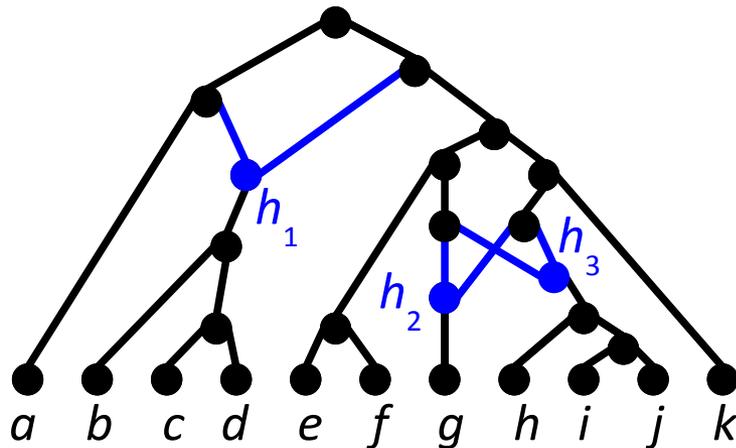


réseau à **une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à une couche de réticulation.

Réseaux phylogénétiques restreints

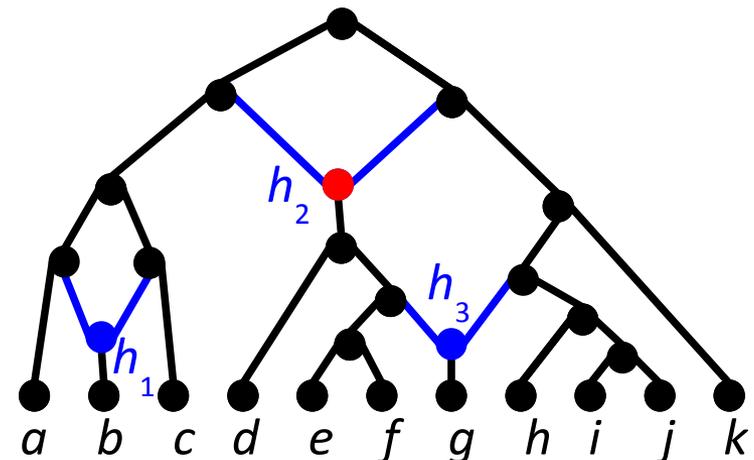
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau à une couche de réticulation.

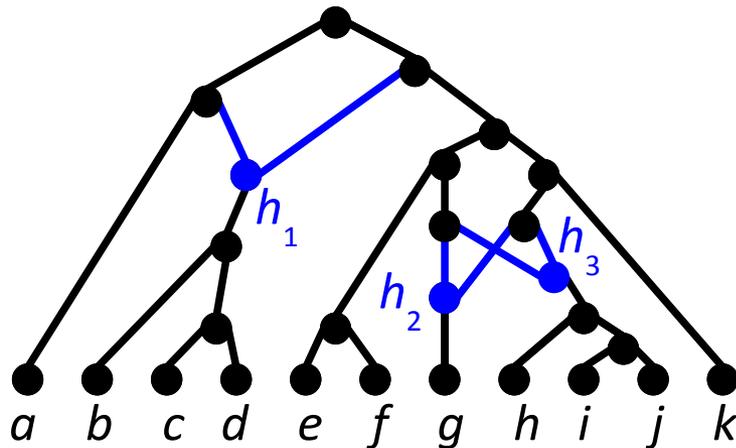
réseau à **une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à deux couches de réticulation.



Réseaux phylogénétiques restreints

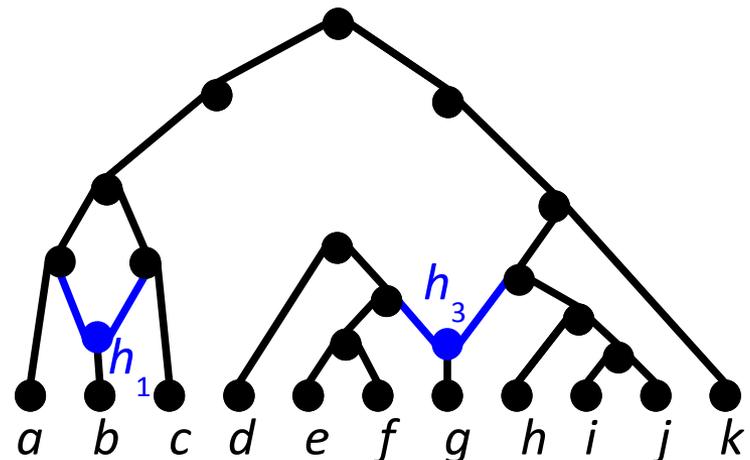
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



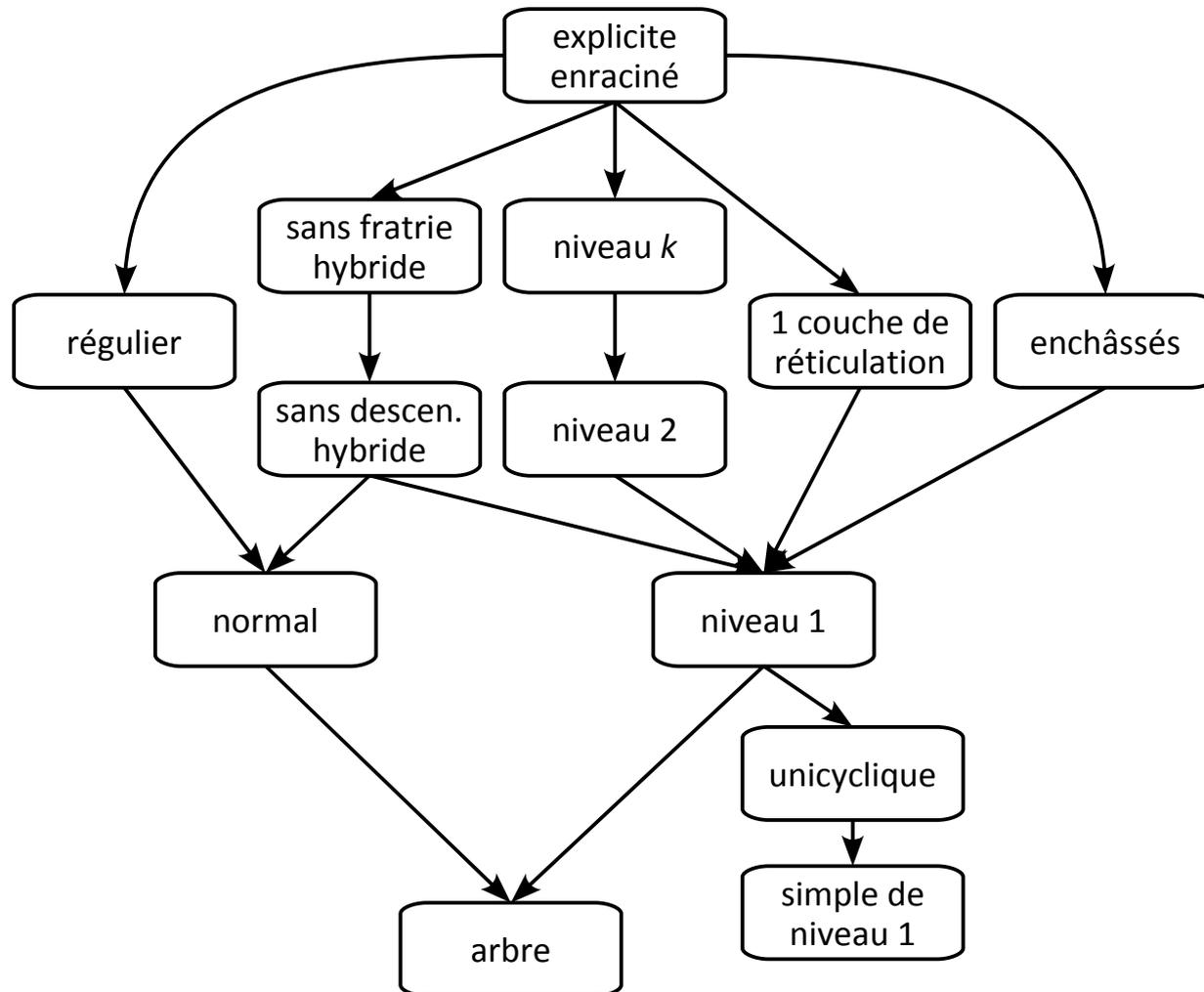
réseau à une couche de réticulation.

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

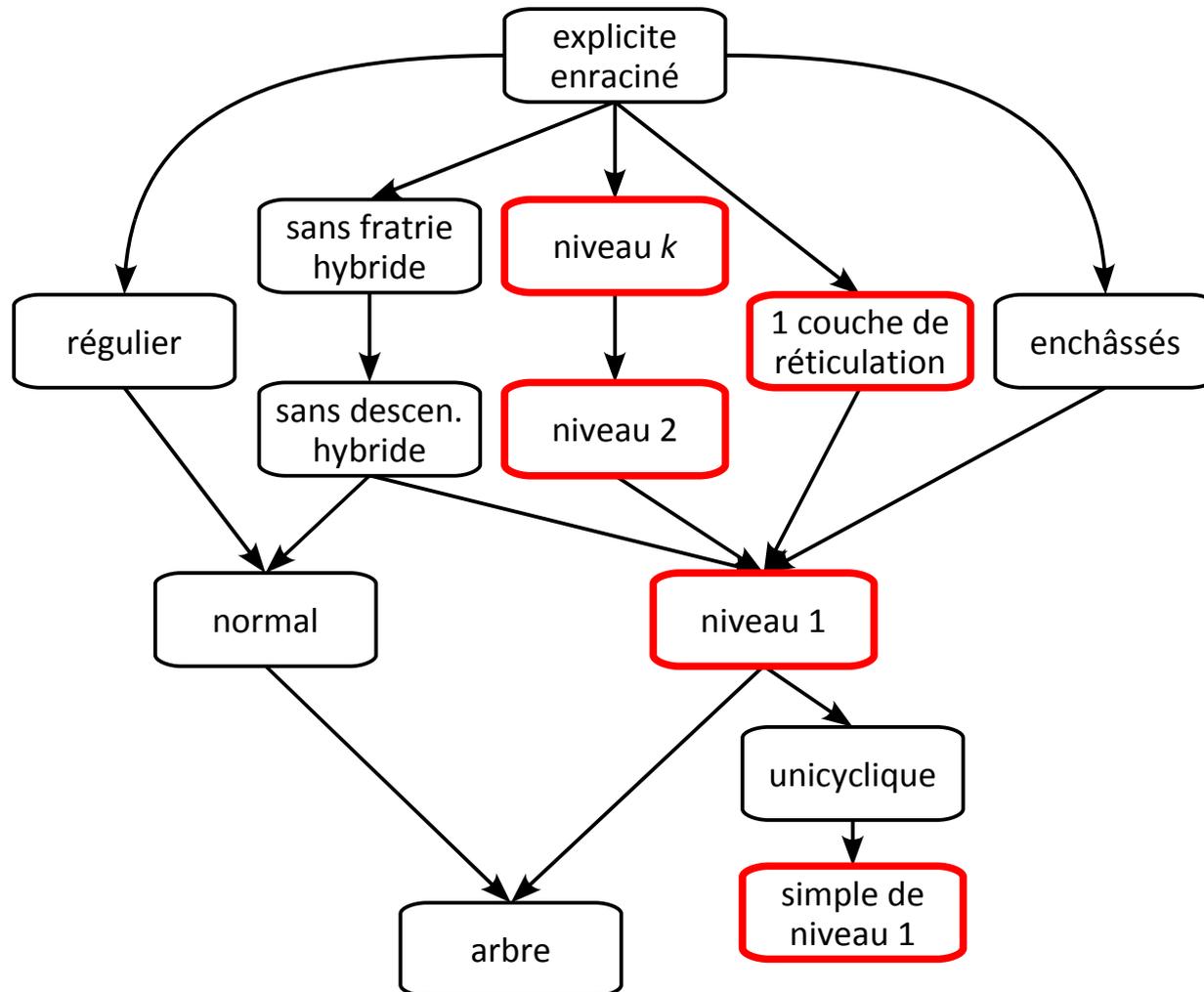
réseau à deux couches de réticulation.



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
- 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

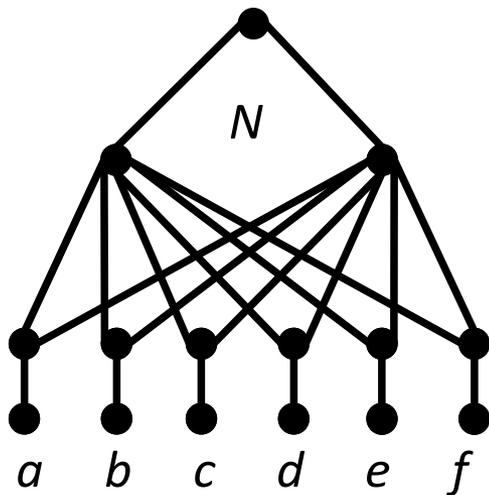


meilleurs résultats mais plus lente pour niveau > 2 .
pour k fixé, certains ensembles de clades contenus
dans aucun réseau de niveau k .

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 .

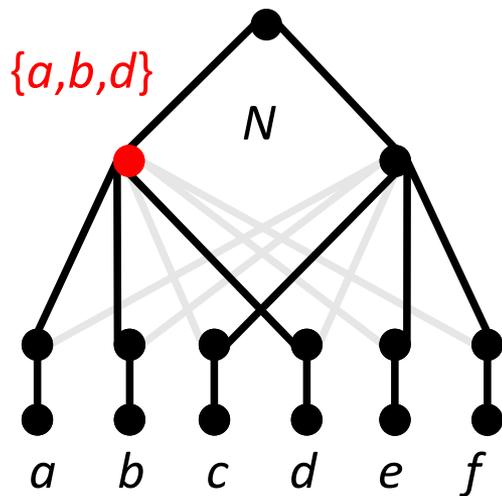


Le réseau à une couche de réticulation N est compatible avec tout clade souple sur $\{a, b, c, d, e, f\}$.

Clades et réseaux à une couche de réticulation

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

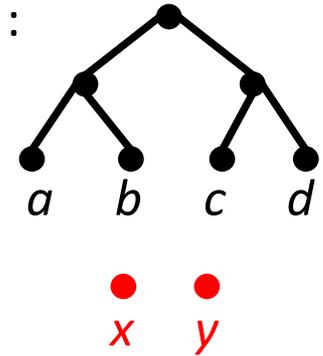
Pour tout ensemble C de clades, il existe un **réseau à une couche de réticulation compatible** avec C .



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

Une approche en deux étapes

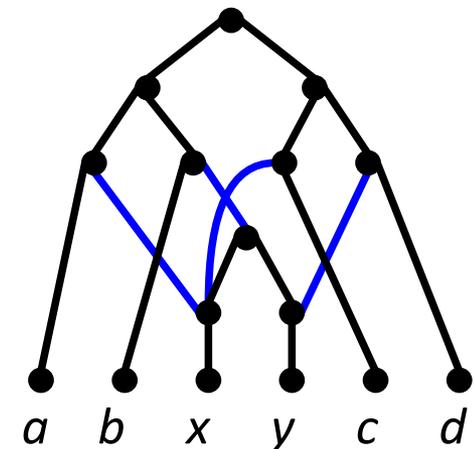
- 1- Trouver un **ensemble minimum de conflits** parmi les clades :
- partie sans conflits ➡ arbre,
 - taxons impliqués dans des **conflits** ➡ sous les réticulations.



MAXIMUM COMPATIBLE SUBSET

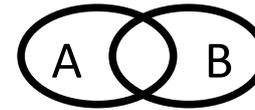
- 2- Attacher à l'arbre les taxons impliqués dans des conflits avec un **nombre minimal d'arcs** :

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

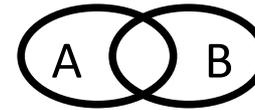
NP-complet sur un graphe connexe, sans taxons “jumeaux”

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

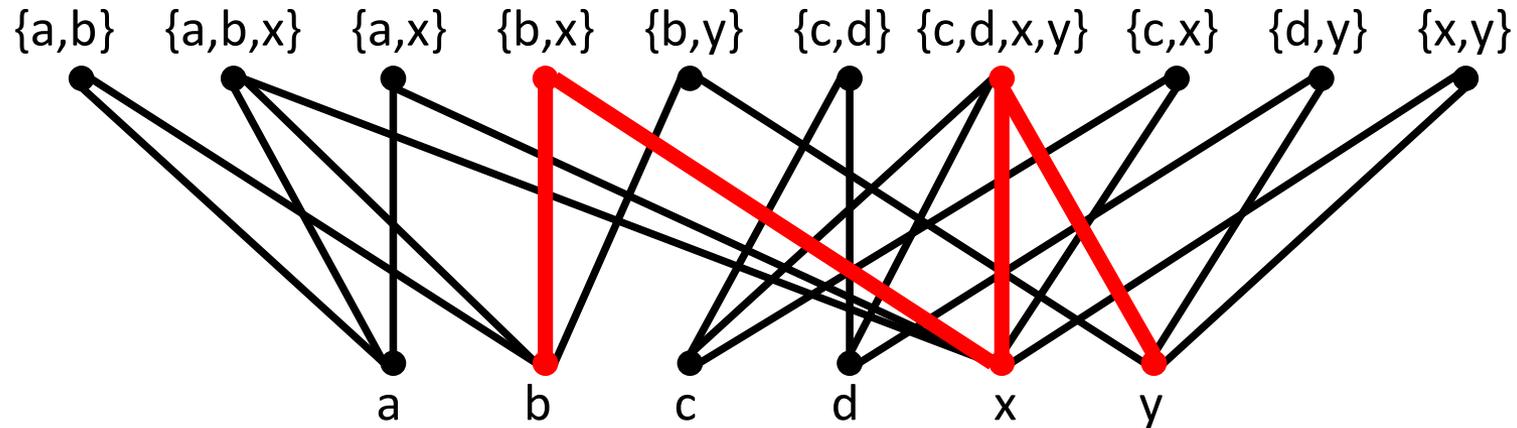
Conflit : clades ni inclus ni disjoints



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

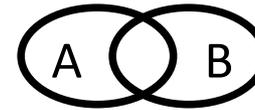
Exemple : $\{\{a,b\},\{a,b,x\},\{a,x\},\{b,x\},\{b,y\},\{c,d\},\{c,d,x,y\},\{c,x\},\{d,y\},\{x,y\}\}$



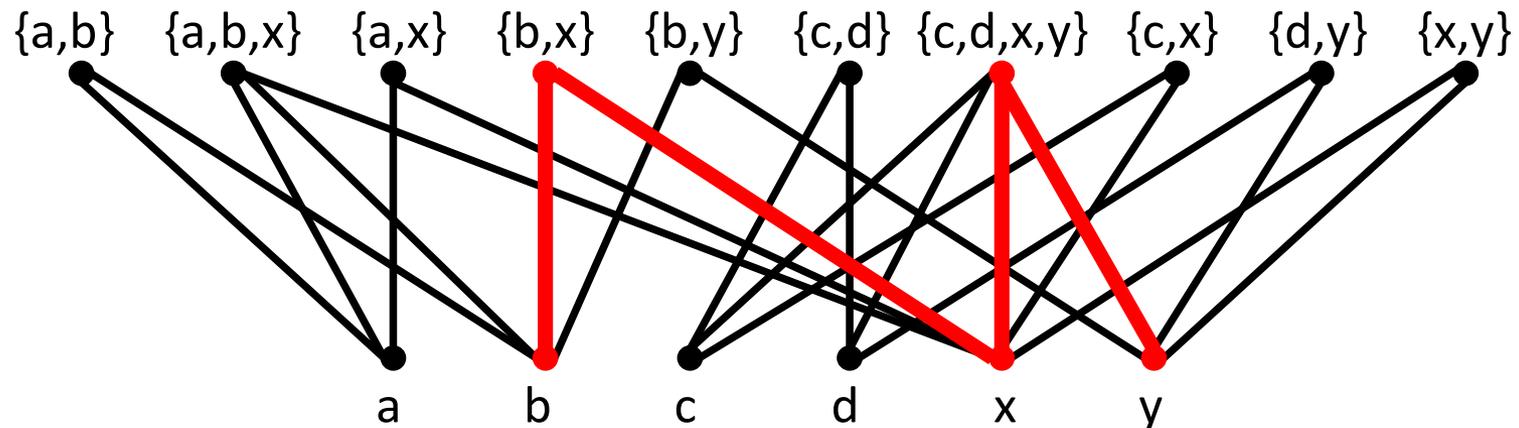
conflit = graphe "M"

L'ensemble minimum de conflits

Conflit : clades ni inclus ni disjoints



Graphe des caractères :



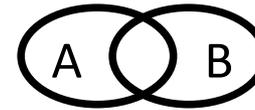
Trouver l'ensemble minimum de conflits



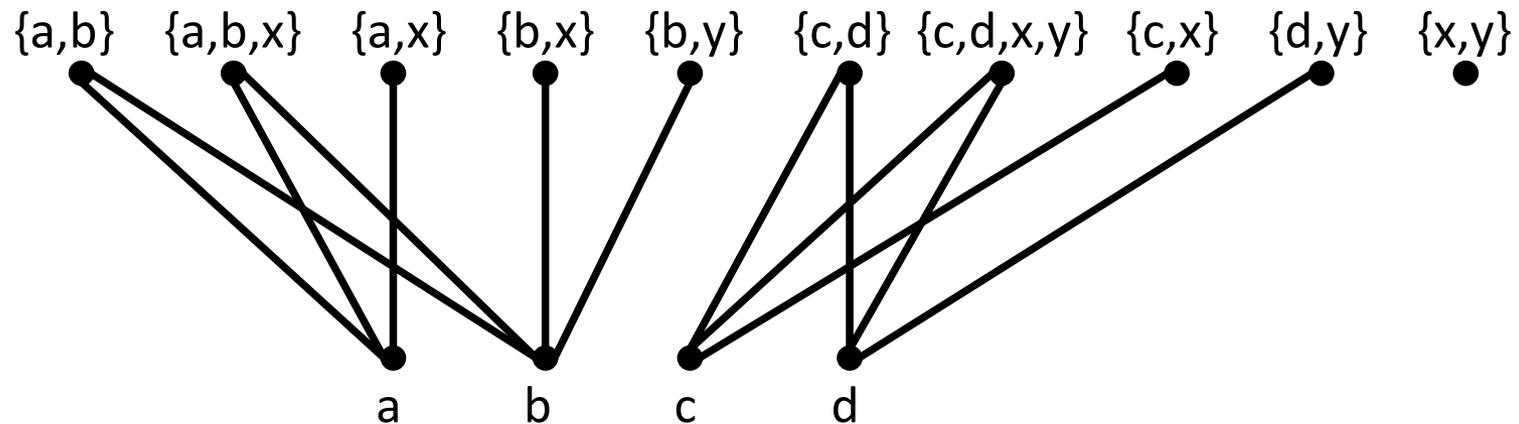
Supprimer le nombre minimum t de sommets-taxons tels que le graphe des caractères est un graphe "sans M"

L'ensemble minimum de conflits

Conflit : clades ni inclus ni disjoints



Graphe des caractères :



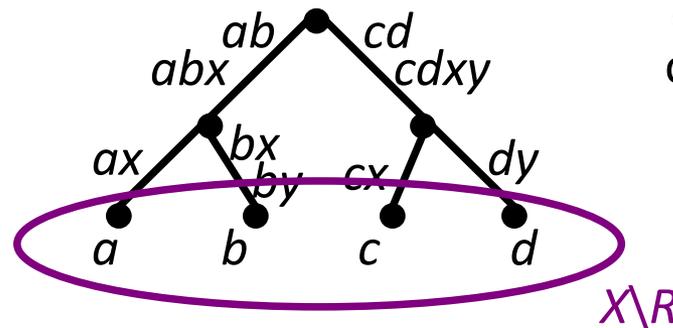
Supprimer le nombre minimum t de sommets-taxons tels que le graphe des caractères est un graphe “sans M” :

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

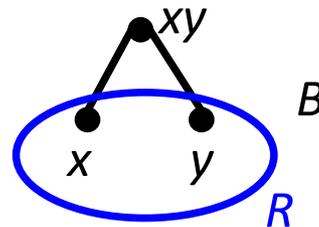
L'attachement minimum

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 .

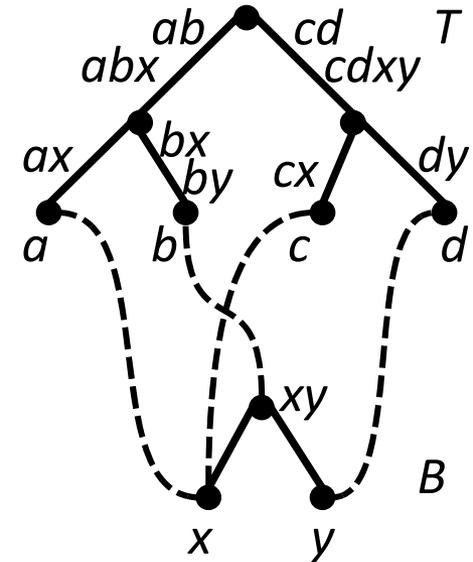
Problème :

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

L'attachement minimum

Problème :

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



NP-complet

W[2]-dur, paramétré par le nombre de liens à ajouter

réduction depuis SetCover

Algorithmes :

- Séparation et évaluation
- Programme linéaire en nombres entiers

implémenté dans Dendroscope 2

Reconstruction depuis les triplets

{arbres}



{triplets}



N'
réseau
de niveau k

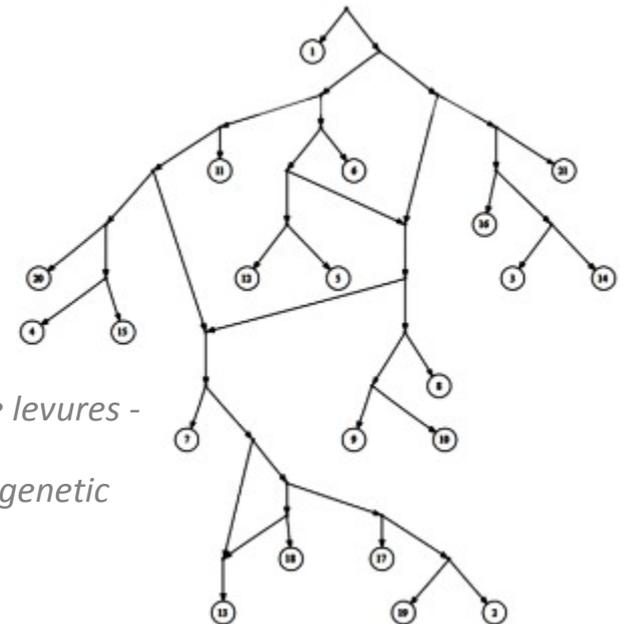
Méthodes exactes rapides pour reconstruire un **réseau de niveau 1 et 2** (s'il en existe un) à partir d'un ensemble dense de **triplets**

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 .

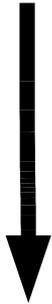
Programme Simplistic



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

Reconstruction depuis les quadruplets

{arbres}



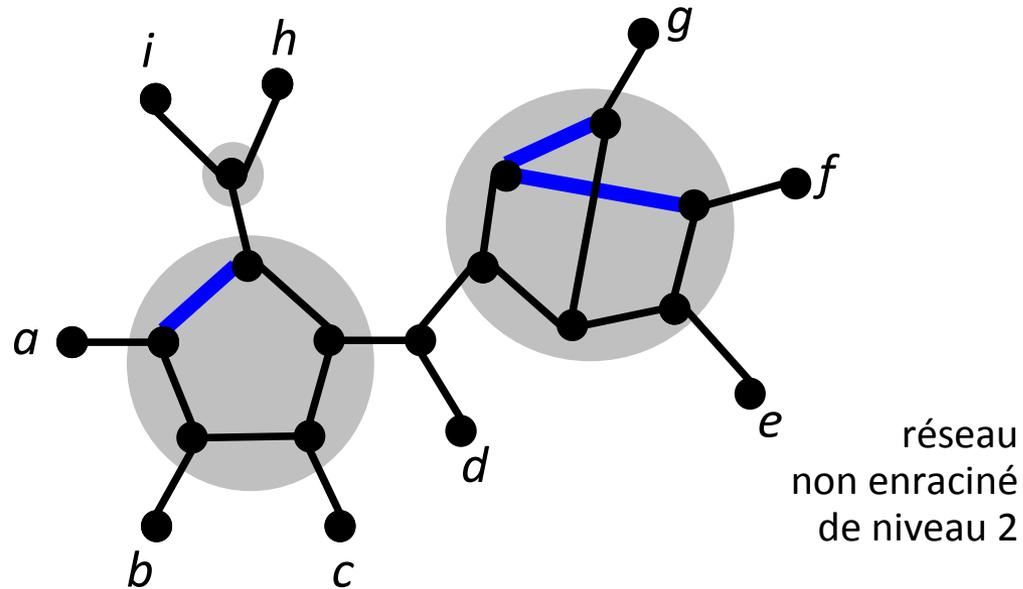
{quadruplets}



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 d'un réseau de niveau k
Niveau **invariant** selon l'orientation

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>Réduction depuis Betweenness</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>

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.

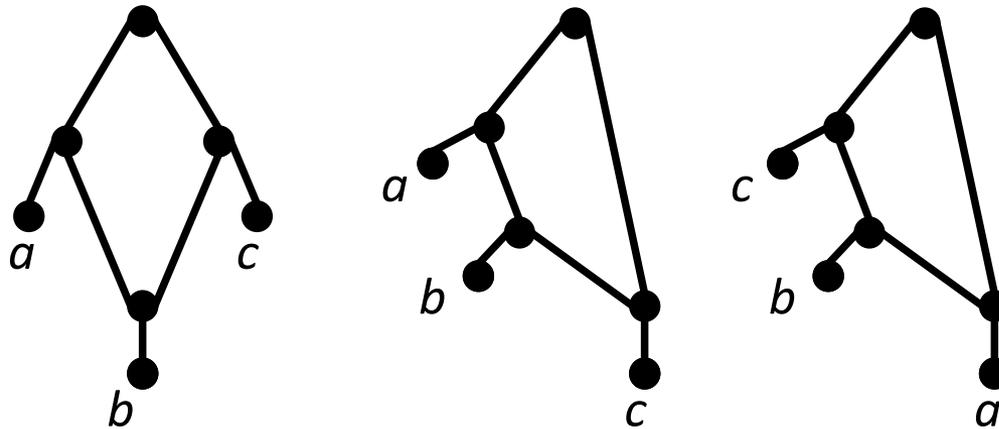
Nombre exponentiel de générateurs de niveau k :

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

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.

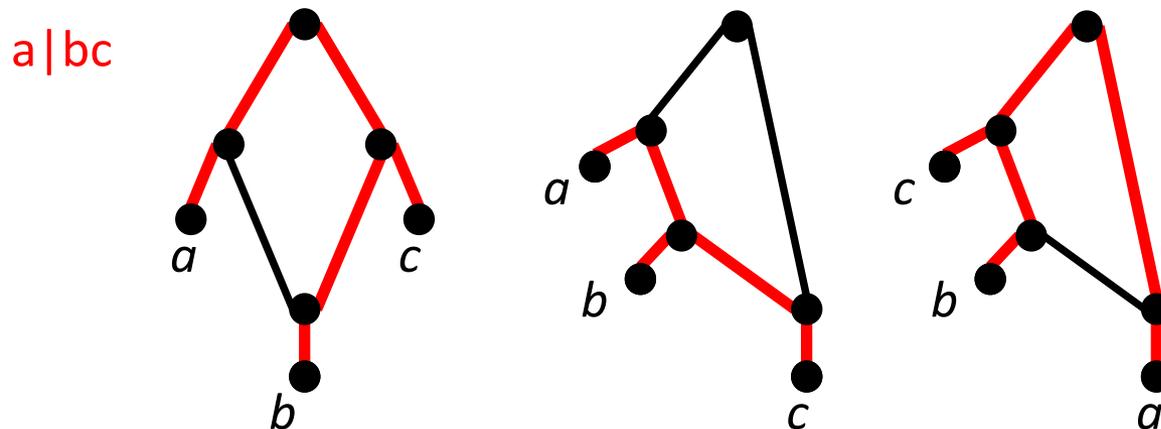


Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 sommets)

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.

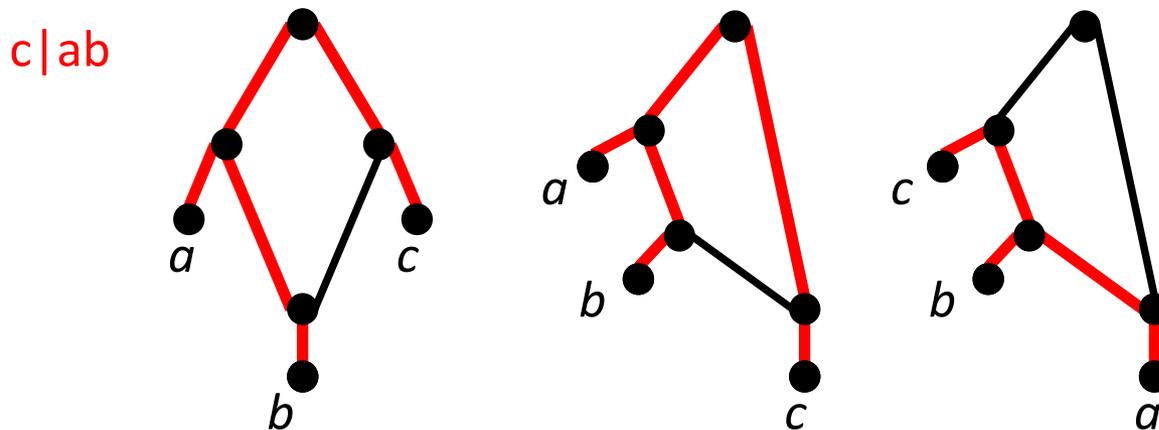


Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 sommets)

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.

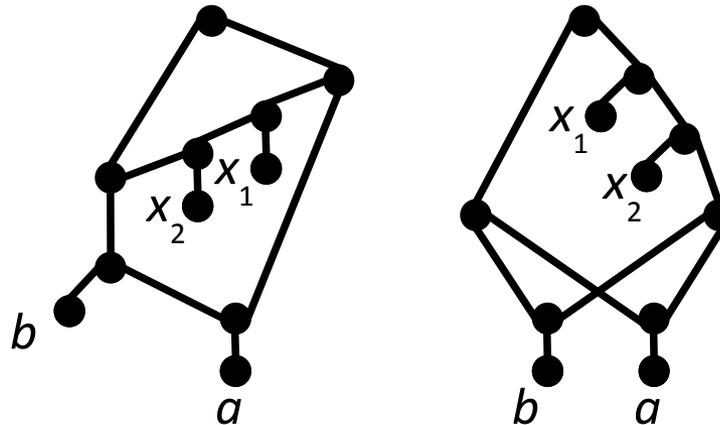


Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 sommets)

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.

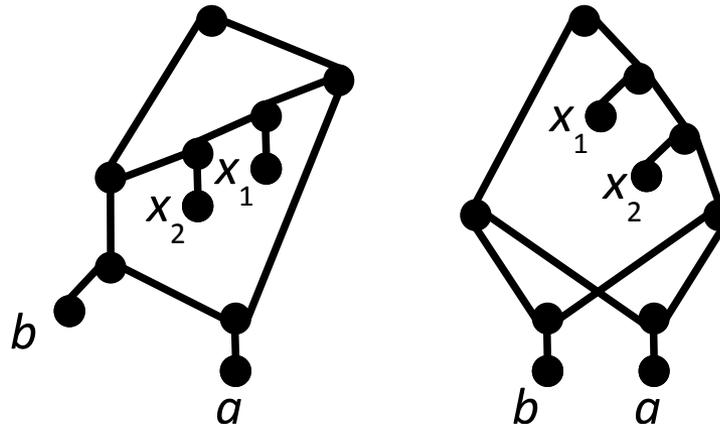


2 réseaux de niveau 2 avec le même ensemble de triplets

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

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 :

$O^*(3^t)$, 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 ou triplets

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

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

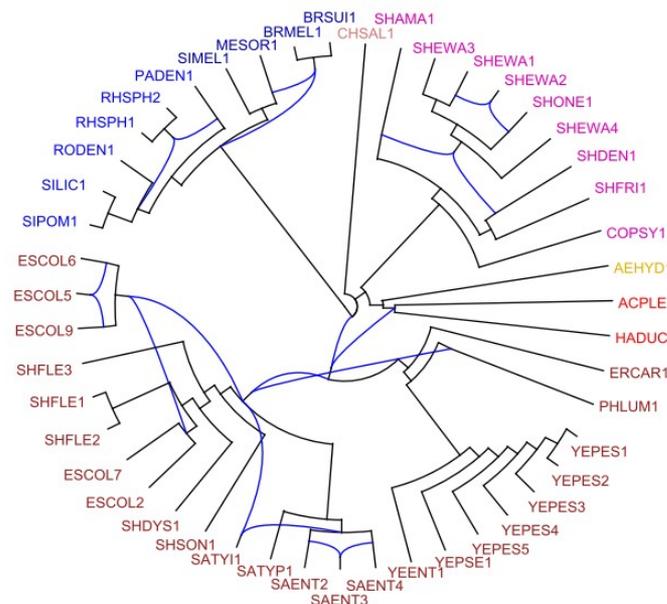
1 Oceanospirillales

6 Rhodobacterales

4 Rhizobiales

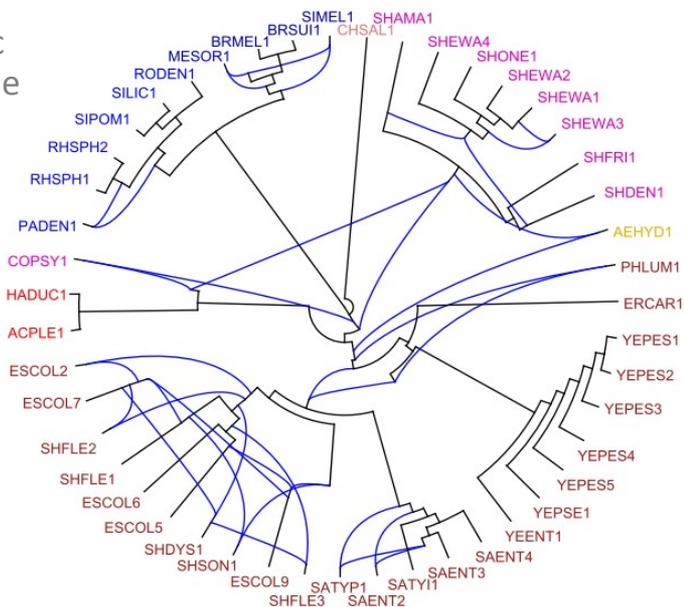


Réseaux contenant les triplets, clades souples,
présents dans au moins 20% des arbres

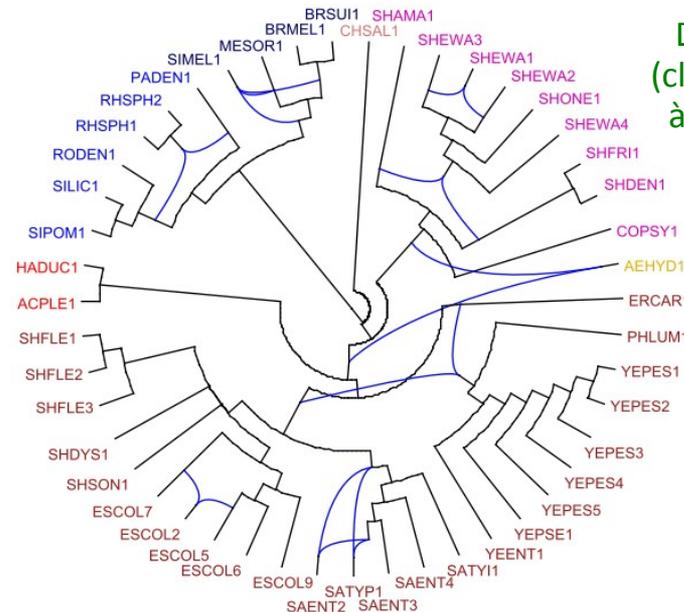


Lev1athan
(heuristique
triplets,
niveau 1)
24 sec.

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 Enterobacterales

2 Pasteurellales

1 Aeromonadales

9 Alteromonadales

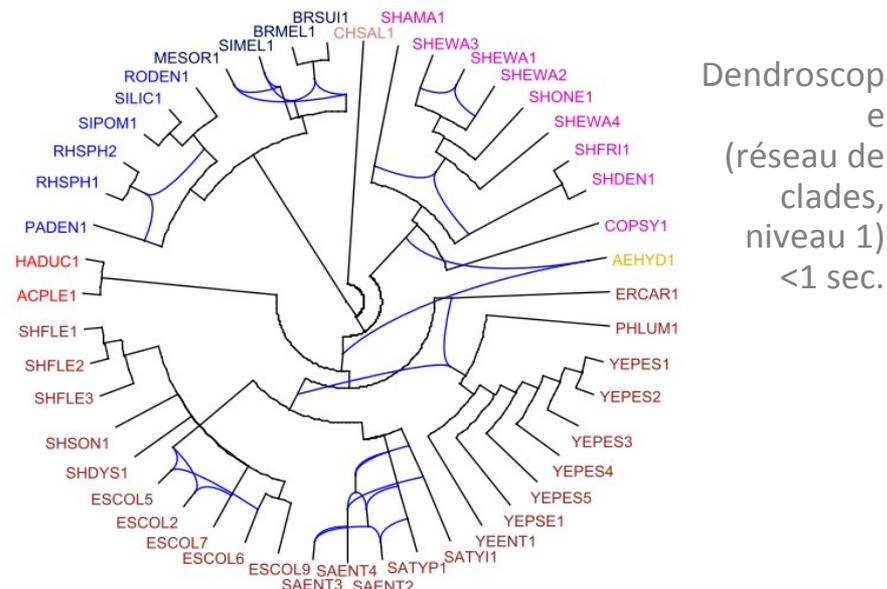
1 Oceanospirillales

6 Rhodobacterales

4 Rhizobiales

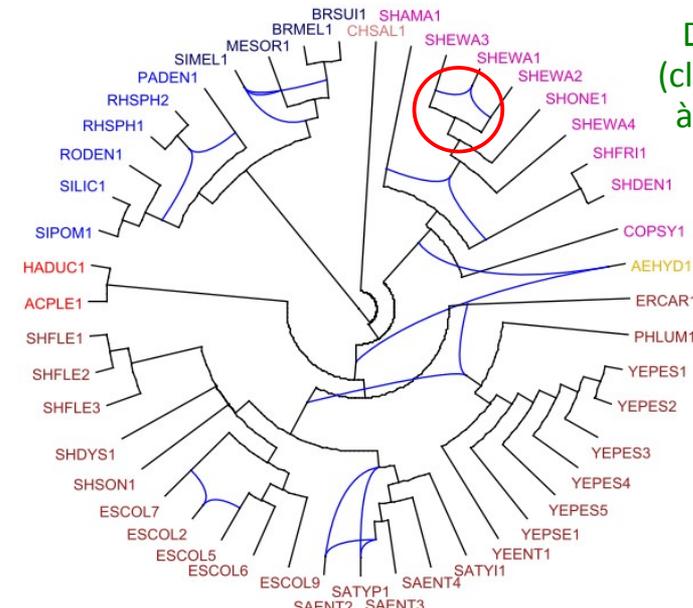
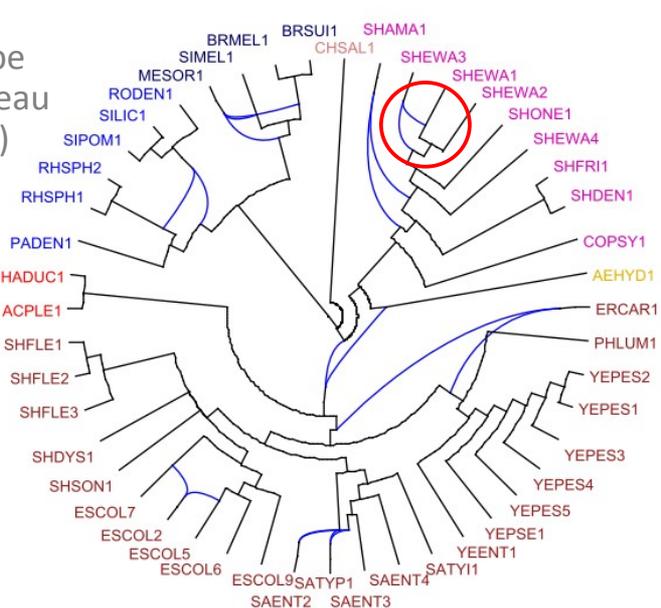


Réseaux contenant les clades souples
présents dans au moins 20% des arbres



Dendroscope
(réseau de
clades,
niveau 1)
<1 sec.

Dendroscope
(clades, réseau
de niveau 7)
2 sec.

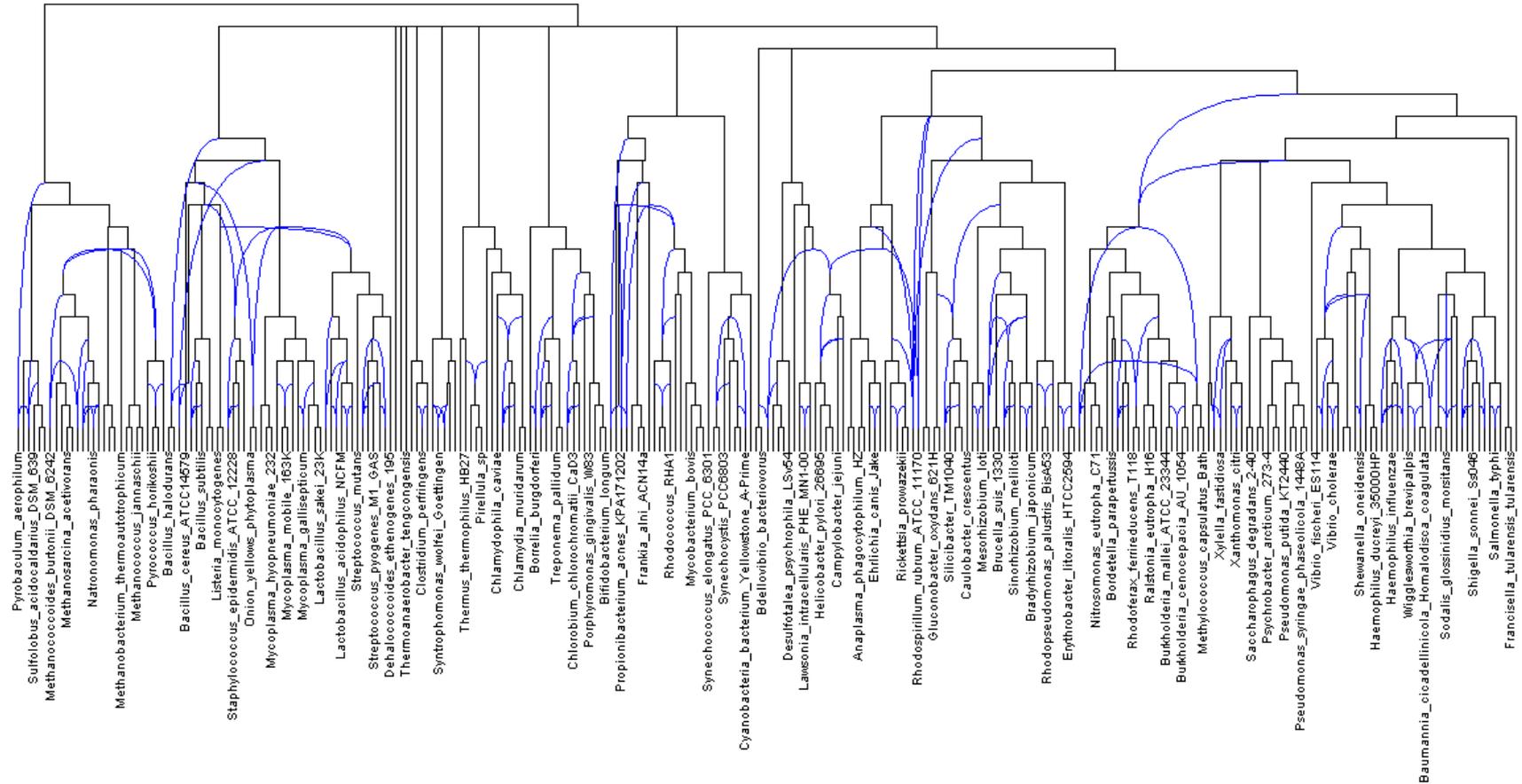


Dendroscope
(clades, réseau
à 1 couche de
réticulation)
<1 sec.

Exemples de résultats

9 arbres sur 279 espèces de procaryotes

Auch et al., 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

Résultats obtenus pendant la thèse



Bioinformatique

Reconstruction de réseaux à une couche de réticulation à partir de clades
 Huson, Rupp, Berry, Gambette & Paul, *BioInf, ISMB 2009*

Structure des réseaux de niveau borné
 Gambette, Berry & Paul, *CPM 2009*

Ambiguïté des réseaux de niveau 1 par rapport à leur ensemble de clades et triplets
Gambette & Huber, soumis à JMB

Quadruplets et réseaux phylogénétiques non enracinés de niveau borné
Gambette, Berry & Paul, en cours de soumission à JBCB

journaux
internationaux

conférences
internationales avec
actes

colloques jeunes
chercheurs

manuscrits

Traitement des langues et des textes

Visualisation de textes par leur nuage arboré, robustesse selon la formule de cooccurrence
 Gambette & Véronis, *IFCS 2009*

Utilisation des nuages arborés dans une analyse contrastée de Cinna et Othon de Corneille
Amstutz & Gambette, *JADT 2010*

Automatisation du calcul de la densité des idées
 Lee, Gambette, Maillé & Thuillier, *RECITAL 2010*

Dégradation de la densité des idées chez des patients atteints de la maladie d'Alzheimer
Lee, Gambette, Barkat-Defradas, *CEDIL 2010*

Algorithmique des graphes

Codage efficace des voisinages dans les graphes de permutation et graphes d'intervalles, parcours en largeur en $O(n)$ des graphes de permutation

Crespelle & Gambette, *IWOCA 2009*

Parcours en largeur en $O(n)$ des graphes trapézoïdaux avec un ordre de priorité quelconque
Crespelle & Gambette, *IPL, 2010*

Contributions logicielles

Bioinformatique

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://atgc.lirmm.fr/phylNet>

Traitement des langues et des textes

Implémentation en Python et Delphi de TreeCloud, logiciel libre de construction de nuages arborés



TreeCloud

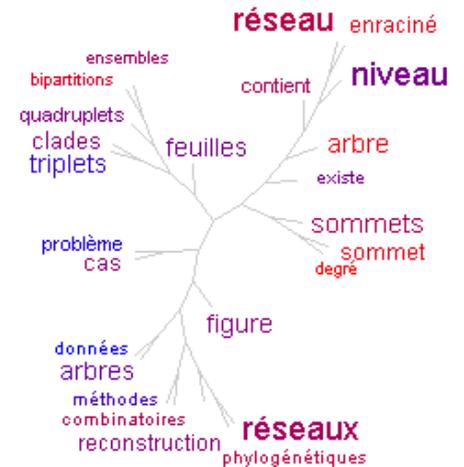
<http://www.treecloud.org>

Implémentation en Python et Delphi de Densidées, logiciel libre de calcul de la densité des idées



Densidées

<http://code.google.com/p/densidees>



Nuage arboré des 25 mots les plus fréquents de cette thèse (rouge au début du manuscrit, bleu à la fin) construit par TreeCloud et SplitsTree

Plan

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

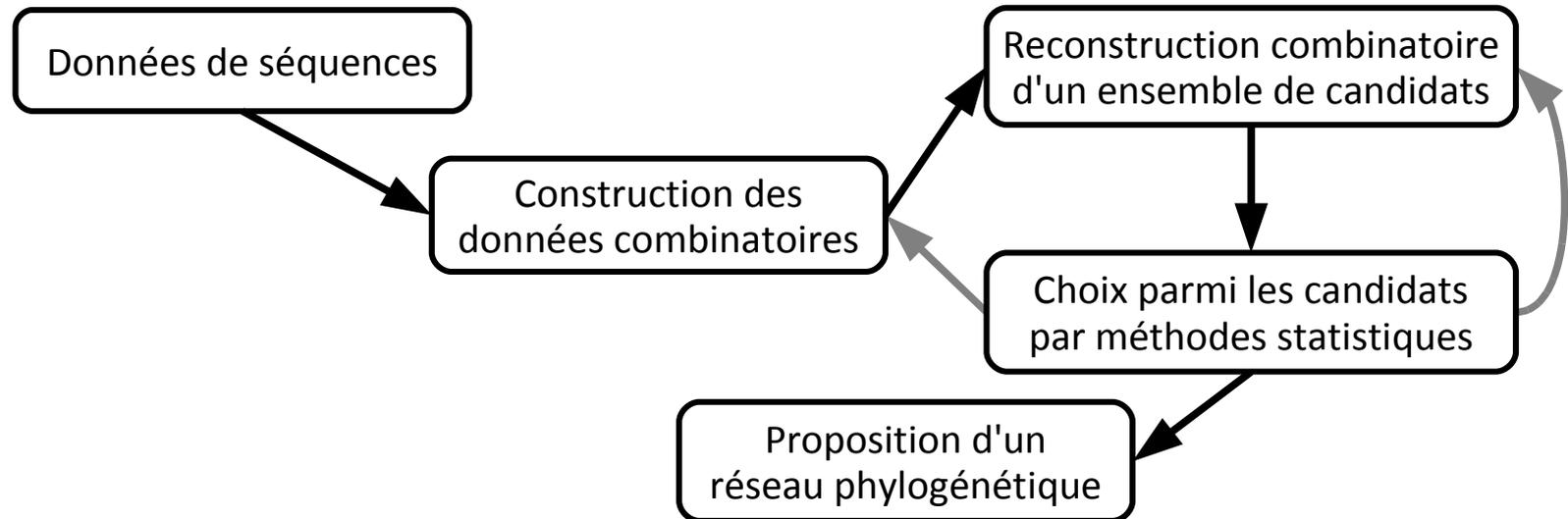
Perspectives de recherche

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



Autres applications des réseaux phylogénétiques :

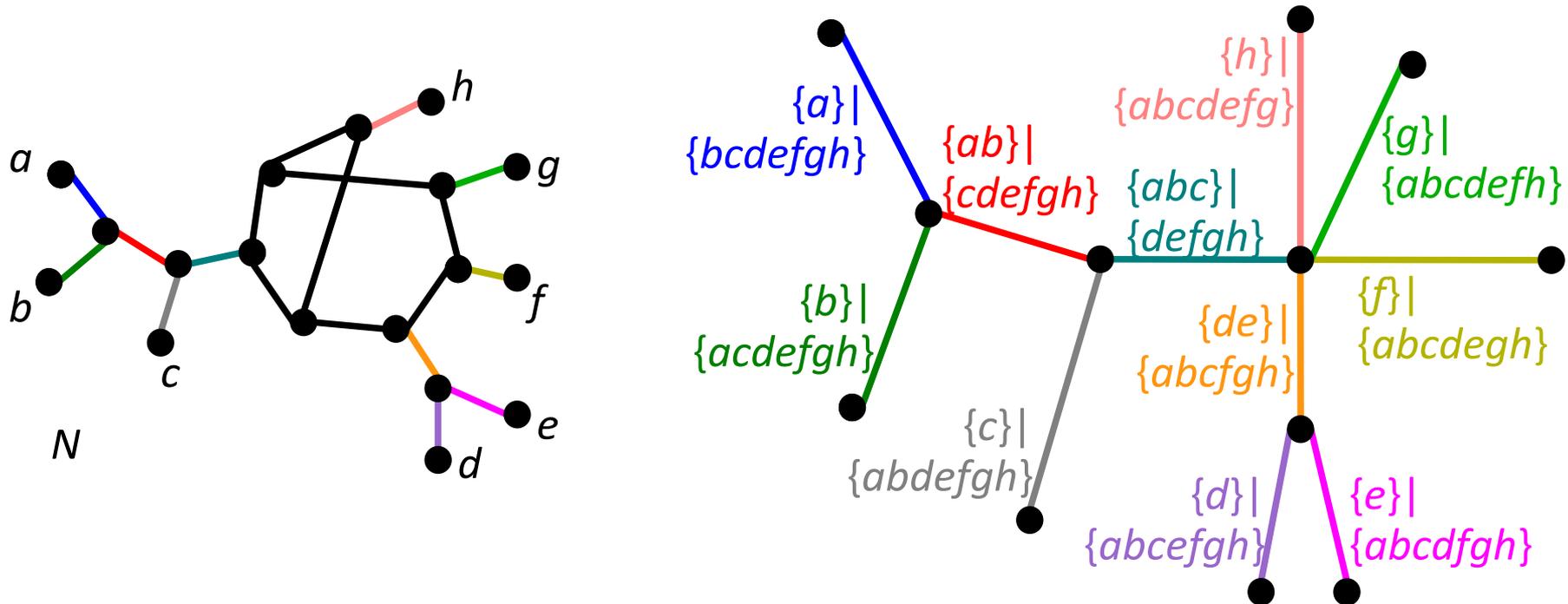
- visualiser la polysémie dans les nuages arborés

SN-bipartitions

$Q(N)$, 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 .



Logiciel Densidées

Densidées

Ce programme calcule la densité des idées d'un texte en français étiqueté par TreeTagger.

mode oral

Version 1.3 - 30/08/2010
Philippe Gambette, Hye-Ran Lee
Logiciel inspiré de CASPR-CPIDR

c'	PRO:DEM	ce	002	PRO:PER	W	il
est	VER:pres	être	002	PRO:PER	W	y
le	DET:ART	le	200	VER:pres	W	pa
jour	NOM	jour	200	NUM	W	P 52
de	PRP	de	002	NOM	W	ans
mon	DET:POS	mon	200	ADV	W	P bientôt
mariage	NOM	mariage	000	PUN	(
(PUN	(200	ADV	W	donc
voilà	ADV	voilà	600	ADV	W	voilà
)	PUN)	000	PUN)	
il	PRO:PER	il	200	ADV	W	P bien
y	PRO:PER	y	701	KON	W	que
a	VER:pres	avoir	002	PRO:DEM	W	ça
52	NUM	@card@	402	VER:aux	W	a
ans	NOM	an	200	VER:pper	W	P été
bientôt	ADV	bientôt	201	DET:ART	W	un
(PUN	(002	NOM	W	mariage
donc	ADV	donc	200	ADV	W	P tout
voilà	ADV	voilà	701	PRP	W	à
)	PUN)	002	NOM	W	fait
bien	ADV	bien	200	ADJ	W	P simple
que	KON	que				
ça	PRO:DEM	cela				
a	VER:pres	avoir				

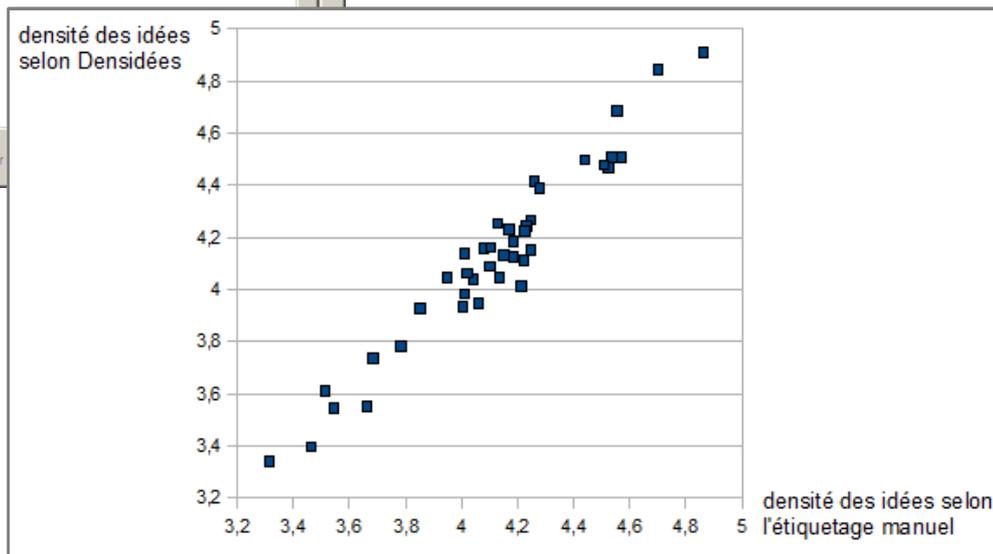
27 mots.
10 propositions.
Densité des idées : 3.704

4 fois la règle 000
9 fois la règle 002
11 fois la règle 200
2 fois la règle 201
1 fois la règle 402
2 fois la règle 600
2 fois la règle 701

Adresse de TreeTagger :
C:\Program Files\TreeTagger\

Adresse de Python :
C:\Python26\python.exe

"C:\Python26\python.exe" "C:\These\2009Densidees\densidees\densidees.py" "C:\Temp\IDtext.txt.tt.txt" oral=1 tag=treetagger



	<i>Densidées</i>	Manuelle
moyenne rangs sains	16,55	16,3
moyenne rangs malades	24,45	24,7
W de Wilcoxon	121	116
p-value	0,03264	0,02272

Logiciel TreeCloud



TreeCloud.org
Nuage Arboré

Create! Downloads Gallery Credits FAQ
Créer! Téléchargements Galerie A propos FAQ

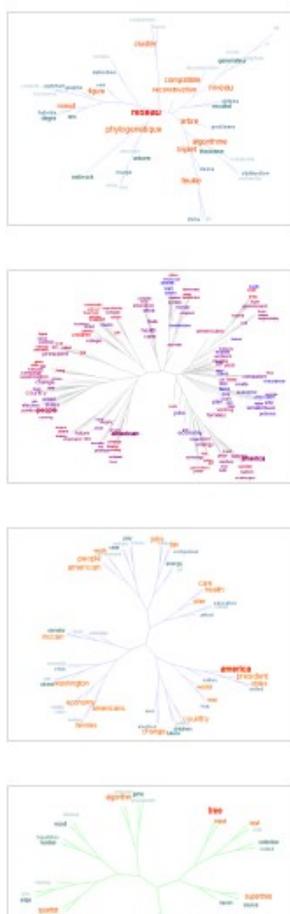
This website helps you to generate **tree clouds** from a text, that is word clouds where the words are arranged on a tree which reflects their semantic proximity inside the text. The first tree cloud appeared on [Jean Véronis's blog](#) in December 2007, you can now [create your own with this website](#), or [with the TreeCloud software](#).

Create your own tree cloud online!

Ce site web vous permet de générer des **nuages arborés** à partir d'un texte, c'est à dire des nuages de mots disposés autour d'un arbre qui indique leur proximité dans le texte. Le premier nuage arboré est apparu sur le [blog de Jean Véronis](#) en décembre 2007, vous pouvez maintenant [créer les vôtres avec ce site web](#), ou [avec le logiciel TreeCloud](#).

Créez vos propres nuages arborés en ligne !

Documents :

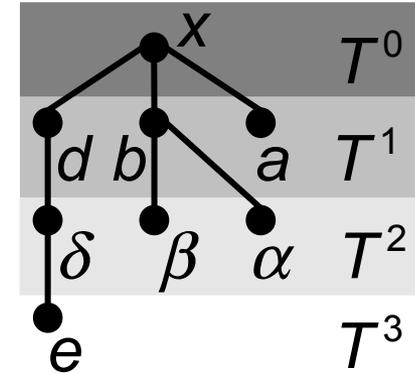
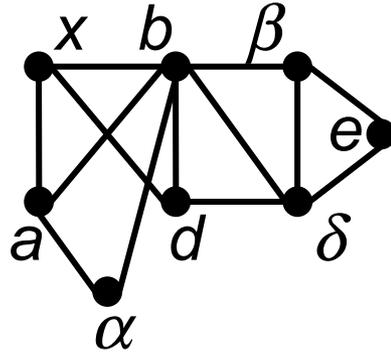




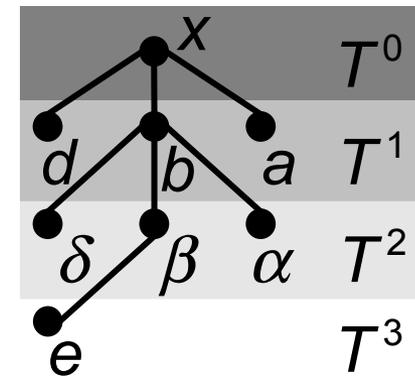
Référencé comme développement de l'Enseignement Supérieur - Recherche dans



Parcours en largeur des graphes trapézoïdaux en $O(n)$



ordre de priorité :
 $x\beta\delta d b e a \alpha$



ordre de priorité :
 $x\beta\delta b d e a \alpha$

