Séminaire Mathématiques, Evolution, Génome
Marseille – 17/03/2010

Reconstruction combinatoire de réseaux phylogénétiques

Philippe Gambette

Plan

- Les réseaux phylogénétiques
- L'arbre en filigrane
- Motivations de l'approche combinatoire
- Méthodes de reconstruction combinatoire
- Limites des méthodes combinatoires
- Perspectives

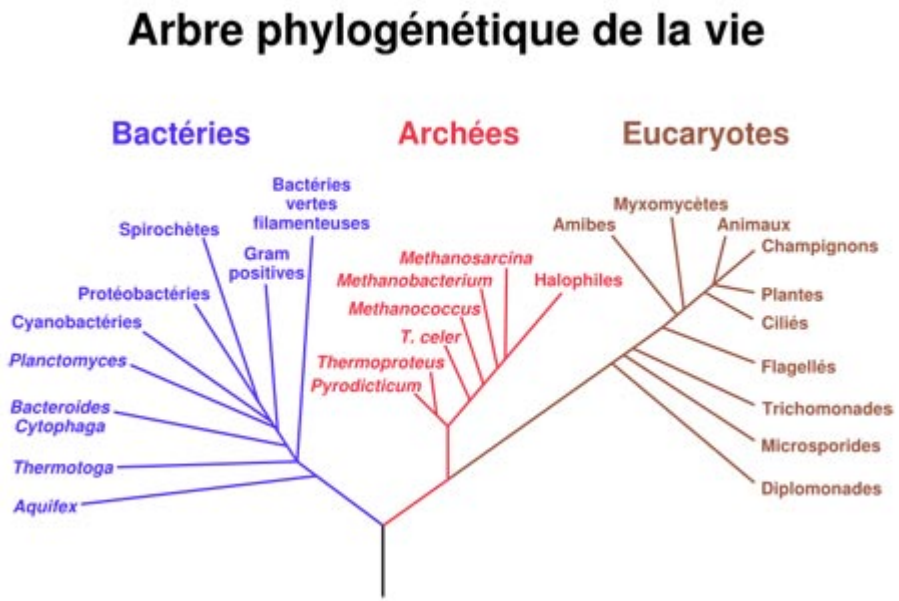
Plan

- **Les réseaux phylogénétiques**
- L'arbre en filigrane
- Motivations de l'approche combinatoire
- Méthodes de reconstruction combinatoire
- Limites des méthodes combinatoires
- Perspectives

Les arbres phylogénétiques

Arbre phylogénétique

Un **arbre phylogénétique** est un arbre schématique qui montre les relations de parentés entre des entités supposées avoir un ancêtre commun.

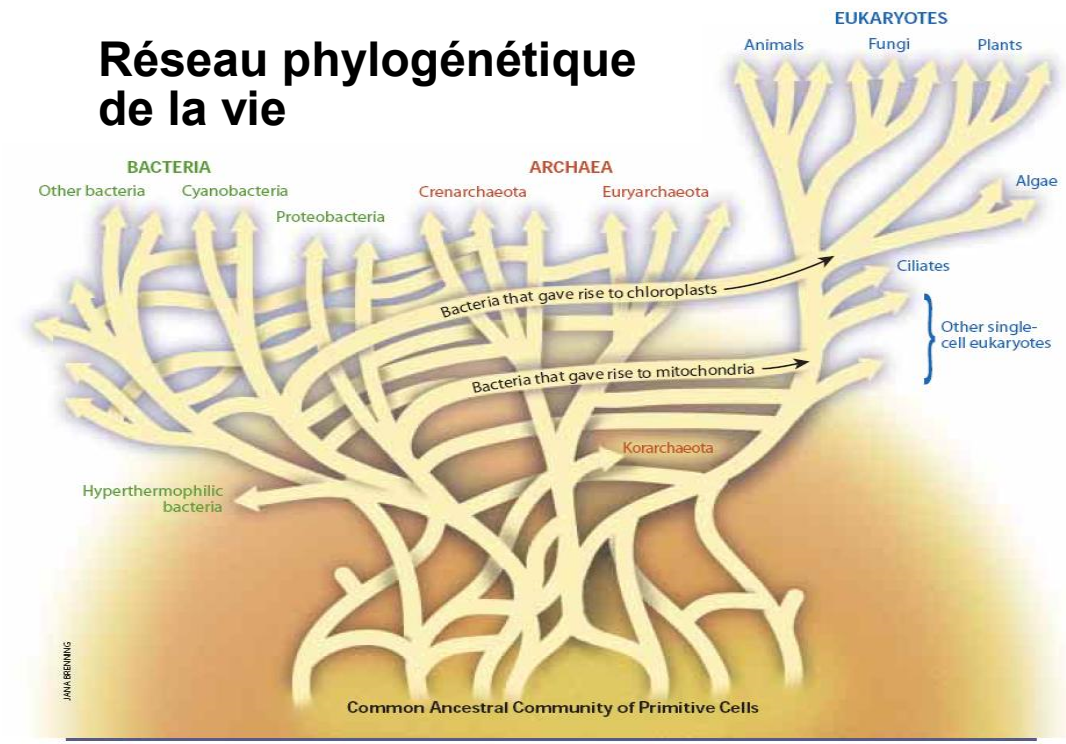


D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

Les réseaux phylogénétiques

Réseau phylogénétique

Un **réseau phylogénétique** désigne un graphe utilisé pour visualiser les relations liées à l'évolution entre des espèces ou des organismes. Il doit être employé quand interviennent des événements d'hybridations, de transferts horizontaux de gènes, ou de recombinaisons génétiques.



Doolittle : Uprooting the Tree of Life, Scientific American (Fév. 2000)

Les réseaux phylogénétiques

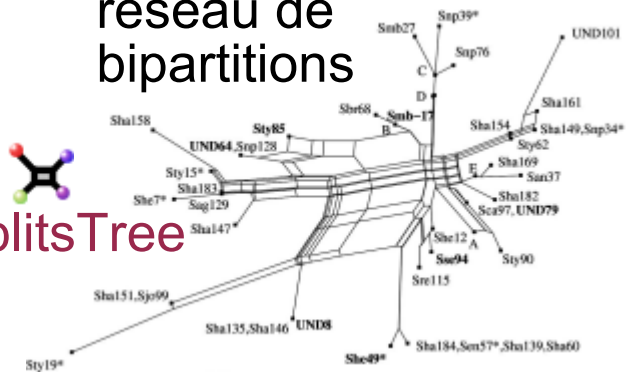
Réseau phylogénétique

Un **réseau phylogénétique** désigne un graphe utilisé pour visualiser les relations liées à l'évolution entre des espèces ou des organismes. Il doit être employé quand interviennent des événements d'hybridations, de transferts horizontaux de gènes, ou de recombinaisons génétiques.

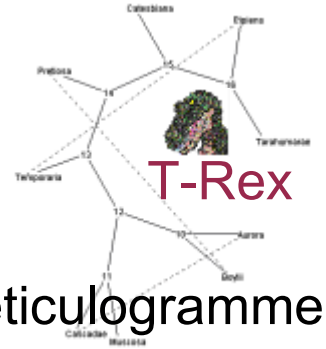
réseau de niveau 2

Level-2

réseau de bipartitions



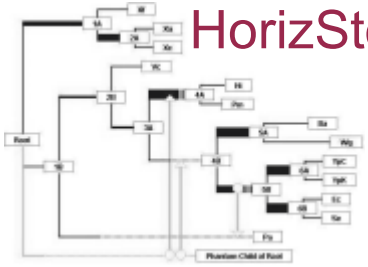
SplitsTree



T-Rex

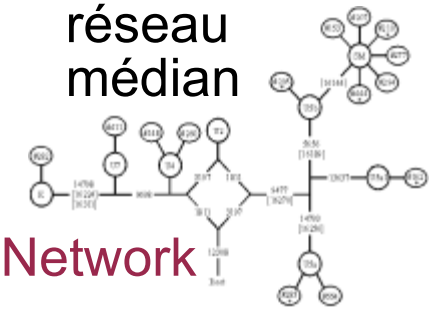
réticulogramme

diagramme de synthèse



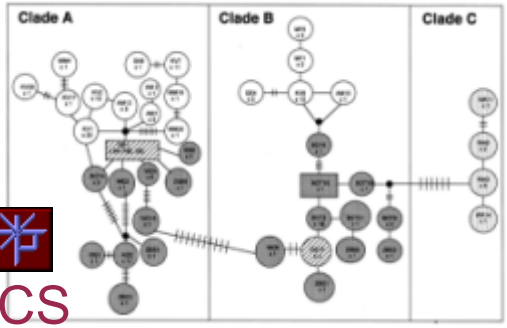
HorizStory

réseau médian



Network

réseau couvrant minimum

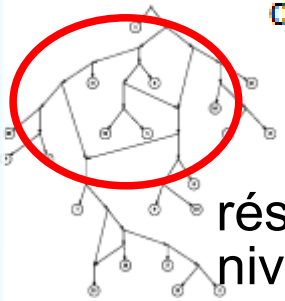


TCS

Les réseaux phylogénétiques

Réseau phylogénétique

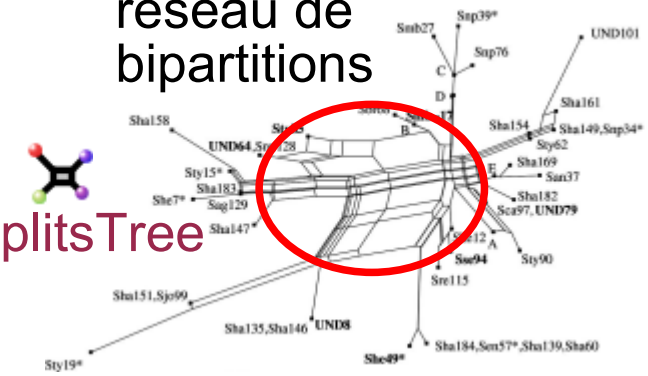
Un **réseau phylogénétique** désigne un graphe utilisé pour visualiser les relations liées à l'évolution entre des espèces ou des organismes. Il doit être employé quand interviennent des événements d'hybridations, de transferts horizontaux de gènes, ou de recombinaisons génétiques.



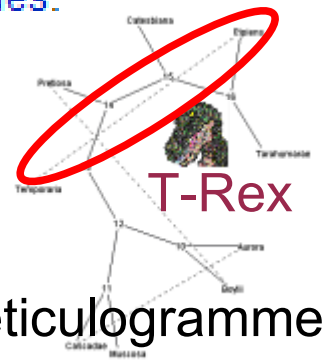
réseau de niveau 2

Level-2

réseau de bipartitions



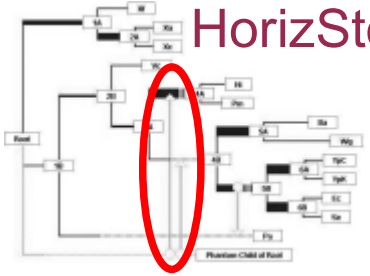
SplitsTree



T-Rex

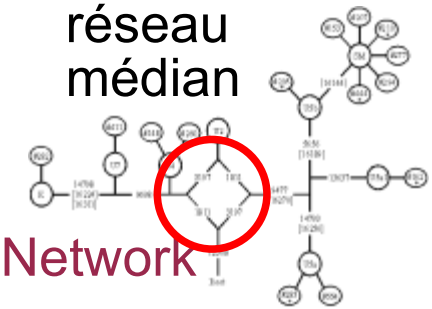
réticulogramme

diagramme de synthèse



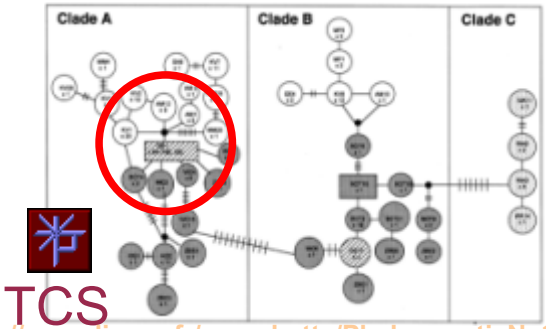
HorizStory

réseau médian



Network

réseau couvrant minimum



TCS

Les réseaux phylogénétiques

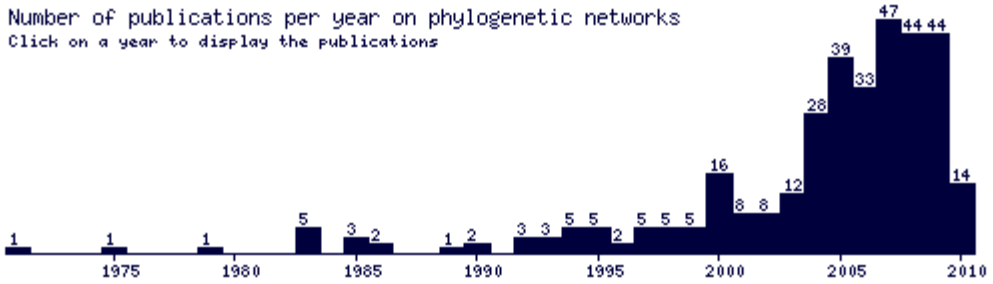
Who is Who in Phylogenetic Networks - Articles, Authors & Programs RSS

Index | Browse — **Contribute!** | My selection

Search: in All Go (word length ≥ 3) — Login

Publications - Index (All 342 publications) — Selection by: Year | Category | Keyword | Author

Selection by Year



Selection by Category

Article (Journal) (193) — **InProceedings** (91) — **InBook** (18)
Book (1) — **PhdThesis** (19) — **MastersThesis** (1)
Misc (19) — **Programs** (41)

Selection by Keyword

abstract-network(25) approximation(8) APX-hard(1) ARG(5) block-realization(1) bootstrap(1) bound(3) branch-and-bound(1) cactus-graph(1) characterization(4) clustering(2) coalescent(7) consensus(8) consistency(2) construction(2) cophylogeny(1) distance-between-networks(20) diversity(1) duplication(7) enumeration(4) evaluation(25) **explicit-network(62)** exponential-algorithm(2) FPT(10) from-clusters(7) from-distances(19) from-multilabeled-tree(3) from-network(7) from-quartets(6) **from-rooted-trees(48)** from-sequences(29) from-species-tree(21) from-splits(10) from-trees(7) from-triplets(14) from-unrooted-trees(7) galled-network(3) galled-tree(31) generation(6) haplotype-network(2) haplotyping(1) heuristic(5) HMM(2) hybridization(26) inapproximability(5) labeling(3) lateral-gene-transfer(25) level-f-phylogenetic-network(15) likelihood(8) lineage-sorting(1) MASN(4) median-network(14) MedianJoining(2) minimum-number(10) minimum-spanning-network(2) mu-distance(2) NeighborNet(10) nested-network(2) netting(3) normal-network(1) NP-complete(23) optimal-realization(2) parsimony(27) perfect(5) **phylogenetic-network(181)** **phylogeny(186)** **polynomial(38)** Program-Arlequin(5) Program-Beagle(3) Program-Bio-PhyloNetwork(4) Program-CombineTrees(2) Program-constNJ(1) Program-Dendroscope(7) Program-EEEP(2) Program-GalledTree(1) Program-HapBound(1)

http://www.lirmm.fr/~gambette/PhylogeneticNetworks

Les réseaux phylogénétiques

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

Index | Browse | Contribute! | My selection

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

Publications of Year << 2008 >>

<< Article (Journal) >>

1

Gabriel Cardona, Mercè Llabrés, Francesc Rosselló and Gabriel Valiente. Metrics for phylogenetic networks II: Nodal and triplets metrics. 2008. [Comment] BIBTEX

Keywords: distance between networks, phylogenetic network, phylogeny. **Note:** Submitted. [Annote]

2

Cuong Than, Derek Ruths and Luay Nakhleh. PhyloNet: A Software Package for Analyzing and Reconstructing Reticulate Evolutionary Relationships. In *BMC Bioinformatics*, Vol. 9(322), 2008. [Comment] BIBTEX

Keywords: Program PhyloNet, software. **Note:** http://dx.doi.org/10.1186/1471-2105-9-322. [Annote]

3

Iyad A. Kanj, Luay Nakhleh, Cuong Than and Ge Xia. Seeing the Trees and Their Branches in the Network is Hard. In *TCS*, Vol. 401:153-164, 2008. [Comment] BIBTEX

Keywords: evaluation, from network, from rooted trees, NP-complete, phylogenetic network, phylogeny.

Note: http://www.cs.rice.edu/~nakhleh/Papers/tcs08.pdf. [Annote]

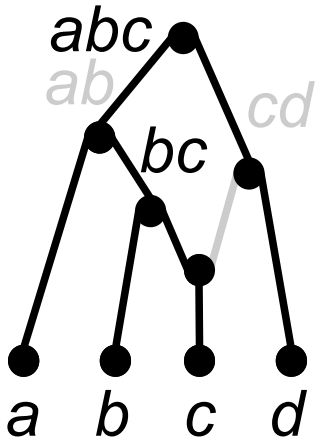
Plan

- Les réseaux phylogénétiques
- **L'arbre en filigrane**
- Motivations de l'approche combinatoire
- Méthodes de reconstruction combinatoire
- Limites des méthodes combinatoires
- Perspectives

L'arbre en filigrane

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

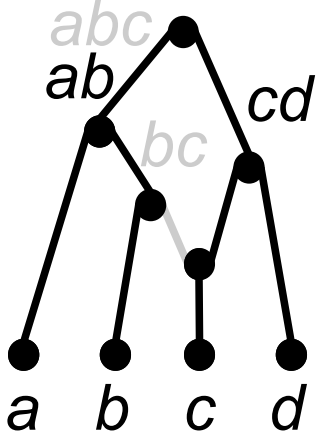
Clades “souples” : groupes monophylétiques dans un arbre de gène inclus dans le réseau



L'arbre en filigrane

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

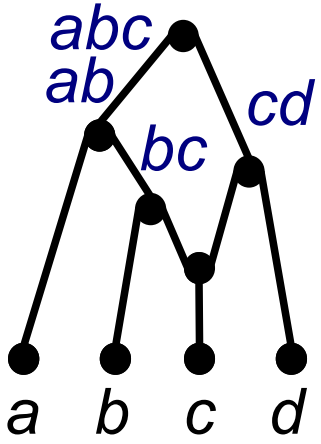
Clades “souples” : groupes monophylétiques dans un arbre de gène inclus dans le réseau



L'arbre en filigrane

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

Clades “souples” : groupes monophylétiques dans un arbre de gène inclus dans le réseau

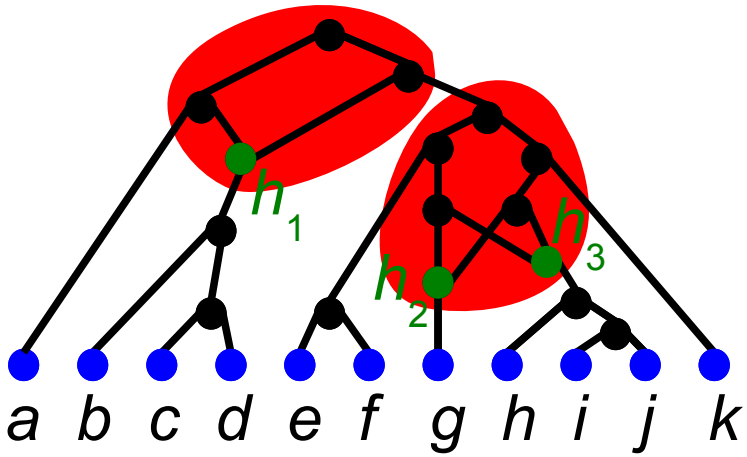


L'ensemble $C(N)$ de **tous les clades souplement compatibles** avec N peut être de taille **exponentielle**.
Test de compatibilité souple **NP-complet**

(Kanj, Nakhleh, Than, Xia, TCS, 2008)

L'arbre en filigrane : réseaux restreints

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

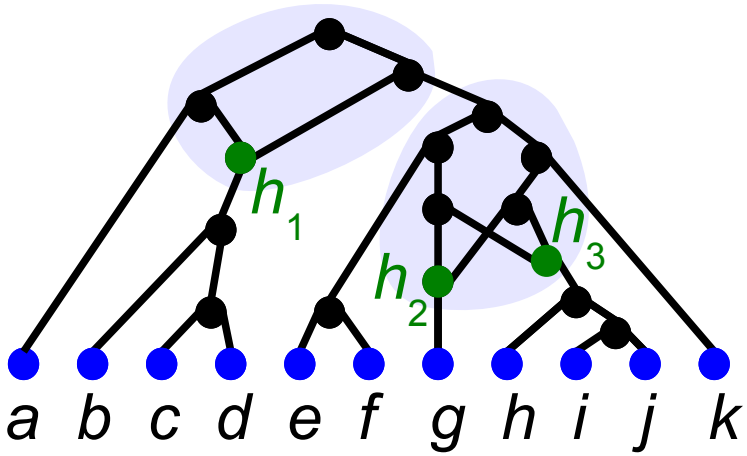


réseau de niveau 2.

niveau = nombre maximum de noeuds de réticulation par partie non arborée (*blob*).

L'arbre en filigrane : réseaux restreints

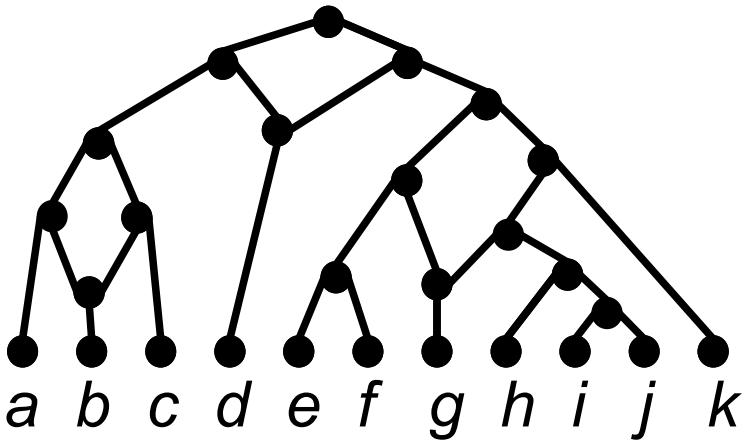
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau de niveau 2.

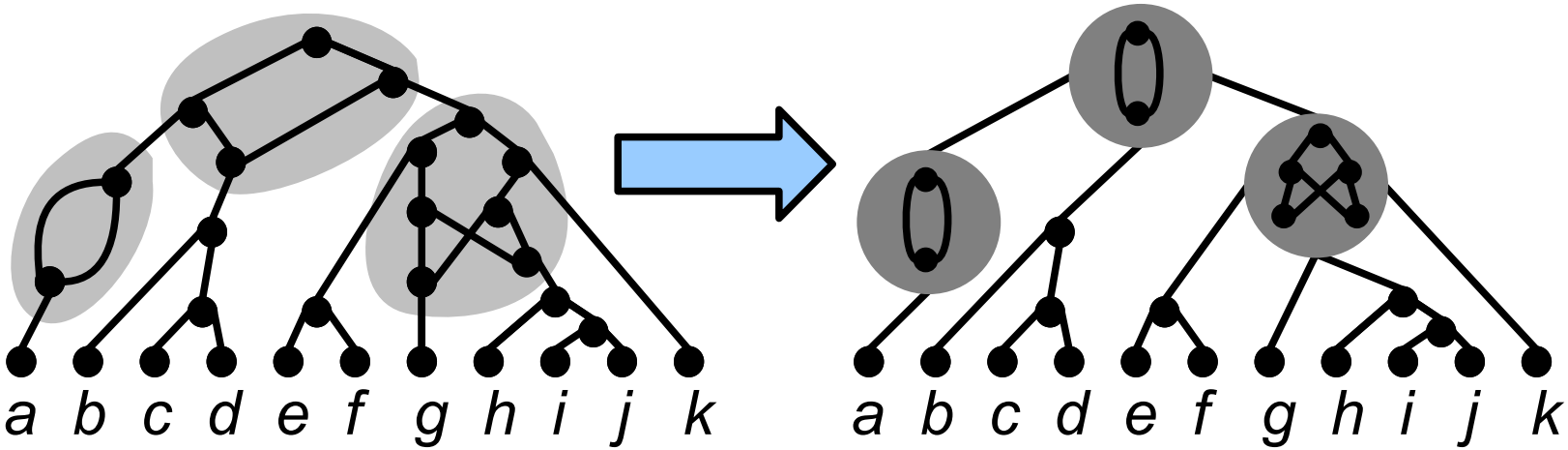
niveau = nombre maximum de noeuds de réticulation par partie non arborée.

réseau de niveau 1
(“*galled tree*”)



Décomposition des réseaux de niveau *k*

Décomposition en blobs :



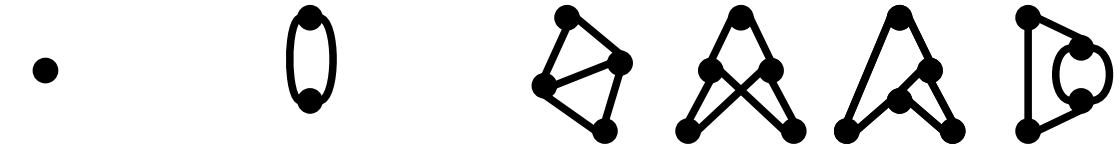
N, un réseau de niveau *k*.

décomposition arborée de *N* en générateurs.

Générateurs introduits par van Iersel & al (Recomb 2008) pour la classe restreinte des réseaux *simples* de niveau *k*.

Décomposition des réseaux de niveau *k*

Un **générateur de niveau *k*** est un réseau de niveau *k* sans isthme (arête dont la suppression déconnecte le réseau).



Les **côtés** d'un générateur sont :
- ses arcs,
- ses sommets de réticulation de degré sortant 0.

Construction des générateurs

Analyse de cas par Van Iersel & al pour trouver les 4 générateurs de niveau 2.
Généralisation par Steven Kelk en un algorithme exponentiel pour trouver les 65 générateurs de niveau 3.

Greetings from The On-Line Encyclopedia of Integer Sequences!

1,4,65 Search Hints

Search: **1, 4, 65**

Displaying 1-2 of 2 results found. page 1

Format: long | short | internal | text Sort: relevance | references | number Highlight: on | off

A041119 Denominators of continued fraction convergents to sqrt(68). +20 2

1, 4, 65, 264, 4289, 17420, 283009, 1149456, 18674305, 75846676, 1232221121, 5004731160, 81307919681, 330236409884, 5365090477825, 21790598321184, 354014663616769, 1437849252788260, 23359602708228929 (list; graph; listen)

OFFSET 0,2

CROSSREFS Cf. A041118.
Sequence in context: A138835 A119601 A058438 this_sequence A015475 A025585 A048828
Adjacent sequences: A041116 A041117 A041118 this_sequence A041120 A041121 A041122

KEYWORD nonn,cofr,easy

AUTHOR njas

A015475 q-Fibonacci numbers for q=4. +20 1

0, 1, 4, 65, 4164, 1066049, 1091638340, 4471351706689, 7325862745403091 6, 4801077413298721817665, 1258573637505038759624004676, 1319710110525284599824799048959041 (list; graph; listen)

OFFSET 0,3

FORMULA a(n) = 4^(n-1) a(n-1) + a(n-2).

CROSSREFS Sequence in context: A119601 A058438 A041119 this_sequence A025585 A048828

Construction des générateurs

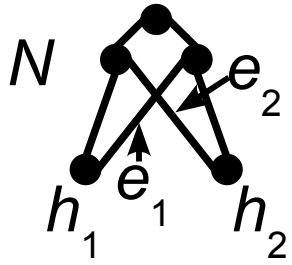
Analyse de cas par Van Iersel & al pour trouver les 4 générateurs de niveau 2.

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

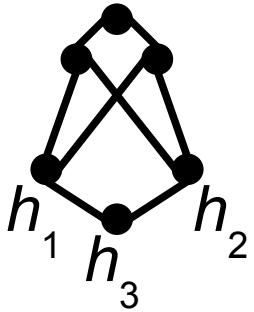
(Gambette, Berry & Paul, CPM 2009)

Construction des générateurs

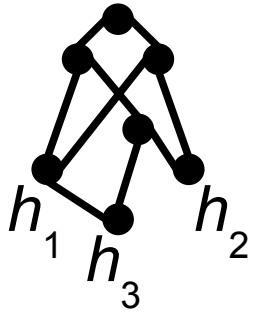
Construction des générateurs de niveau k+1 à partir de ceux de niveau k :



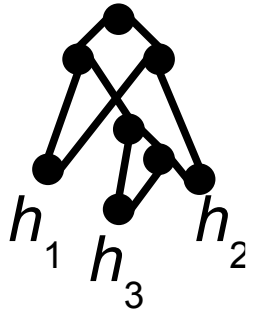
Règle R_1 :



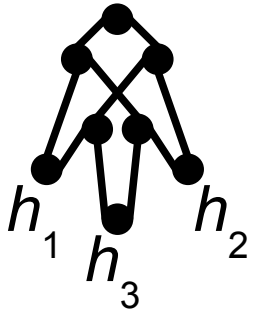
$R_1(N,h_1,h_2)$



$R_1(N,h_1,e_2)$



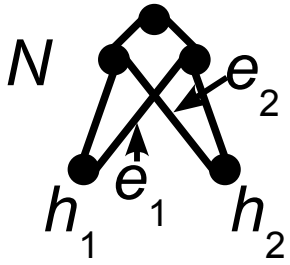
$R_1(N,e_2,e_2)$



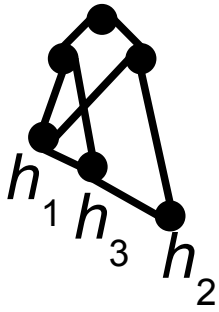
$R_1(N,e_1,e_2)$

Construction des générateurs

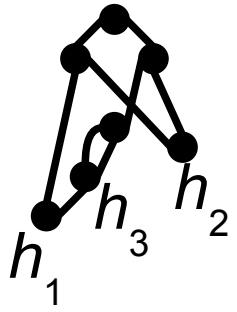
Construction des générateurs de niveau k+1 à partir de ceux de niveau k :



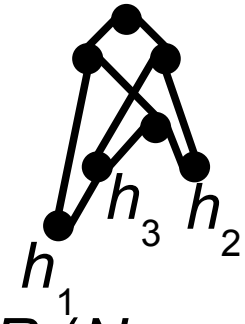
Règle R_2 :



$R_2(N,h_1,e_2)$



$R_2(N,e_1,e_1)$

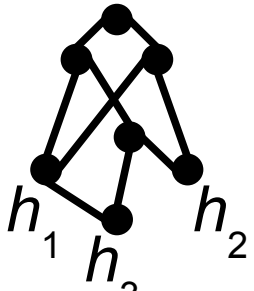


$R_2(N,e_2,e_1)$

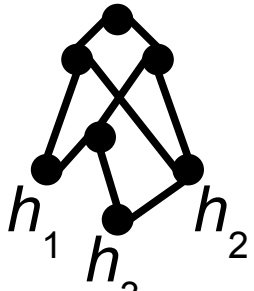
Construction des générateurs

Problème !

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



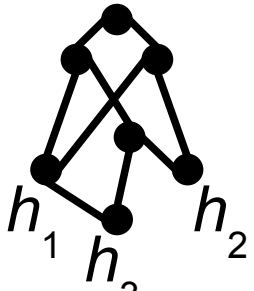
$R_1(N,h_1,e_2)$



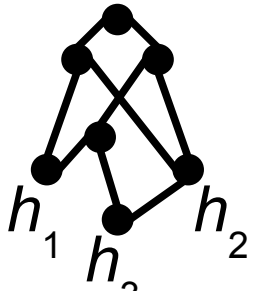
$R_1(N,h_2,e_1)$

Construction des générateurs

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



$R_1(N,h_1,e_2)$



$R_1(N,h_2,e_1)$

→ comptage difficile !

Borne supérieure

R_1 et R_2 peuvent être appliquées sur au plus toutes les paires de côtés.
Un générateur de niveau k a au plus $5k$ côtés :

$$g_{k+1} < 50\ k^2\ g_k$$

Borne supérieure :

$$g_k < k!^2\ 50^k$$

Corollaire théorique :
Il existe un algorithme polynomial pour construire l'ensemble des générateurs de niveau k+1 depuis l'ensemble des générateurs de niveau k.

Corollaire pratique :

$$g_4 < 28350$$

→ on peut énumérer tous les générateurs de niveau 4.

Nombre de générateurs de niveau *k*

On peut énumérer tous les générateurs de niveau 4.

Isomorphisme de graphes de degré maximal borné : polynomial

(Luks, FOCS 1980)

Algorithme pratique ?
Simple algorithme exponentiel de backtrack suffisant pour le niveau 4 :

parcourir les deux graphes en parallèle depuis leur racine et identifier leurs sommets : $O(n2^{n-k})$

→ $g_4 = 1993$
→ $g_5 > 71000$

Nombre de générateurs de niveau *k*

Greetings from The On-Line Encyclopedia of Integer Sequences!

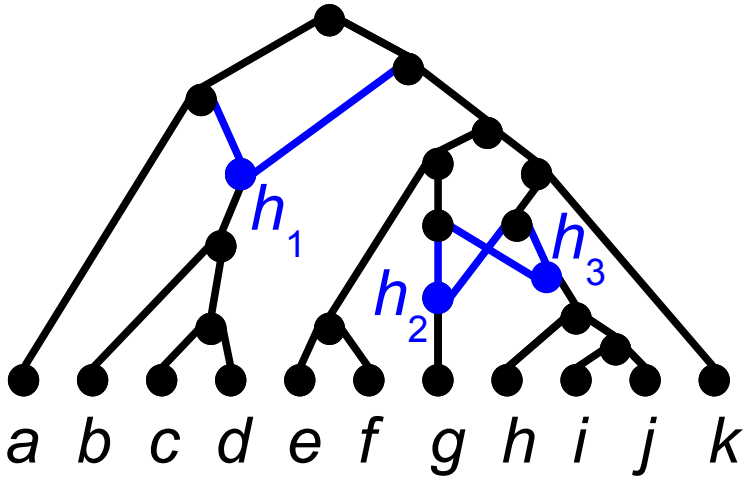
1,4,65,1993 Search Hints

Search: **1, 4, 65, 1993**

I am sorry, but the terms do not match anything in the table.

L'arbre en filigrane : réseaux restreints

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

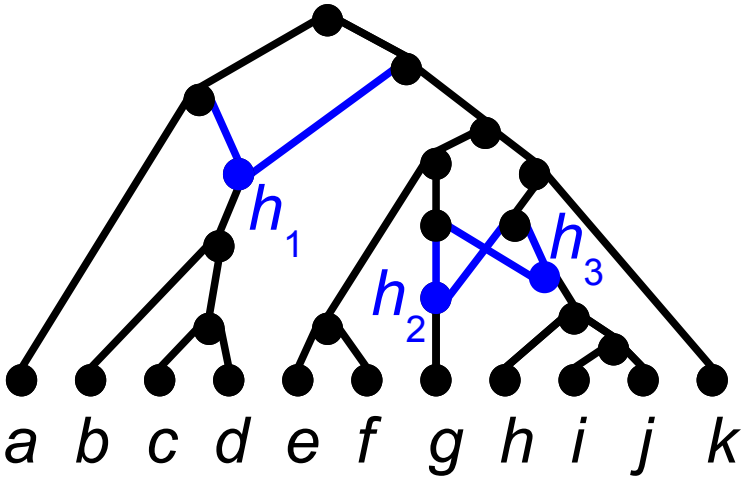


réseau à une couche de réticulation.

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

L'arbre en filigrane : réseaux restreints

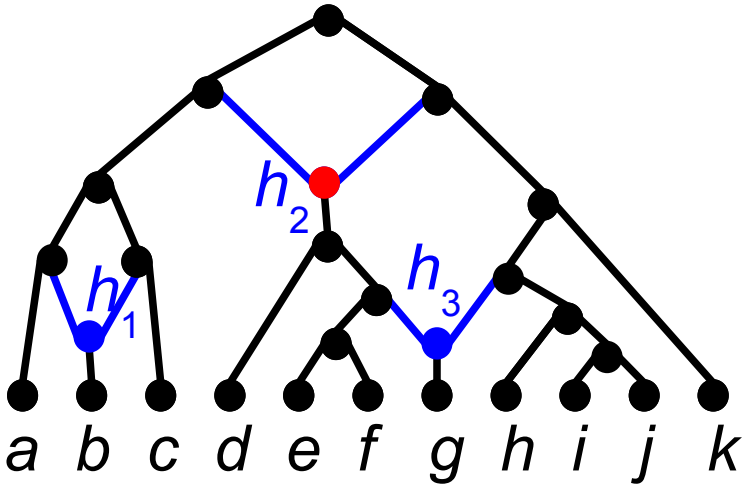
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



réseau à une couche de réticulation.

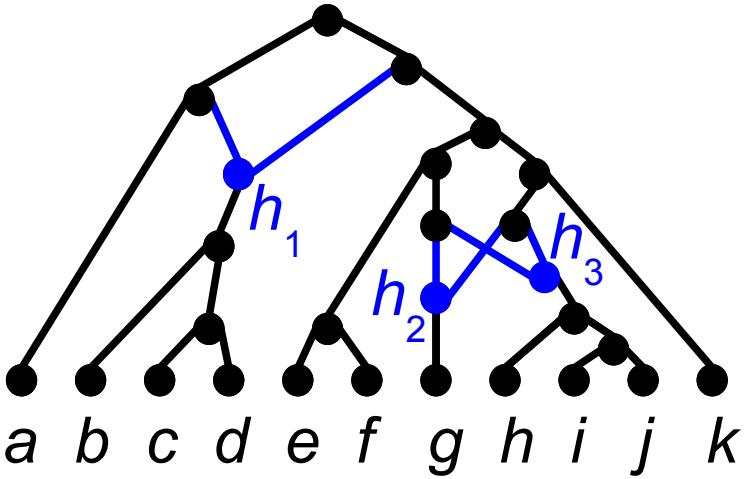
réseau **à une couche de réticulation** (“*galled network*”) : la suppression d'un noeud de réticulation déconnecte le réseau.

réseau à deux couches de réticulation.



L'arbre en filigrane : réseaux restreints

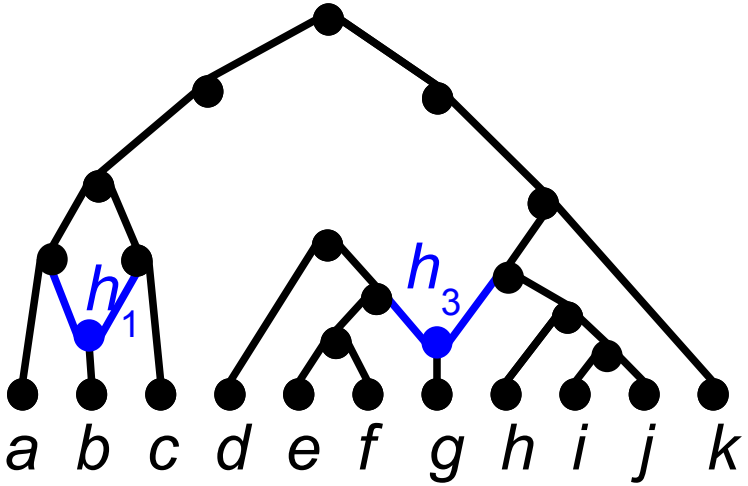
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



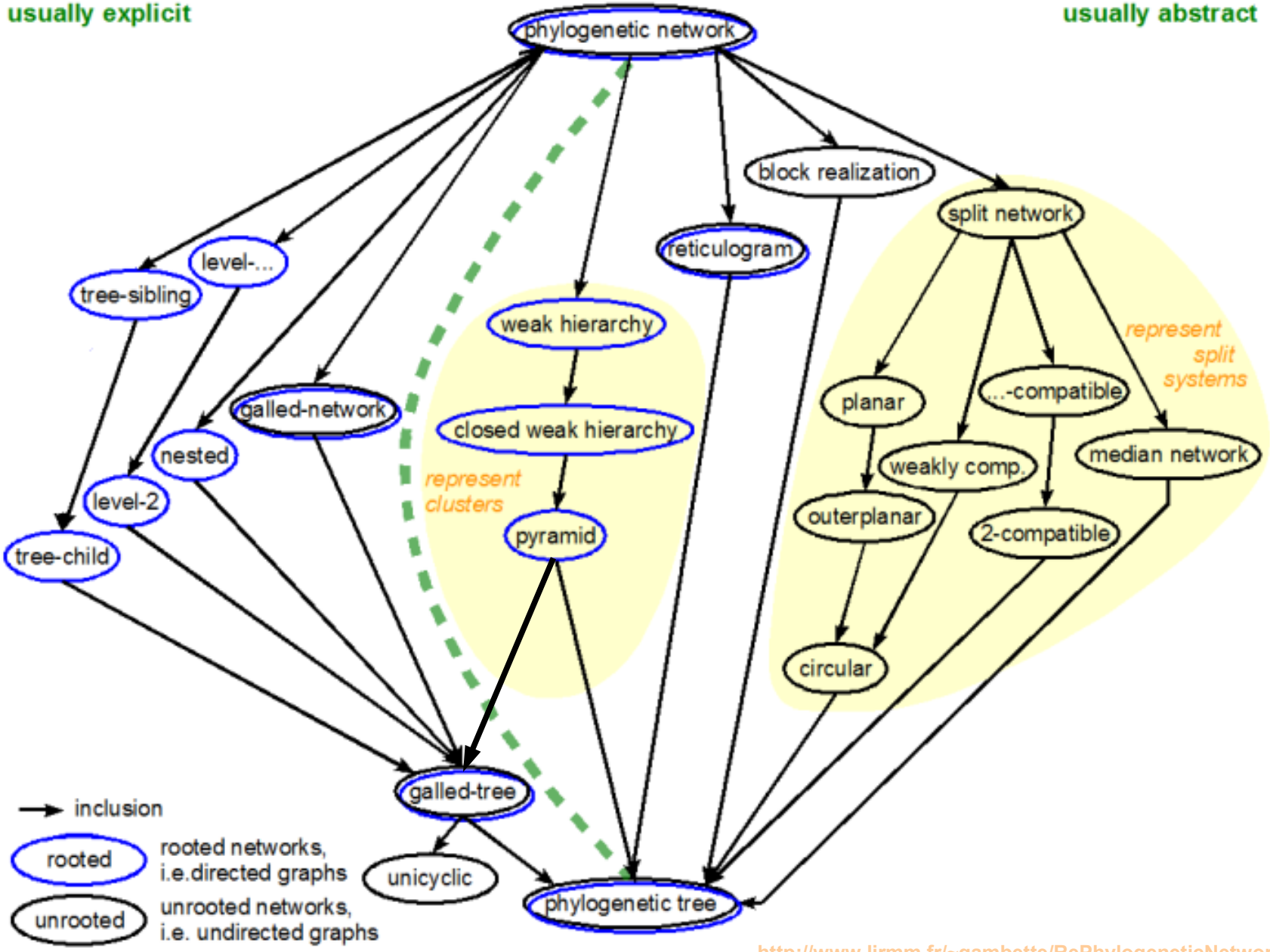
réseau à une couche de réticulation.

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

réseau à deux couches de réticulation.

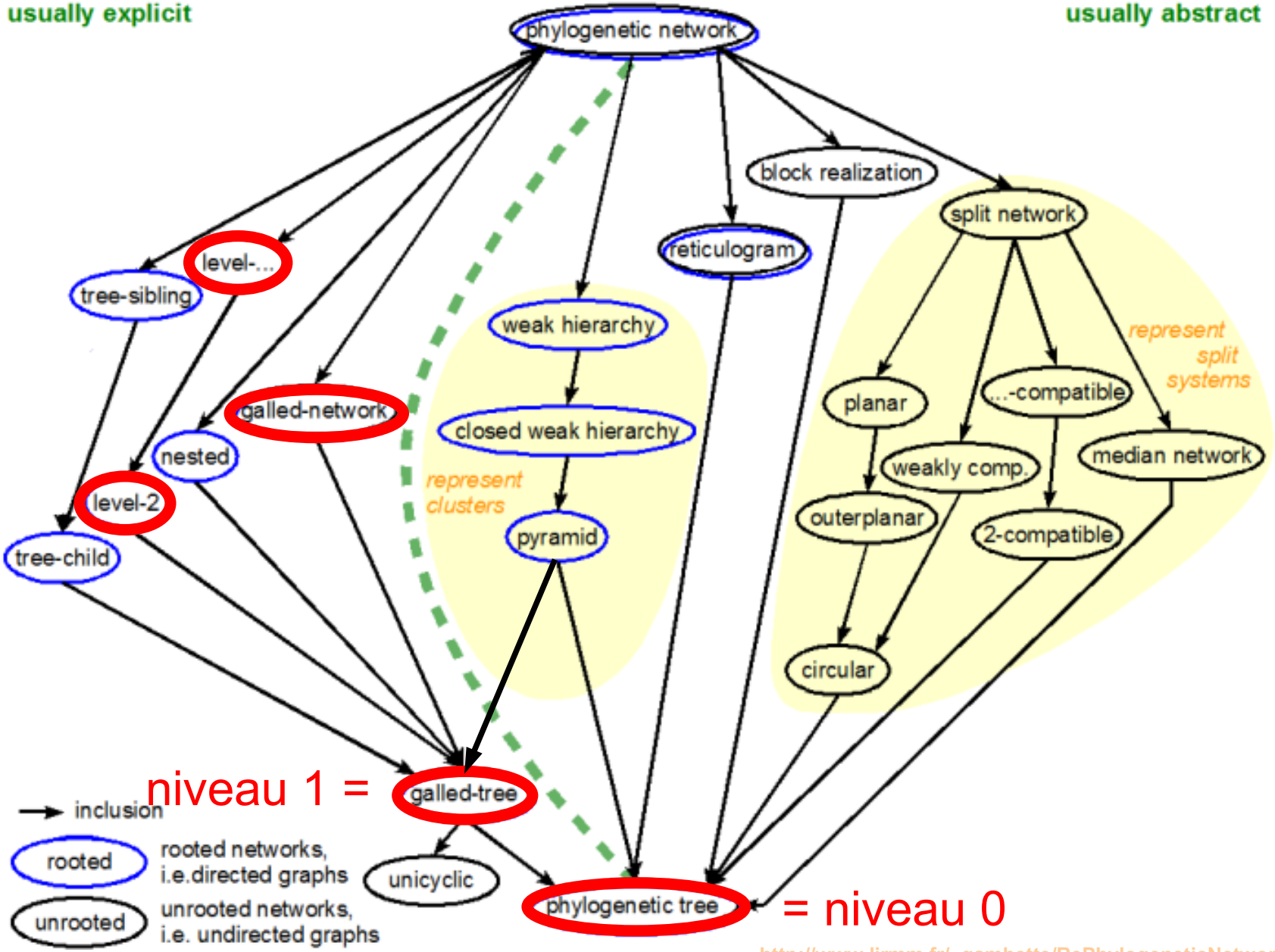


Hiérarchie de sous-classes de réseaux



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

Hiérarchie de sous-classes de réseaux

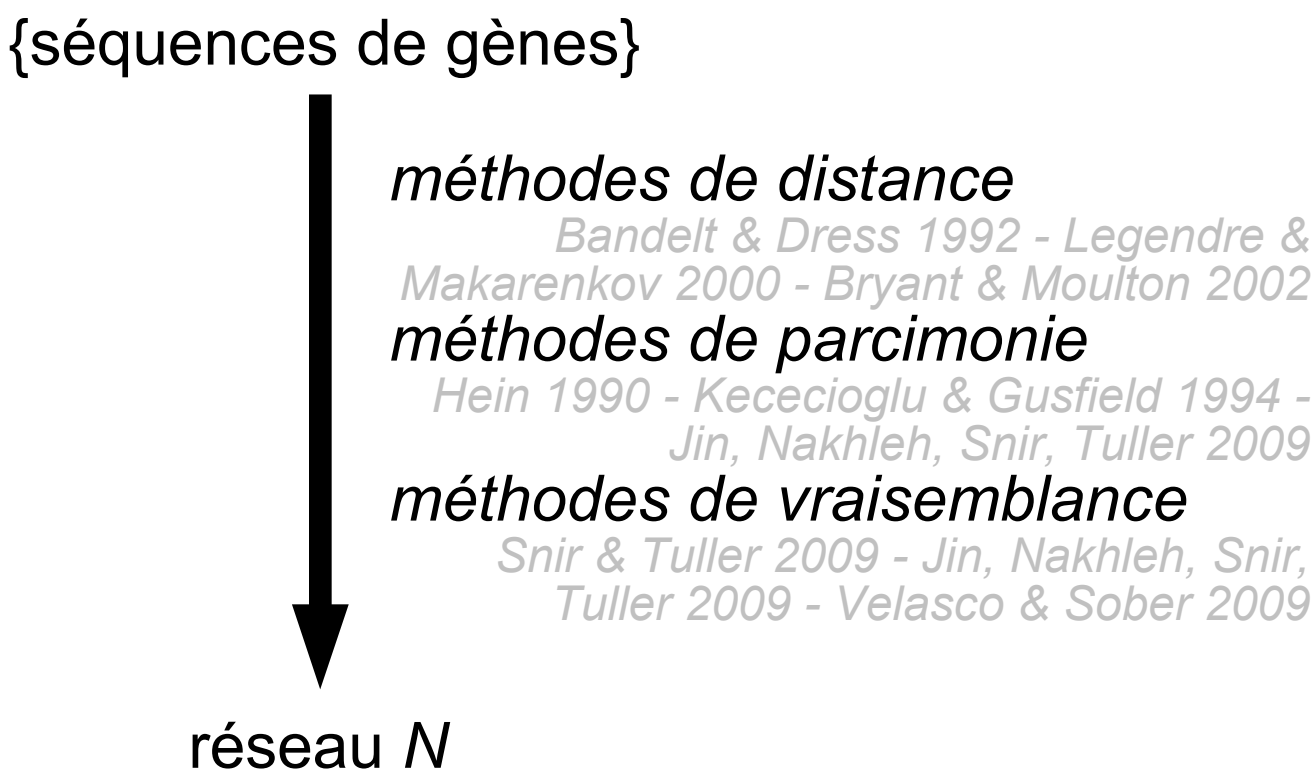


http://www.lirmm.fr/~gambette/RePhylogeneticNetworks.php

Plan

- Les réseaux phylogénétiques
- L'arbre en filigrane
- **Motivations de l'approche combinatoire**
- Méthodes de reconstruction combinatoire
- Limites des méthodes combinatoires
- Perspectives

Reconstruction de réseaux



Reconstruction de réseaux

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

{séquences de gènes}

méthodes de distance

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

méthodes de parcimonie

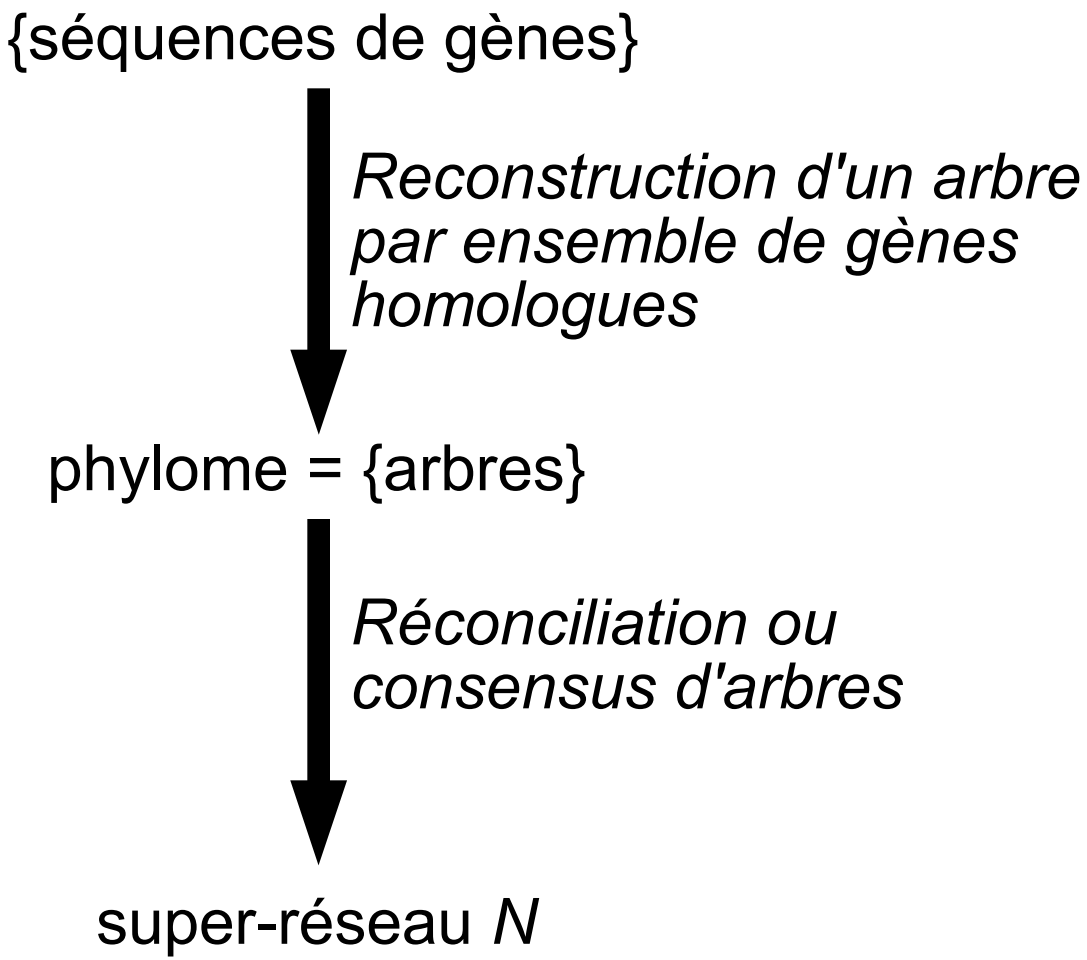
Hein 1990 - Kececioglu & Gusfield 1994 - Jin, Nakhleh, Snir, Tuller 2009

méthodes de vraisemblance

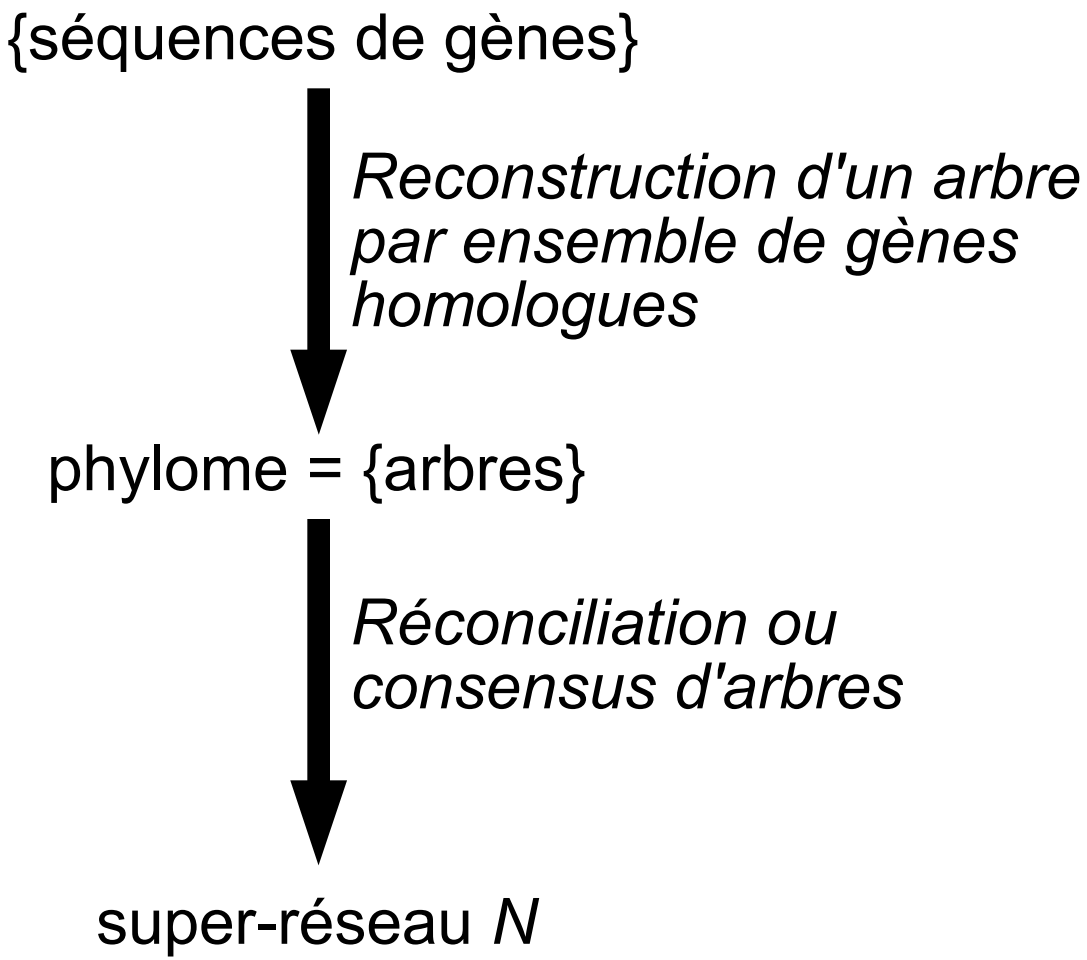
Snir & Tuller 2009 - Jin, Nakhleh, Snir, Tuller 2009 - Velasco & Sober 2009

réseau *N*

Reconstruction combinatoire de réseaux



Reconstruction combinatoire de réseaux



Problème : le consensus d'arbres est un problème NP-complet pour 2 arbres

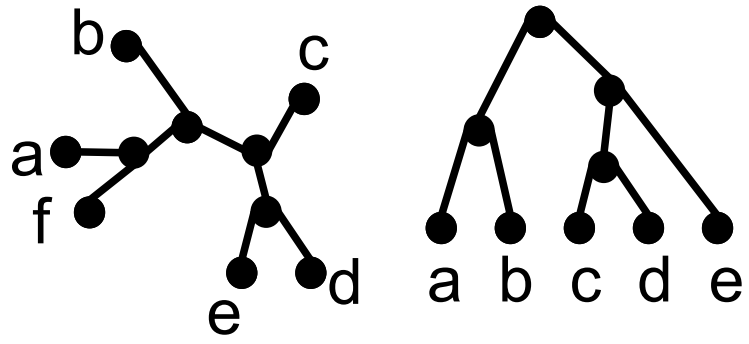
Triplets/quadruplets, splits/clades

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

Idée :
reconstuire un réseau contenant tous les :

triplets
quadruplets
clades
splits

des arbres en entrée ?

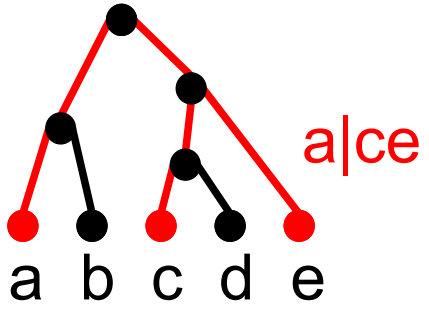


Motivations algorithmiques !

Triplets/quadruplets, splits/clades

Idée :
reconstuire un réseau contenant tous les :

triplets

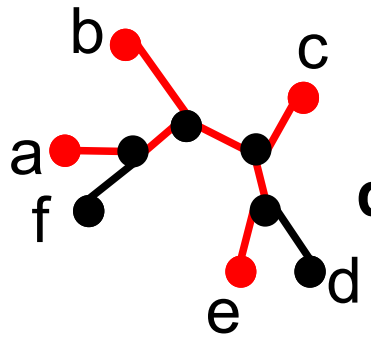


des arbres en entrée ?

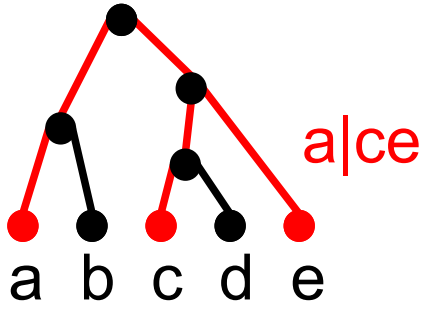
Triplets/quadruplets, splits/clades

Idée :
reconstuire un réseau contenant tous les :

ab|ce



triplets



a|ce

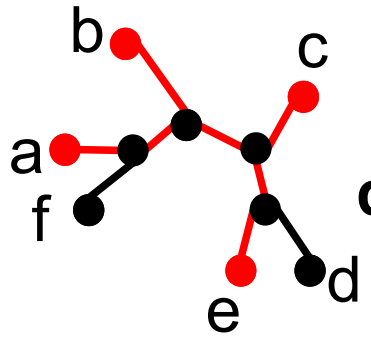
quadruplets

des arbres en entrée ?

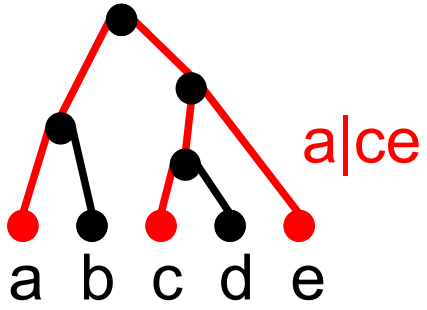
Triplets/quadruplets, splits/clades

Idée :
reconstuire un réseau contenant tous les :

ab|ce

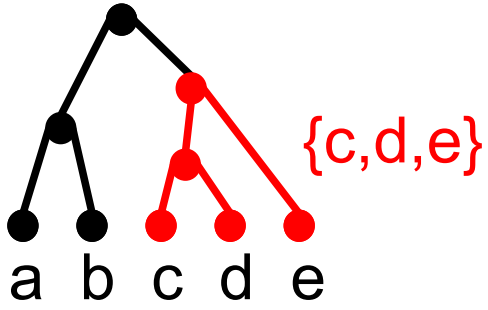


triplets



quadruplets

clades

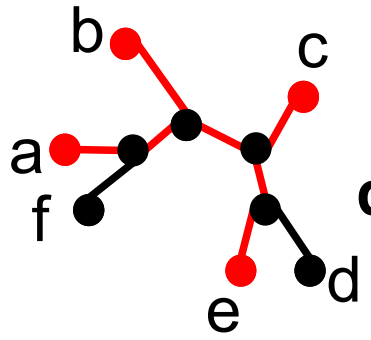


des arbres en entrée ?

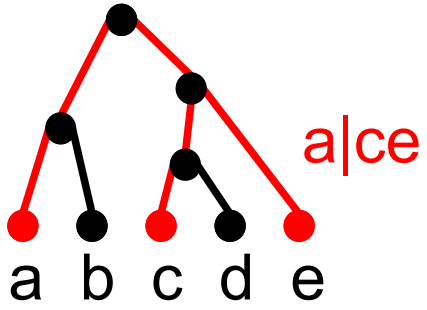
Triplets/quadruplets, splits/clades

Idée :
reconstuire un réseau contenant tous les :

ab|ce

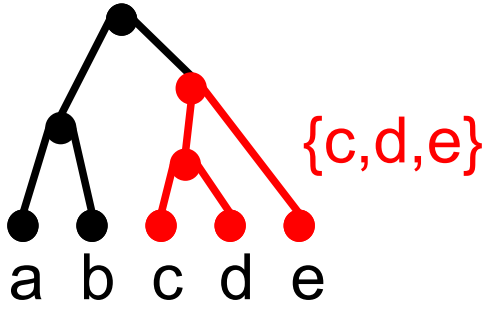


triplets

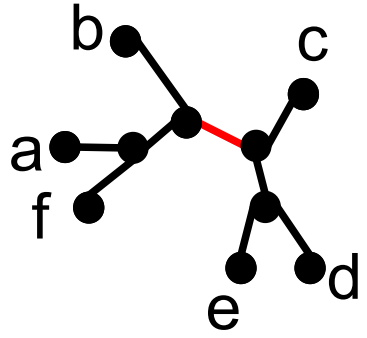


quadruplets

clades



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



splits

des arbres en entrée ?

Triplets/quadruplets, splits/clades

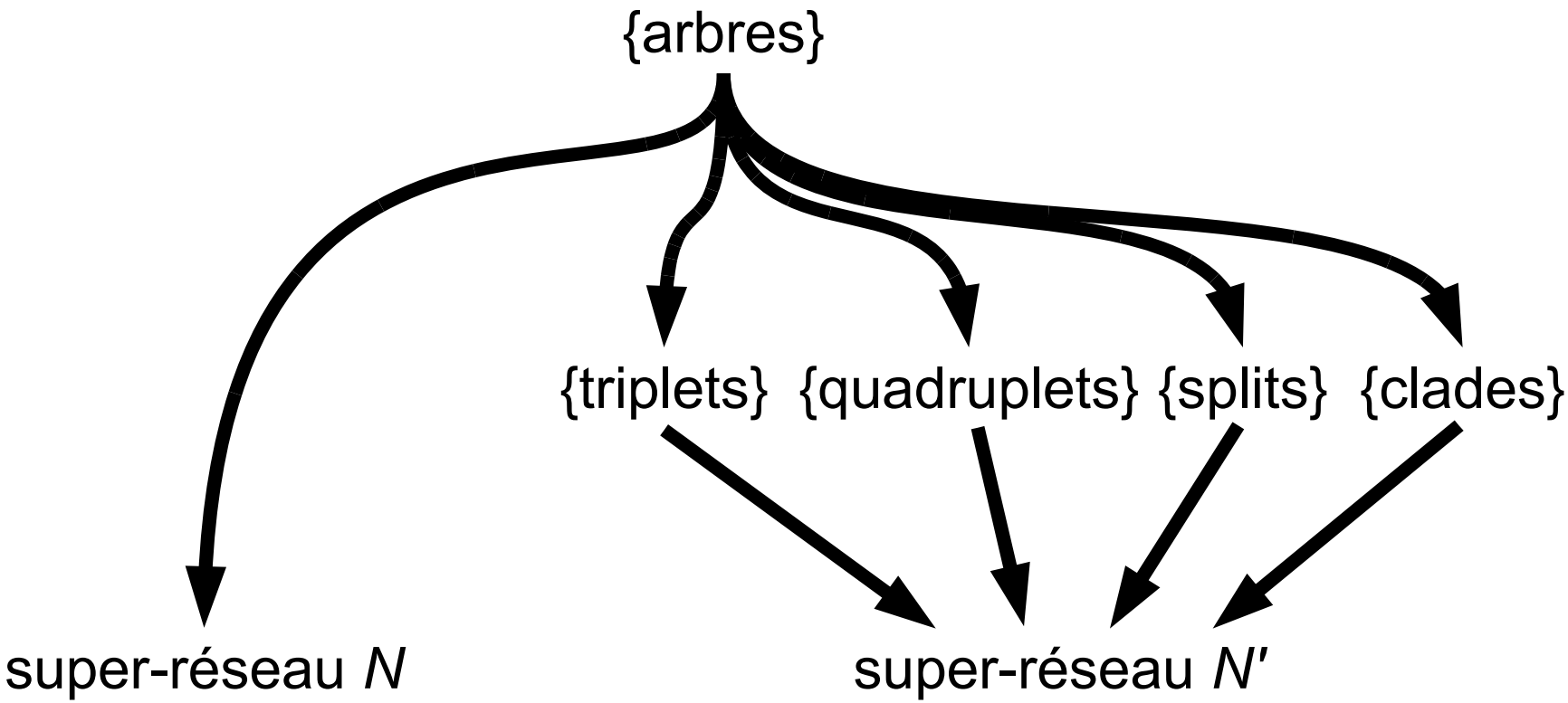
Idée :
modifier le type de données à traiter

{arbres}

super-réseau *N*

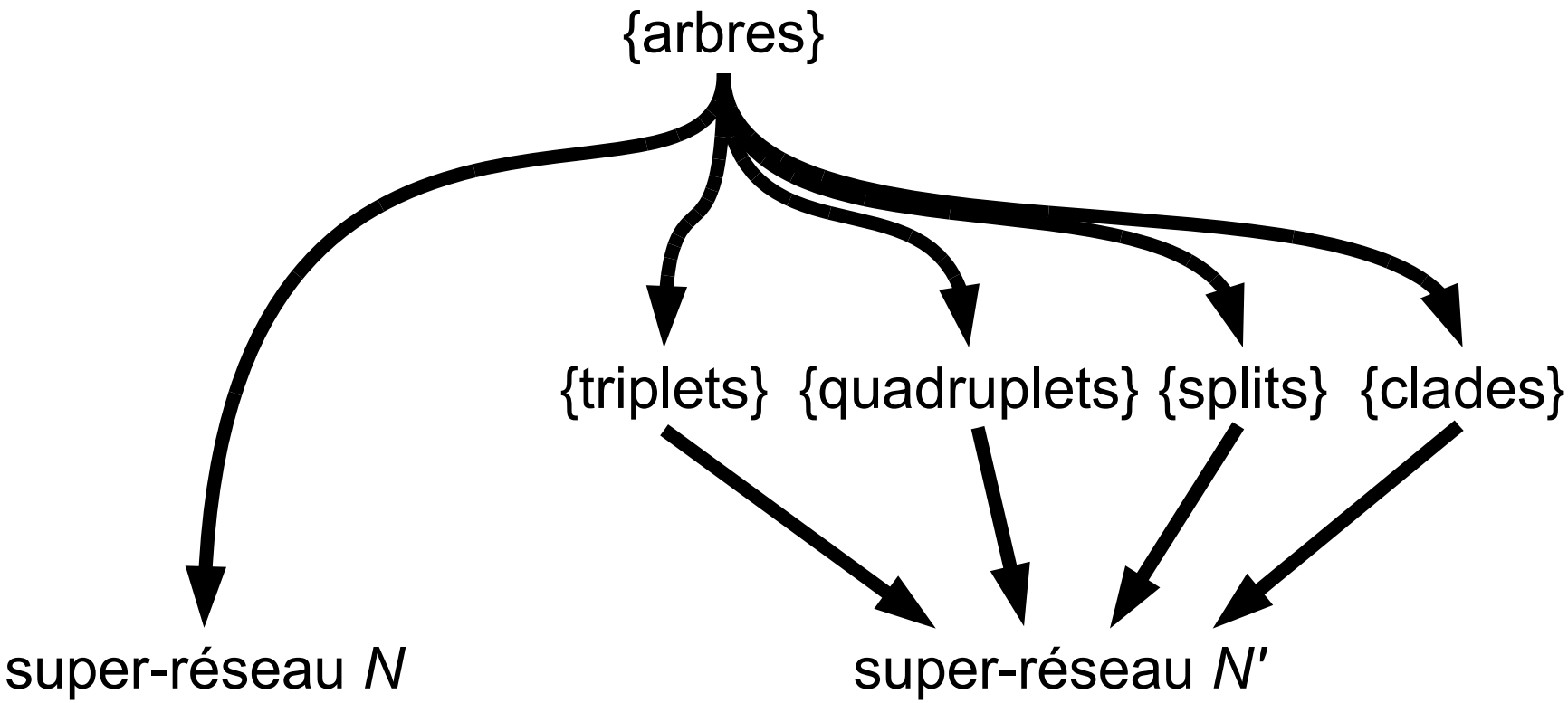
Triplets/quadruplets, splits/clades

Idée :
modifier le type de données à traiter



Triplets/quadruplets, splits/clades

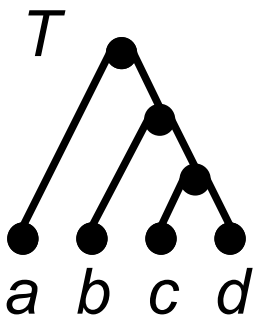
Idée :
modifier le type de données à traiter

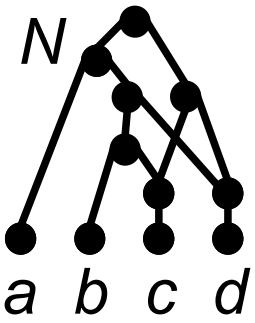


N'=N ?

Triplets/quadruplets, splits/clades

Un réseau compatible avec l'ensemble de **tous les triplets d'un arbre *T*** n'est **pas forcément compatible avec *T***.

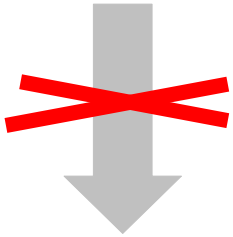




compatible avec
{*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas avec *T*

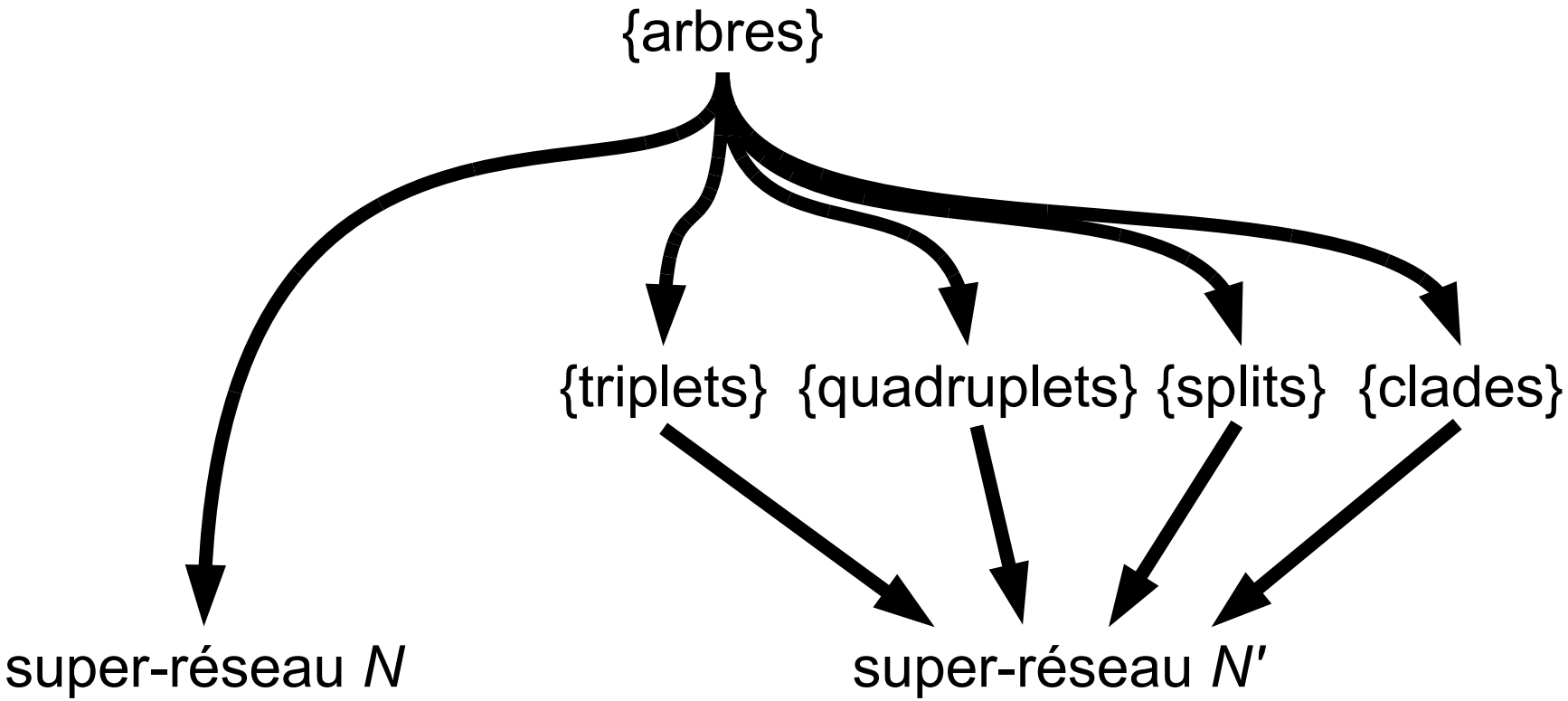
compatible avec
{*abcd*,*bcd*,*cd*,*a*,*b*,*c*}
mais pas avec *T*

compatible avec les clades d'un arbre *T*



compatible avec *T*.

Triplets/quadruplets, splits/clades



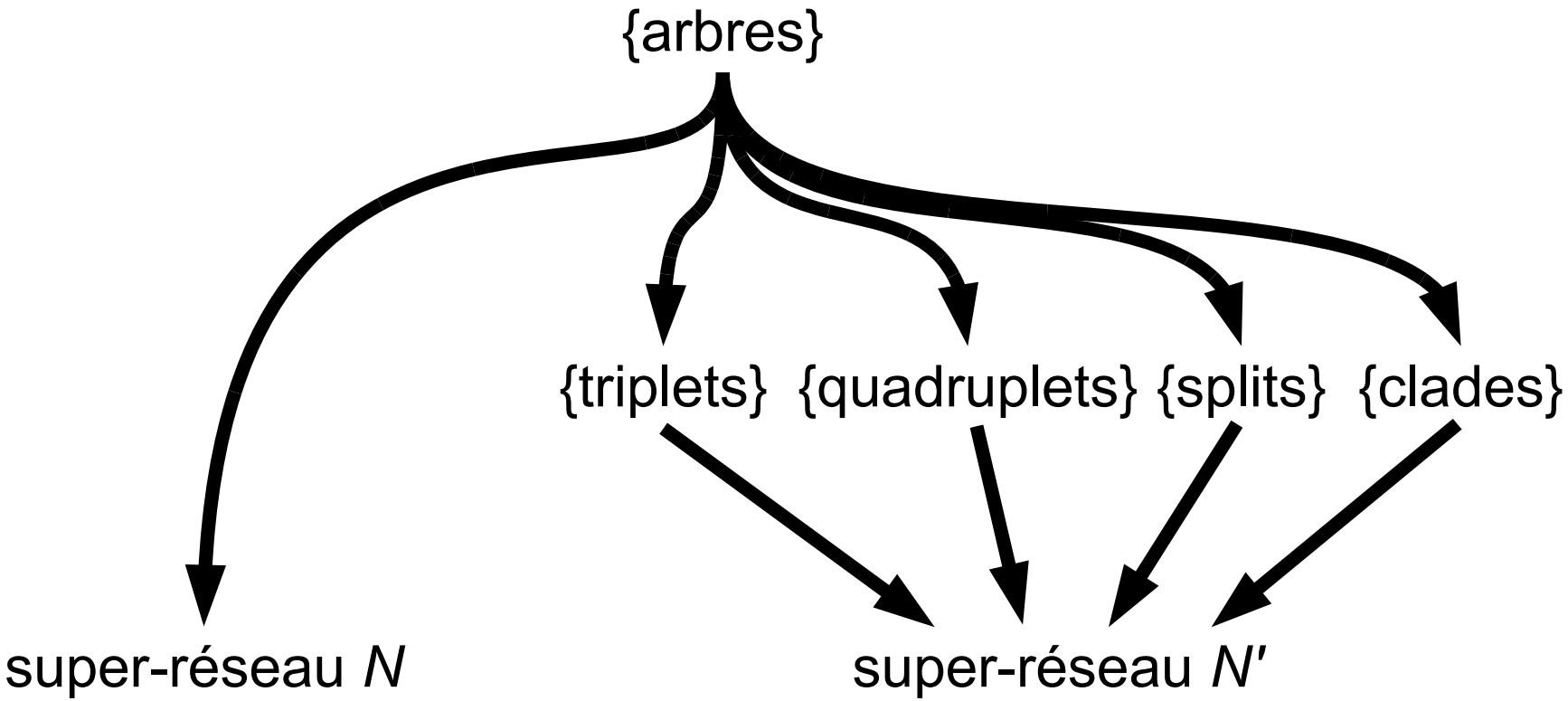
***N'*=*N* ?**

Pas nécessairement, mais :

***N'* complexe** ⟹ ***N* complexe**

N contient également les triplets, quadruplets...

Triplets/quadruplets, splits/clades



N'=N ?

Pas nécessairement, mais :
N' peut être intéressant en soi...

Plan

- Les réseaux phylogénétiques
- L'arbre en filigrane
- Motivations de l'approche combinatoire
- **Méthodes de reconstruction combinatoire**
- Limites des méthodes combinatoires
- Perspectives

Reconstruction depuis les triplets

{arbres}

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**

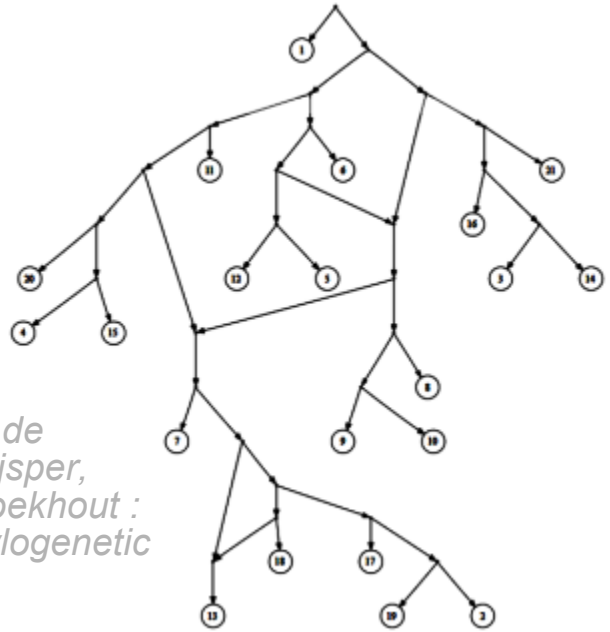
(Jansson, Nguyen & Sung, SODA'05 : $O(n^3)$ pour niveau 1)
(van Iersel, Kelk & al, RECOMB'08 : $O(n^8)$ pour niveau 2)

{triplets}

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

Programme Simplistic

N' réseau de niveau 1 ou 2

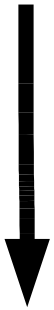


Réseau phylogénétique de levures - Van Iersel, Keijsper, Kelk, Stougie, Hagen Boekhout : Constructing level-2 phylogenetic networks from triplets. Recomb 2008

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

Reconstruction depuis les quadruplets

{arbres}

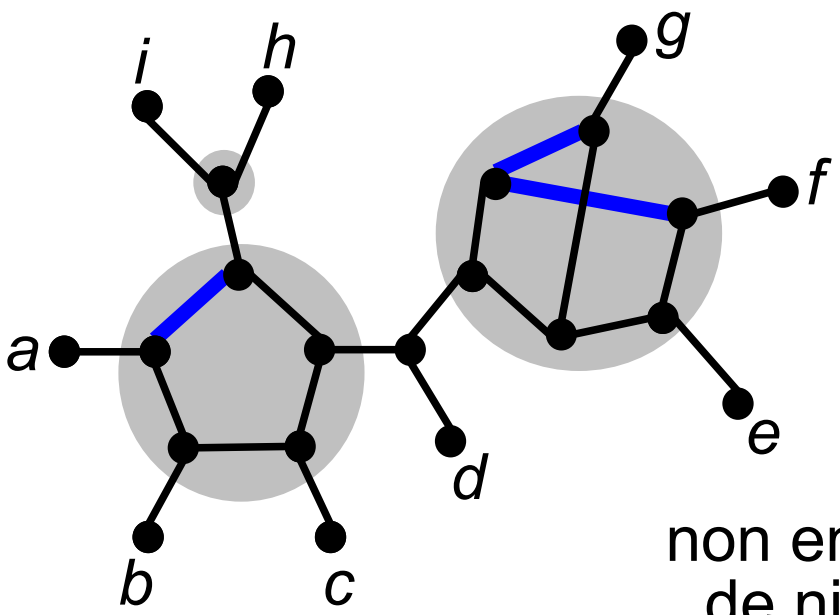


{quadruplets}

N' réseau
de niveau 1

Méthode exacte pour reconstruire un **réseau non enraciné de niveau 1** (s'il en existe un) à partir d'un ensemble dense de **quadruplets**

(Gambette, Berry, Paul, soumis : $O(n^{10})$)



réseau
non enraciné
de niveau 2

Reconstruction depuis les clades souples

{arbres}

{clades}

N' réseau "galled network"

Consensus de clades souples : **Dendroscope**

(Huson, Dezulian, Franz, Rausch, Richter & Rupp, 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)

2 étapes :
- choix du plus gros sous-ensemble de taxons où les clades sont compatibles avec un arbre
- ajout du minimum de réticulations pour connecter les autres taxons

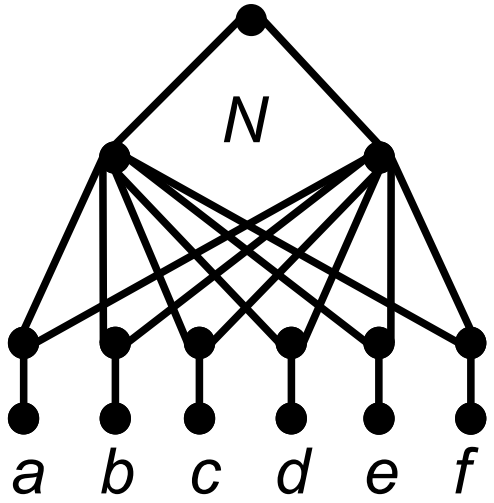
Méthode exacte de reconstruction de **réseaux de niveau *k*** à partir de **clades souples**

(Iersel, Kelk, Rupp & Huson, soumis, 2009)

Clades et réseaux à une couche de rét.

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.

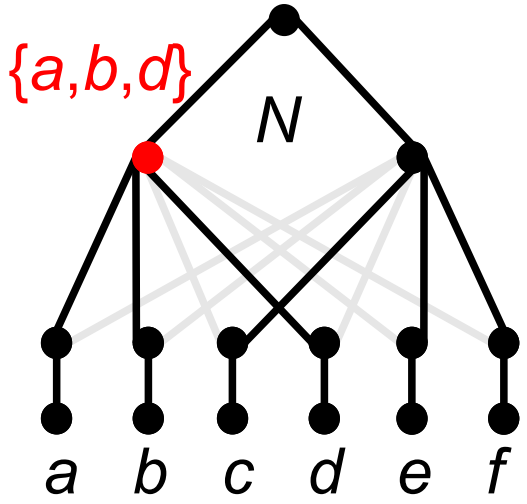


Le réseau à une couche de réticulation *N*
est compatible avec
tout clade souple sur {*a*,*b*,*c*,*d*,*e*,*f*}.

Clades et réseaux à une couche de rét.

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.

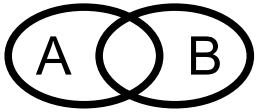
MAXIMUM COMPATIBLE SUBSET

2- Attacher les taxons impliqués dans des conflits à l'arbre 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 clades.

NP-complet dans le cas général

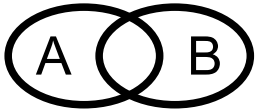
Steel & Hamel, AML, 1996

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

réduction depuis le cas général

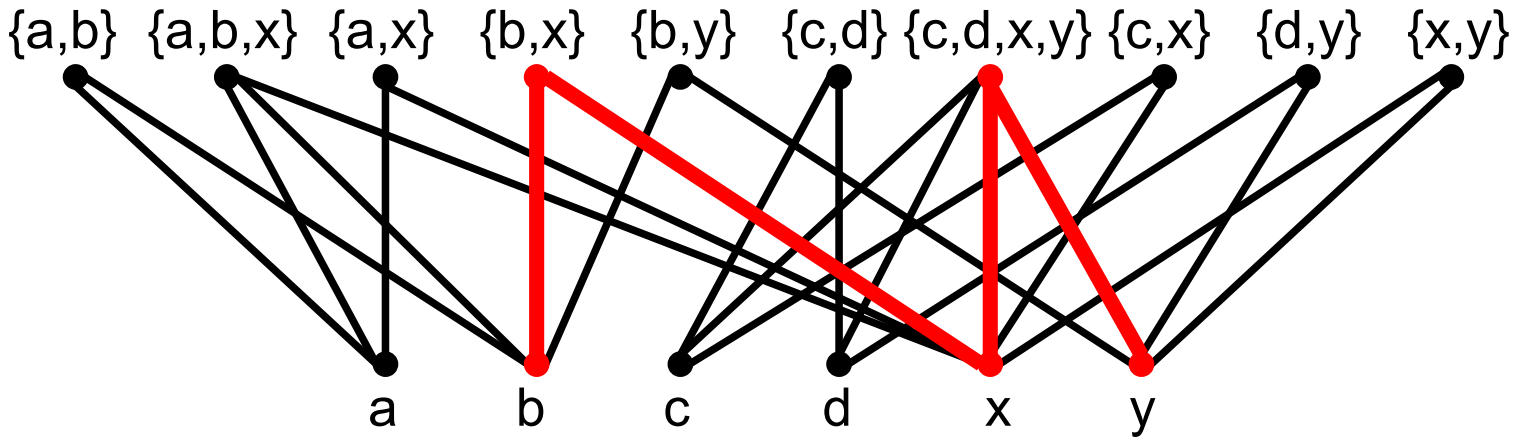
L'ensemble minimum de conflits

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

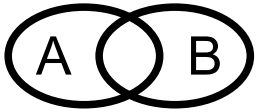
Exemple : {{a,b},{a,b,x},{a,x},{b,x},{b,y},{c,d},{c,d,x,y},{c,x},{d,y},{x,y}}



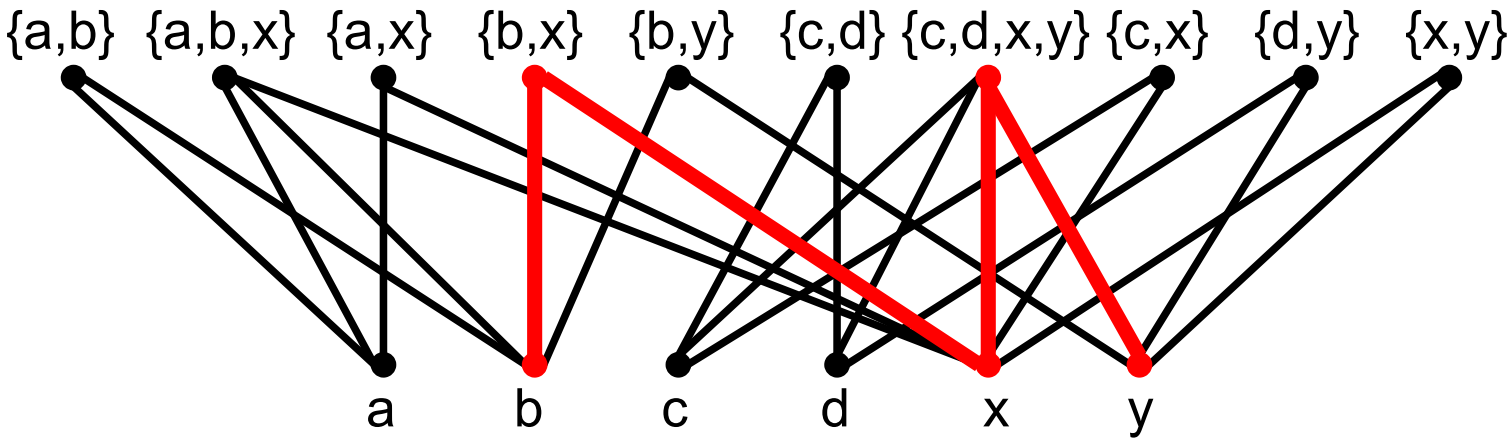
conflit = graphe “M”

L'ensemble minimum de conflits

Conflit : clades ni inclus ni disjoints



Graphe des caractères :



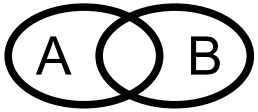
Trouver l'ensemble minimum de conflits

⟺

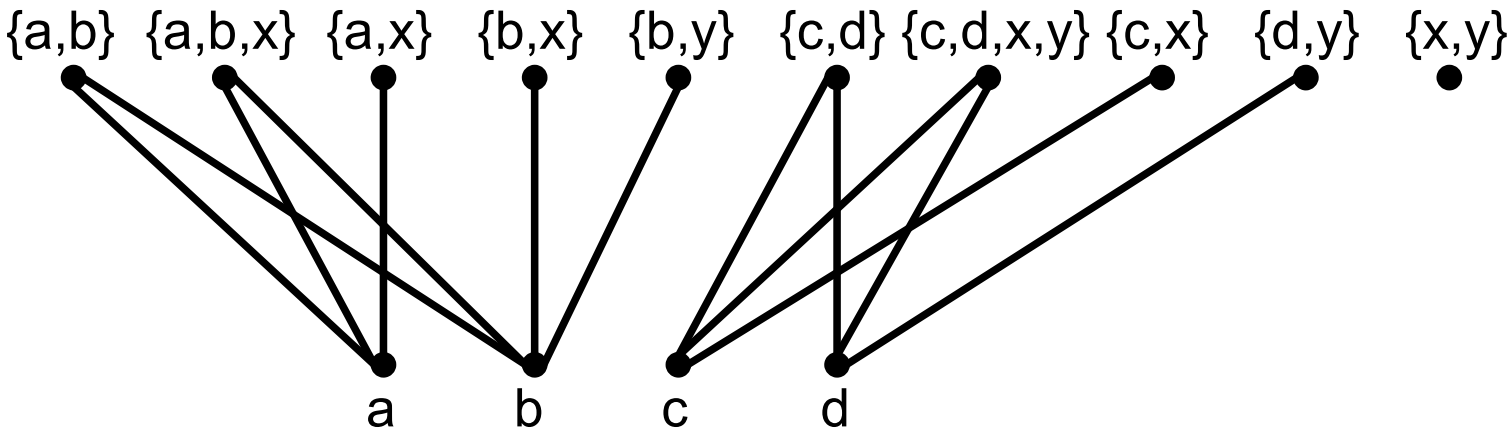
Supprimer le nombre minimum *t* de sommets-taxons tels que le graphe des caractères est un graphe “sans M”

L'ensemble minimum de conflits

Conflit : clades ni inclus ni disjoints



Graphe des caractères :

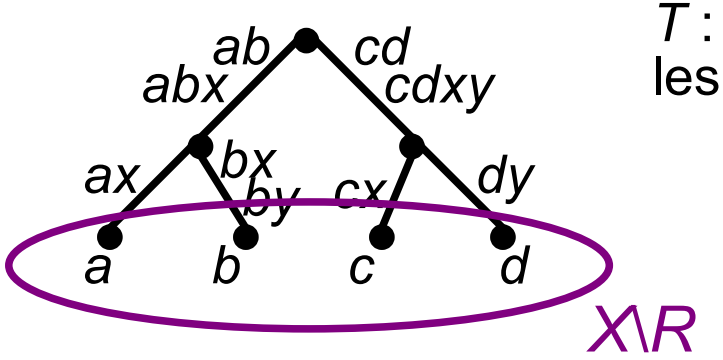


Supprimer le nombre minimum *t* de sommets-taxons tels que le graphe des caractères est un graphe “sans M” :

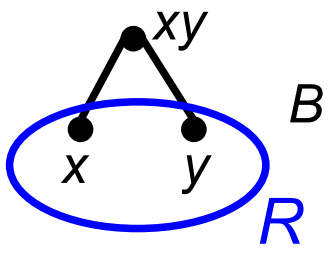
- algorithme FPT basique en $O^*(3^t)$
- algorithme FPT 3-Hitting-Set en $O^*(2{,}076^t)$

L'attachement minimum

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



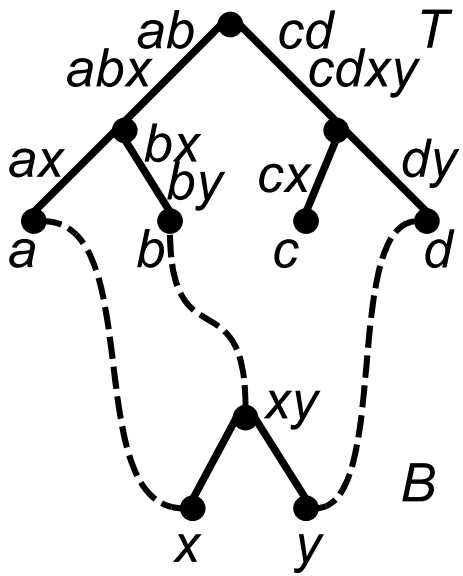
T : arbre représentant les clades sur *X**R*



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

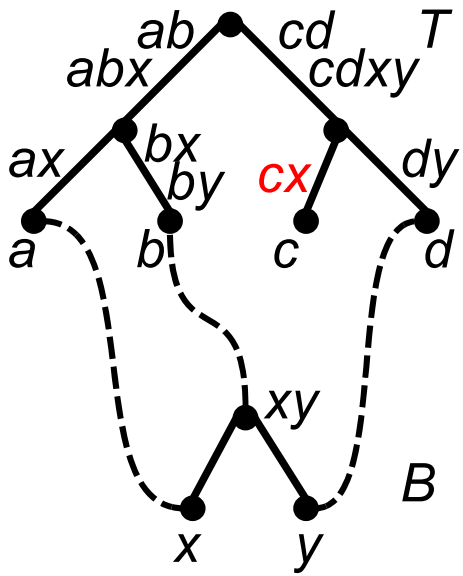
Problème :
Attacher *T* à *B* avec le **minimum de liens**.

L'attachement minimum



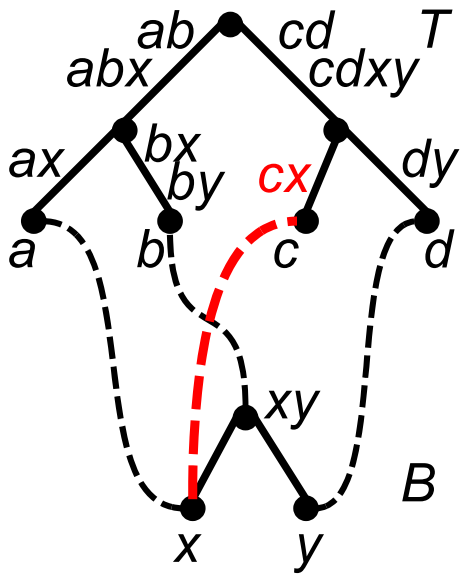
C1 - Satisfaction des clades de *T* :
Pour tout arc *e* de l'arbre *T* et tout taxon *r* de *R* contenu dans un clade de *e*, il existe un lien depuis un des descendants de *e* dans *T* vers le noeud correspondant à *r* dans *B*, ou un noeud de *B* correspondant à un clade maximal qui contient *r.*

L'attachement minimum



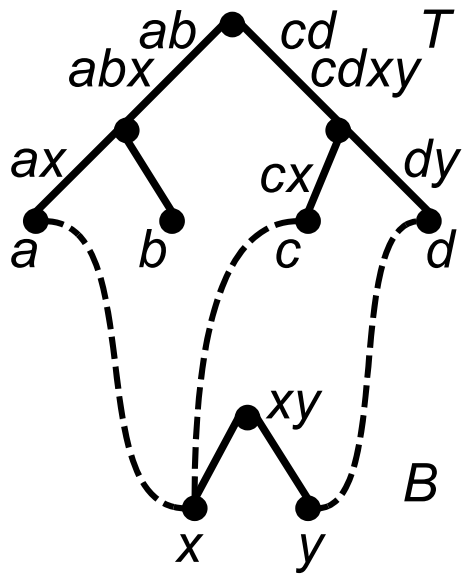
C1 - Satisfaction des clades de *T* :
Pour tout arc *e* de l'arbre *T* et tout taxon *r* de *R* contenu dans un clade de *e*, il existe un lien depuis un des descendants de *e* dans *T* vers le noeud correspondant à *r* dans *B*, ou un noeud de *B* correspondant à un clade maximal qui contient *r.*

L'attachement minimum



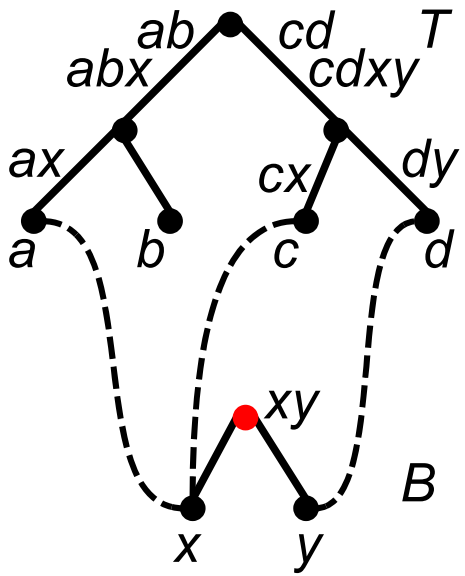
C1 - Satisfaction des clades de *T* :
Pour tout arc *e* de l'arbre *T* et tout taxon *r* de *R* contenu dans un clade de *e*, il existe un lien depuis un des descendants de *e* dans *T* vers le noeud correspondant à *r* dans *B*, ou un noeud de *B* correspondant à un clade maximal qui contient *r.*

L'attachement minimum



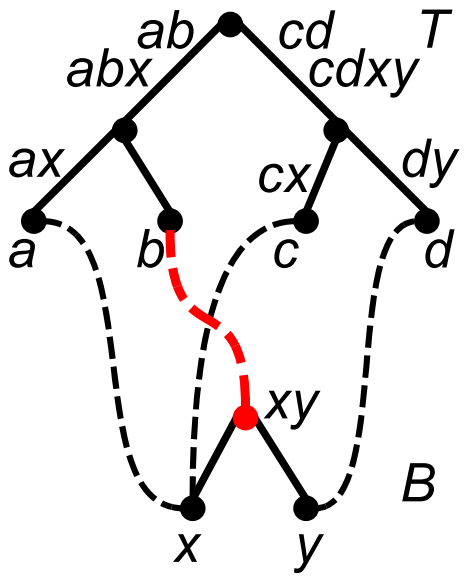
C2 - Satisfaction de la paternité des noeuds de *B* :
Tout noeud de *B* correspondant à plus d'un taxon est relié à un noeud de *T* par un lien exactement.

L'attachement minimum



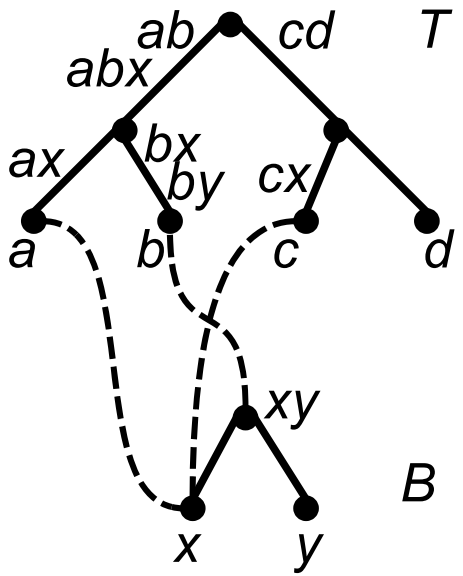
C2 - Satisfaction de la paternité des noeuds de *B* :
Tout noeud de *B* correspondant à plus d'un taxon est relié à un noeud de *T* par un lien exactement.

L'attachement minimum



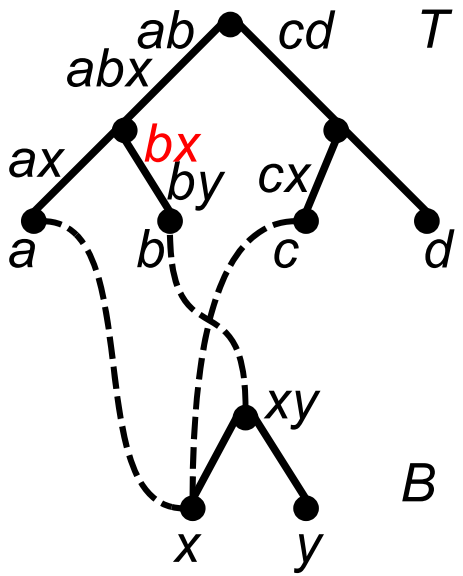
C2 - Satisfaction de la paternité des noeuds de *B* :
Tout noeud de *B* correspondant à plus d'un taxon est relié à un noeud de *T* par un lien exactement.

L'attachement minimum



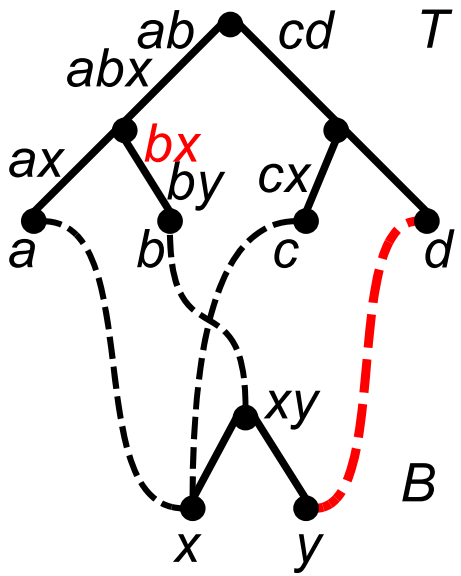
C3 – Absence de parasites des clades de *T* dans *B* :
Pour tout arc *e* de *T*, si un clade correspondant à *e* ne contient pas un taxon *r* de *R*, il existe un chemin d'un noeud qui ne descend pas de *e* vers le noeud associé à *r*.

L'attachement minimum



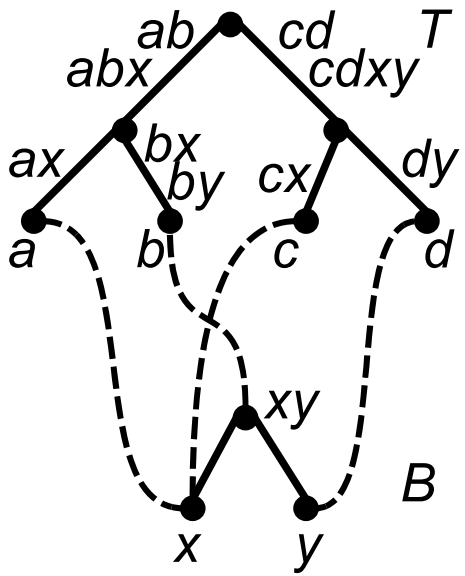
C3 – Absence de parasites des clades de *T* dans *B* :
Pour tout arc *e* de *T*, si un clade correspondant à *e* ne contient pas un taxon *r* de *R*, il existe un chemin d'un noeud qui ne descend pas de *e* vers le noeud associé à *r*.

L'attachement minimum



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

L'attachement minimum



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

L'attachement minimum

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

NP-complet

réduction depuis SetCover

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

réduction depuis SetCover

Algorithmes :
- Séparation et évaluation

(branch-and-bound, implémenté dans Dendroscope 2)

- Programme linéaire en nombres entiers

Reconstruction depuis les clades souples

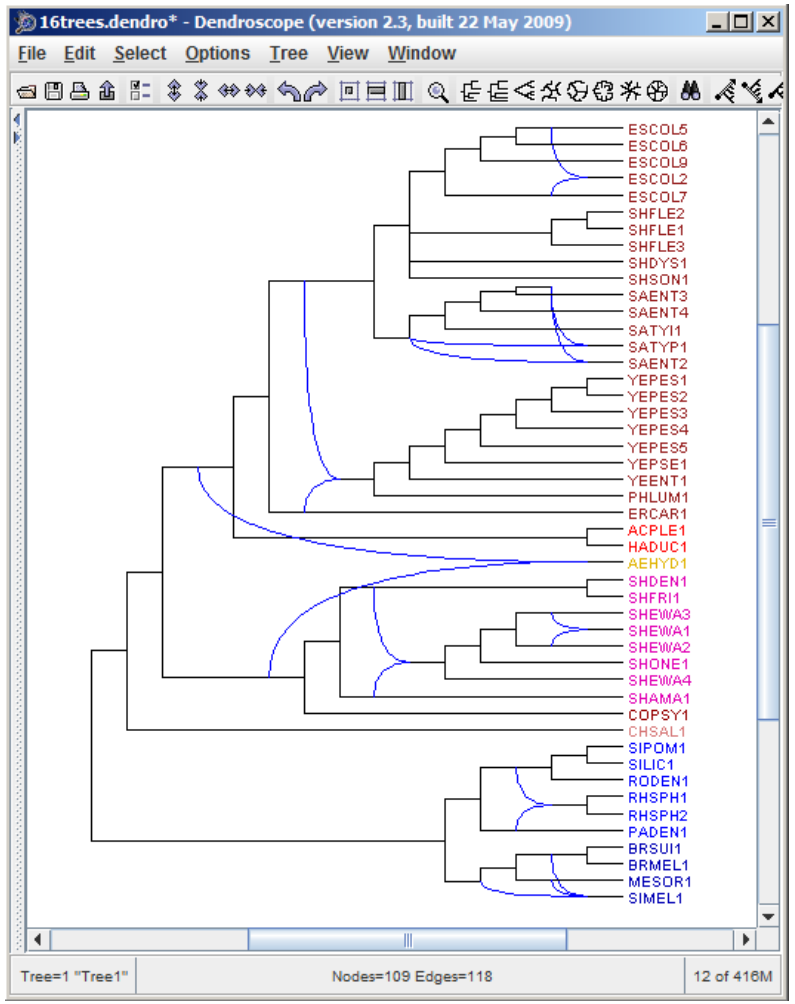
{arbres}

{clades}

N' réseau à une couche de réticulation

16 arbres de gènes de 46 espèces de bactéries
Réseau “galled network” des clades apparaissant dans 20% des arbres :

Dendroscope



Enterobacteriales
Pasteurellales
Aeromonadales
Alteromonadales
Oceanospirillales
Rhodobacterales
Rhizobiales

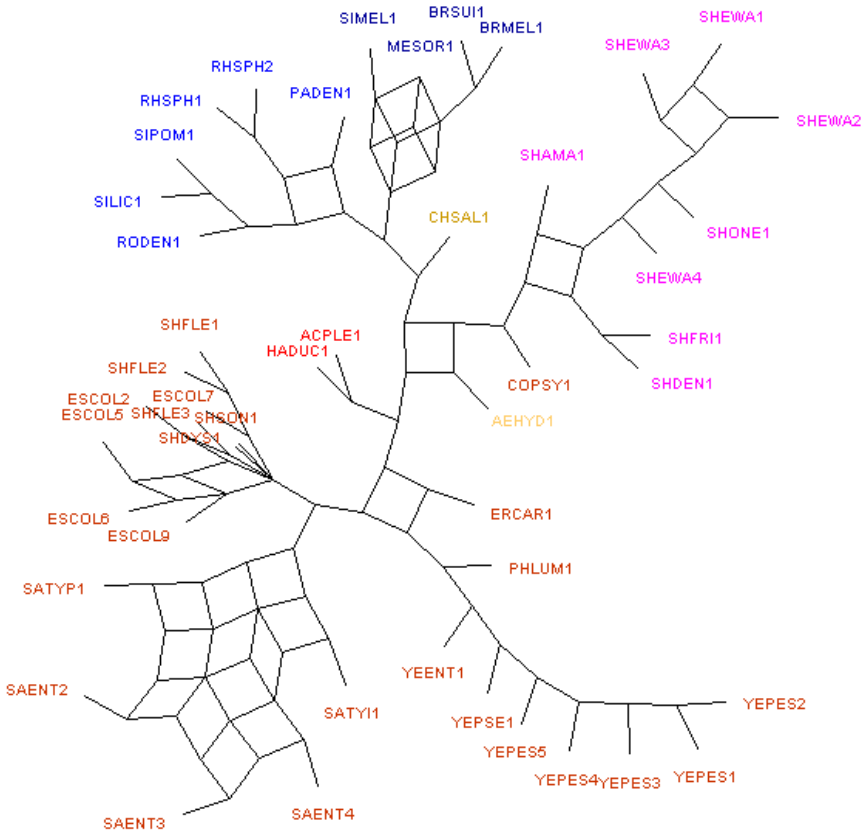
Reconstruction depuis les clades souples

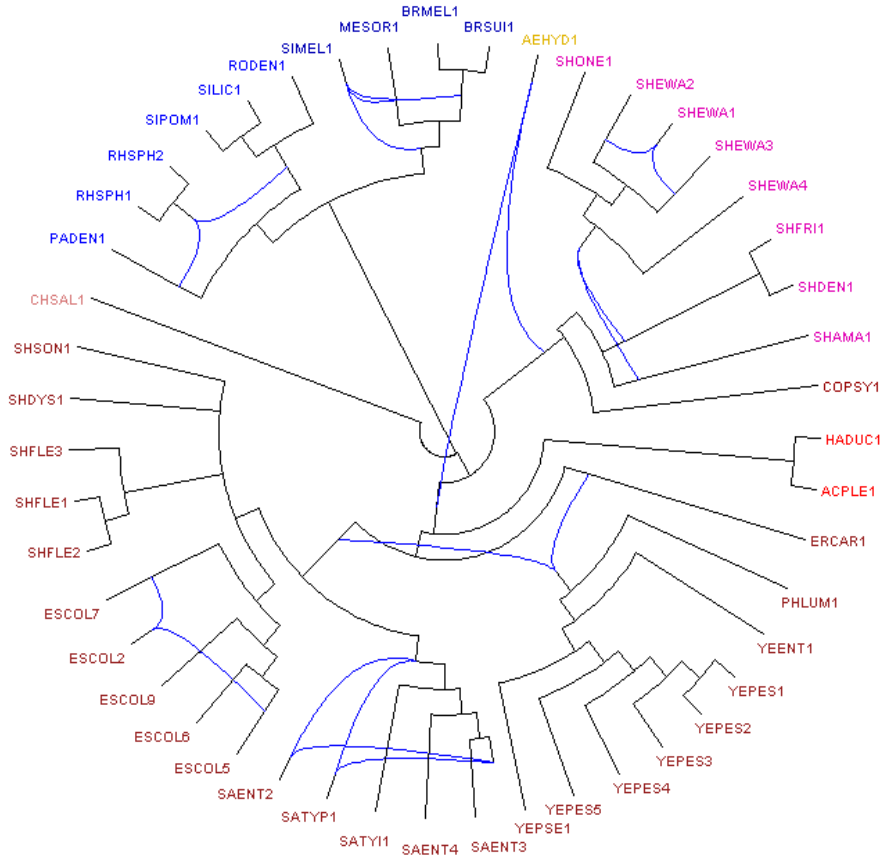
{arbres}

SplitsTree **Dendroscope**

Filtre : présence dans 20% des arbres

{splits} {clades}





Données de la base HOGENOM, projet ANR PhylAriane

Plan

- Les réseaux phylogénétiques
- L'arbre en filigrane
- Motivations de l'approche combinatoire
- Méthodes de reconstruction combinatoire
- **Limites des méthodes combinatoires**
- Perspectives

Limites des méthodes combinatoires

Restrictions sur les réseaux reconstruits :
- n'empêchent pas l'**explosion combinatoire**
- **trop contraignantes** pour certains modèles

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

Confrontation à des données réelles :
- gérer le **bruit**
- gérer le **silence**

Limites des méthodes combinatoires

Restrictions sur les réseaux reconstruits :
- n'empêchent pas l'**explosion combinatoire**
- **trop contraignantes** pour certains modèles

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

Confrontation à des données réelles :
- gérer le **bruit**
- gérer le **silence**

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

Borne inférieure du nombre de générateurs

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 du nombre de générateurs

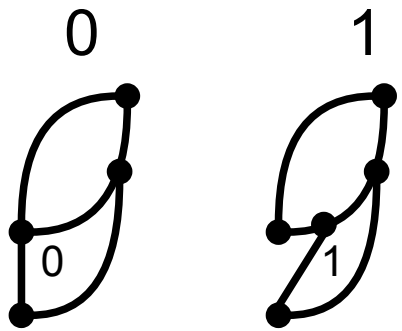
Borne inférieure :

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

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.

$k = 1$



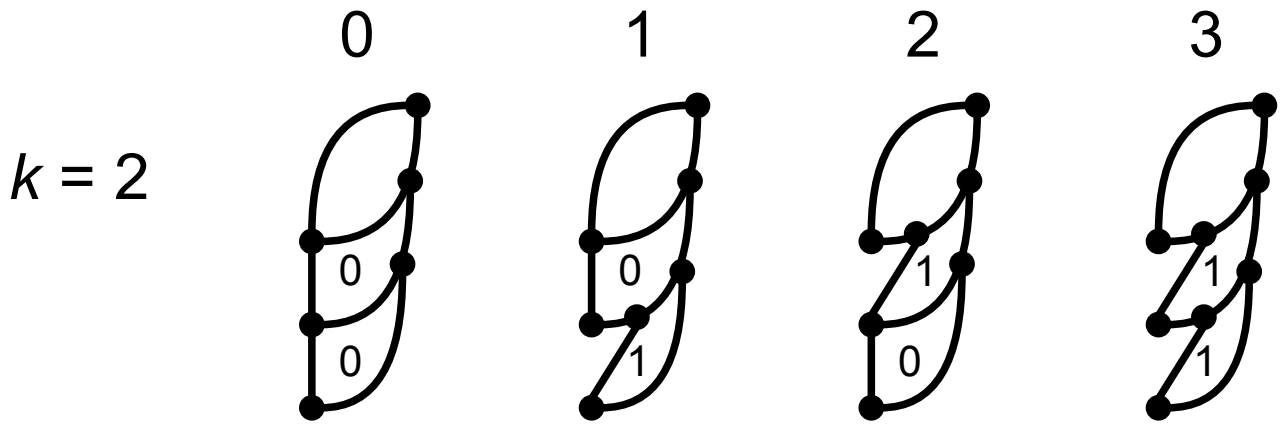
Borne inférieure du nombre de générateurs

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.



Limites des méthodes combinatoires

Restrictions sur les réseaux reconstruits :
- n'empêchent pas l'**explosion combinatoire**
- **trop contraignantes** pour certains modèles

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

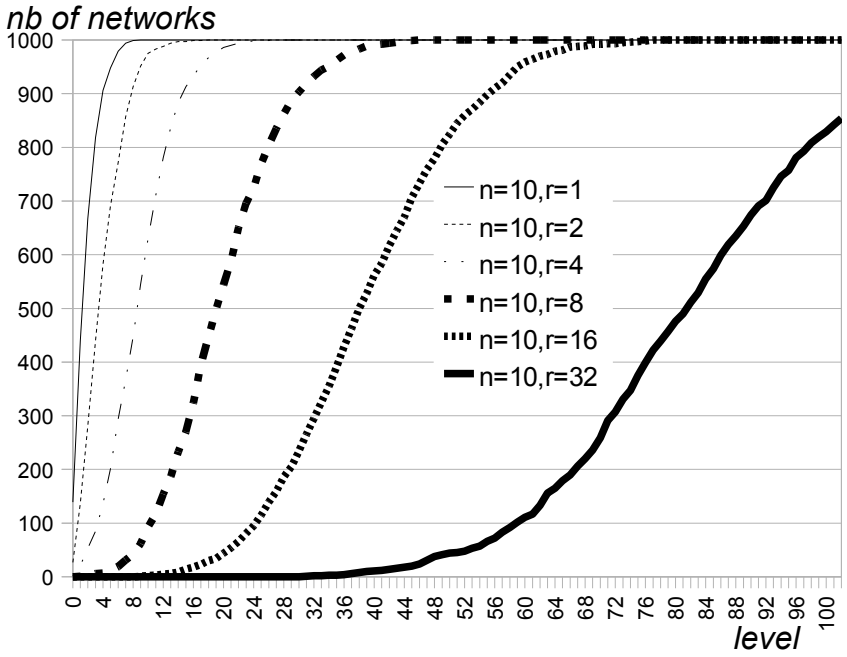
Confrontation à des données réelles :
- gérer le **bruit**
- gérer le **silence**

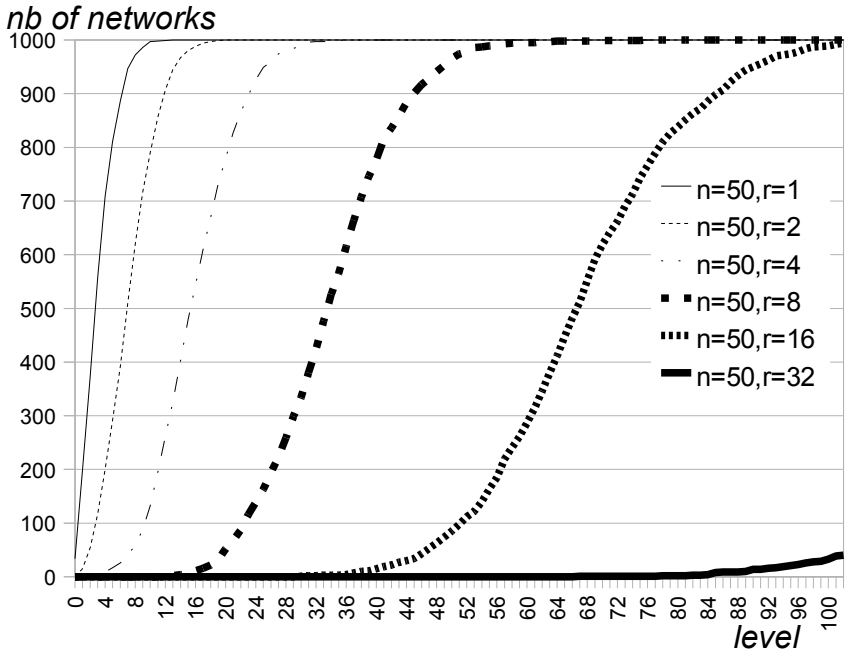
Simulations de réseaux de niveau *k*

Simulation de 1000 réseaux phylogénétiques selon le modèle coalescent avec recombinaison.

Arenas, Valiente, Posada 2008
Programme Recodon

Combien sont de niveau 1, 2, 3 ?

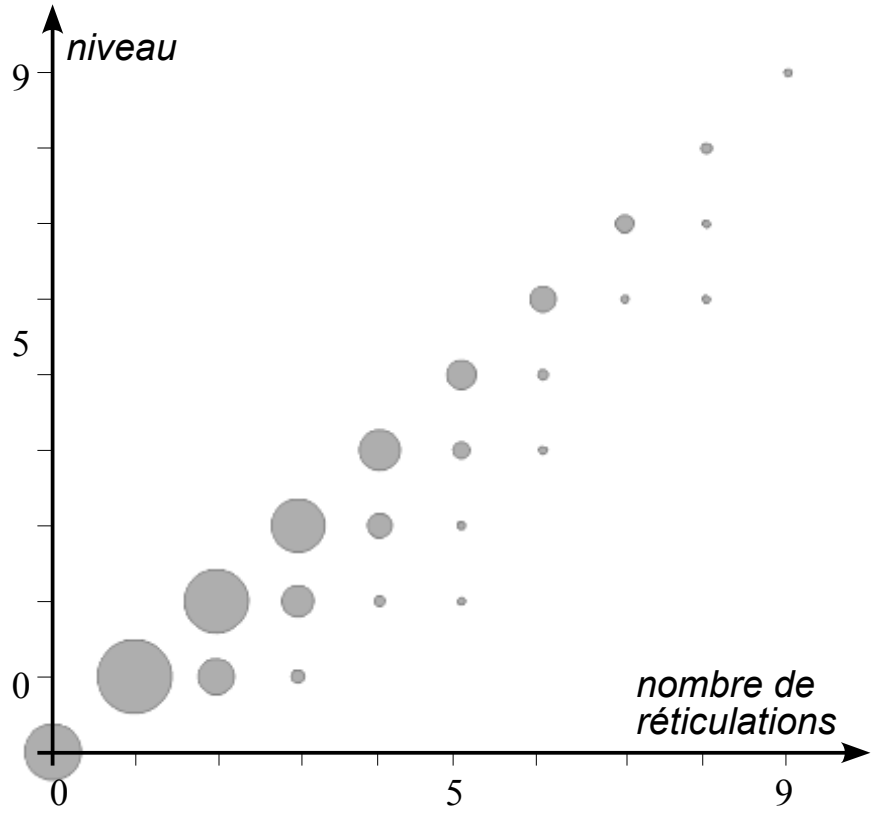




Simulations de réseaux de niveau *k*

Simulation de 1000 réseaux phylogénétiques selon le modèle coalescent avec recombinaison.

Lien entre le niveau et le nombre de réticulations :



Limites des méthodes combinatoires

Restrictions sur les réseaux reconstruits :
- n'empêchent pas l'**explosion combinatoire**
- **trop contraignantes** pour certains modèles

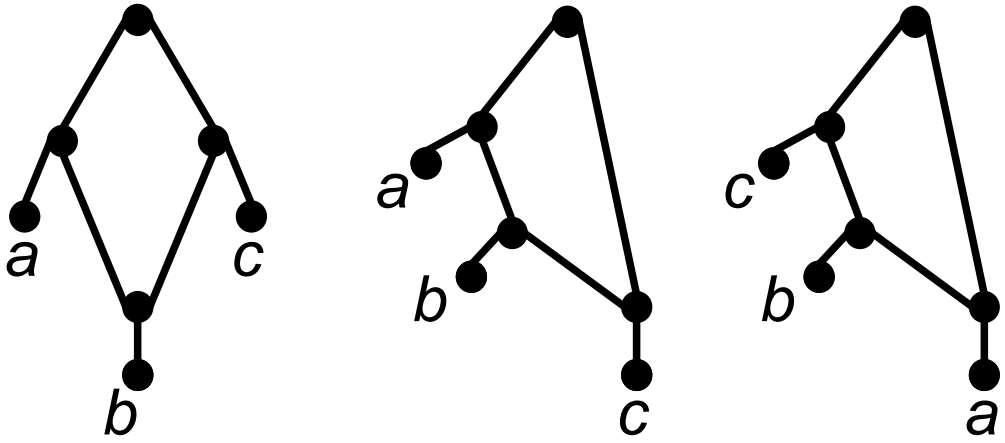
Ambiguïté de la reconstruction, même à partir de données complètes et correctes.

Confrontation à des données réelles :
- gérer le **bruit**
- gérer le **silence**

Limites des approches combinatoires

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

Gambette & Huber, 2009

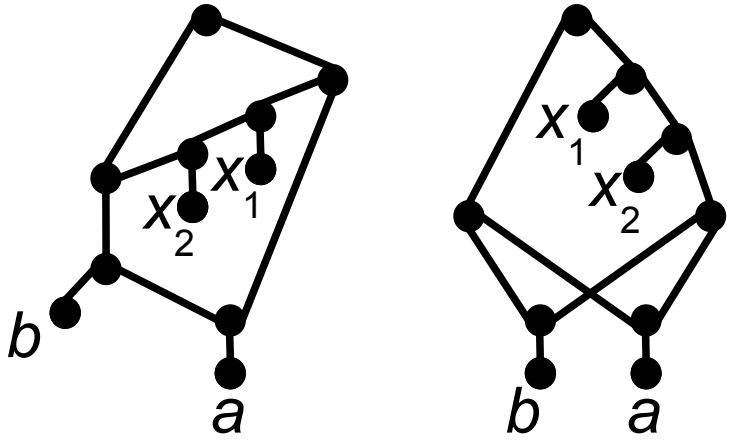


Caractérisation des réseaux de niveau 1 ayant exactement le même ensemble d'arbres, de triplets, de clades.

Limites des approches combinatoires

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

Gambette & Huber, 2009

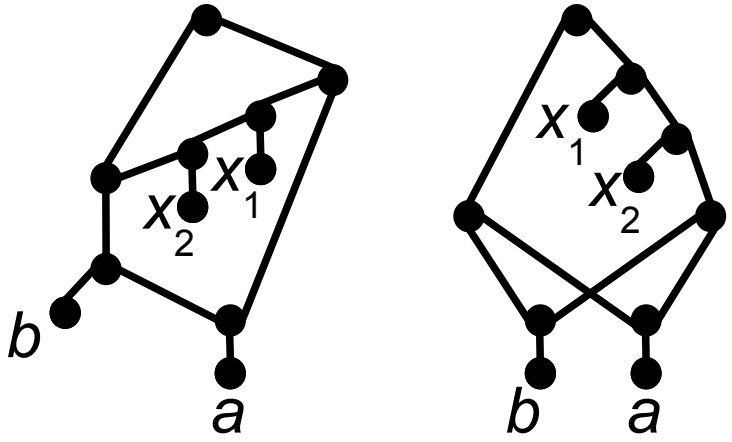


2 réseaux de niveau 2 avec le même ensemble de triplets

Limites des approches combinatoires

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

Gambette & Huber, 2009



2 réseaux de niveau 2 avec le même ensemble de triplets

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

Limites des méthodes combinatoires

Restrictions sur les réseaux reconstruits :
- n'empêchent pas l'**explosion combinatoire**
- **trop contraignantes** pour certains modèles

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

Confrontation à des données réelles :
- gérer le **bruit**
- gérer le **silence**

Gestion du bruit dans les données

Approches de **filtres** :
Ne considérer que les clades, triplets, avec un bon support.

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

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

- arbres à partir de triplets

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

- réseaux de niveau 1 à partir de triplets

(en cours...)

Limites des méthodes combinatoires

Restrictions sur les réseaux reconstruits :
- n'empêchent pas l'**explosion combinatoire**
- **trop contraignantes** pour certains modèles

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

Confrontation à des données réelles :
- gérer le **bruit**
- gérer le **silence**

Gestion du 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**

Interface de sélection de taxons et de familles de gènes pour trouver :

un grand nombre d'arbres de gènes
sur un grand nombre de taxons

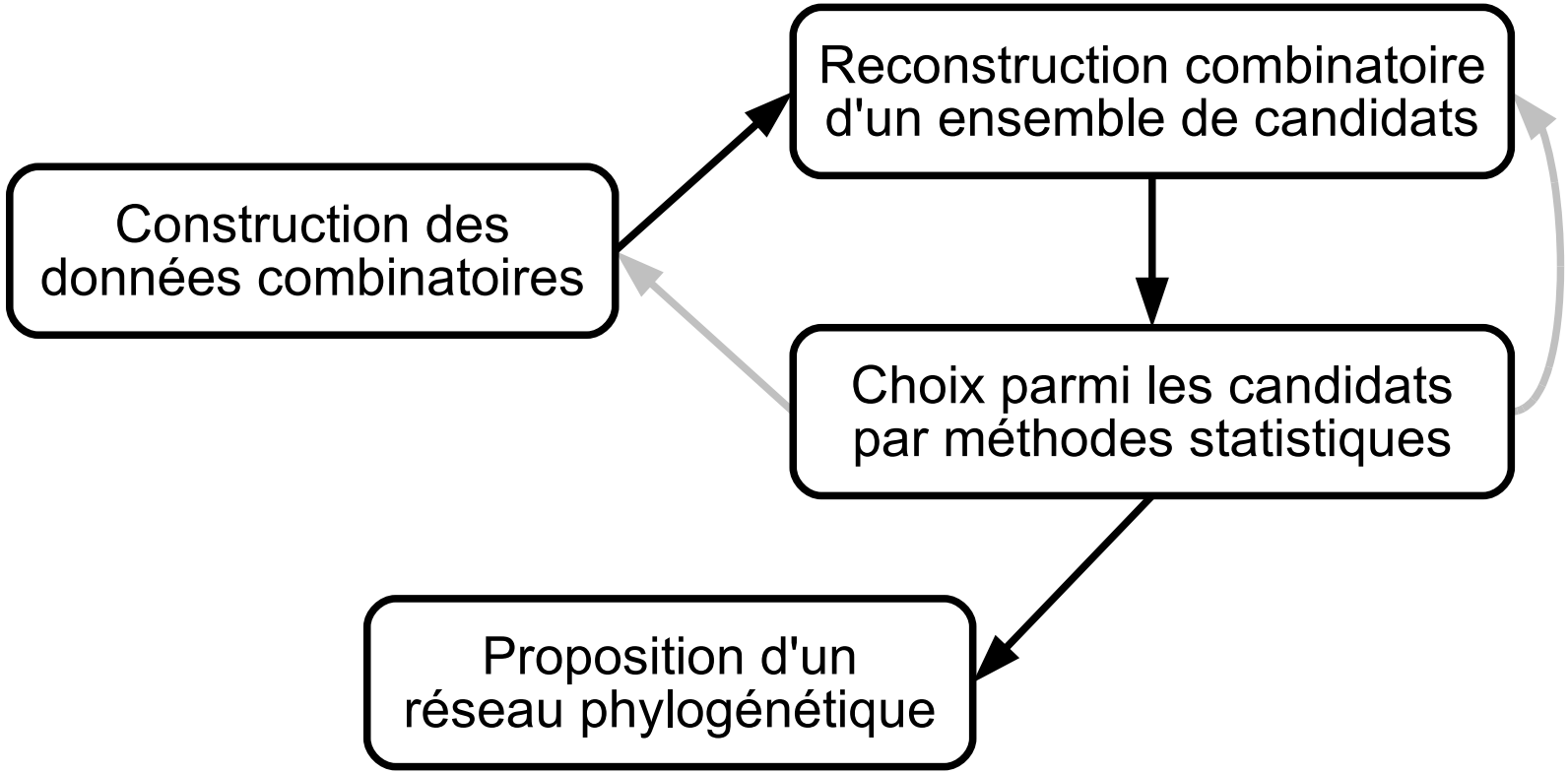
(en cours : rectangles maximaux,
sélection heuristique sur nuages arborés...)

Plan

- Les réseaux phylogénétiques
- L'arbre en filigrane
- Motivations de l'approche combinatoire
- Méthodes de reconstruction combinatoire
- Limites des méthodes combinatoires
- **Perspectives**

Perspectives

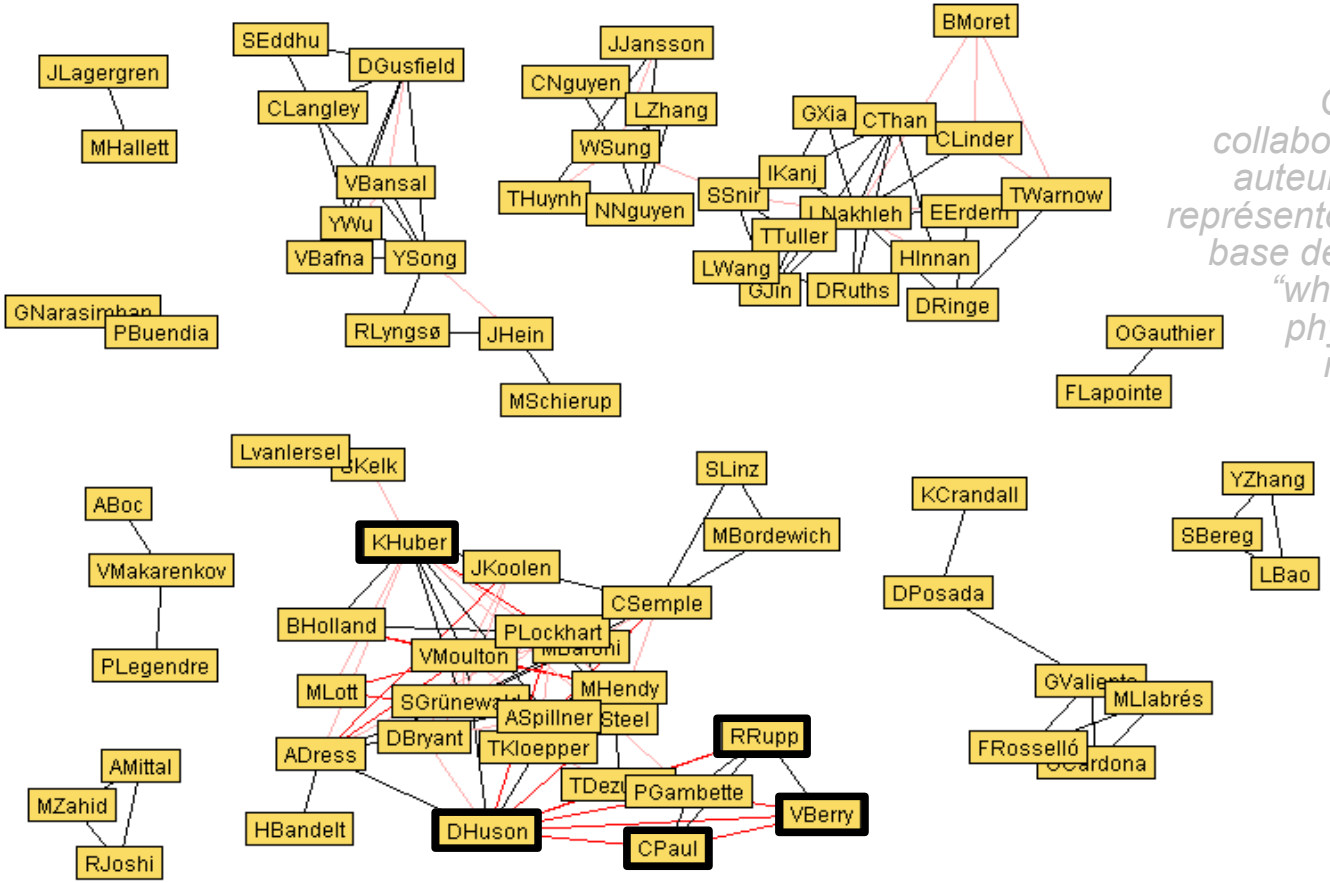
Reconstruction du réseau par un dialogue entre méthodes combinatoires et méthodes statistiques



Des questions ?

Merci pour votre attention !

Co-auteurs des résultats présentés :
- Vincent Berry, Christophe Paul (LIRMM, Université Montpellier 2)
- Katharina Huber (Université East Anglia, UK)
- Daniel Huson, Regula Rupp (Université de Tübingen, Allemagne)



Graphe de collaboration des auteurs les plus représentés dans la base de données "who's who in phylogenetic networks".