Réunion PhyI'ARIANE – LIRMM – 25/11/2008

Des réseaux phylogénétiques pour visualiser les différences entre arbres de gènes

Philippe Gambette

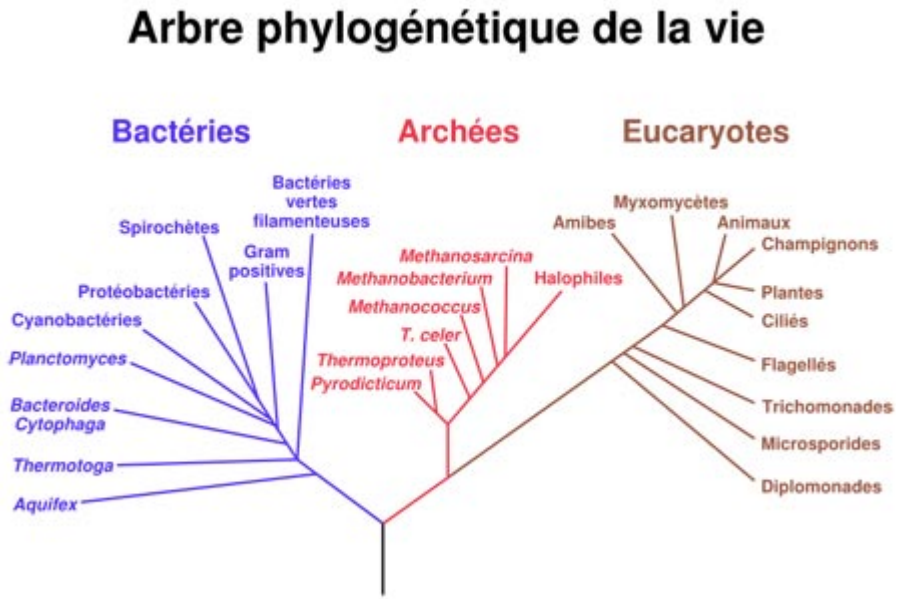
Plan

- Les types de réseaux phylogénétiques
- Reconstruction depuis des arbres
- Des sous-classes de réseaux
- Les arbres et leurs quadruplets/triplets, splits/clusters
- Reconstruction depuis des arbres multi-étiquetés
- Autres phénomènes à considérer

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.

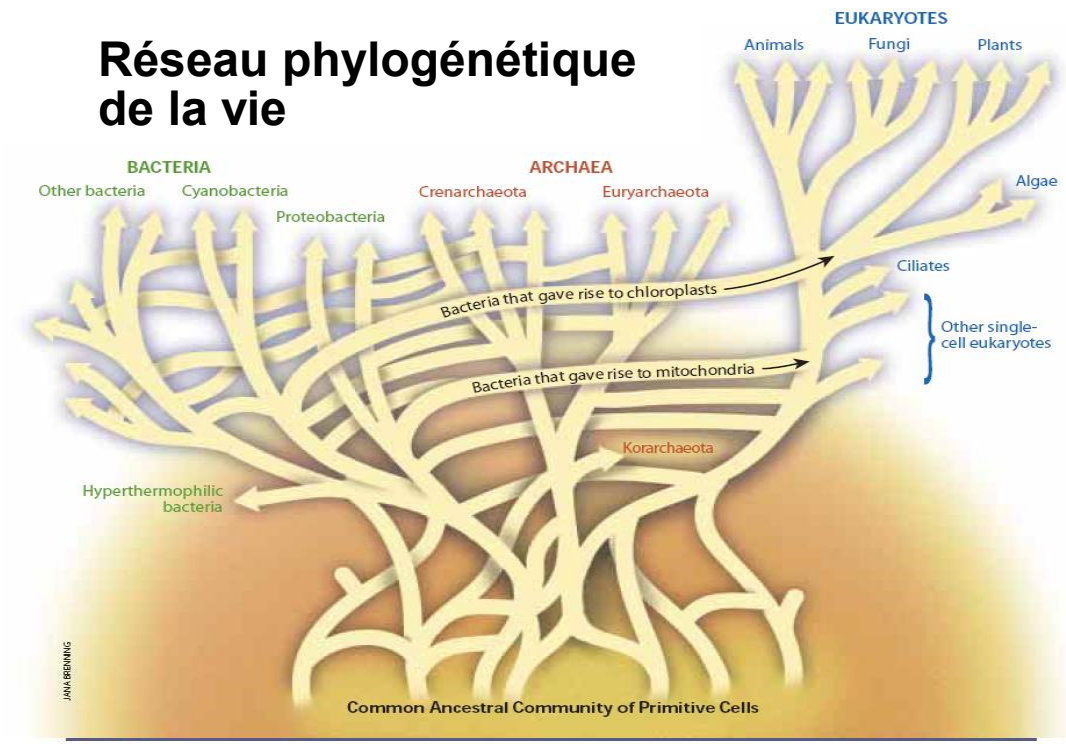


D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

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

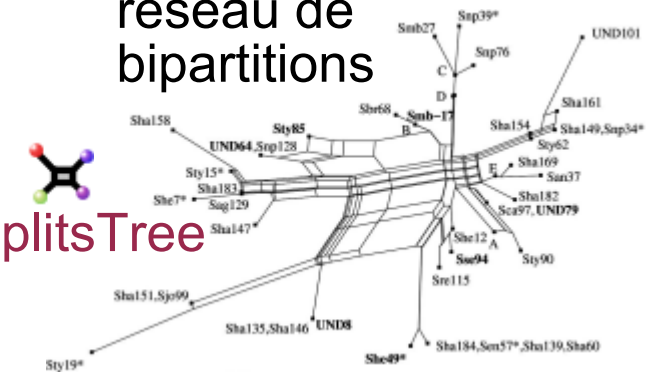
Réseau phylogénétique

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.

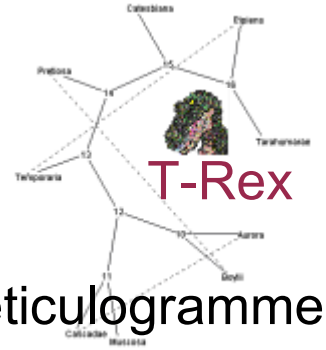
réseau de niveau 2

Level-2

réseau de bipartitions



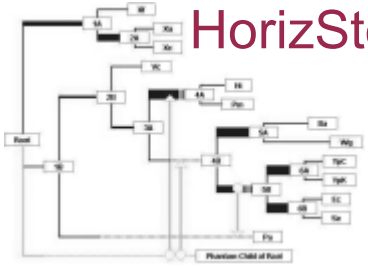
SplitsTree



T-Rex

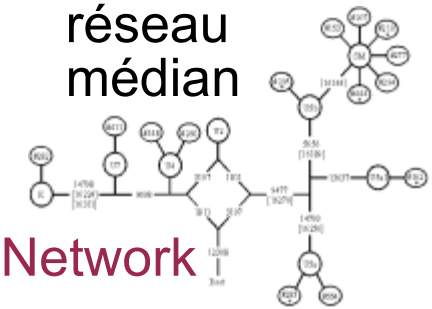
réticulogramme

diagramme de synthèse



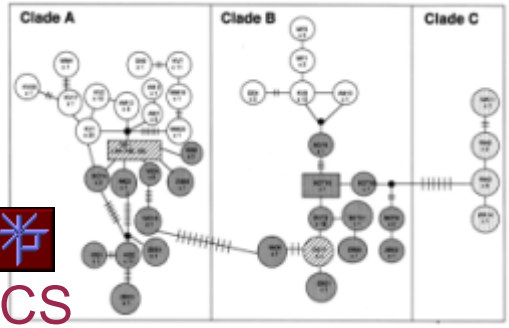
HorizStory

réseau médian



Network

réseau couvrant minimum

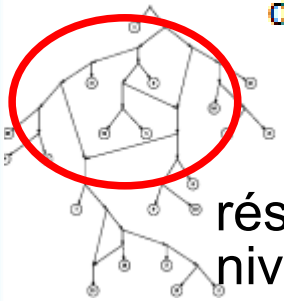


TCS

Les réseaux phylogénétiques

Réseau phylogénétique

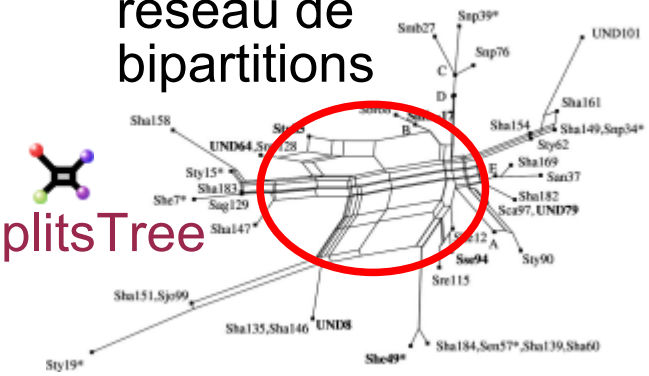
Un **réseau phylogénétique** désigne un graphe utilisé pour visualiser les relations liées à l'évolution entre des espèces ou des organismes. Il doit être employé quand interviennent des événements d'hybridations, de transferts horizontaux de gènes, ou de recombinaisons génétiques.



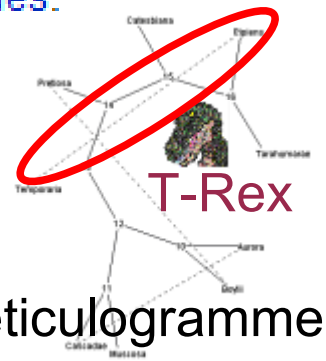
réseau de niveau 2

Level-2

réseau de bipartitions



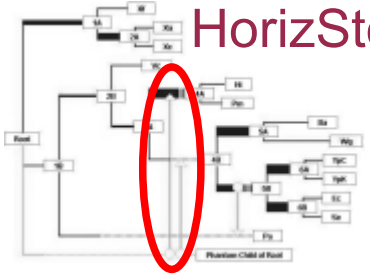
SplitsTree



T-Rex

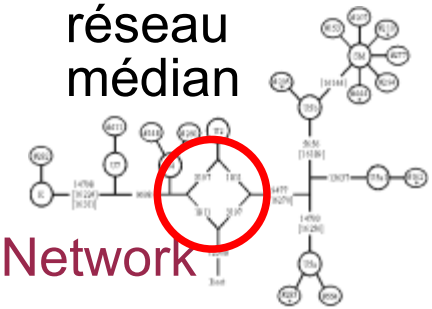
réticulogramme

diagramme de synthèse



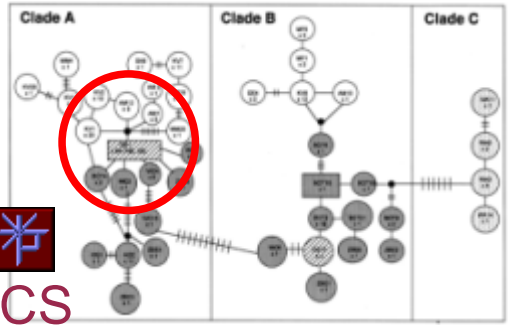
HorizStory

réseau médian



Network

réseau couvrant minimum



TCS

Réseaux abstraits ou explicites

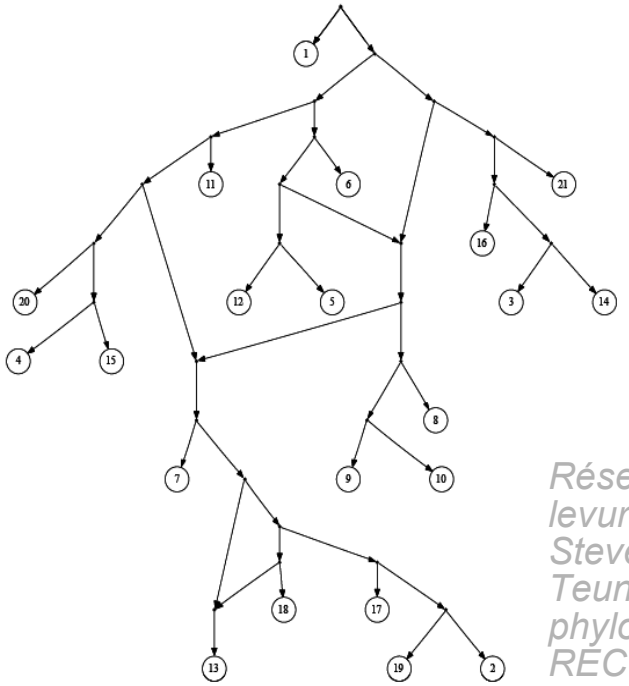
Un **réseau phylogénétique explicite** est un réseau phylogénétique dont tous les noeuds correspondent à des événements biologiques précis.

Un **réseau phylogénétique abstrait** reflète des signaux phylogénétiques sans nécessairement représenter explicitement des événements biologiques.

Réseaux abstraits ou explicites

Un **réseau phylogénétique explicite** est un réseau phylogénétique dont tous les noeuds correspondent à des événements biologiques précis.

Un **réseau phylogénétique abstrait** reflète des signaux phylogénétiques sans nécessairement représenter explicitement des événements biologiques.

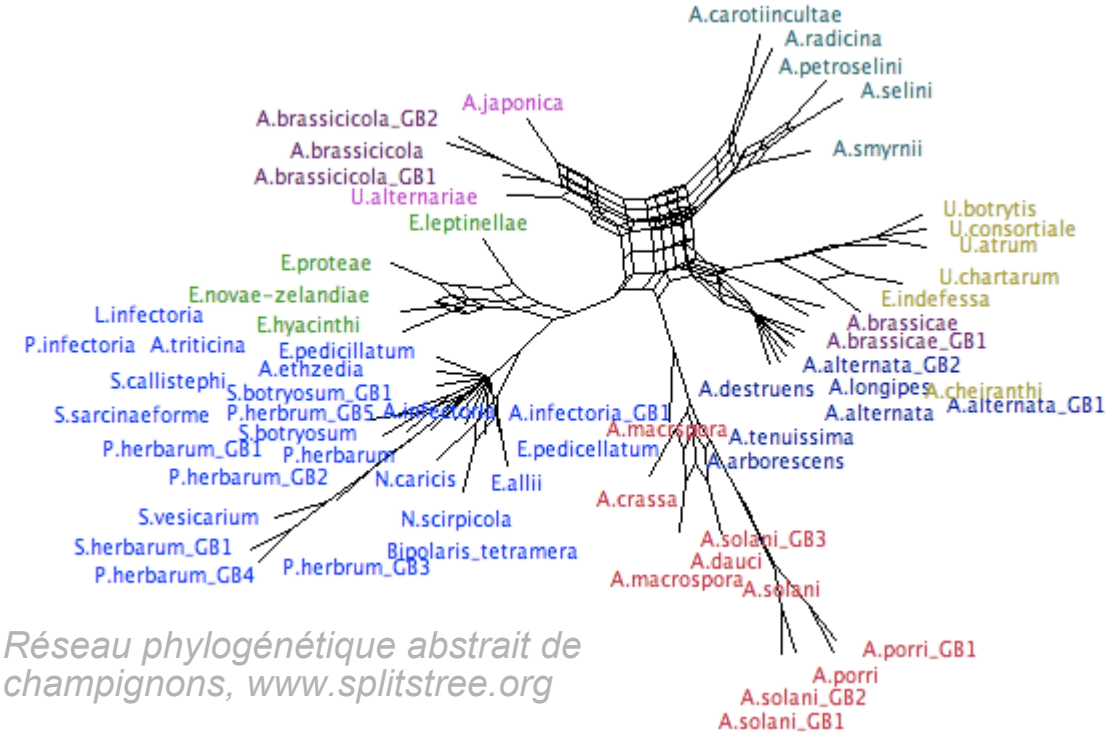


Réseau phylogénétique explicite de levures, Leo van Iersel, Judith Keijsper, Steven Kelk, Leen Stougie, Ferry Hagen, Teun Boekhout : Constructing level-2 phylogenetic networks from triplets. RECOMB'08

Réseaux abstraits ou explicites

Un **réseau phylogénétique explicite** est un réseau phylogénétique dont tous les noeuds correspondent à des événements biologiques précis.

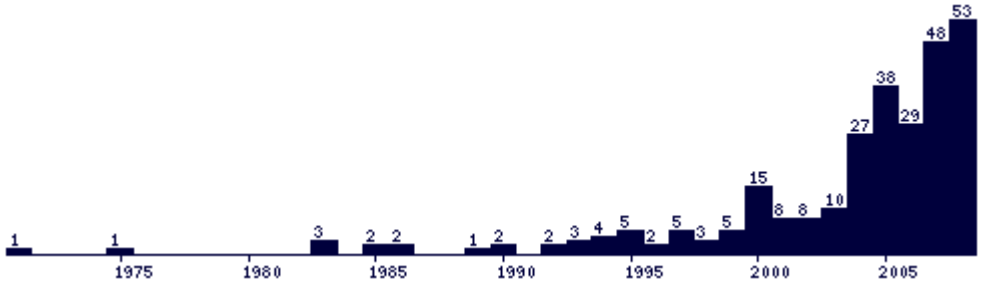
Un **réseau phylogénétique abstrait** reflète des signaux phylogénétiques sans nécessairement représenter explicitement des événements biologiques.



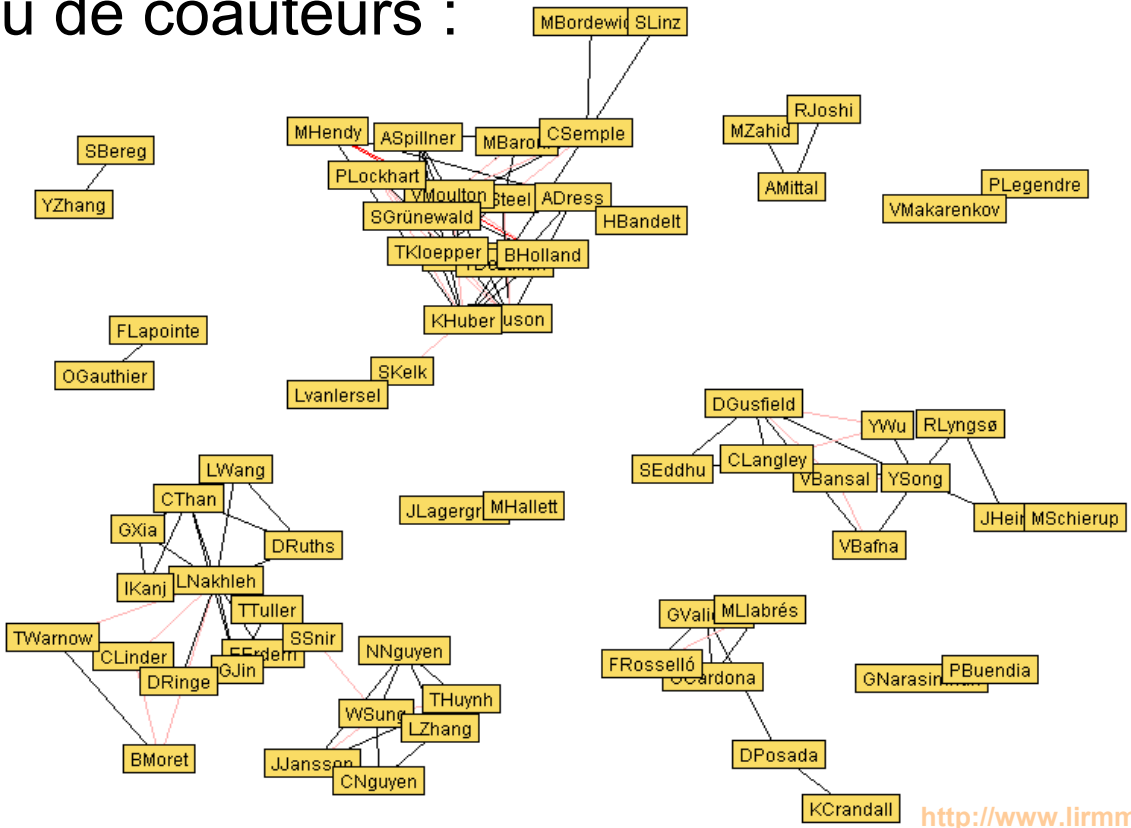
Réseau phylogénétique abstrait de champignons, www.splitstree.org

Réseaux phylogénétiques : sujet d'actualité

Publications sur les réseaux phylogénétiques :



Réseau de coauteurs :



Réseaux phylogénétiques : sujet d'actualité

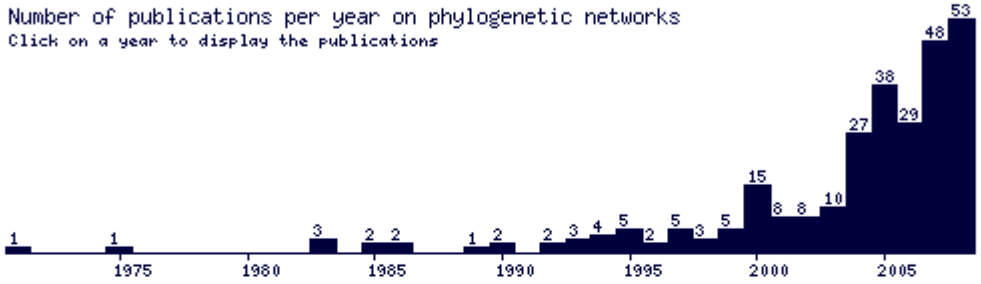
Who is Who in Phylogenetic Networks - Articles, Authors & Programs RSS

Index | Browse | Contribute! | My selection

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

Publications - Index (All 277 publications) Selection by: Year | Category | Keyword | Author

Selection by Year



Selection by Category

Article (Journal) (159) **InProceedings** (77) **InBook** (13)
PhdThesis (11) **Misc** (17) **Programs** (37)

Selection by Keyword

abstract-network(21) approximation(4) APX-hard(1) ARG(5) block-realization(1) bootstrap(1) bound(3) branch-and-bound(1) cactus-graph(1) characterization(3) clustering(2) coalescent(5) consensus(8) consistency(2) construction(2) distance-between-networks(19) diversity(1) enumeration(1) evaluation(22) explicit-network(23) FPT(6) from-clusters(3) from-distances(18) from-multilabeled-tree(2) from-network(5) from-quartets(5) from-rooted-trees(26) from-sequences(17) from-splits(10) from-trees(7) from-triplets(10) from-unrooted-trees(7) galled-network(2) galled-tree(26) generation(5) haplotype-network(2) haplotyping(1) heuristic(4) HMM(1) hybridization(16) labeling(3) lateral-gene-transfer(13) level-f-phylogenetic-network(7) likelihood(7) lineage-sorting(1) MASN(4) median-network(14) MedianJoining(2) minimum-number(8) minimum-spanning-network(2) mu-distance(1) NeighborNet(9) nested-network(2) netting(3) normal-network(1) NP-complete(13) optimal-realization(2) parsimony(10) perfect(5) **phylogenetic-network(151)** **phylogeny(155)** polynomial(26) Program-Arlequin(5) Program-Beagle(3) Program-Bio-PhyloNetwork(4) Program-CombineTrees(2) Program-Dendroscope(5) Program-EEEP(2) Program-GalledTree(1) Program-HapBound(1) Program-HorizStory(2) Program-HybridNumber(1) Program-LatTrans(3) Program-Level2(1) Program-Marlon(1) Program-

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

Réseaux phylogénétiques : sujet d'actualité

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 types de réseaux phylogénétiques
- **Reconstruction depuis des arbres**
- Des sous-classes de réseaux
- Les arbres et leurs quadruplets/triplets, splits/clusters
- Reconstruction depuis des arbres multi-étiquetés
- Autres phénomènes à considérer

Reconstruction depuis des arbres

Problème :
reconstruire l'arbre des espèces à partir d'arbres de gènes ?

Méthodes de **superarbre**

(MRP – Baum & Ragan 1992,
MinCut – Semple & Steel 2000,
AncestralBuild – Berry & Bryant 2006,
PhySIC_IST – Scornavacca, Berry, Douzery & Ranwez 2008,
...)

Méthodes de **super-réseau**
- combiner **deux** arbres **enracinés** avec le minimum *r* de réticulations ?

NP-complet (Bordewich & Semple, 2007)
FPT en *r* (Bordewich, Linz, St John & Semple, 2007,
algorithme en cours d'implémentation)

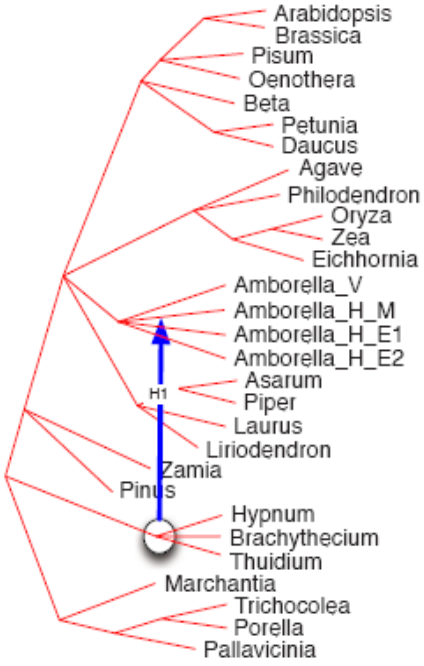
- tester si un arbre est contenu dans un réseau ?

NP-complet, **FPT** (Kanj, Nakhleh, Than, Xia, TCS, 2008)

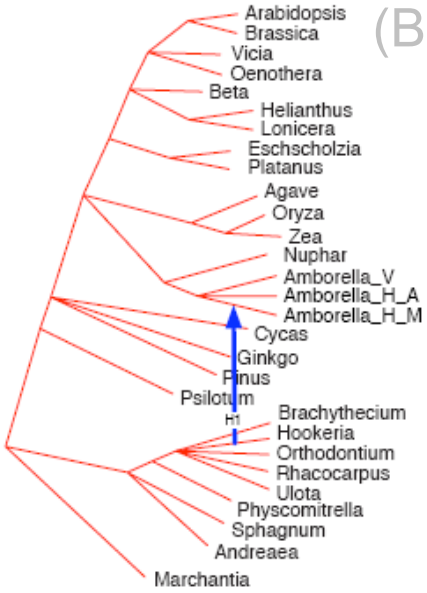
Reconstruction depuis des arbres

Problème :
reconstruire l'arbre des espèces à partir d'arbres de gènes ?

Méthodes **SPR** pour les transferts horizontaux



The *cox2* gene data set



The *nad5* gene data set

(Beiko & Hamilton, EEEP, BMCEB, 2006
Than, Jin & Nakhleh,
RIATA-HGT, RecombCG'08)

Attention : $d_{SPR}(T,T') \leqslant h(T,T')$

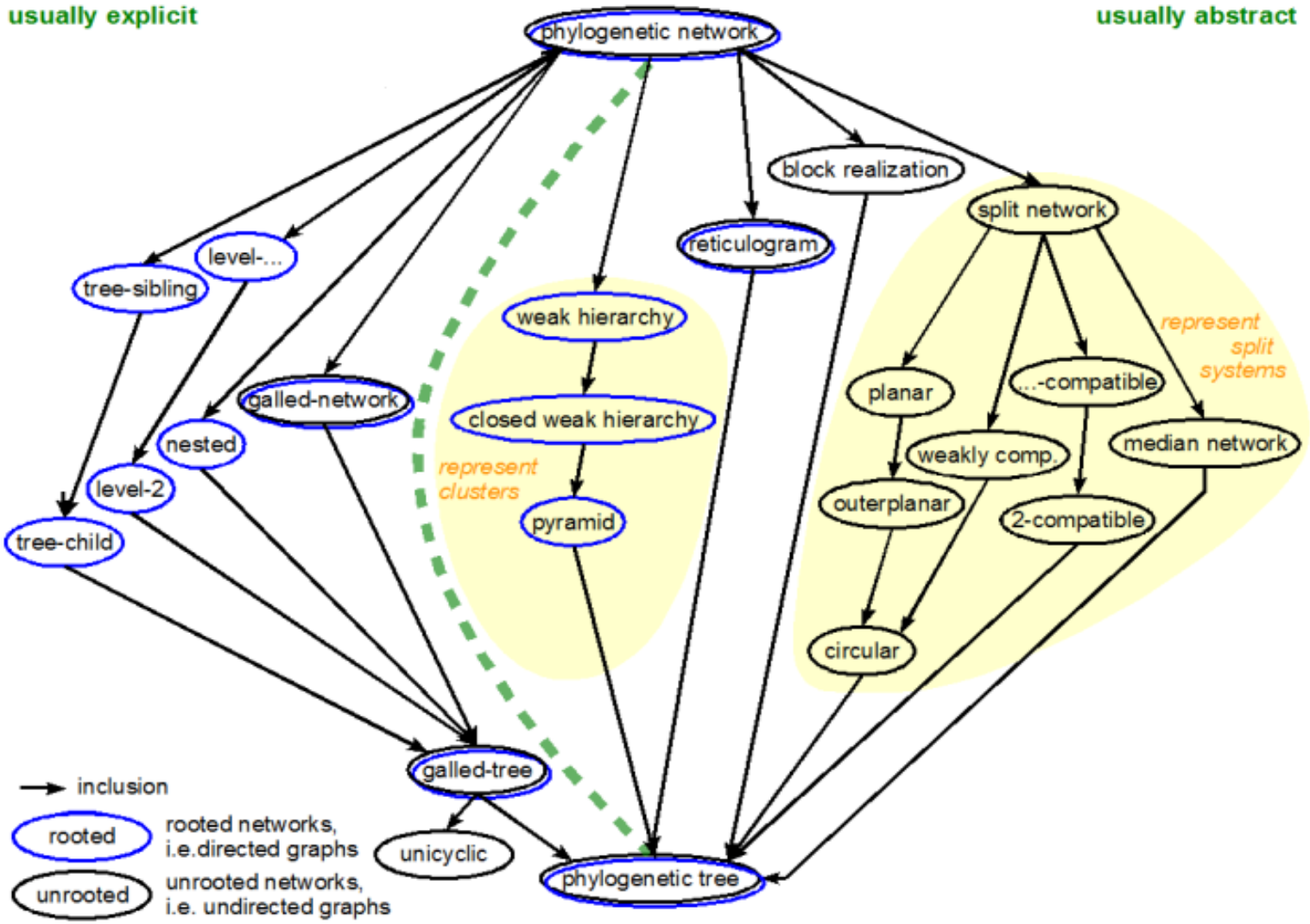
(Baroni, Grünewald, Moulton & Semple, 2005)

Plan

- Les types de réseaux phylogénétiques
- Reconstruction depuis des arbres
- **Des sous-classes de réseaux**
- Les arbres et leurs quadruplets/triplets, splits/clusters
- Reconstruction depuis des arbres multi-étiquetés
- Autres phénomènes à considérer

Hiérarchie de sous-classes de réseaux

Réconciliation d'arbres en un réseau : **problème difficile**
plus simple sur certaines sous-classes de réseaux ?



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

Hiérarchie de sous-classes de réseaux

Problème :
restrictions sur les réseaux **pertinentes** ?

Table 1: Number of simulated networks falling in each class as a function of the recombination rate $\rho = 0, 1, 2, 4, 8, 16, 32$, for sample size $n = 10$.

Network class	Recombination rate 0	1	2	4	8	16	32
Regular	1,000	200	58	5	0	0	0
Tree-sibling	1,000	832	514	151	14	0	0
Tree-child	1,000	560	205	39	1	0	0
Galled-trees	1,000	440	137	21	1	0	0
Trees	1,000	139	27	1	0	0	0

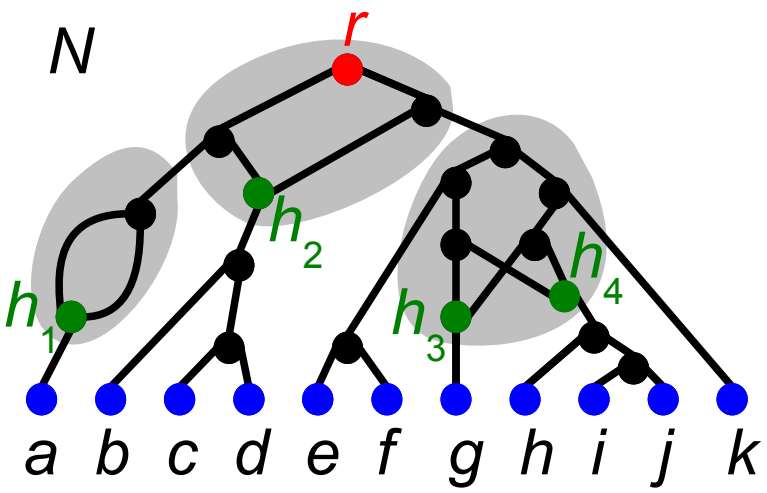
Table 2: Number of simulated networks falling in each class as a function of the recombination rate $\rho = 0, 1, 2, 4, 8, 16, 32$, for sample size $n = 50$.

Network class	Recombination rate 0	1	2	4	8	16	32
Regular	1,000	57	1	0	0	0	0
Tree-sibling	1,000	784	469	101	2	0	0
Tree-child	1,000	463	126	9	0	0	0
Galled-trees	1,000	161	5	0	0	0	0
Trees	1,000	34	0	0	0	0	0

Arenas, Valiente, Posada : Characterization of Phylogenetic Reticulate Networks based on the Coalescent with Recombination, Molecular Biology and Evolution, to appear.

Réseaux phylogénétiques de niveau *k*

Un **réseau phylogénétique de niveau *k*** peut être vu comme un arbre de blobs, où chaque blob contient **au plus *k*** sommets réticulés



N a niveau 2.

Un **blob** est une composante **biconnexe** maximale du graphe non orienté sous-jacent, c'est à dire un sous-graphe maximal non déconnecté par la suppression d'un sommet quelconque.

Plan

- Les types de réseaux phylogénétiques
- Reconstruction depuis des arbres
- Des sous-classes de réseaux
- **Les arbres et leurs quadruplets/triplets, splits/clusters**
- Reconstruction depuis des arbres multi-étiquetés
- Autres phénomènes à considérer

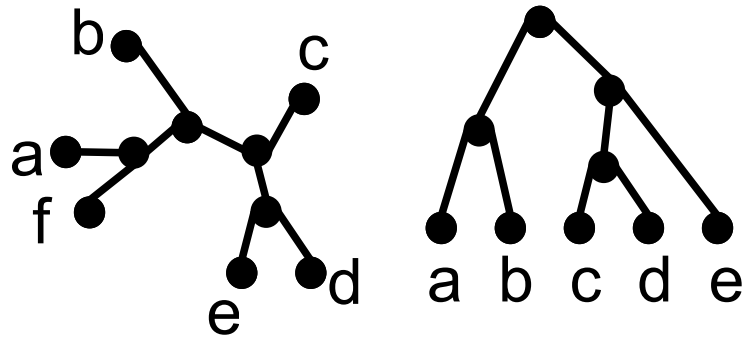
Triplets/quadruplets, splits/clusters

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

Idée :
reconstuire un réseau contenant tous les :

triplets
quadruplets
clusters
splits

des arbres en entrée ?

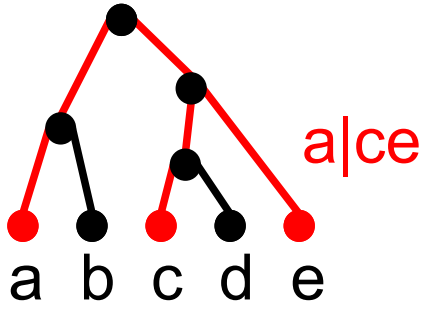


Motivations algorithmiques !

Triplets/quadruplets, splits/clusters

Idée :
reconstuire un réseau contenant tous les :

triplets

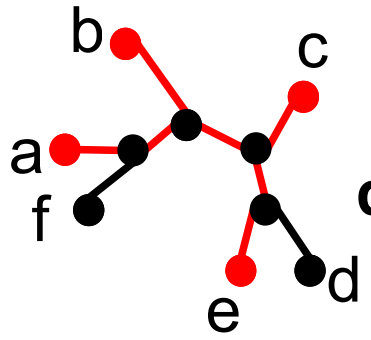


des arbres en entrée ?

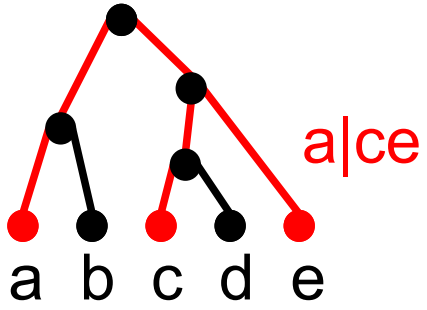
Triplets/quadruplets, splits/clusters

Idée :
reconstuire un réseau contenant tous les :

ab|ce



triplets



a|ce

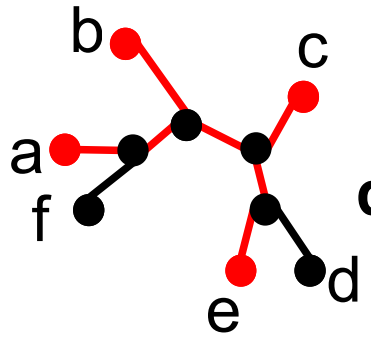
quadruplets

des arbres en entrée ?

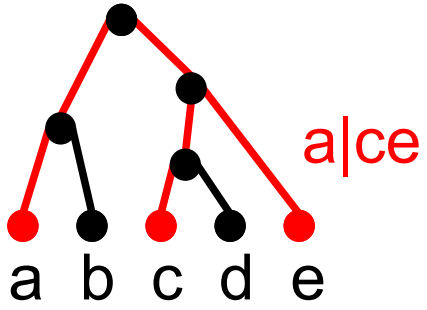
Triplets/quadruplets, splits/clusters

Idée :
reconstuire un réseau contenant tous les :

ab|ce

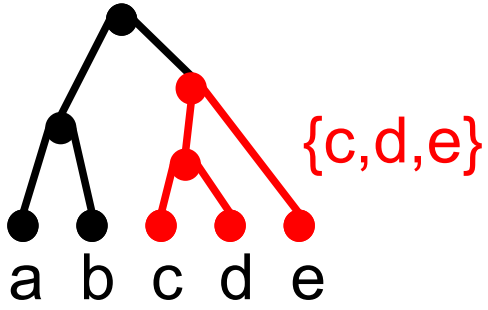


triplets



quadruplets

clusters

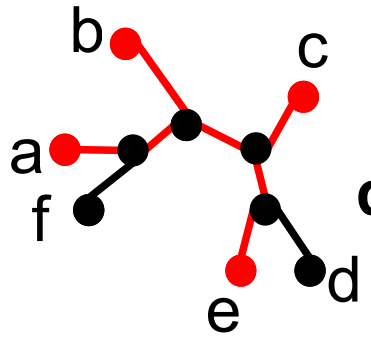


des arbres en entrée ?

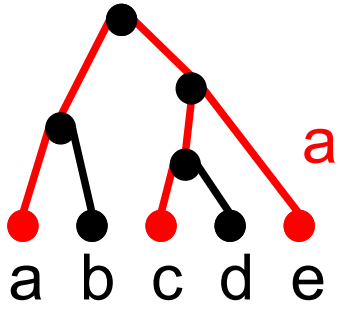
Triplets/quadruplets, splits/clusters

Idée :
reconstuire un réseau contenant tous les :

ab|ce



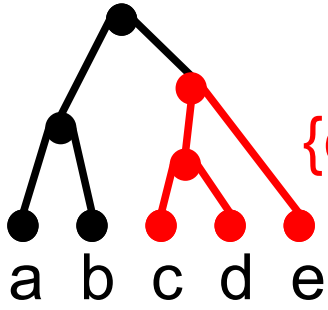
triplets



a|ce

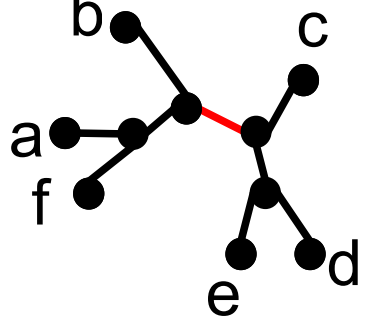
quadruplets

clusters



{c,d,e}

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



splits

des arbres en entrée ?

Reconstruction depuis les triplets

Motivations biologiques pour les triplets :

+ topologies fiables

Pour 5 taxons ou plus, pour tout arbre, il existe une taille de population et des longueurs de branches telles que l'arbre des gènes le plus vraisemblable est différent de l'arbre des espèces.

(Degnan & Rosenberg, 2006)

+ permettent un consensus plus fiable.

(Degnan, DeGiorgio, Bryant & Rosenberg, 2008)

+ les méthodes de triplets donnent de bons résultats.

(Ranwez, Berry & al, PhySIC, 2007)

- peu d'algorithmes implémentés de façon ergonomique

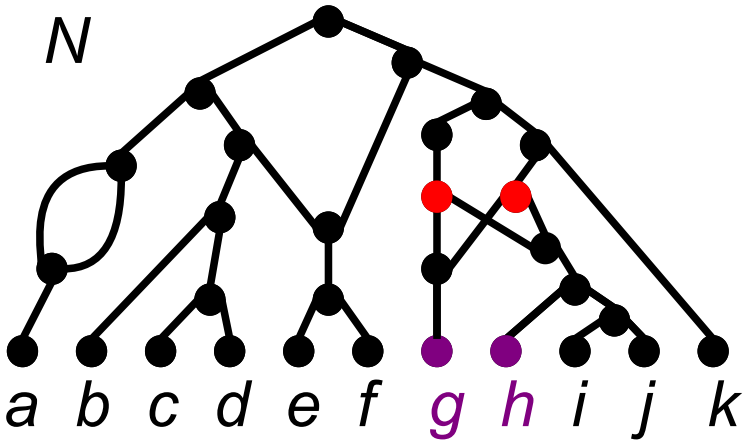
- doutes sur la qualité des triplets : phylogénies fiables = nombreux taxons ?

(Philippe, 1993)

Triplets et réseaux

Un triplet *x*|*yz* est **compatible** avec un réseau phylogénétique *N* de niveau *k* si:
- *N* contient deux noeuds *u* et *v*
- et des chemins intérieurement disjoints deux à deux :
  - de *u* à *y*,
  - de *u* à *z*,
  - de v à *u*,
  - et de *v* à *x*.

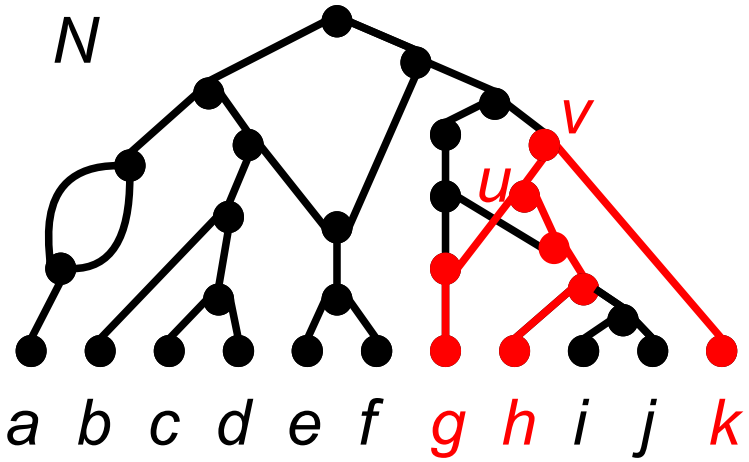
Pas d'unicité du plus petit ancêtre commun dans un réseau !



Triplets et réseaux

Un triplet *x*|*yz* est **compatible** avec un réseau phylogénétique *N* de niveau *k* si:
- *N* contient deux noeuds *u* et *v*
- et des chemins intérieurement disjoints deux à deux :
  - de *u* à *y*,
  - de *u* à *z*,
  - de v à *u*,
  - et de *v* à *x*.

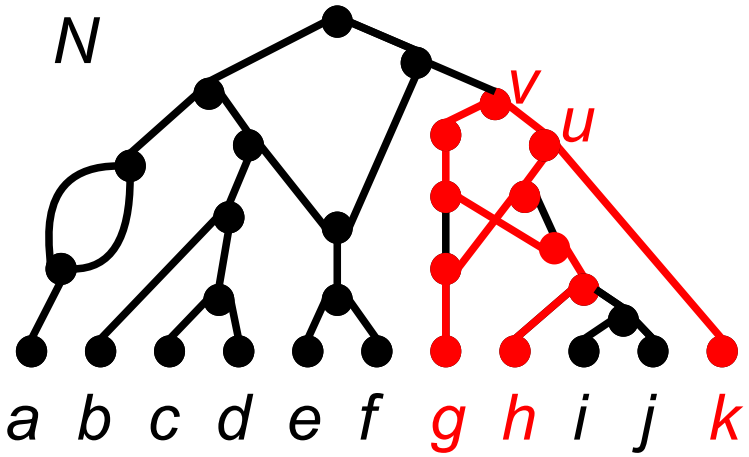
k|*gh* **compatible** avec *N*.



Triplets et réseaux

Un triplet *x*|*yz* est **compatible** avec un réseau phylogénétique *N* de niveau *k* si:
- *N* contient deux noeuds *u* et *v*
- et des chemins intérieurement disjoints deux à deux :
  - de *u* à *y*,
  - de *u* à *z*,
  - de v à *u*,
  - et de *v* à *x*.

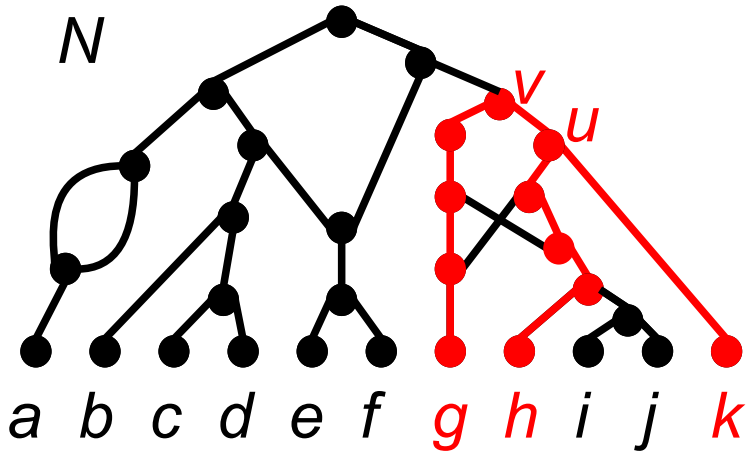
h|*gk* **compatible** avec *N*.



Triplets et réseaux

Un triplet *x*|*yz* est **compatible** avec un réseau phylogénétique *N* de niveau *k* si:
- *N* contient deux noeuds *u* et *v*
- et des chemins intérieurement disjoints deux à deux :
  - de *u* à *y*,
  - de *u* à *z*,
  - de v à *u*,
  - et de *v* à *x*.

g|*hk* **compatible** avec *N*.

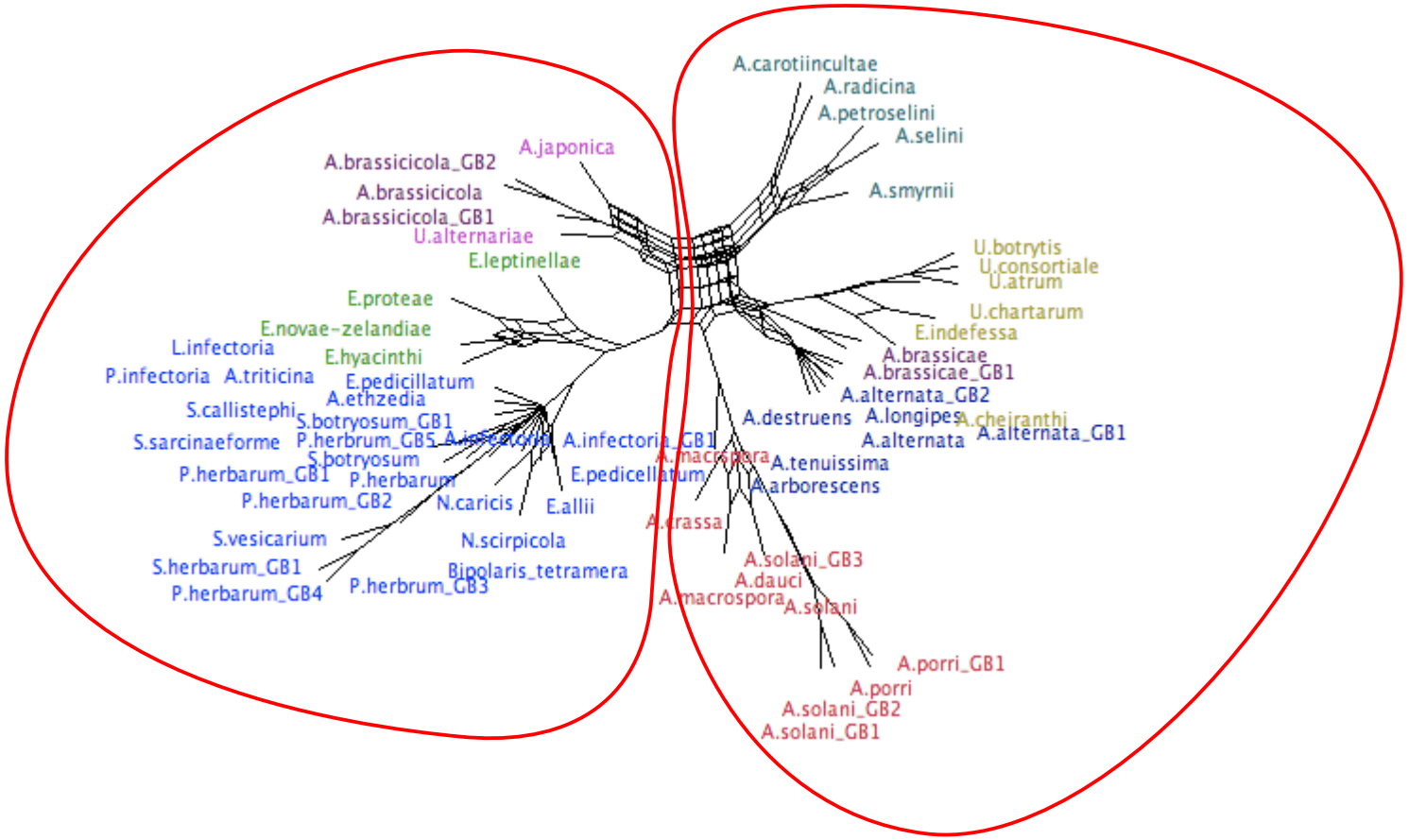


L'ensemble $\mathcal{T}(N)$ de **tous les triplets compatibles** avec un réseau *N* de niveau *k* peut être construit en $O(|\mathcal{T}(N)|) = O(n^3)$

(programmation dynamique, Byrka, Gawrychowski, Huber, Kelk, 2008)

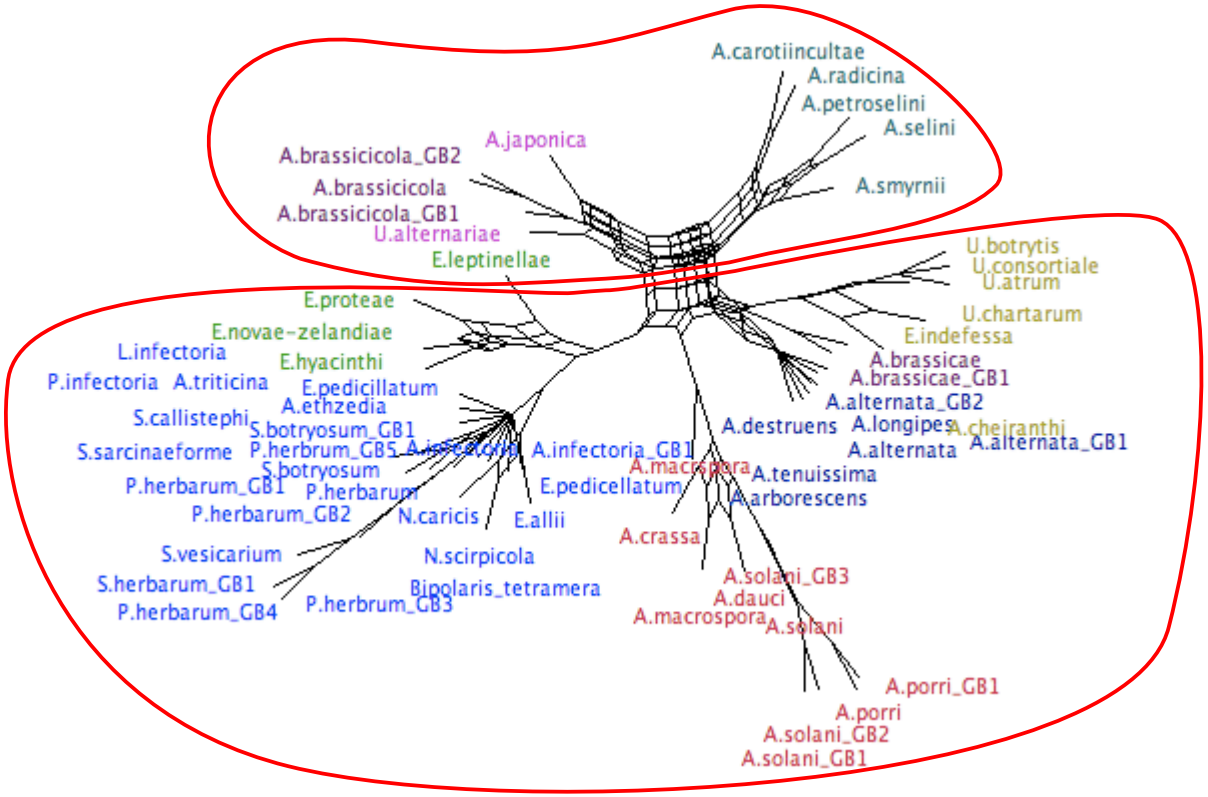
Splits et réseaux

Un ensemble de splits est représenté par un **split network**



Splits et réseaux

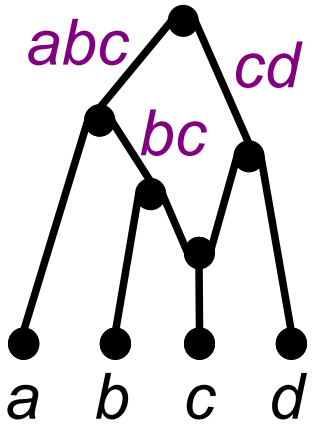
Un ensemble de splits est représenté par un **split network**



Clusters pleins / souples et réseaux

X cluster **pleinement compatible** avec *N* (**hardwired**) si *X* est l'ensemble des feuilles sous un noeud de *N*.

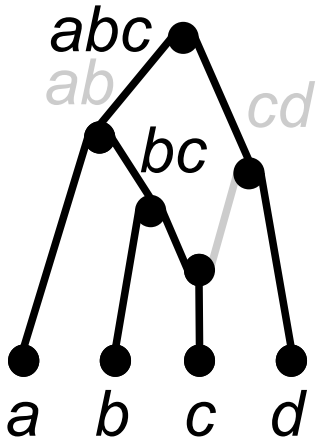
X cluster **souplement compatible** avec *N* (**softwired**) s'il existe un arbre dans *N* sur les taxons *L* tel que *X* est l'ensemble des feuilles sous un noeud de *N*.



Clusters pleins / souples et réseaux

X cluster **pleinement compatible** avec *N* (**hardwired**) si *X* est l'ensemble des feuilles sous un noeud de *N*.

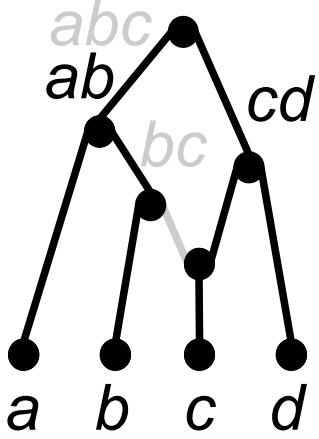
X cluster **souplement compatible** avec *N* (**softwired**) s'il existe un arbre dans *N* sur les taxons *L* tel que *X* est l'ensemble des feuilles sous un noeud de *N*.



Clusters pleins / souples et réseaux

X cluster **pleinement compatible** avec *N* (**hardwired**) si *X* est l'ensemble des feuilles sous un noeud de *N*.

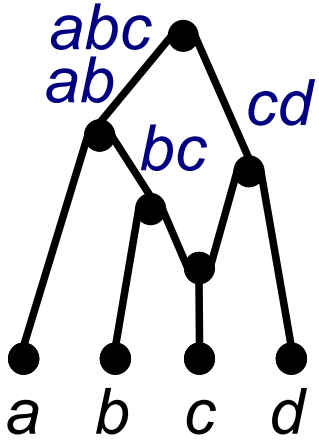
X cluster **souplement compatible** avec *N* (**softwired**) s'il existe un arbre dans *N* sur les taxons *L* tel que *X* est l'ensemble des feuilles sous un noeud de *N*.



Clusters pleins / souples et réseaux

X cluster **pleinement compatible** avec *N* (**hardwired**) si *X* est l'ensemble des feuilles sous un noeud de *N*.

X cluster **souplement compatible** avec *N* (**softwired**) s'il existe un arbre dans *N* sur les taxons *L* tel que *X* est l'ensemble des feuilles sous un noeud de *N*.



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

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

Triplets/quadruplets, splits/clusters

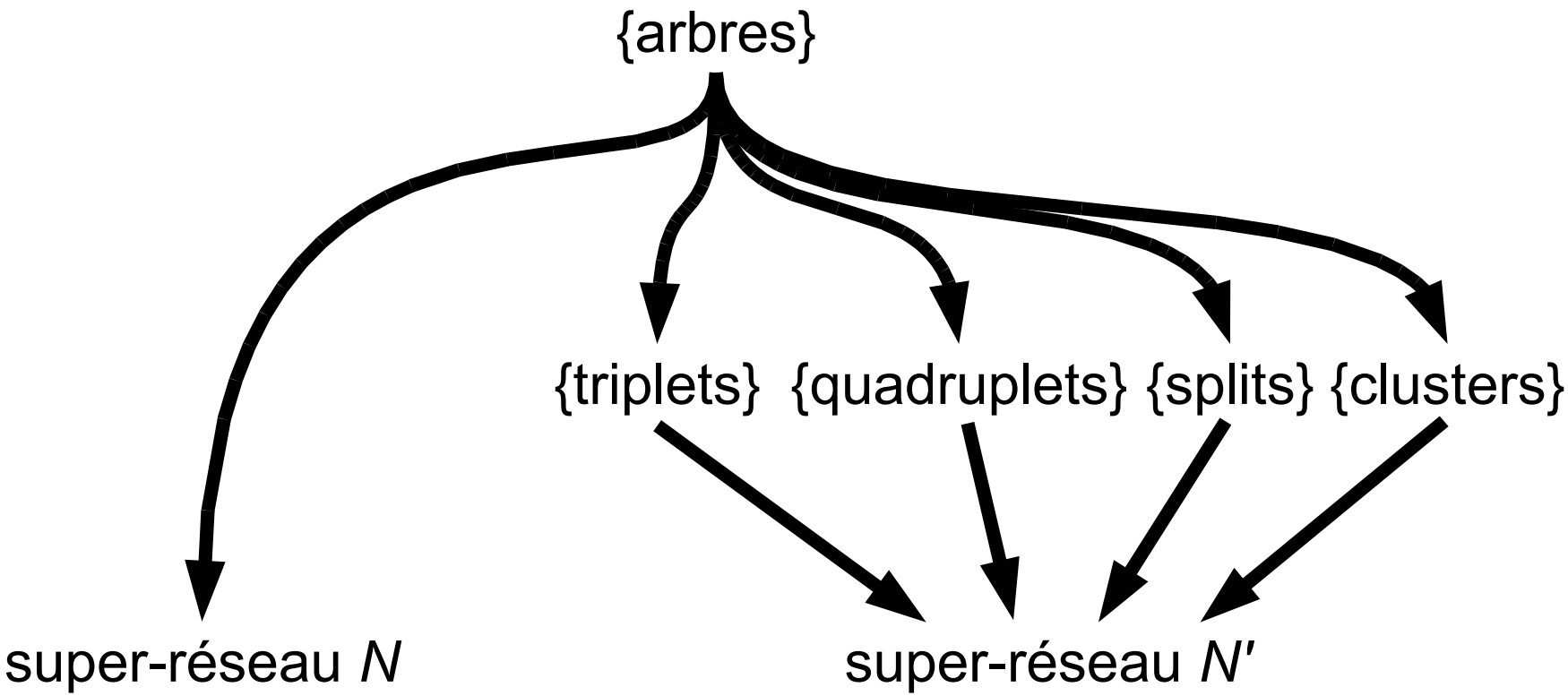
Idée :
modifier le type de données à traiter

{arbres}

super-réseau *N*

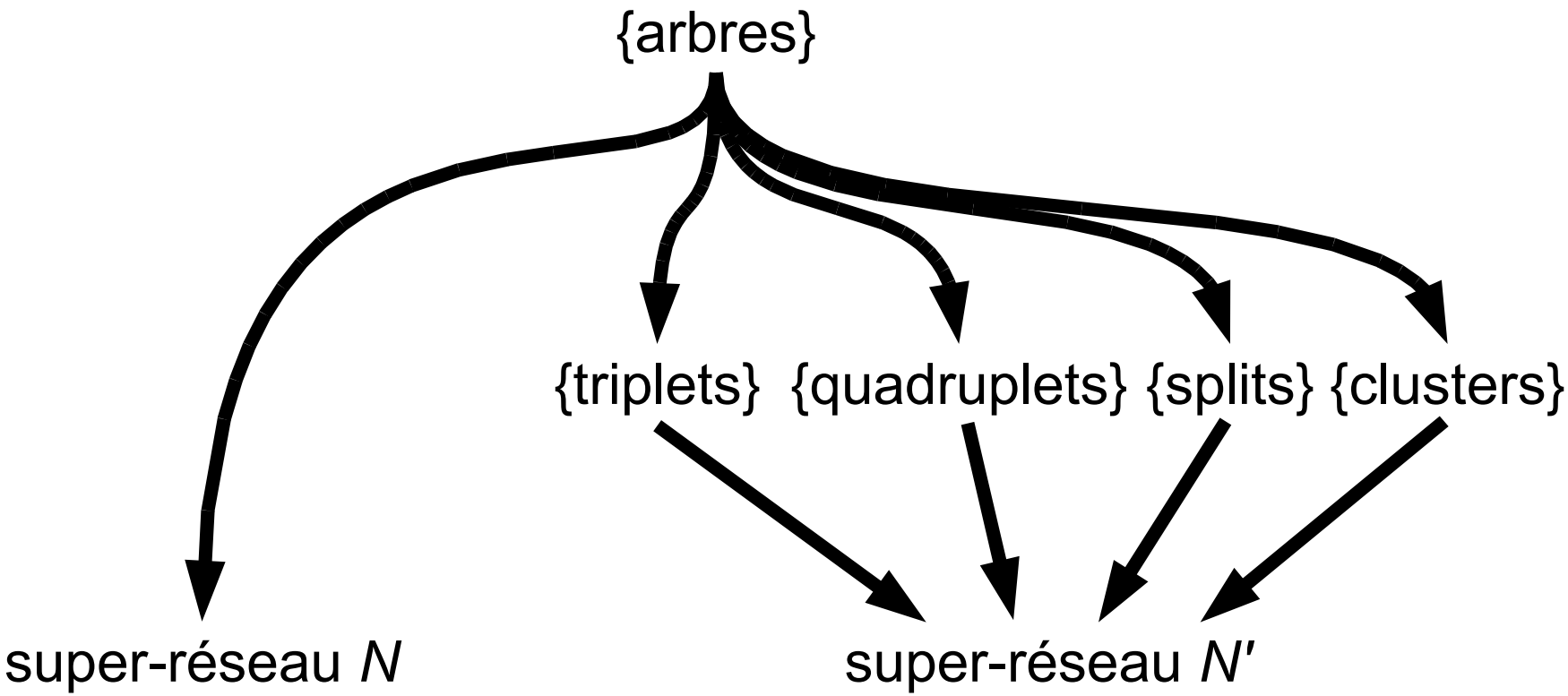
Triplets/quadruplets, splits/clusters

Idée :
modifier le type de données à traiter



Triplets/quadruplets, splits/clusters

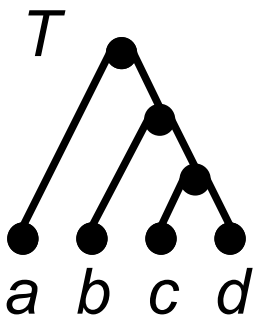
Idée :
modifier le type de données à traiter

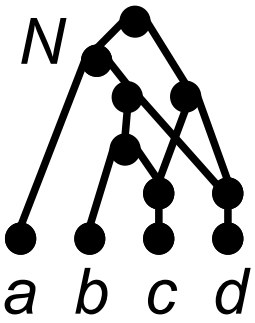


N'=N ?

Compatibilité avec cluster / triplets

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

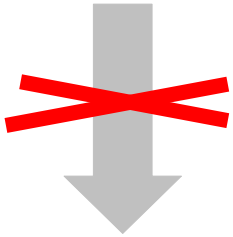




compatible avec
{*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas avec *T*

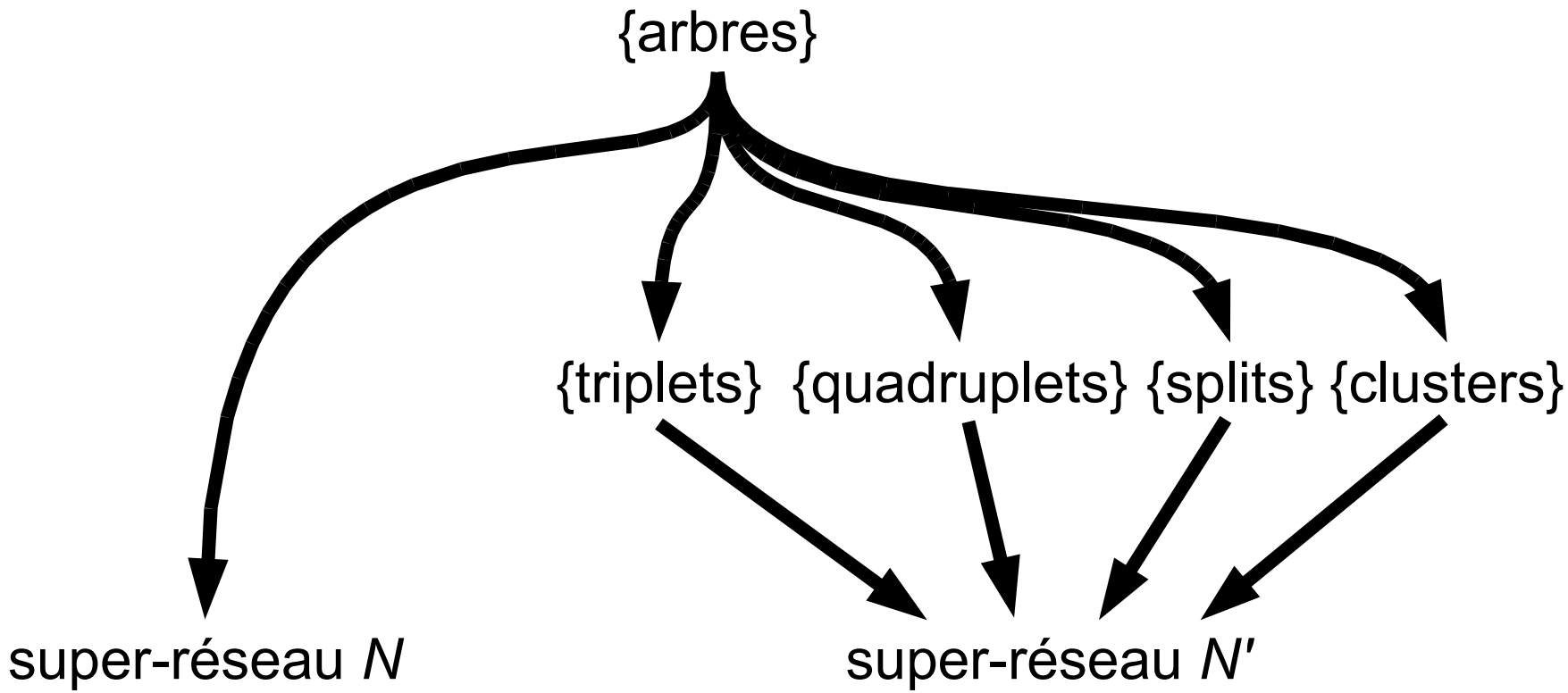
compatible avec
{*abcd*,*bcd*,*cd*,*a*,*b*,*c*}
mais pas avec *T*

compatible avec les clusters d'un arbre *T*



compatible avec *T*.

Triplets/quadruplets, splits/clusters



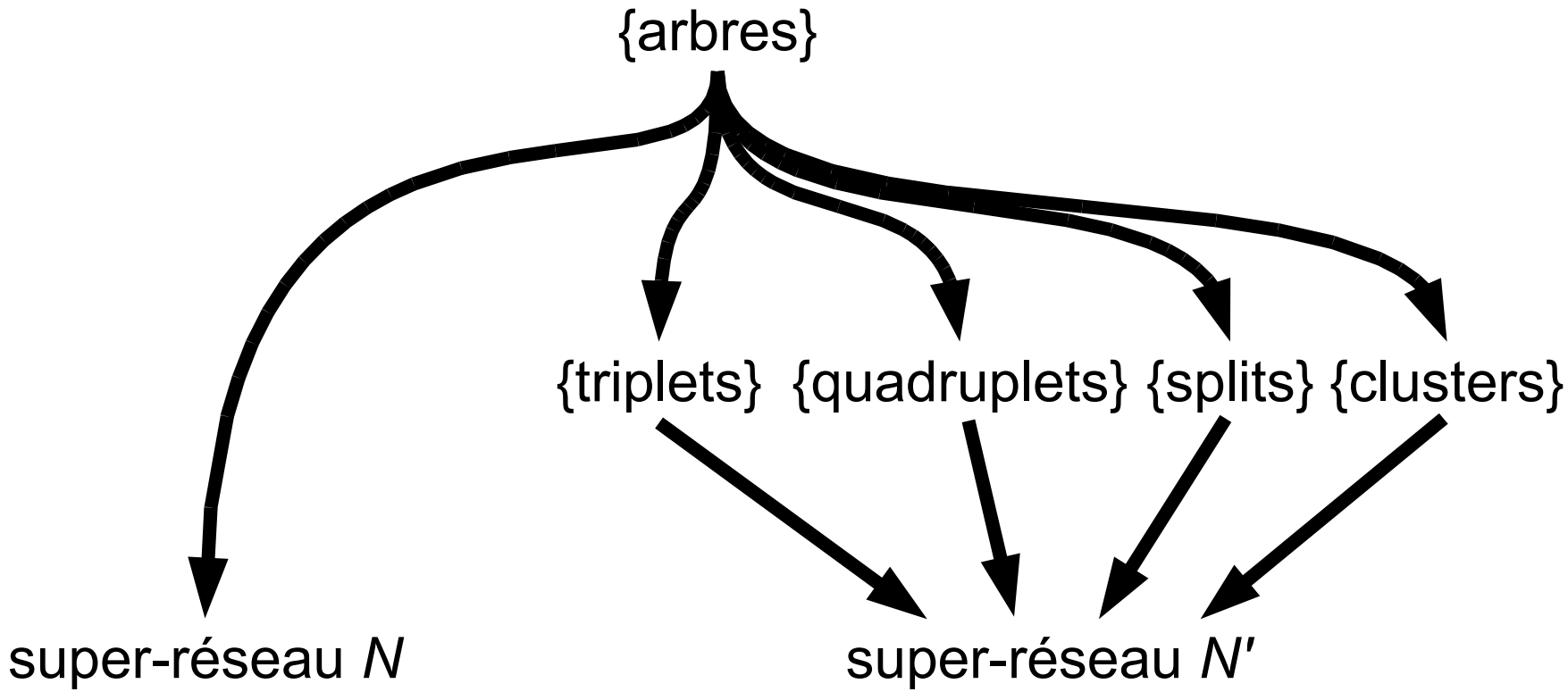
***N'*=*N* ?**

Pas nécessairement, mais :

***N'* complexe** ⟹ ***N* complexe**

N contient également les triplets, quadruplets...

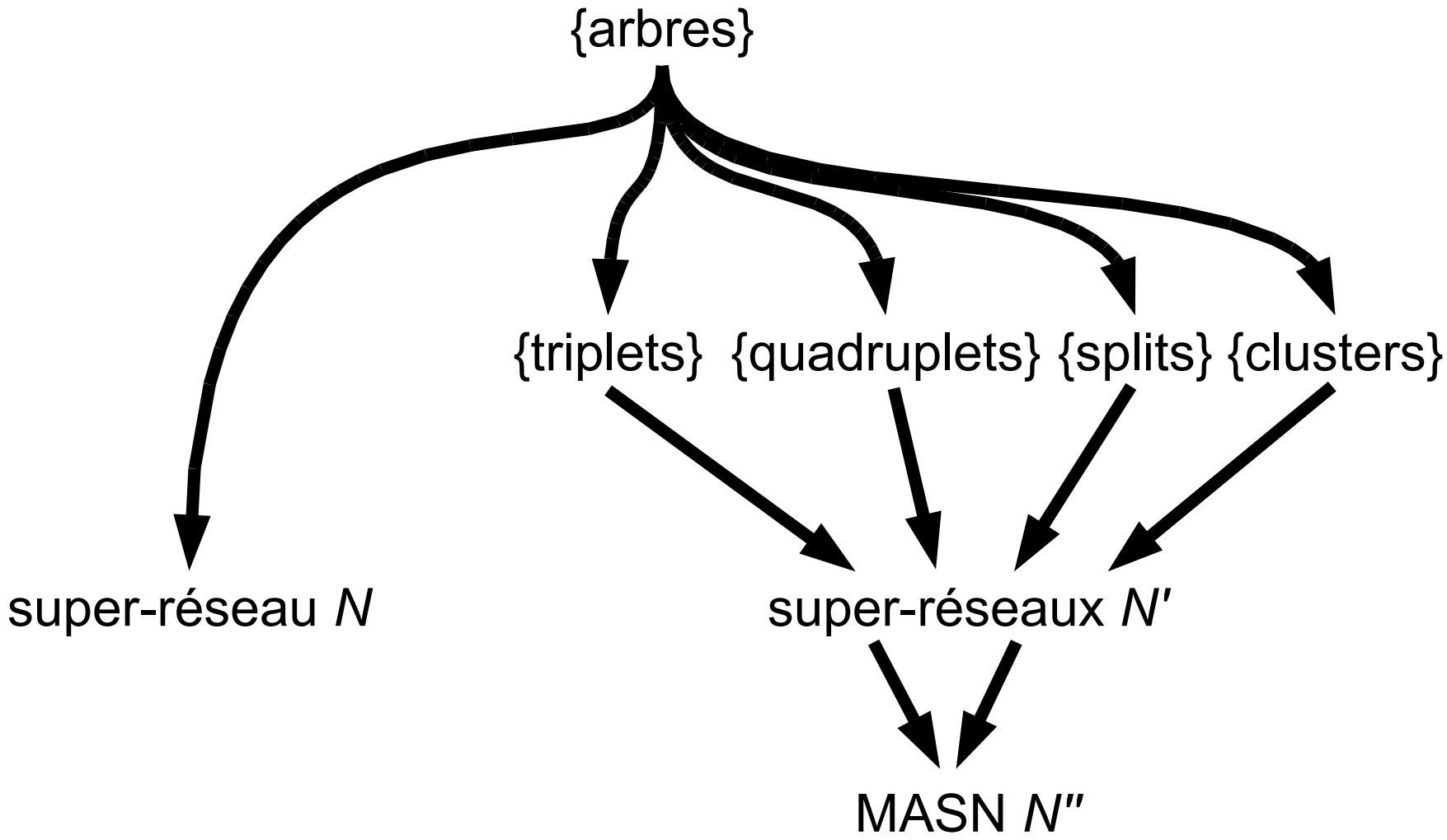
Triplets/quadruplets, splits/clusters



***N'=N* ?**

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

Triplets/quadruplets, splits/clusters



Reconstruction depuis les splits

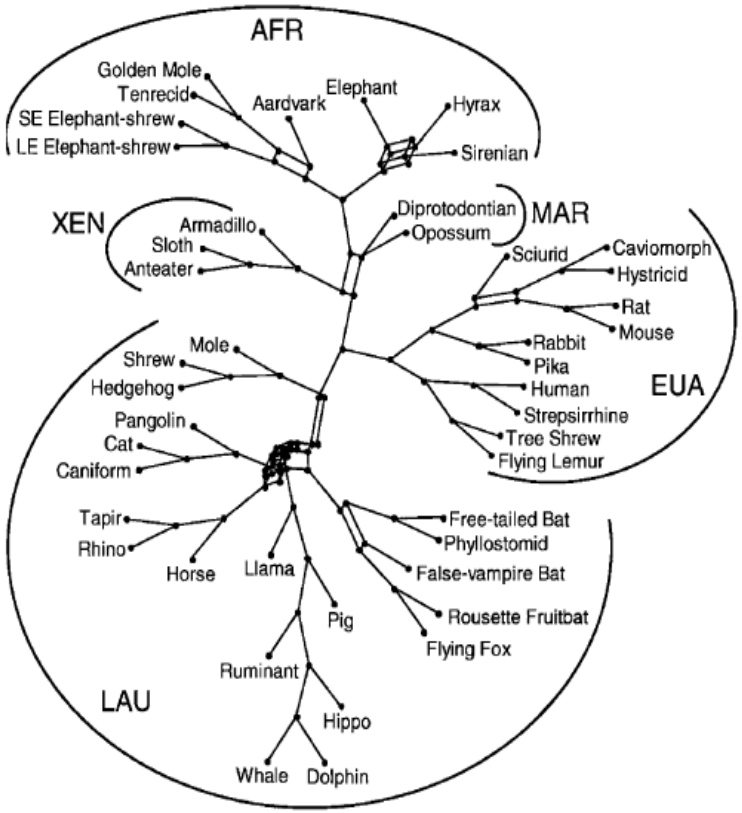
{arbres}

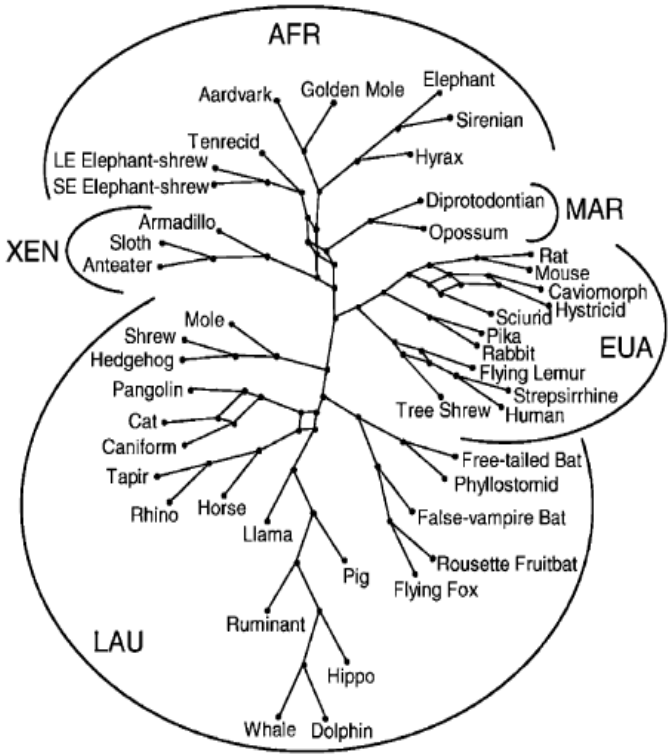
{splits}

N'

Construction de réseaux de consensus “split networks”

SplitsTree 4





Dress & Huson, TCBB, 2004

Holland, Huber, Moulton, Lockhart, MBE 2004

Holland, Delsuc & Moulton, Systematic Biology, 2005

Reconstruction depuis les splits

{arbres}

sur divers ensembles de taxons

clôture

{splits}

N'

Z-clôture

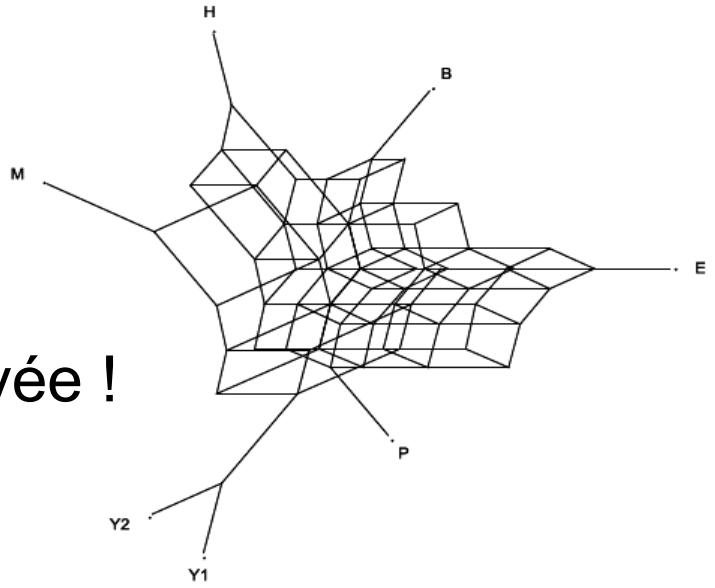
(Huson, Dezulian, Klöpper & Steel, TCBB, 2004)

$$\frac{A}{B} Z \frac{C}{D} \rightarrow \frac{A}{B \cup D}, \frac{A \cup C}{D} \quad (Z \neq \emptyset)$$

Y/M-clôture

(Grünewald, Huber & Wu, BMB, 2008)

Problème :
dimension trop élevée !



Reconstruction depuis les splits

{arbres}

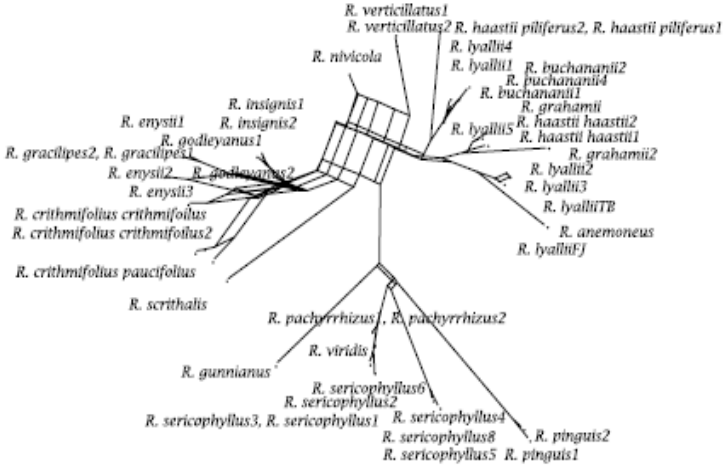
Filtres

(Huson, Steel & Whitfield, WABI'06)
(Whitfield, Cameron, Huson & Steel, Systematic Biology 2009)

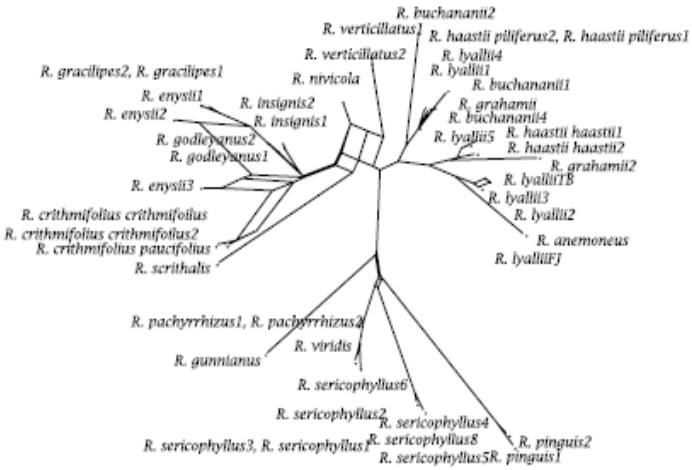
filtre

{splits}

N'



(c)



(d)

splits de 2 arbres

splits de distortion <2
(score d'homoplasie)

Reconstruction depuis les triplets

{arbres}

NP-complet dans le cas général, pour niveau 1

(Jansson, Nguyen & Sung, SODA'05)

Polynomial pour niveau 1 et 2 si l'ensemble de triplets est **dense**.

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

{triplets}

(Jansson, Nguyen & Sung, SODA'05 : O(n^3) pour niveau 1)

(van Iersel, Kelk & al, RECOMB'08 : O(n^8) pour niveau 2)

Simplistic

N'

Ouvert pour les niveaux >2

Reconstruction depuis les clusters

{arbres}

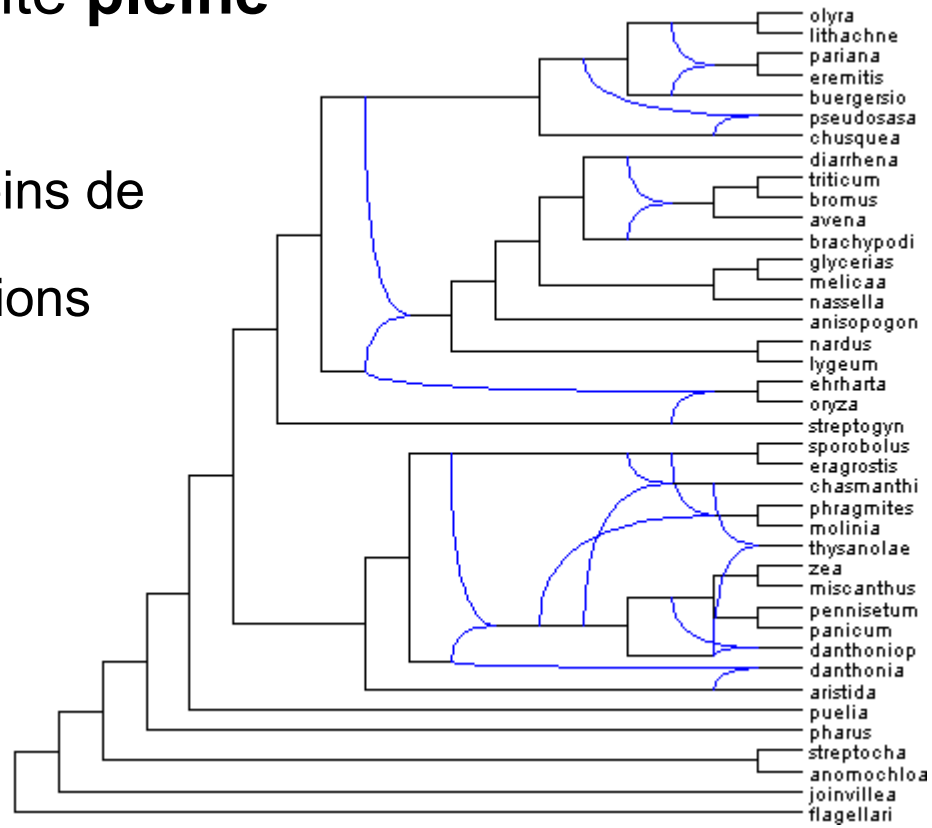
Consensus de clusters :
Dendroscope

Compatibilité **pleine**

{clusters}

Réseau de
clusters pleins de
2 arbres :
11 réticulations

N'



(Rupp & Huson, WABI'08)

Reconstruction depuis les clusters

{arbres}

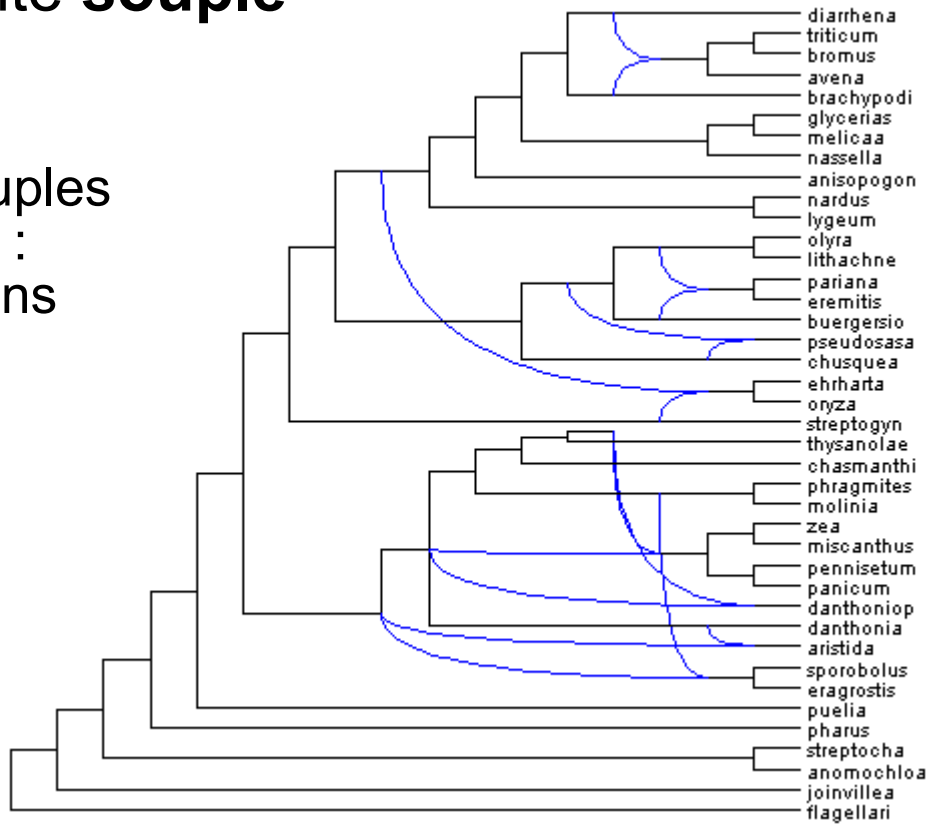
Consensus de clusters :
Dendroscope

Compatibilité **souple**

{clusters}

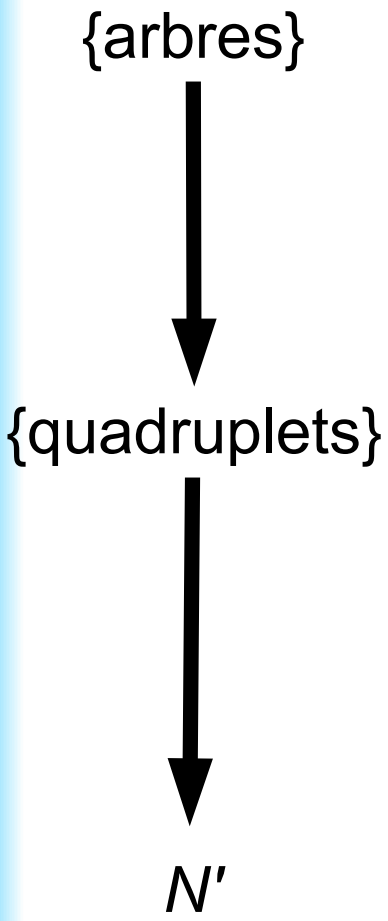
Réseau de clusters souples de 2 arbres : 8 réticulations

N' réseau galled network



(Rupp, Huson, Berry, Gambette & Paul)

Reconstruction depuis les quadruplets



Reconstruction depuis les quadruplets

{arbres}

↓

{quadruplets}

↓

{splits}

↓

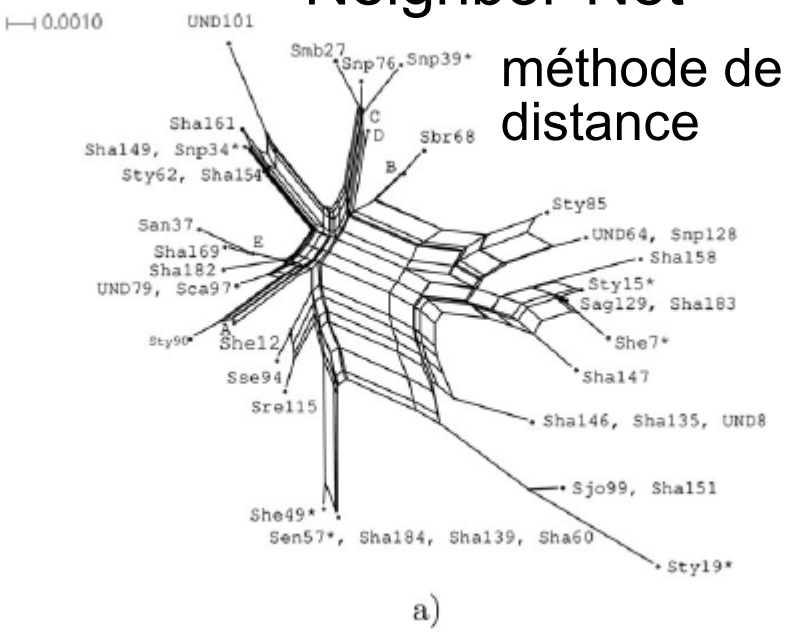
N'

Consensus de quadruplets :
Q-Net

(Grünewald, Forslund, Dress & Moulton, 2006)

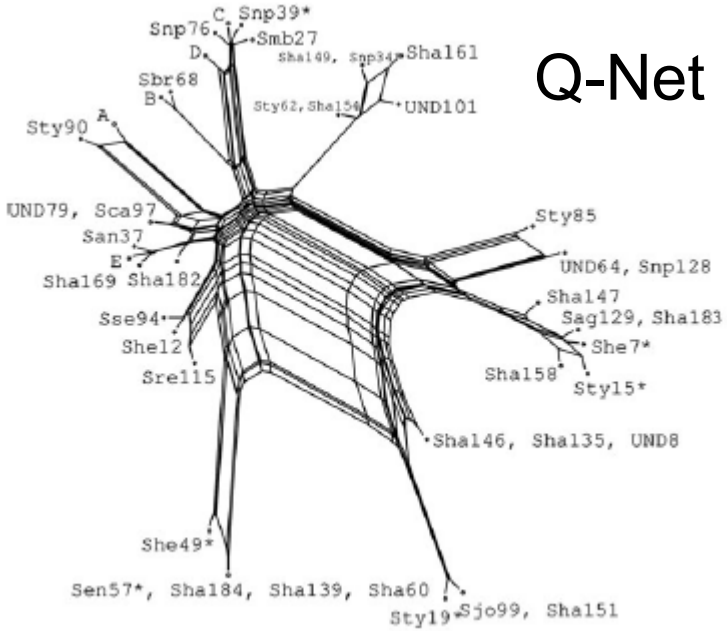
Neighbor-Net

méthode de distance



a)

Q-Net



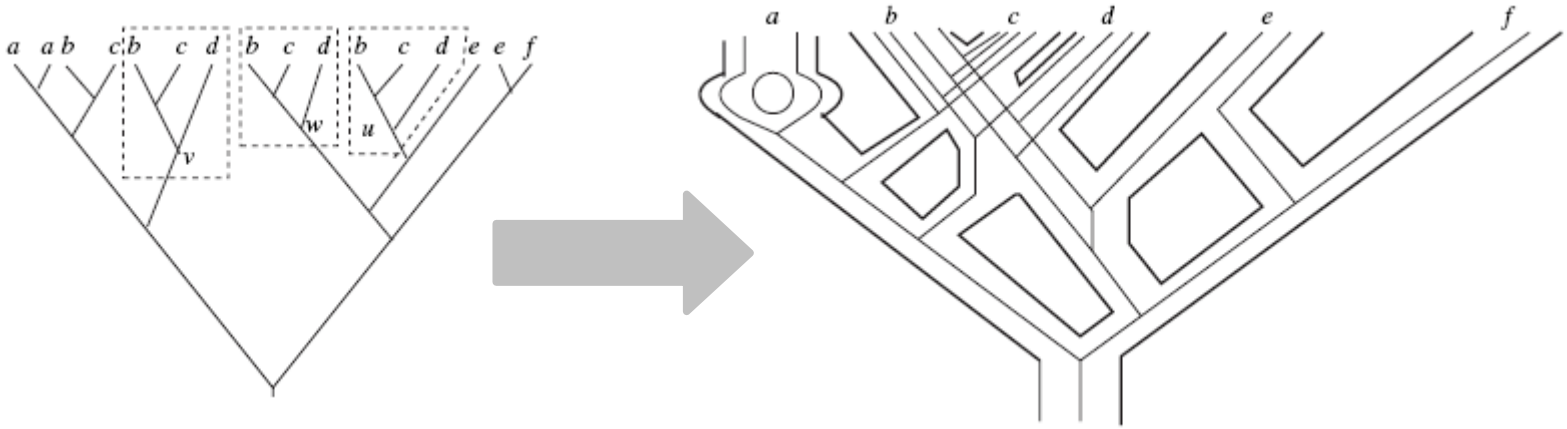
Plan

- Les types de réseaux phylogénétiques
- Reconstruction depuis des arbres
- Des sous-classes de réseaux
- Les arbres et leurs quadruplets/triplets, splits/clusters
- **Reconstruction depuis des arbres multi-étiquetés**
- Autres phénomènes à considérer

Reconstruction depuis des arbres MUL

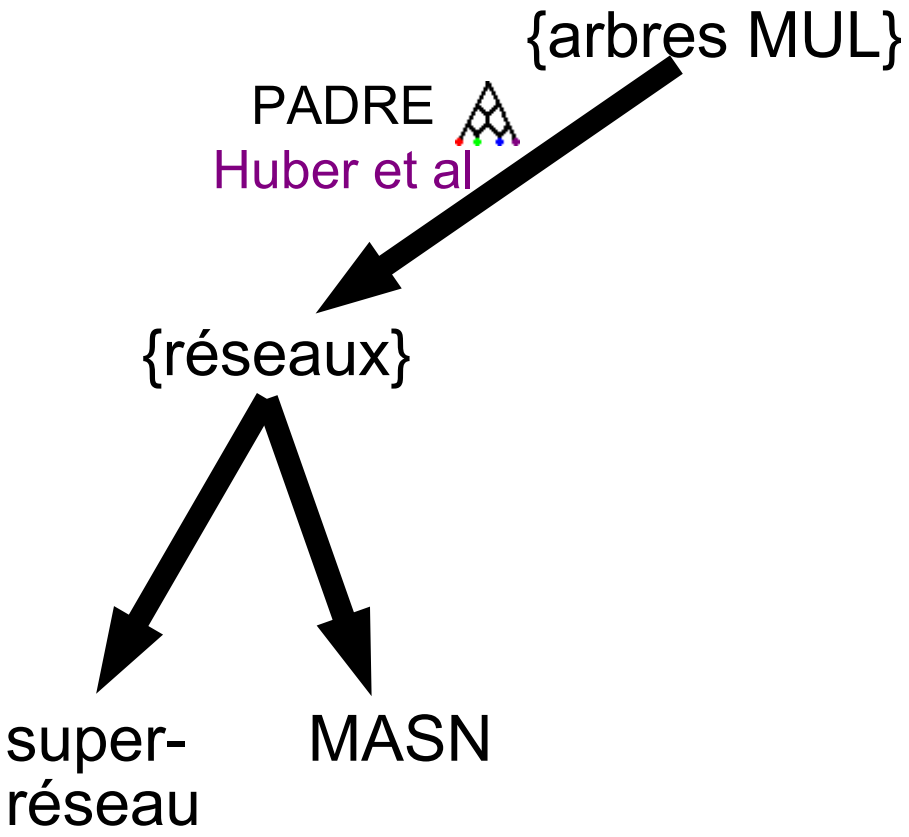
Un réseau sans étiquettes multiples expliquant **UN arbre MUL** peut être reconstruit en $O(n \log n)$

(en partant de sous-arbres isomorphes maximaux,
Huber, Oxelman, Lott, Moulton, MBE 2007,
en utilisant la caractérisation de Huber & Moulton, JMB 2006)



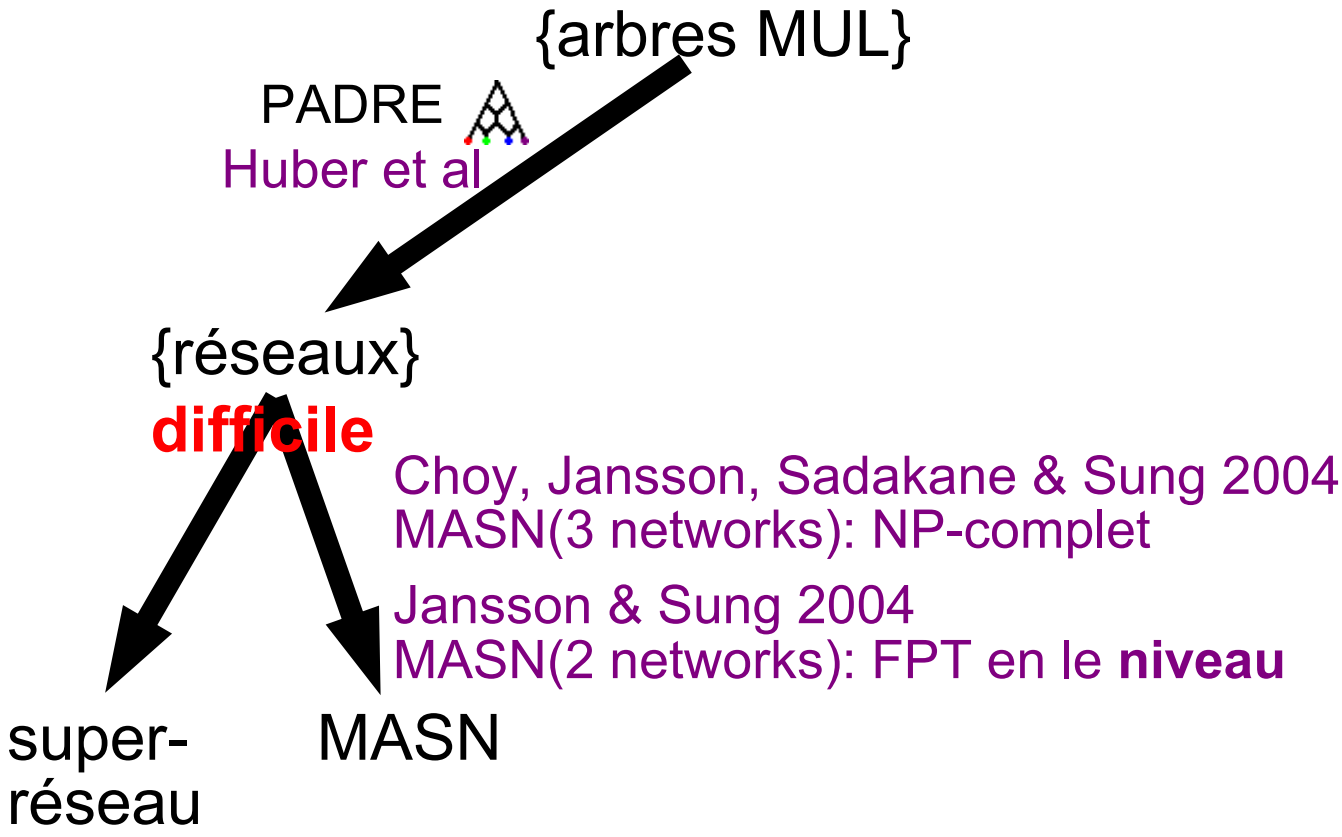
Reconstruction depuis des arbres MUL

Approches possibles :



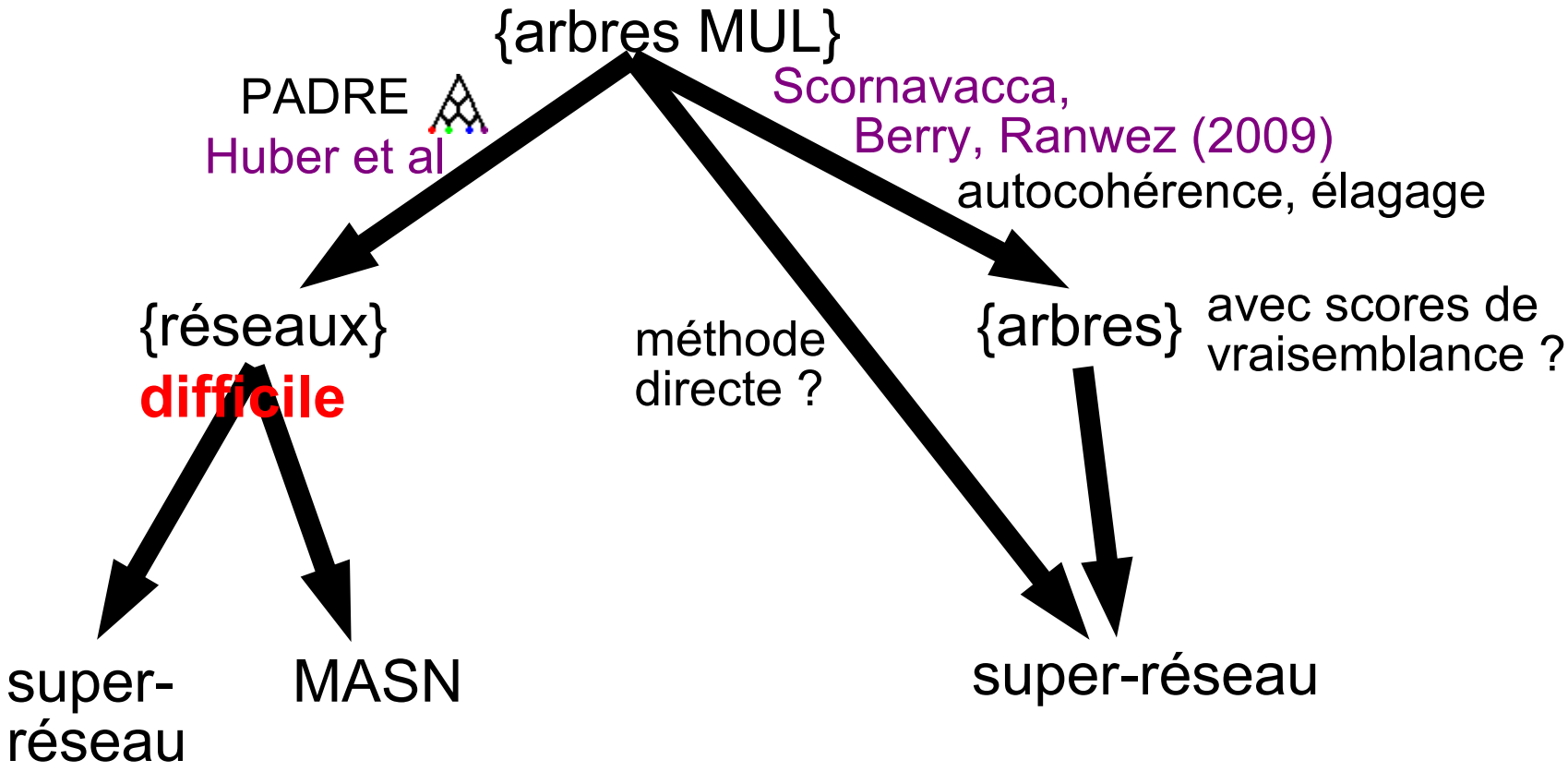
Reconstruction depuis des arbres MUL

Approches possibles :



Reconstruction depuis des arbres MUL

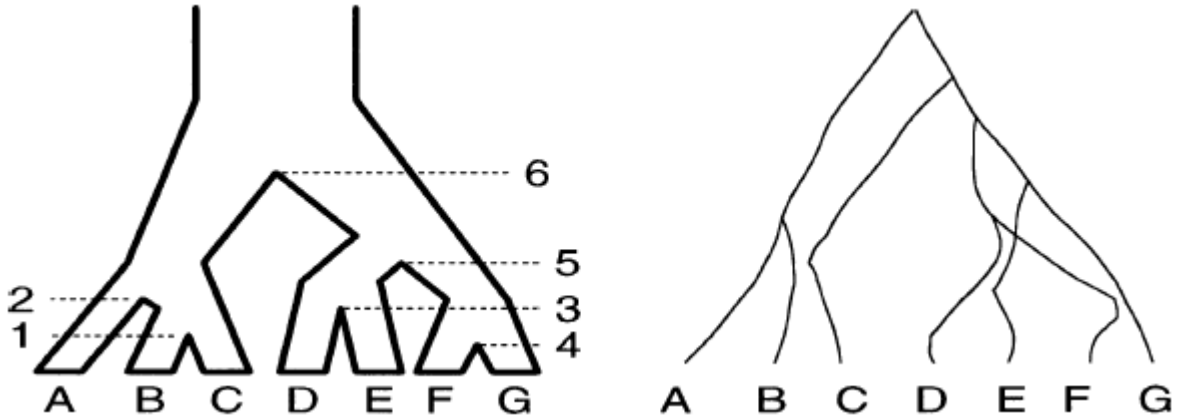
Approches possibles :



Plan

- Les types de réseaux phylogénétiques
- Reconstruction depuis des arbres
- Des sous-classes de réseaux
- Les arbres et leurs quadruplets/triplets, splits/clusters
- Reconstruction depuis des arbres multi-étiquetés
- **Autres phénomènes à considérer**

Hybridisation et tri de lignées



Méthode de **consensus** :
Q-clôture ou Z-clôture, puis filtres

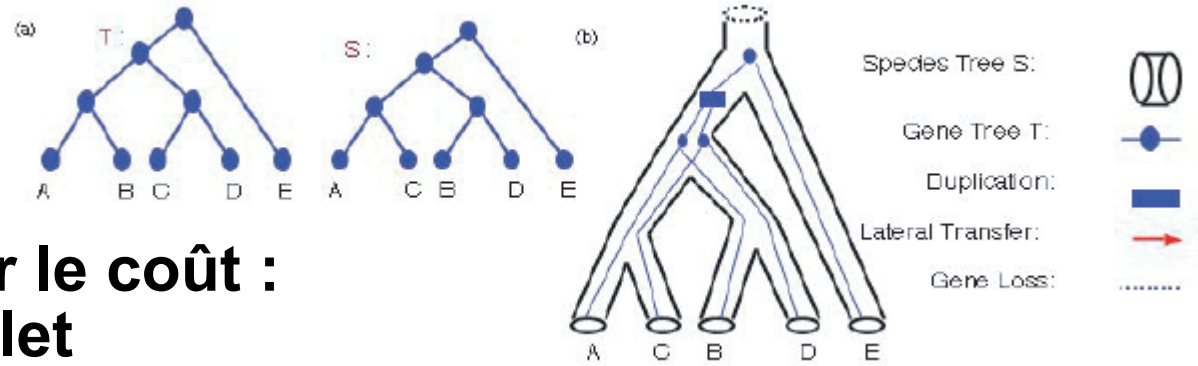
(Holland, Benthin, Lockhart, Moulton & Huber, BMCEB 2008)

Transfert de gènes et duplication/perte

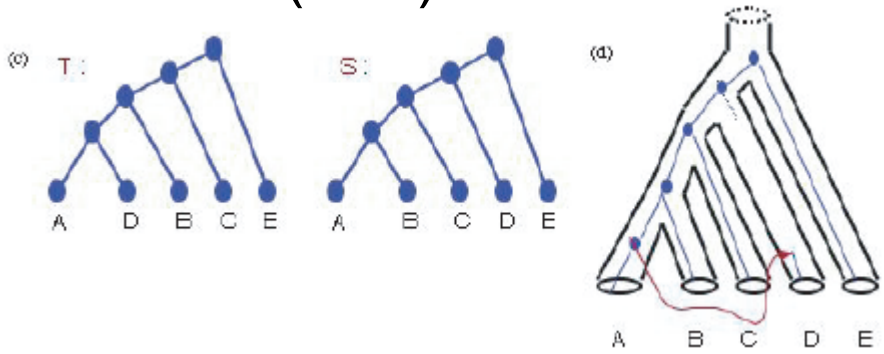
Problème : et les transferts horizontaux ?
Un **arbre principal** !

Scores de **parsimonie** pour divers scénarios faisant intervenir transferts horizontaux et duplications :

(Hallett, Lagergren & Tofigh, RECOMB 2004)



Minimiser le coût :
NP-complet
FPT en le coût c : $O(3^c n^2)$



Vernot & al, 2007

Discussion

Merci pour votre attention !