Colloque MODEMAVE - LAREMA
31/05/2012 - Angers

# *Méthodes combinatoires de reconstruction de réseaux phylogénétiques*

**Philippe Gambette**

# Plan

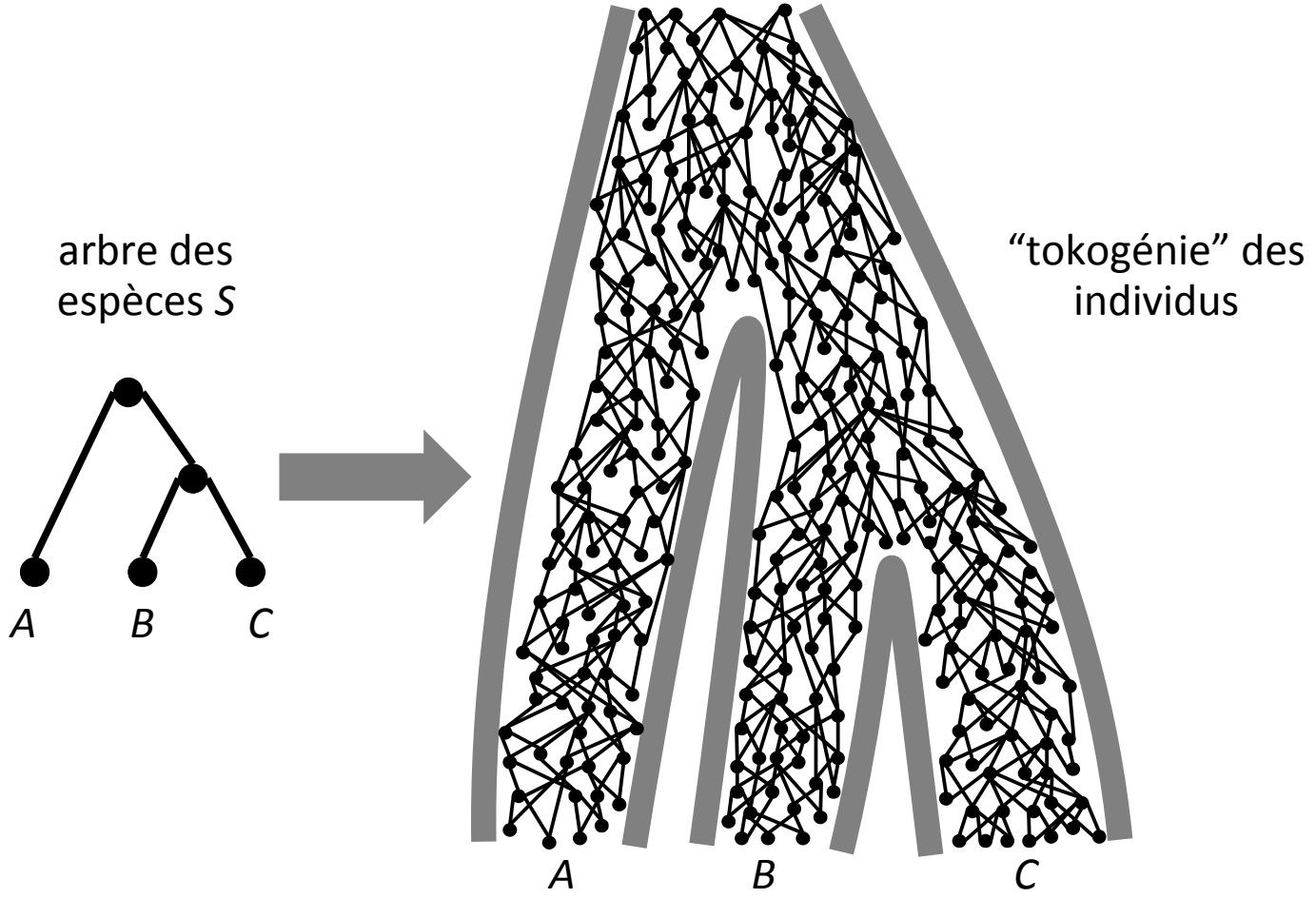
- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Méthodes combinatoires de reconstruction
- Utilisation pratique
- Illustrations
- Perspectives

# Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Méthodes combinatoires de reconstruction
- Utilisation pratique
- Illustrations
- Perspectives

# Les arbres phylogénétiques

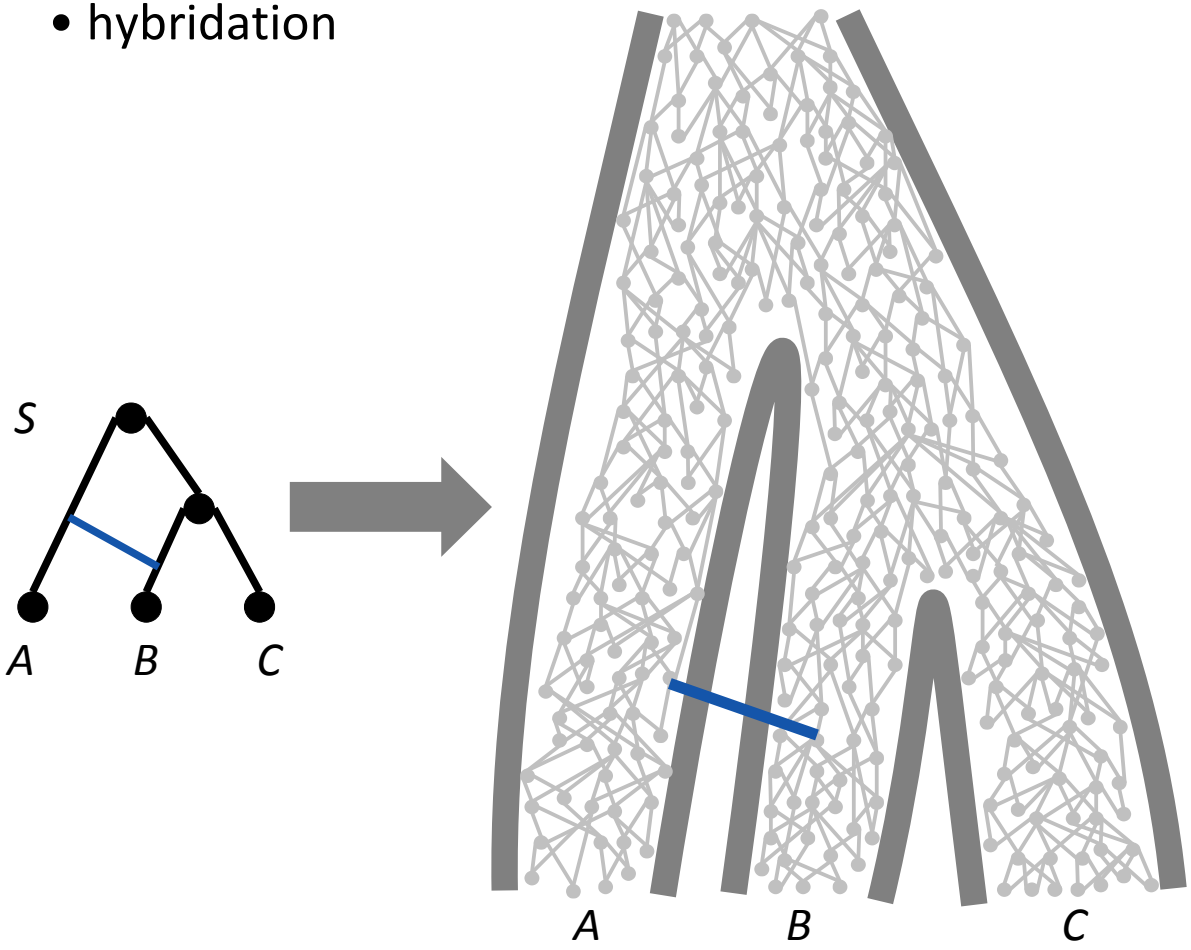
**Arbre phylogénétique** d'un ensemble d'espèces



# Transferts de matériel génétique

**Transferts** de matériel génétique entre espèces coexistantes :

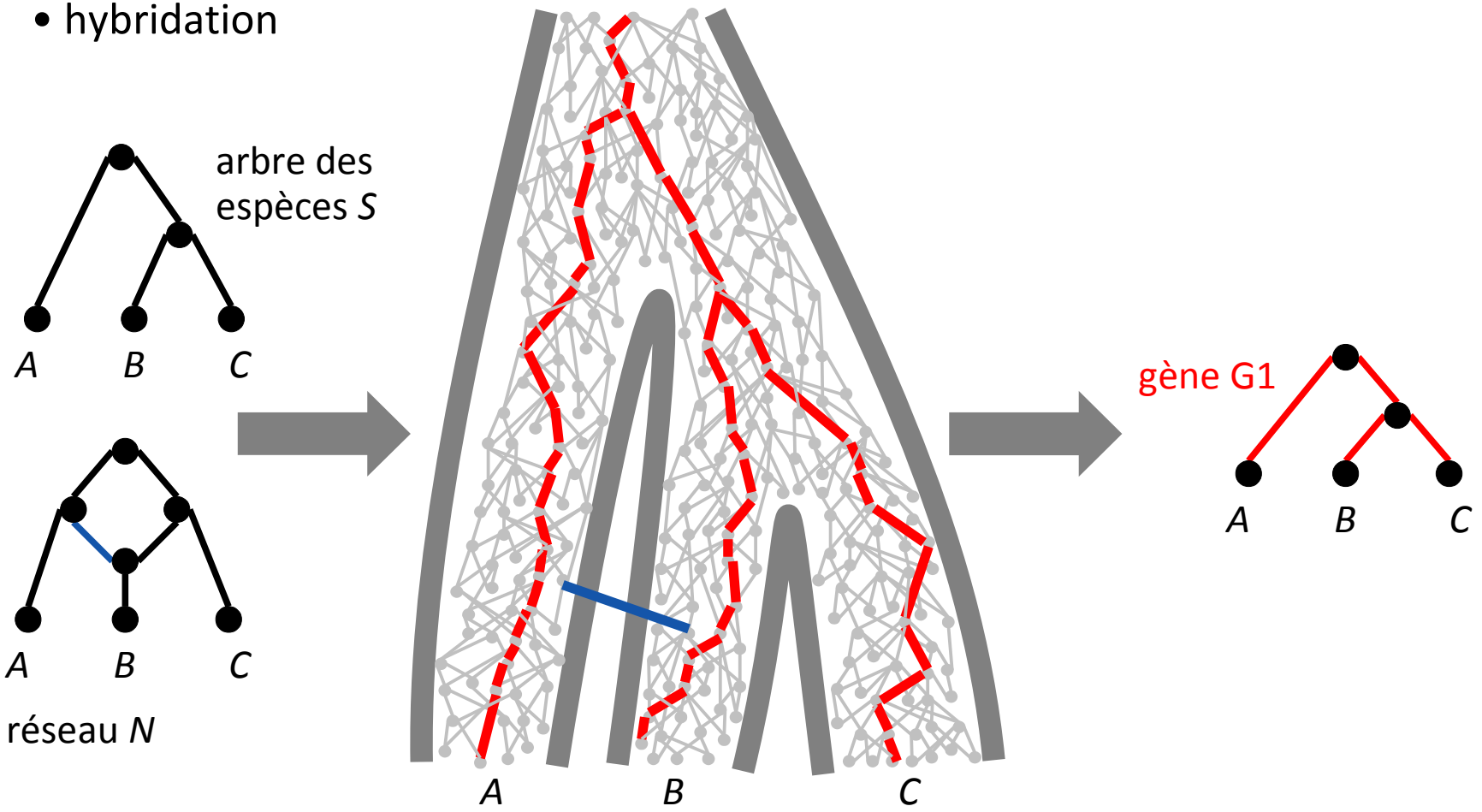
- transfert horizontal
- hybridation



# Transferts de matériel génétique

**Transferts** de matériel génétique entre espèces coexistantes :

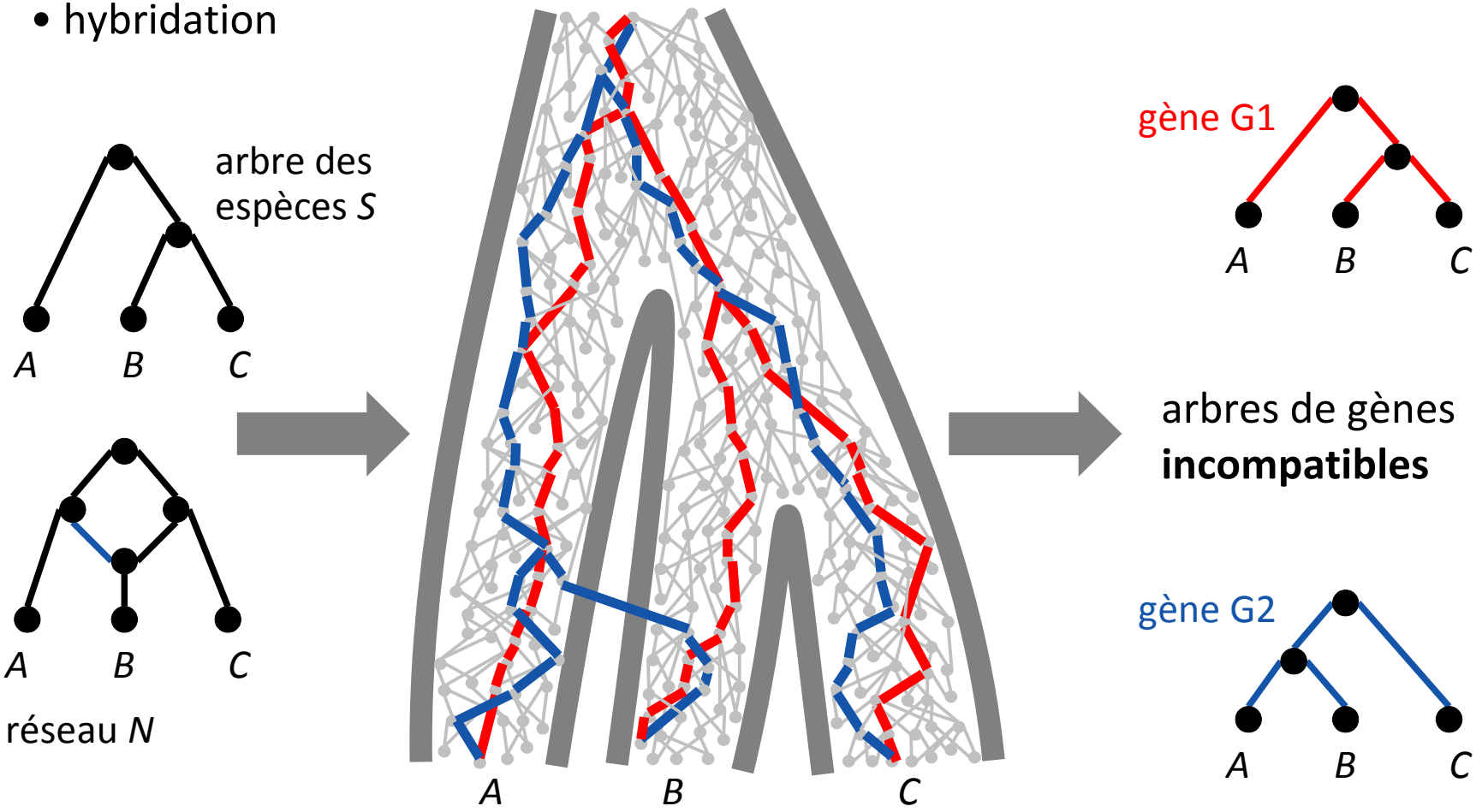
- transfert horizontal
- hybridation



# Transferts de matériel génétique

**Transferts** de matériel génétique entre espèces coexistantes :

- transfert horizontal
- hybridation

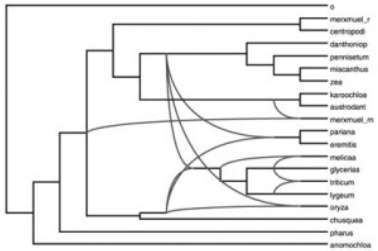


# Les réseaux phylogénétiques

**Réseau phylogénétique** : réseau représentant des données d'évolution

- réseaux phylogénétiques **explicites**

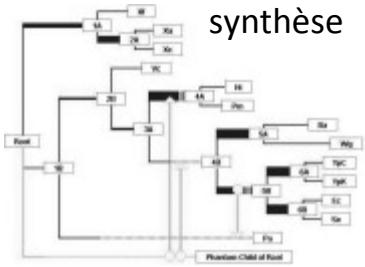
**modélisation de l'évolution**



réseau à une couche de réticulation

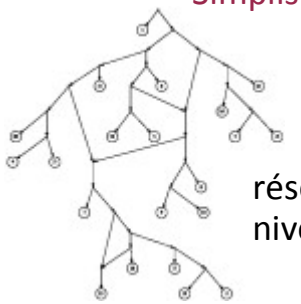
Dendroscope

diagramme de synthèse



HorizStory

Simplistic

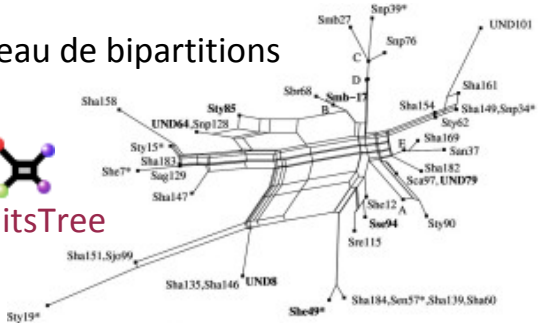


réseau de niveau 2

- réseaux phylogénétiques **abstraits**

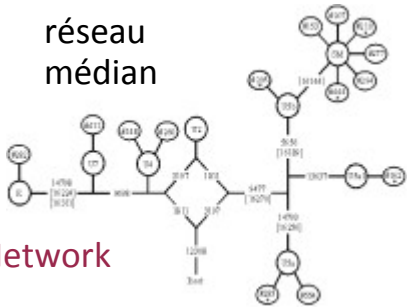
**classification, visualisation de données**

réseau de bipartitions



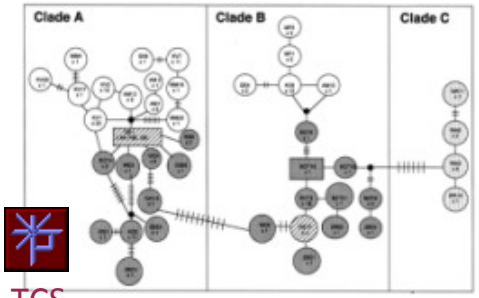
SplitsTree

réseau médian



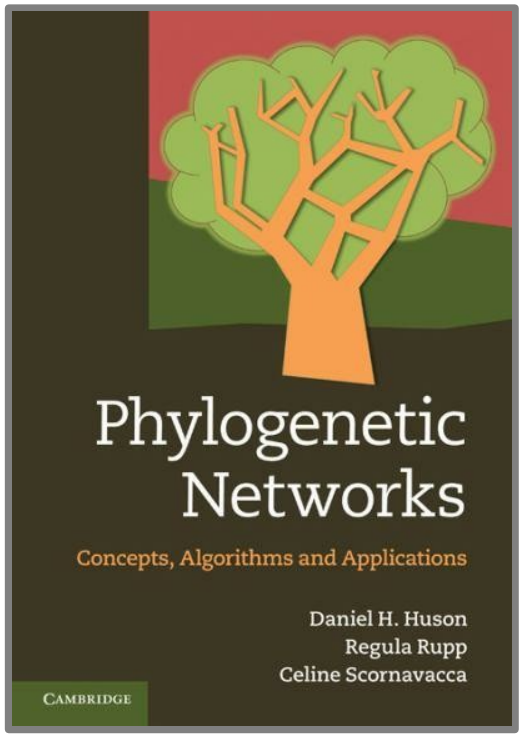
Network

réseau couvrant minimum



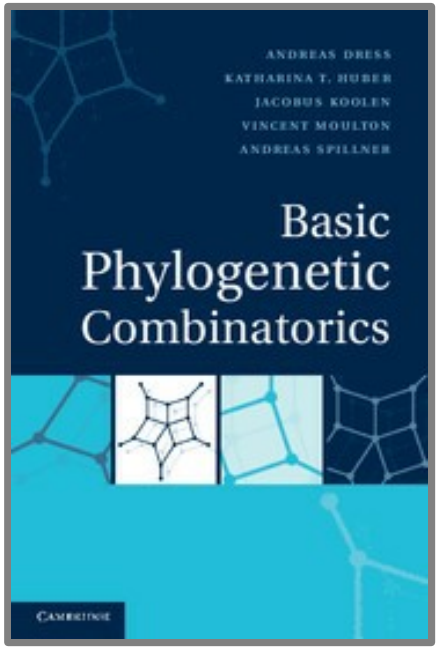

TCS

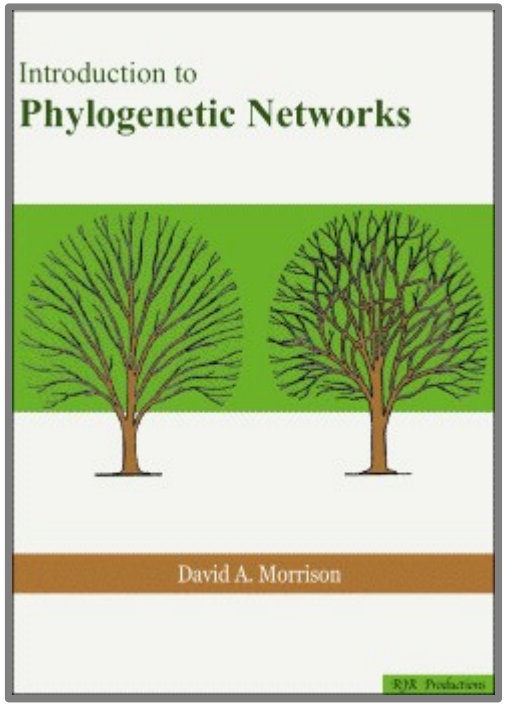
# Les réseaux phylogénétiques




Huson, Rupp, Scornavacca, 2011

Dress, Huber, Koolen, Moulton, Spillner, 2012




Morrison, 2011

Lorentz center

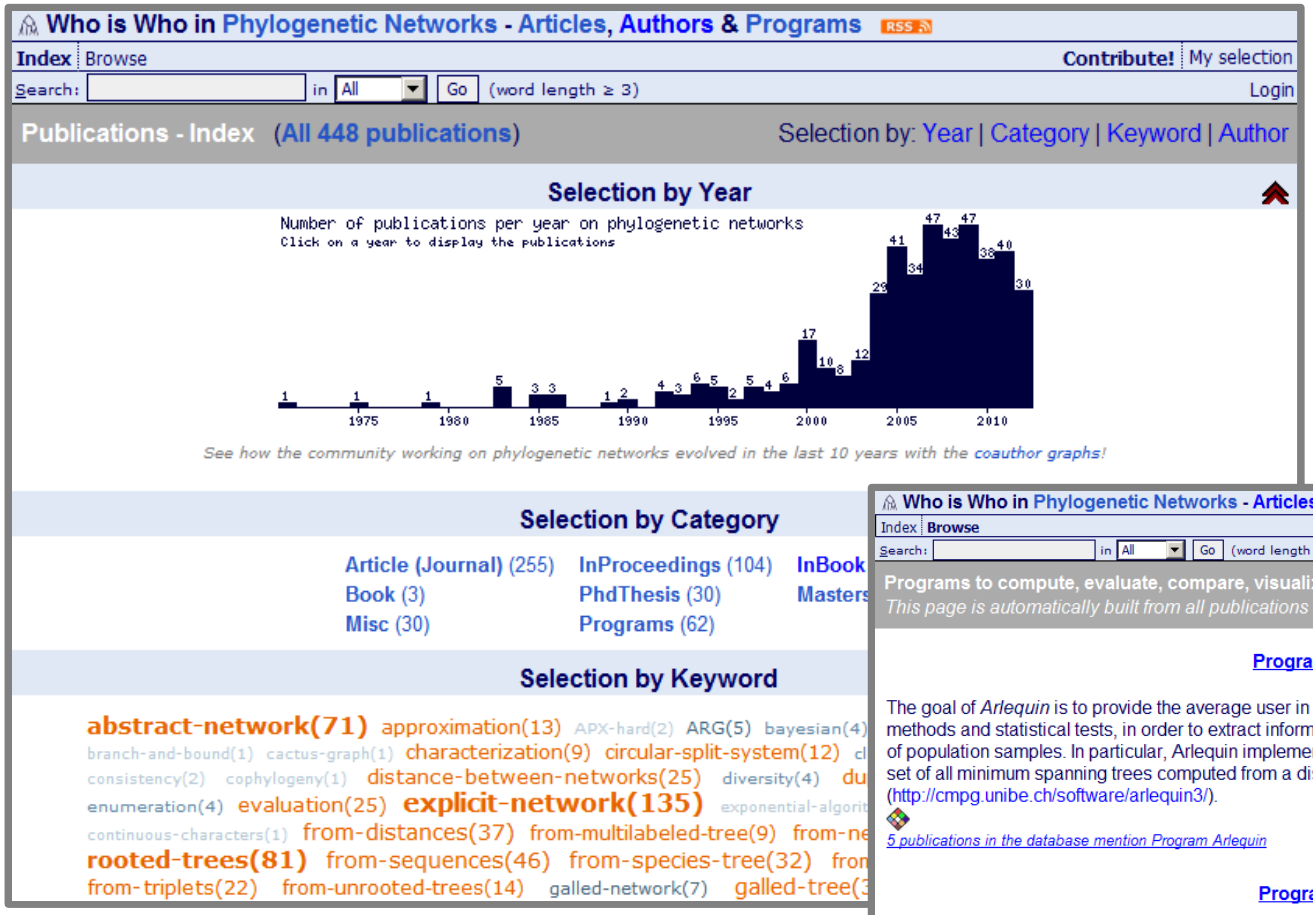
**The Future of Phylogenetic Networks**

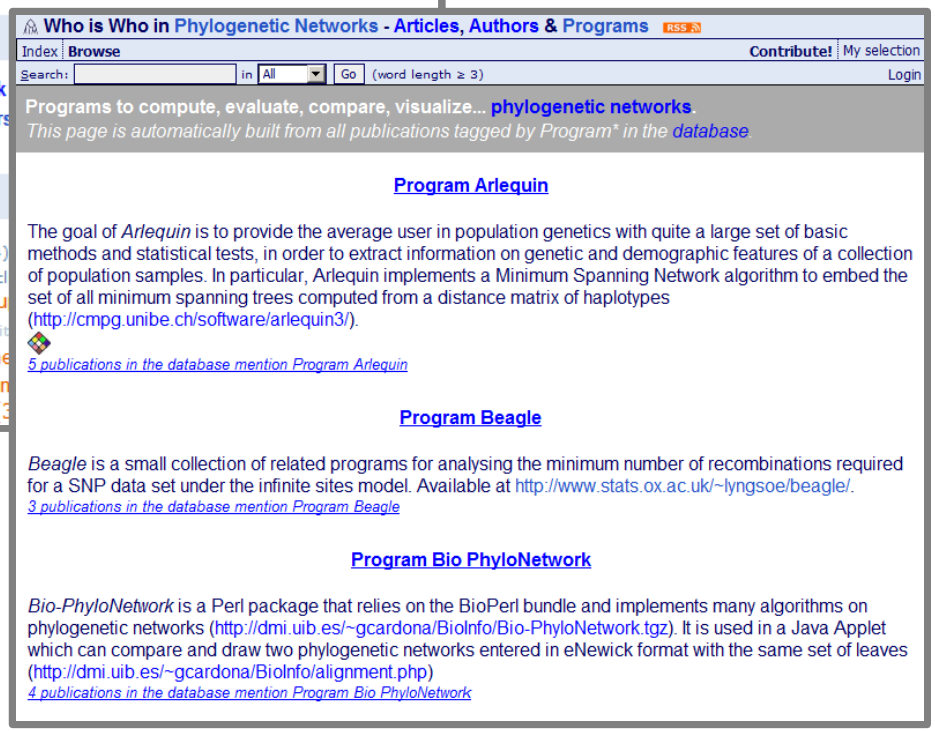
15 – 19 October 2012
Leiden, Netherlands

Scientific organizers:

- Leo van Iersel (Amsterdam)
- Steven Kelk (Maastricht)
- David Morrison (Uppsala)
- Leen Stougie (Amsterdam)

# Programmes sur les réseaux phylogénétiques



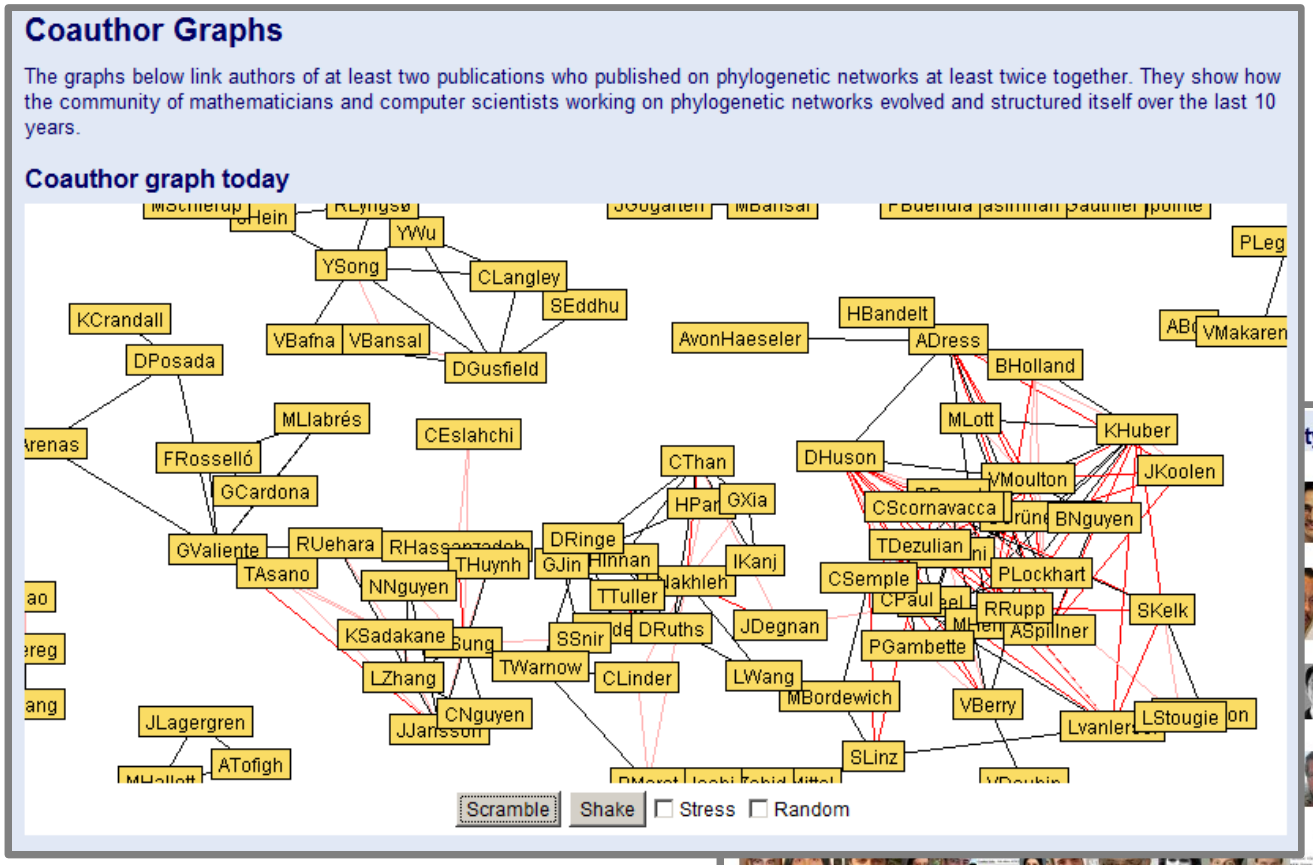




Who is Who in Phylogenetic Networks, **Articles**, Authors & **Programs**

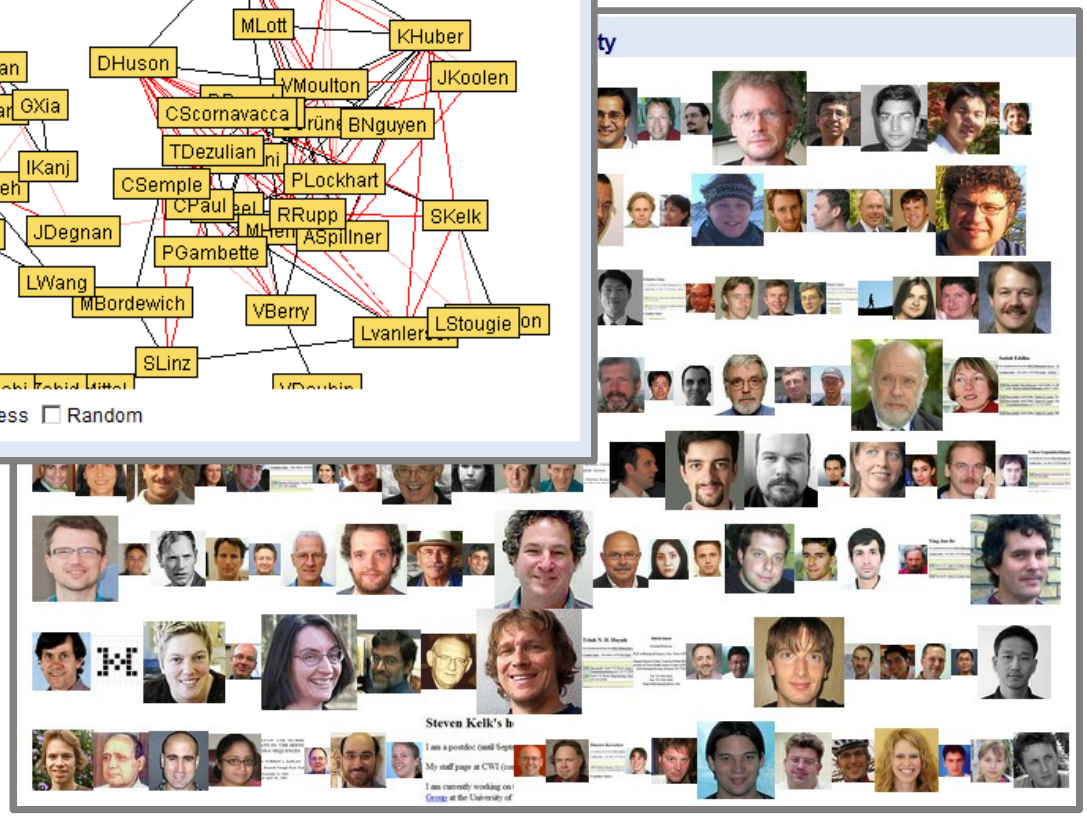
Basé sur BibAdmin
par Sergiu Chelcea
+ nuages de mots, histogramme des dates, liste des journaux, graphes de co-auteurs, définition des mots-clés.

http://www.atgc-montpellier.fr/phylnet

# Programmes sur les réseaux phylogénétiques




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

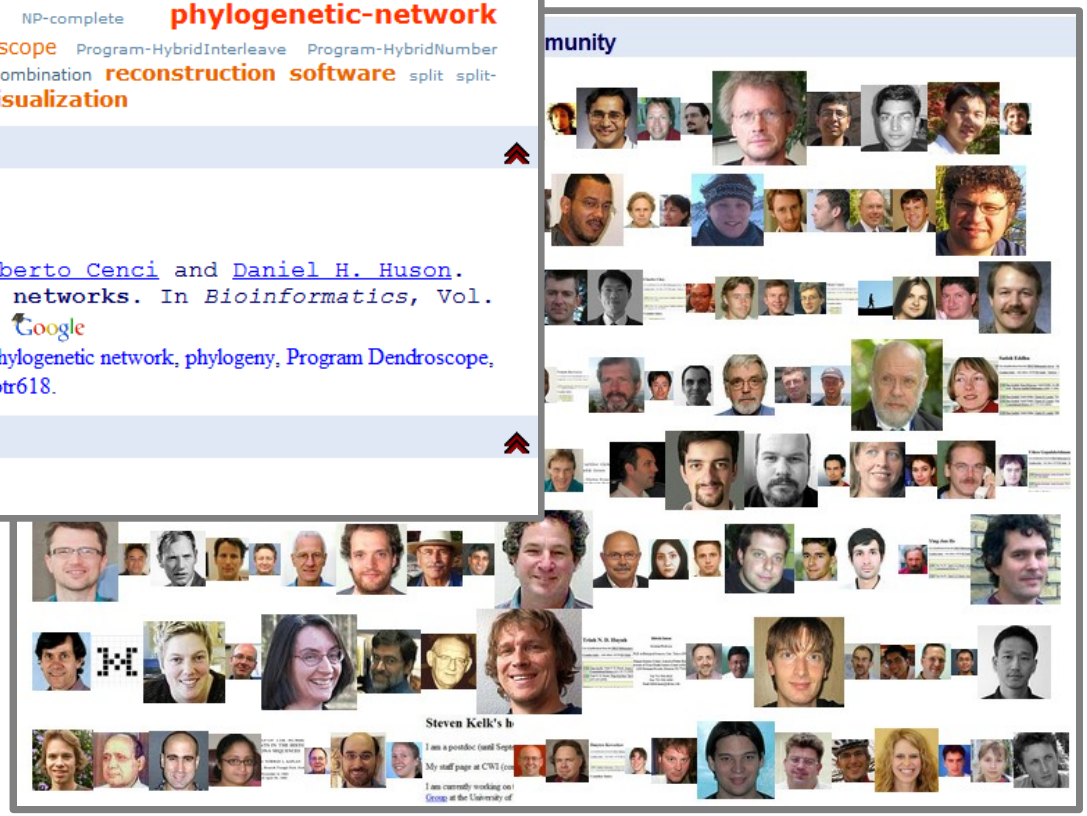


Basé sur BibAdmin
par Sergiu Chelcea
+ nuages de mots, histogramme des dates, liste des journaux, graphes de co-auteurs, définition des mots-clés.

http://www.atgc-montpellier.fr/phylnet

# Programmes sur les réseaux phylogénétiques


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




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

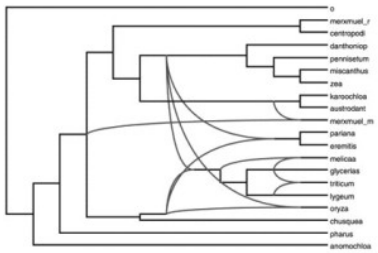
http://www.atgc-montpellier.fr/phylnet

# Les réseaux phylogénétiques

**Réseau phylogénétique** : réseau représentant des données d'évolution

- réseaux phylogénétiques **explicites**

**modélisation de l'évolution**



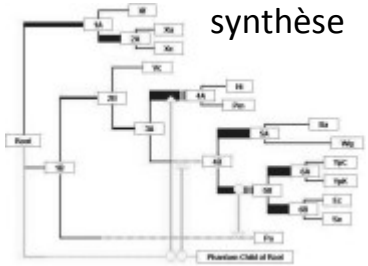
réseau à une couche de réticulation

Dendroscope

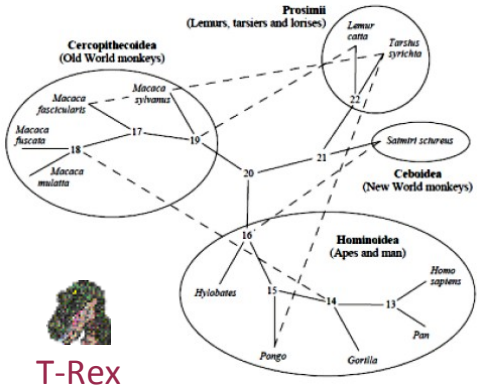
Simplistic

réseau de niveau 2

diagramme de synthèse



HorizStory

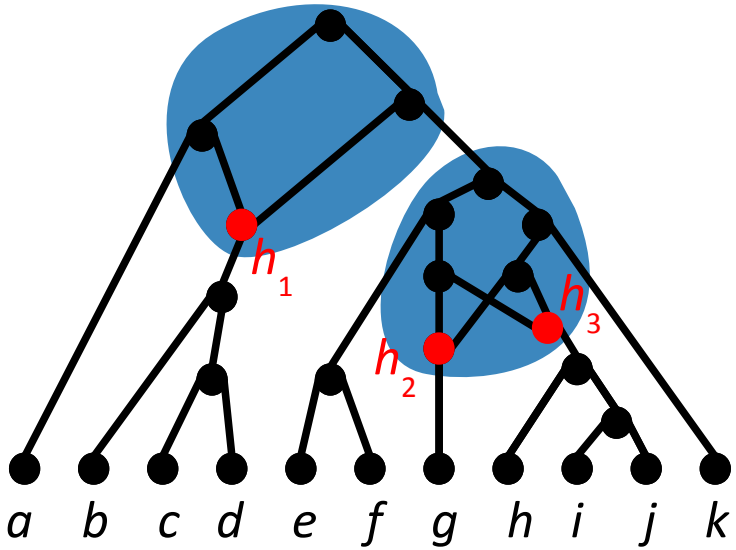



T-Rex

réticulogramme

# Les réseaux phylogénétiques

**Réseau phylogénétique explicite enraciné :**

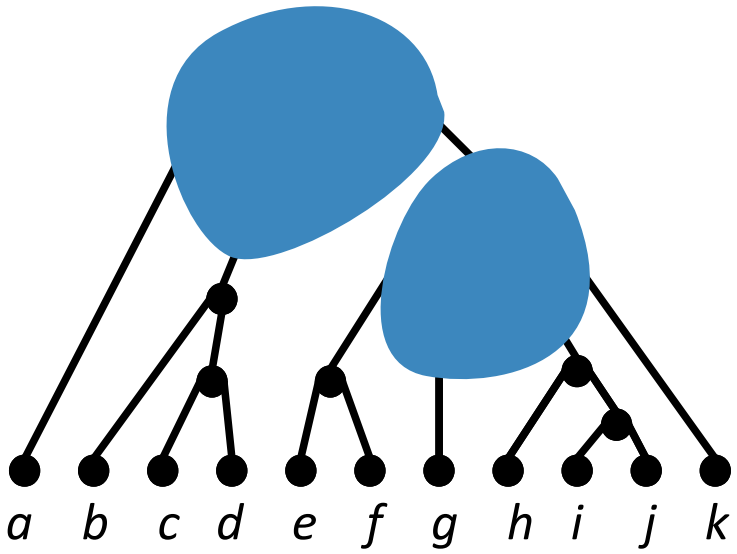



Sommets à plus d'un parent :
***réticulations / hybridations***

Partie **non arborée** : ***blob***.

# Les réseaux phylogénétiques

**Réseau phylogénétique explicite enraciné :**




Partie **non arborée** : ***blob***.

**Structure arborée des blobs** exploitée par les algorithmes :
- DIVISER : attribuer les données au blob auquel elles correspondent
- RESOUDRE : résoudre le problème sur chaque blob
- COMBINER : combiner les réseaux reconstruits sur chaque blob

# Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Méthodes combinatoires de reconstruction
- Utilisation pratique
- Illustrations
- Perspectives

# Reconstruction de réseaux phylogénétiques

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

G1 G2

{séquences de gènes}

*méthodes de distance*

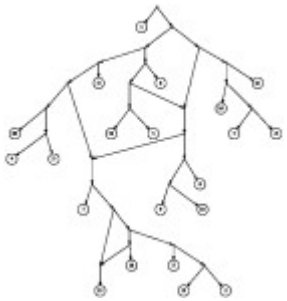
*Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002 - Chan, Jansson, Lam & Yiu 2006*

*méthodes de parcimonie*

*Hein 1990 - Kececioglu & Gusfield 1994 - Jin, Nakhleh, Snir, Tuller 2009 - Park, Jin & Nakhleh 2010 - Kannan & Wheeler, 2012*

*méthodes de vraisemblance*

*Snir & Tuller 2009 - Jin, Nakhleh, Snir, Tuller 2009 - Velasco & Sober 2009 - Meng & Kubatko 2009*



réseau *N*

# Reconstruction de réseaux phylogénétiques

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

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

G1 G2

{séquences de gènes}

*méthodes de distance*

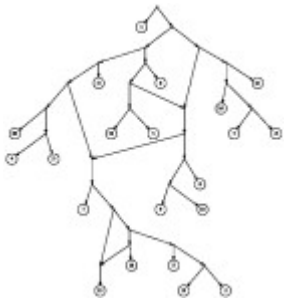
*Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002 - Chan, Jansson, Lam & Yiu 2006*

*méthodes de parcimonie*

*Hein 1990 - Kececioglu & Gusfield 1994 - Jin, Nakhleh, Snir, Tuller 2009 - Park, Jin & Nakhleh 2010 - Kannan & Wheeler, 2012*

*méthodes de vraisemblance*

*Snir & Tuller 2009 - Jin, Nakhleh, Snir, Tuller 2009 - Velasco & Sober 2009 - Meng & Kubatko 2009*



réseau *N*

# Reconstruction de réseaux phylogénétiques

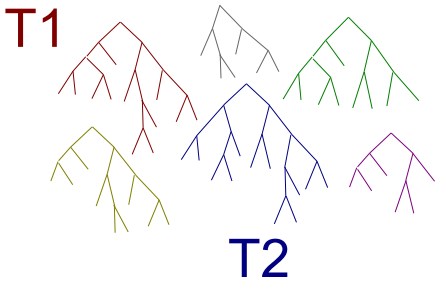
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

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

Guindon & Gascuel, SB, 2003




{arbres}

Base HOGENOM

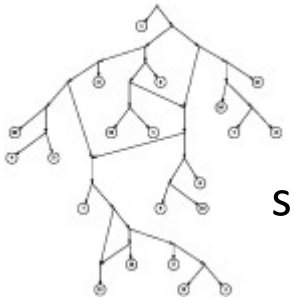
Phyl-ARIANE

ANR

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

*Réconciliation ou consensus d'arbres*

réseau explicite



super-réseau optimal *N*

- contient les arbres en entrée
- a le moins de réticulations

# Reconstruction de réseaux phylogénétiques

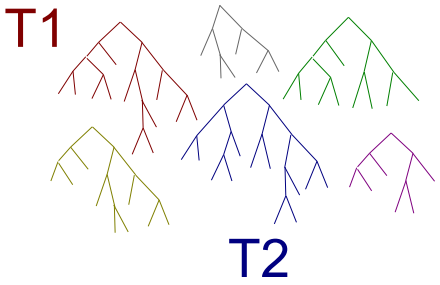
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

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

Guindon & Gascuel, SB, 2003




{arbres}

Base HOGENOM

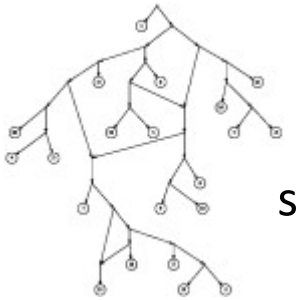
Phyl-ARIANE

ANR

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

*Réconciliation ou consensus d'arbres*

réseau explicite



super-réseau optimal *N*

**Problème : la réconciliation d'arbres est un problème difficile**

(NP-complet pour 2 arbres avec le minimum d'hybridations)

Bordewich & Semple, DAM, 2007

# Reconstruction de réseaux phylogénétiques

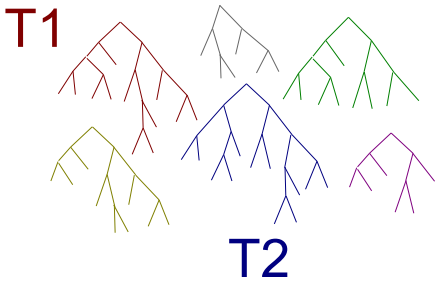
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

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

Guindon & Gascuel, SB, 2003




{arbres}

Base HOGENOM

Phyl-ARIANE

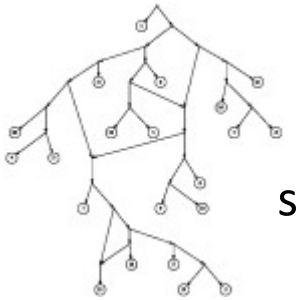
ANR

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

> 500 espèces, >70 000 arbres

*Réconciliation ou consensus d'arbres*

réseau explicite



super-réseau optimal *N*

**Problème : la réconciliation d'arbres est un problème difficile**

(NP-complet pour 2 arbres avec le minimum d'hybridations)

Bordewich & Semple, DAM, 2007

# Reconstruction de réseaux phylogénétiques

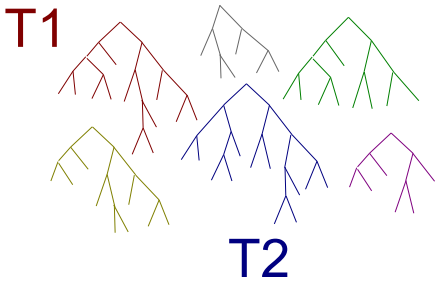
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

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

Guindon & Gascuel, SB, 2003




{arbres}

Base HOGENOM

Phyl-ARIANE

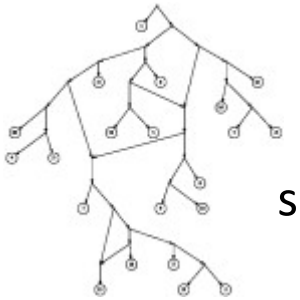
ANR

Penel, Arigon, Dufayard, Sertier, Daubin, Duret, Gouy & Perrière, BMC Bioinf 10(supp. 6):S3, 2009

> 1400 espèces, >290 000 arbres (v6)

*Réconciliation ou consensus d'arbres*

réseau explicite



super-réseau optimal *N*

**Problème : la réconciliation d'arbres est un problème difficile**

(NP-complet pour 2 arbres avec le minimum d'hybridations)

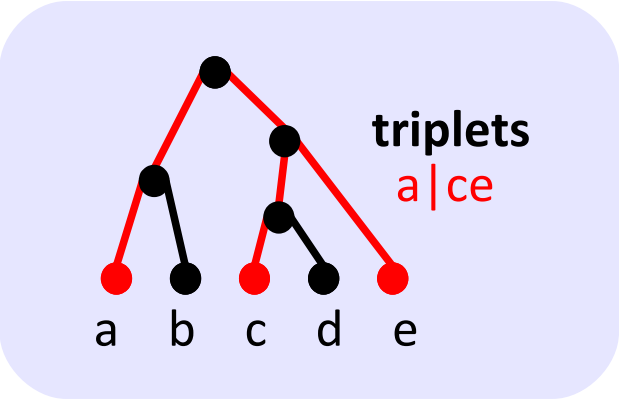
Bordewich & Semple, DAM, 2007

# Triplets et clades

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

**difficile**.

**Idée :**
reconstuire un réseau contenant tous les :



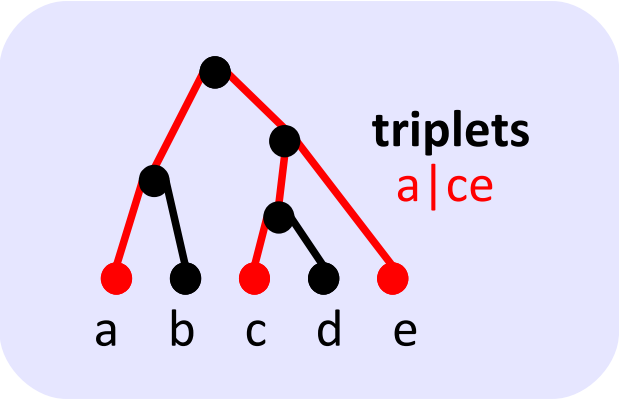

des arbres en entrée ?

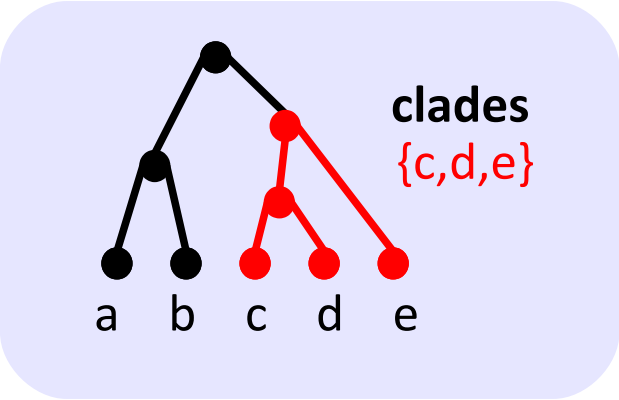
# Triplets et clades

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

**difficile**.

**Idée :**
reconstuire un réseau contenant tous les :



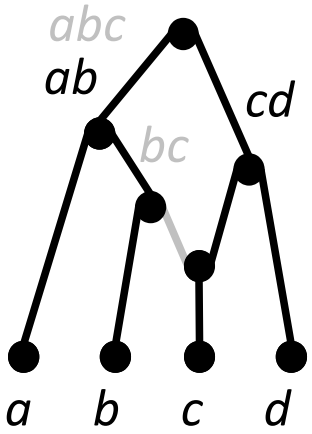




des arbres en entrée ?

# Clades (souples)

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

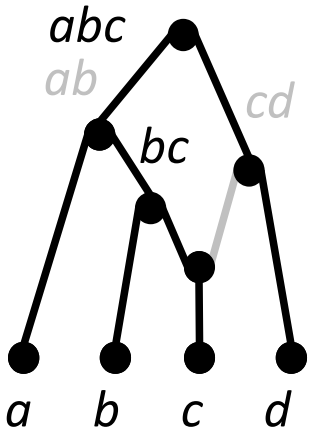
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



# Clades (souples)

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

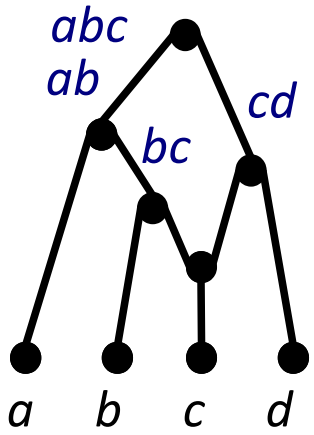
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



# Clades stricts et souples

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

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

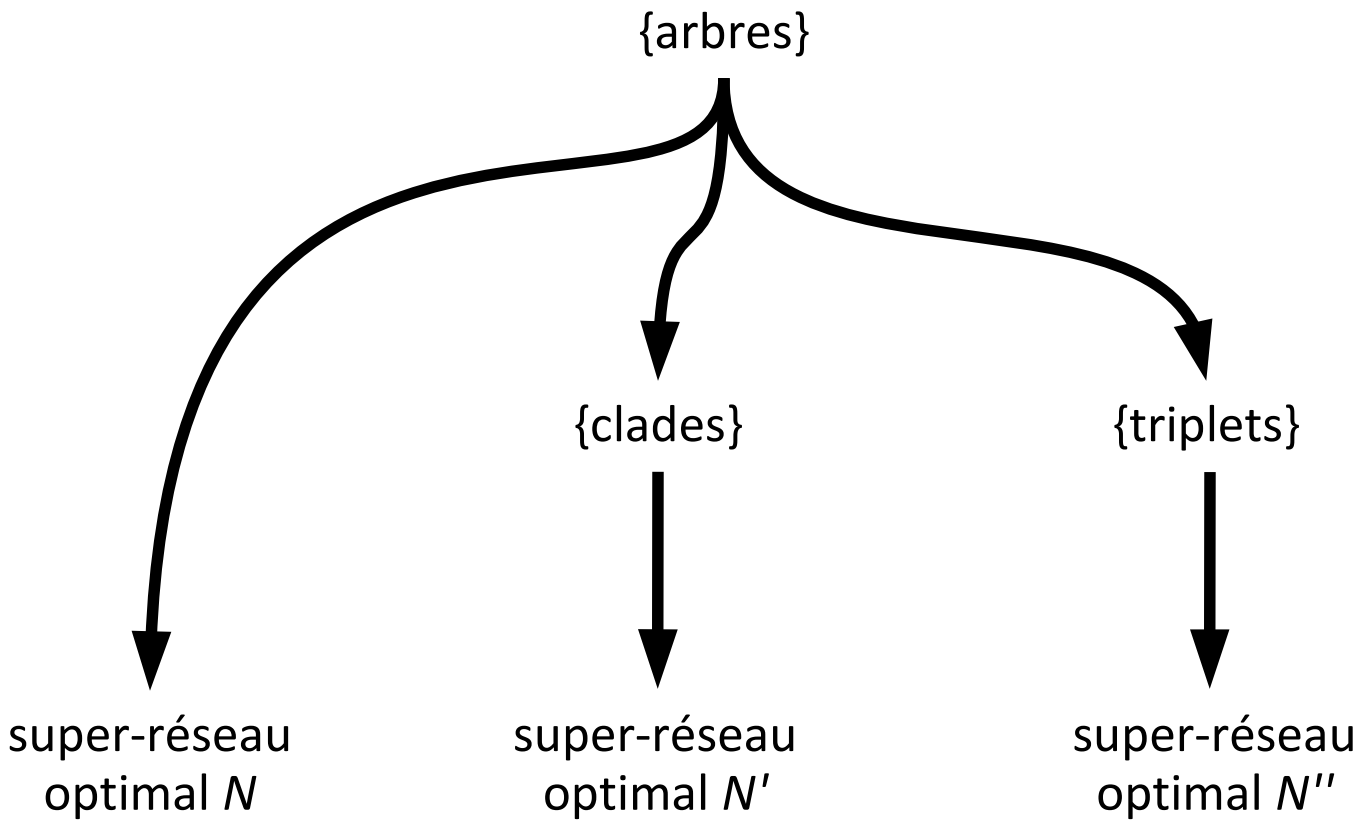



L'ensemble *S*(*N*) de **tous les clades souplement compatibles** avec *N* peut être de taille **exponentielle.**
Tester si un **clade souple** appartient à un réseau : **NP-complet.**

Kanj, Nakhleh, Than & Xia, TCS, 2008

# Reconstruction combinatoire de réseaux phylogénétiques

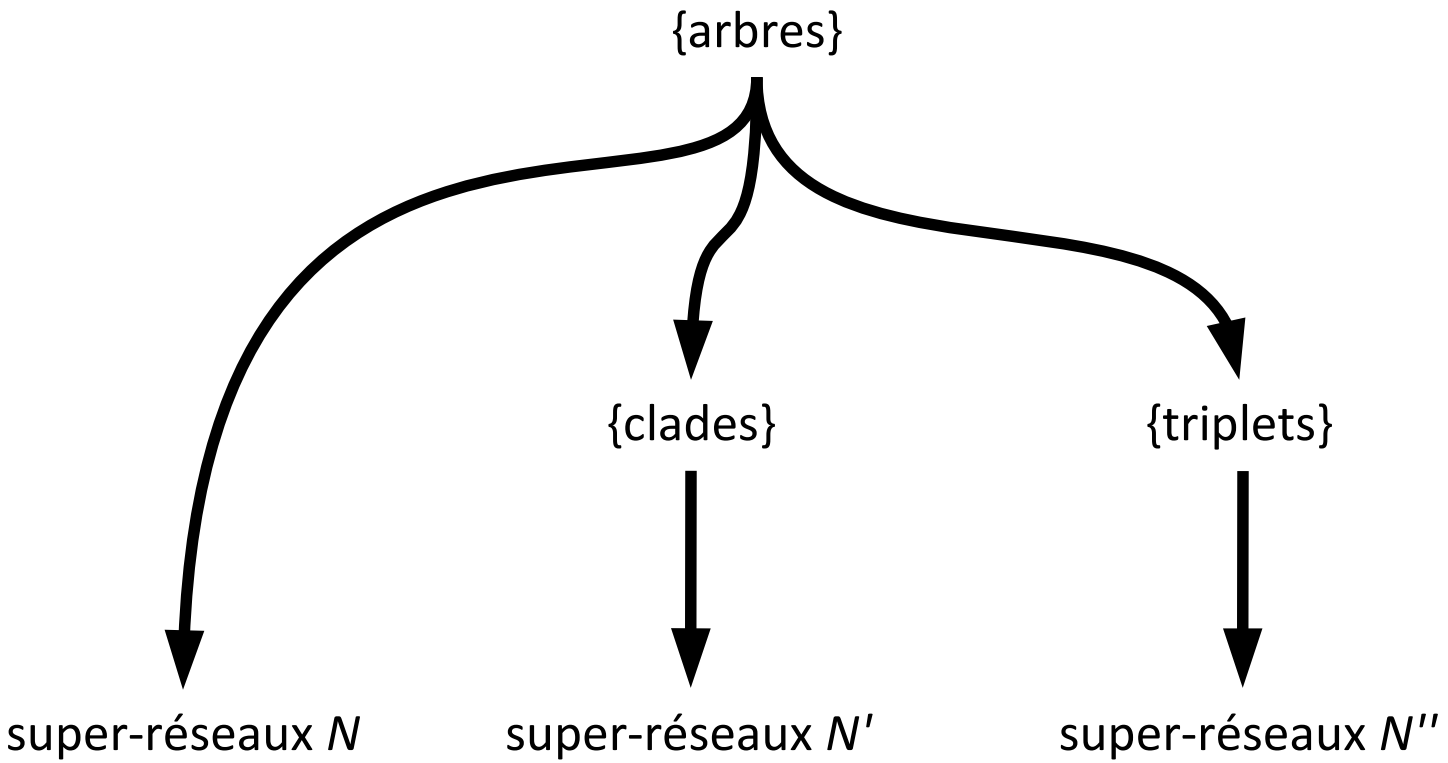
**Idée :**
modifier le type de données à traiter




$$N = N' = N'' \ ?$$

# Reconstruction combinatoire de réseaux phylogénétiques

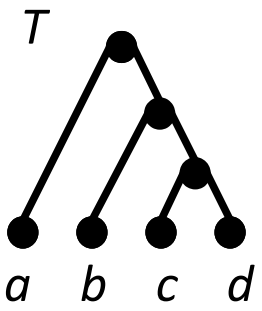
**Idée :**
modifier le type de données à traiter

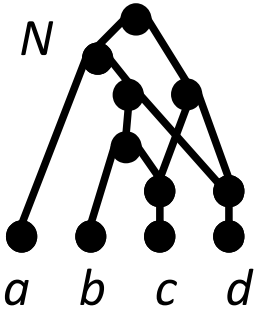



$$\{ N \} \subseteq \{ N' \} \subseteq \{ N'' \}$$

# Reconstruction combinatoire de réseaux phylogénétiques

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



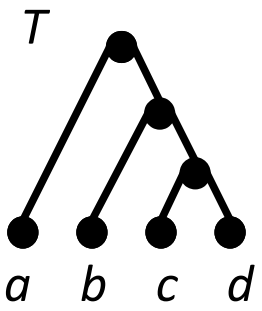



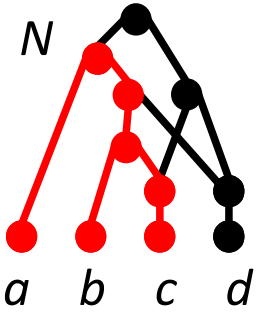

contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

# Reconstruction combinatoire de réseaux phylogénétiques

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



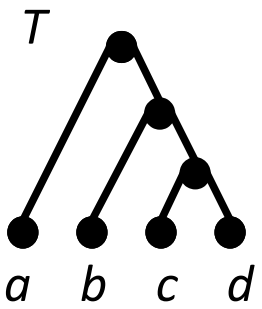



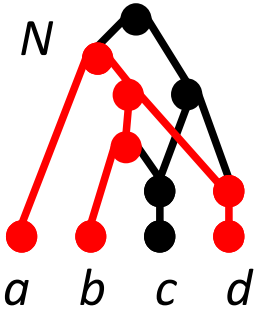

contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

# Reconstruction combinatoire de réseaux phylogénétiques

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



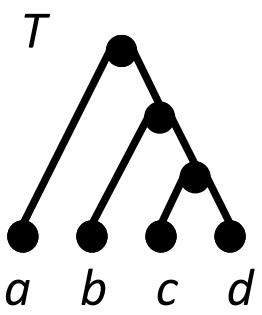



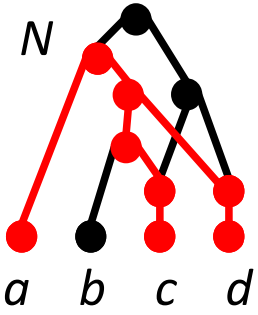

contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

# Reconstruction combinatoire de réseaux phylogénétiques

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



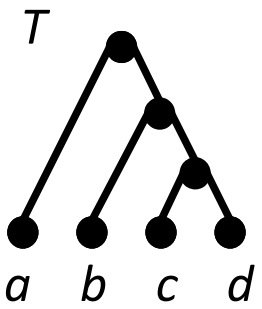



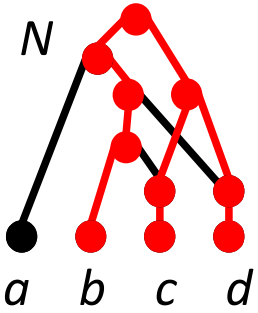

contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

# Reconstruction combinatoire de réseaux phylogénétiques

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



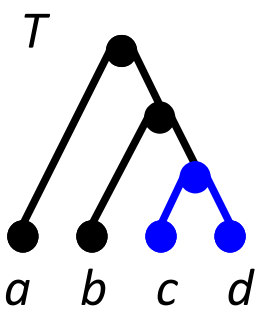



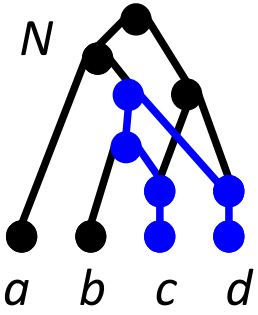

contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

# Reconstruction combinatoire de réseaux phylogénétiques

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



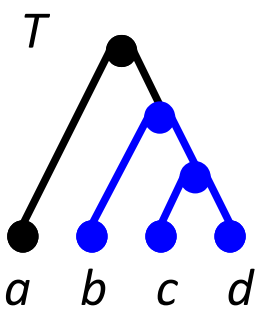



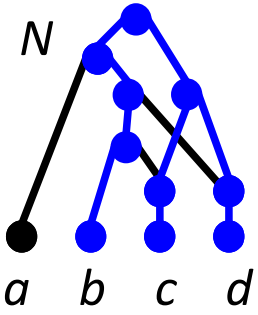

contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

# Reconstruction combinatoire de réseaux phylogénétiques

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



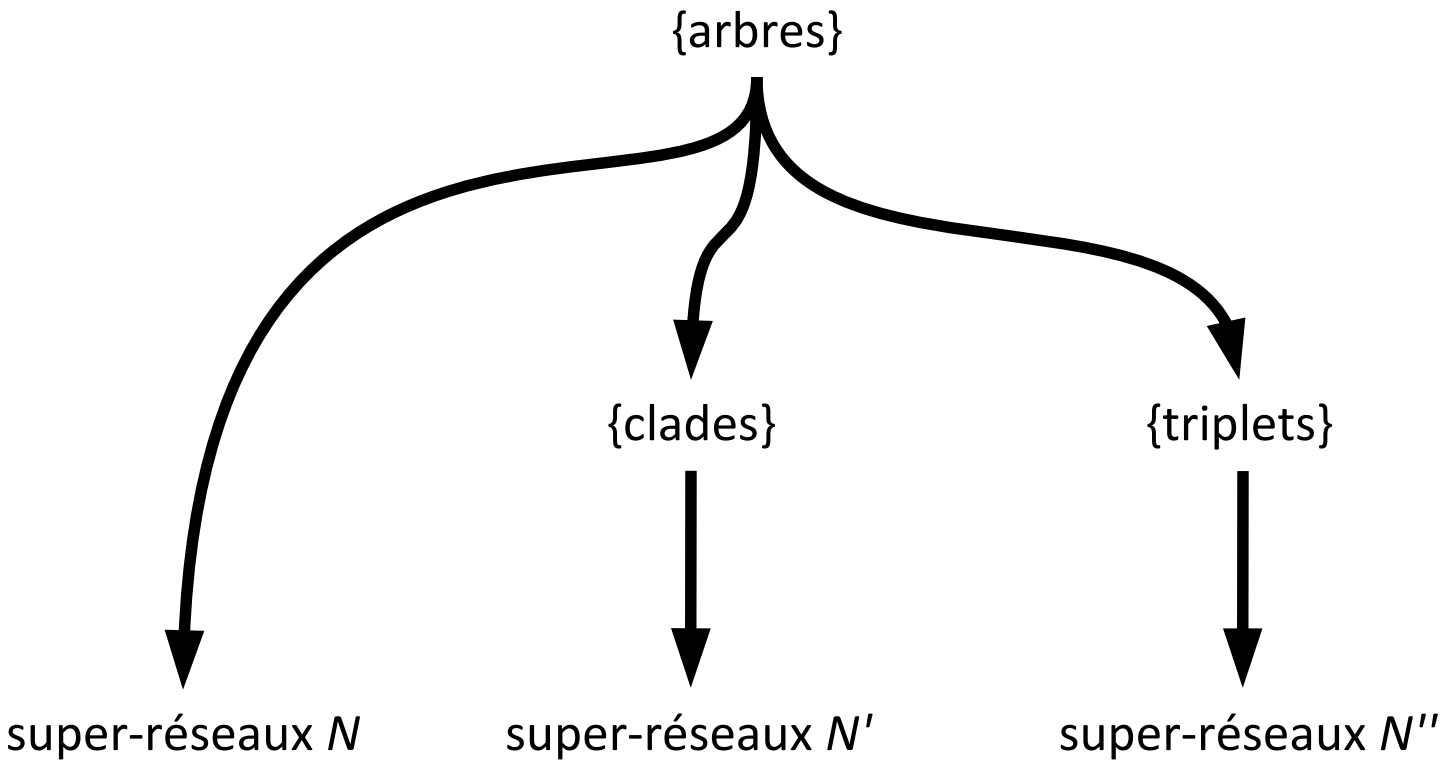




contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

# Reconstruction combinatoire de réseaux phylogénétiques

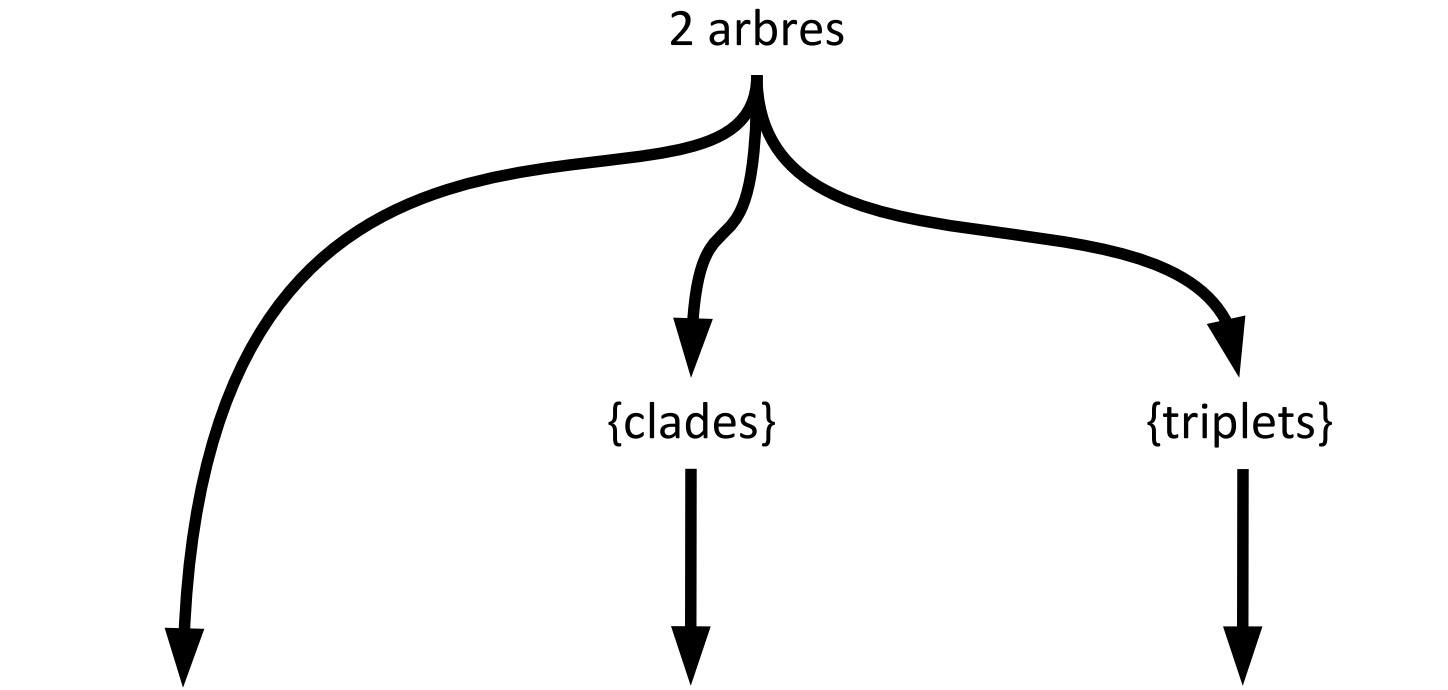
**Idée :**
modifier le type de données à traiter




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

# Reconstruction combinatoire de réseaux phylogénétiques

**Idée :**
modifier le type de données à traiter




$$\{ N \} = \{ N' \} = \{ N'' \}$$

Van Iersel & Kelk, Journal of Theoretical Biology, 2011

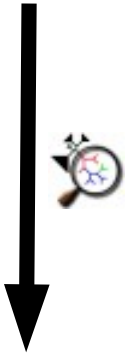
# Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Méthodes combinatoires de reconstruction
- Utilisation pratique
- Illustrations
- Perspectives

# Reconstruction depuis les clades souples

{arbres}

{clades}

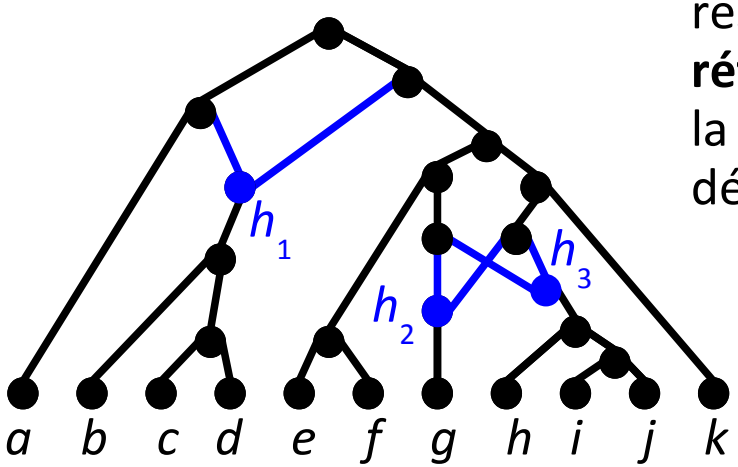


*N'*

réseau à 1 couche de réticulation

Méthode exacte rapide de reconstruction de **réseaux à 1 couche de réticulation** à partir de **clades souples**

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




réseau à une couche de réticulation.

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

# Reconstruction depuis les clades souples

{arbres}

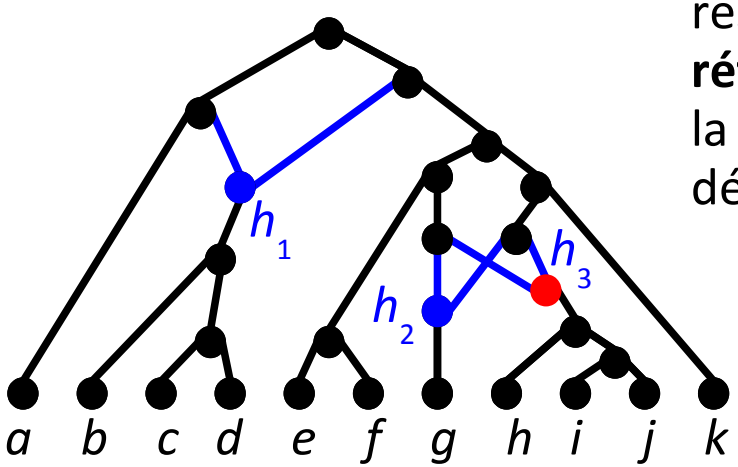
{clades}

*N'*

réseau à 1 couche de réticulation

Méthode exacte rapide de reconstruction de **réseaux à 1 couche de réticulation** à partir de **clades souples**

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




réseau à une couche de réticulation.

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

# Reconstruction depuis les clades souples

{arbres}

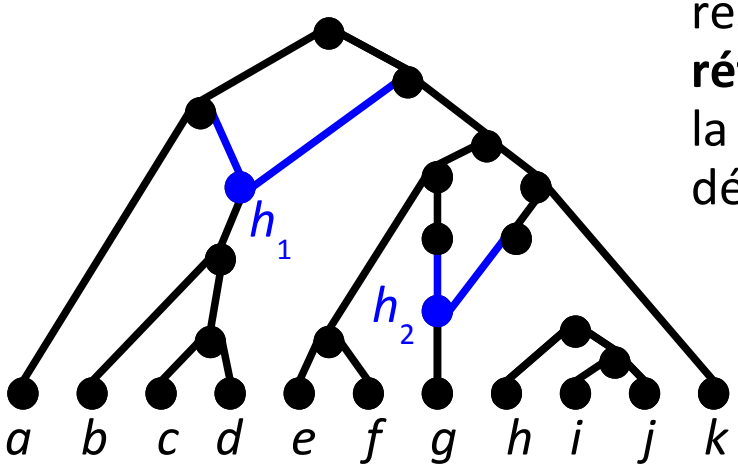
{clades}

$N'$

réseau à 1 couche de réticulation

Méthode exacte rapide de reconstruction de **réseaux à 1 couche de réticulation** à partir de **clades souples**

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




réseau à une couche de réticulation.

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

# Reconstruction depuis les clades souples

{arbres}

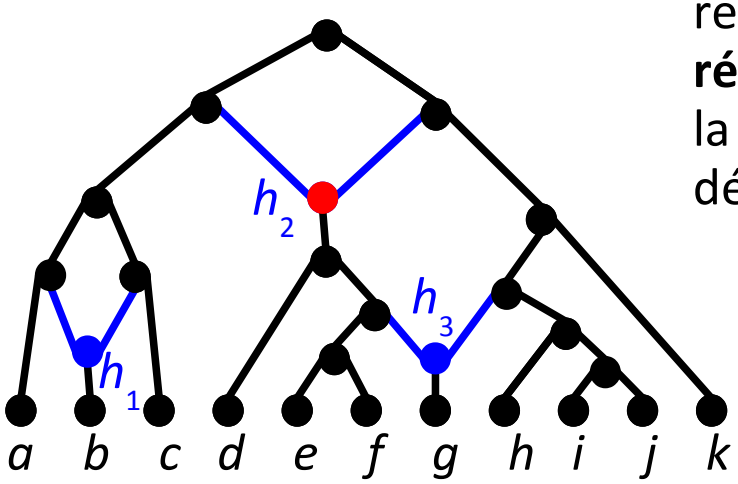
{clades}

*N'*

réseau à 1 couche de réticulation

Méthode exacte rapide de reconstruction de **réseaux à 1 couche de réticulation** à partir de **clades souples**

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




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

réseau à **deux** couches de réticulation.

# Reconstruction depuis les clades souples

{arbres}

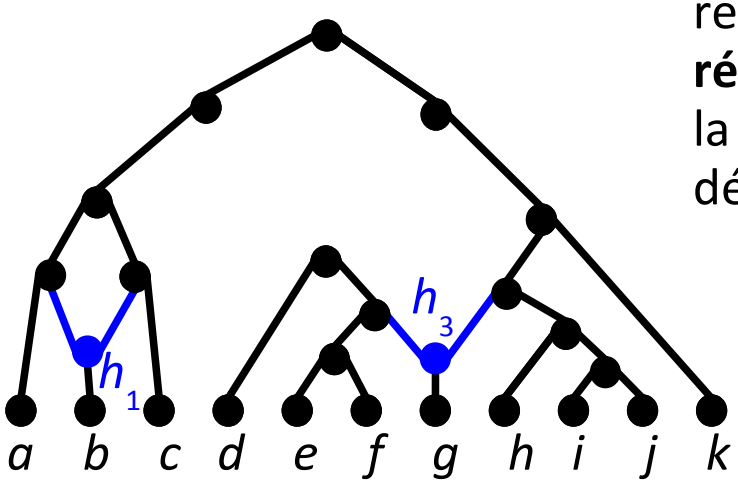
{clades}

*N'*

réseau à 1 couche de réticulation

Méthode exacte rapide de reconstruction de **réseaux à 1 couche de réticulation** à partir de **clades souples**

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




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

réseau à **deux** couches de réticulation.

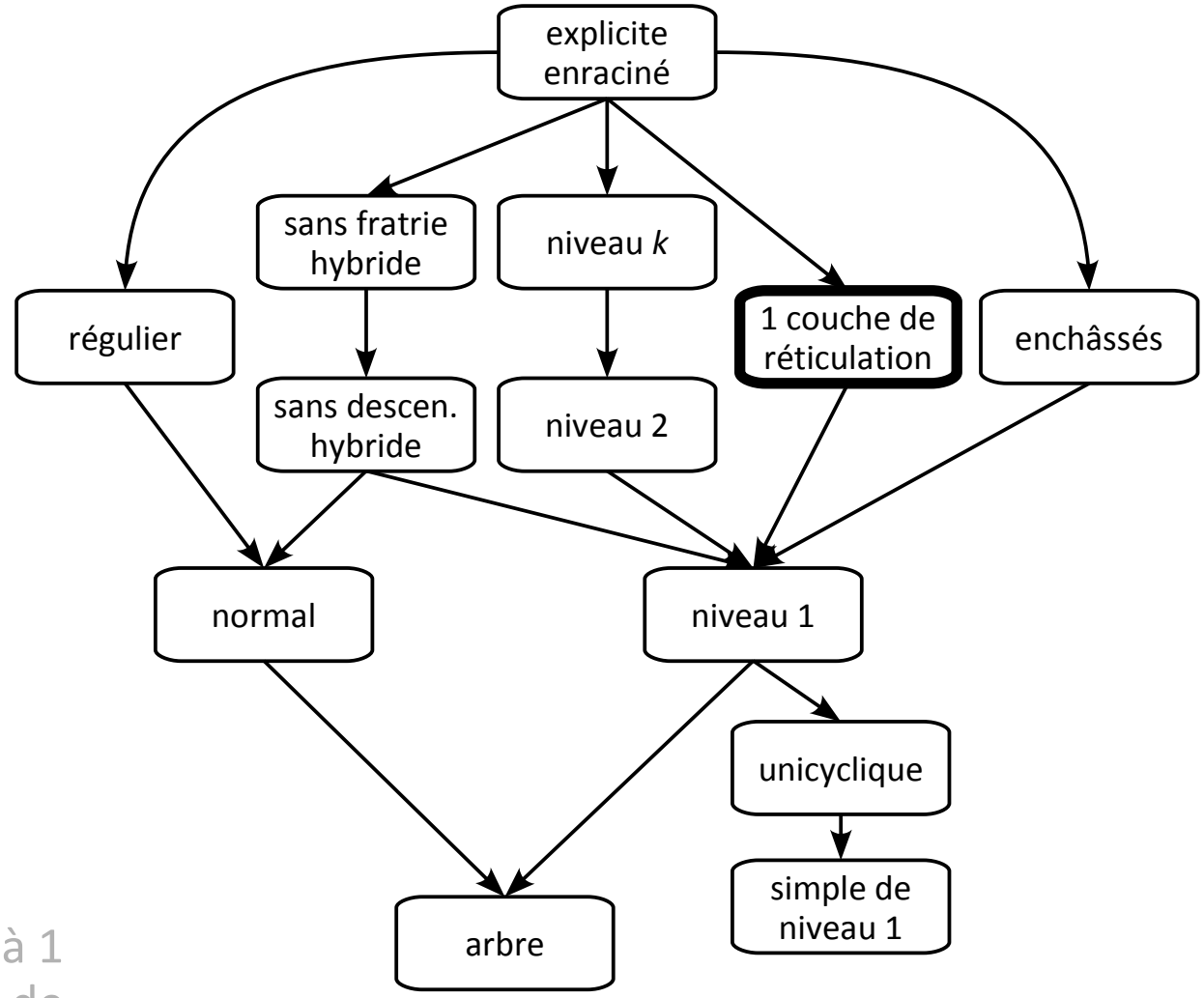
# Classes de réseaux phylogénétiques restreints

{arbres}

{clades}

*N'*

réseau à 1 couche de réticulation




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

# Reconstruction depuis les clades souples

{arbres}

{clades}

*N'*

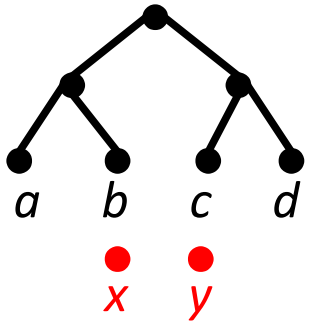
réseau à 1 couche de réticulation

Méthode exacte rapide de reconstruction de **réseaux à 1 couche de réticulation** à partir de **clades souples**

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

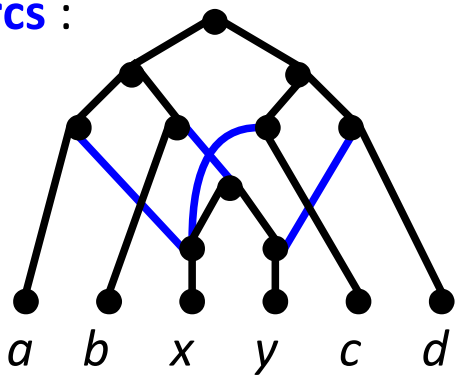
Etape 1- Résoudre les **conflits entre clades** :
- partie sans conflits ➡ arbre,
- minimum de taxons impliqués dans des **conflits** ➡ sous les réticulations.

MAXIMUM COMPATIBLE SUBSET




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

MINIMUM ATTACHMENT



# Reconstruction depuis les clades souples

{arbres}

{clades}

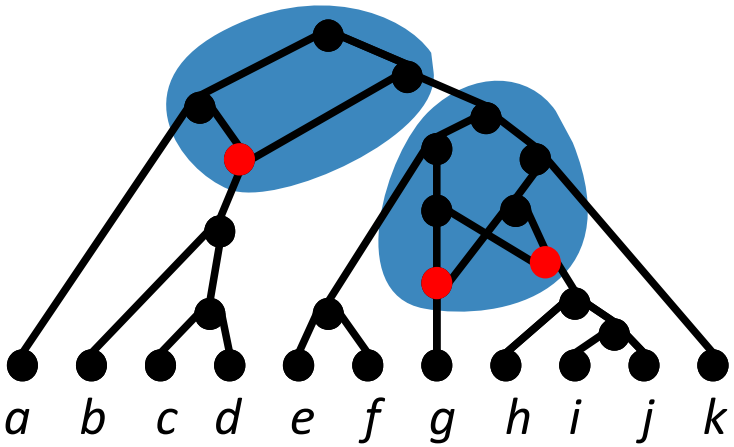
*N'*

réseau de niveau *k*

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

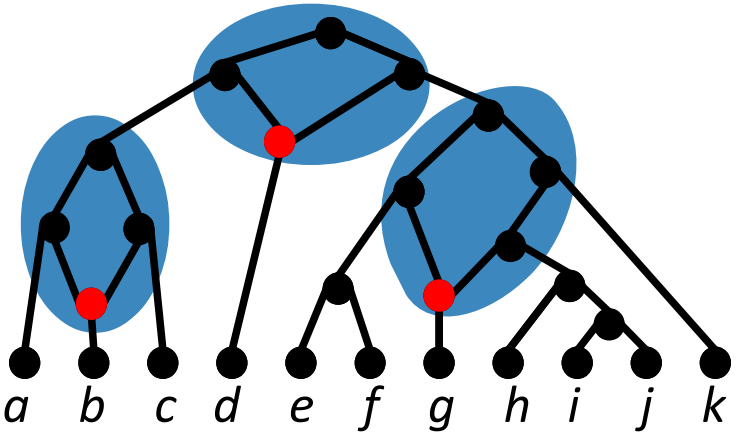
Iersel, Kelk, Rupp & Huson, ISMB 2010

↳ moins de réticulations, mais plus lente pour niveau > 2.




réseau de niveau 2

**niveau** =
nombre maximum d'**hybridations** par **blob**.




réseau de niveau 1
(“*galled tree*”)

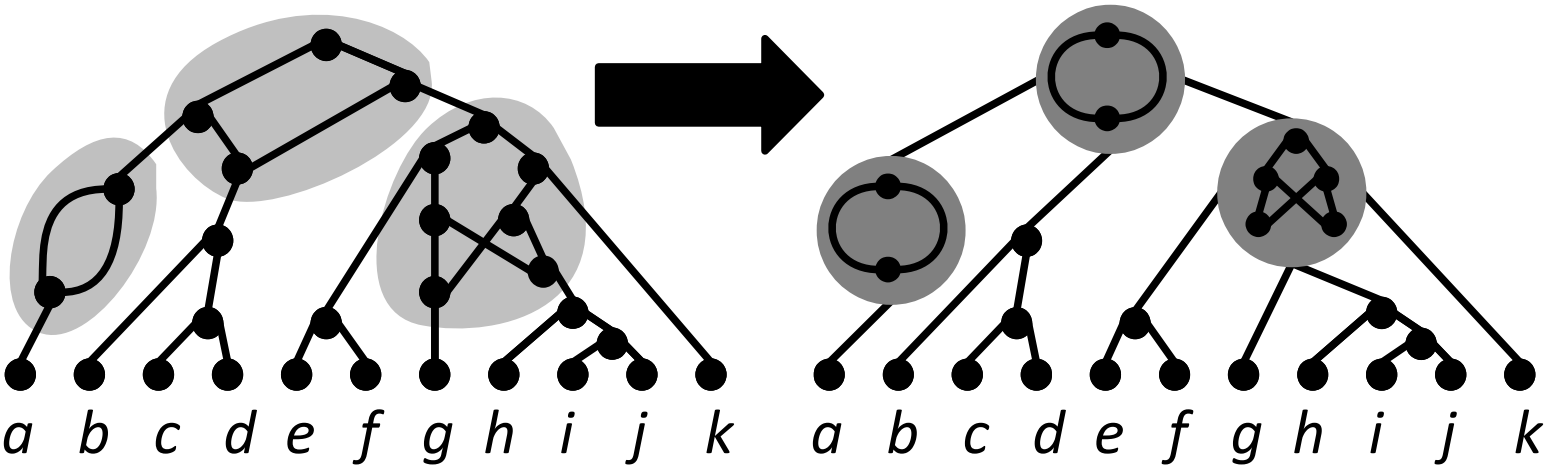
# Réseaux de niveau *k*

{arbres}

{clades}

*N'*

réseau de niveau *k*




*N*, réseau de niveau *k*

*N* décomposé en arbre de motifs de graphes simples : **générateurs**

Générateurs de niveau 1, 2, 3, 4, 5 :
1, 4, 65, 1993, 91454

(croissance exponentielle en *k*)

Gambette, Berry & Paul, CPM 2009

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

# Réseaux de niveau *k*

{arbres}

**Réseaux non enracinés de niveau 1**

Formule explicite pour *n* feuilles, *c* cycles, *m* arêtes dans les feuilles

Semple & Steel, *TCBB*, 2006

+ évaluation asymptotique pour *n* feuilles : $\approx 0.207\ (1.890)^n\ n^{n-1}$

{clades}

**Réseaux enracinés de niveau 1 :**

Formule explicite pour *n* feuilles, *c* cycles, *m* arêtes dans les feuilles

+ évaluation asymptotique pour *n* feuilles : $\approx 0.134\ (2.943)^n\ n^{n-1}$

**Réseaux non enracinés de niveau 2 :**

Formule explicite pour *n* feuilles :

$$(n-1)!\sum_{\substack{0\le s\le q\le p\le k\le i\le n-1\\ j=n-1-i-k-p-q-s\ge 0\\ i\neq 0}} \binom{n+i-1}{i}\binom{4i+j-1}{j}\binom{i}{k}\binom{k}{p}\binom{p}{q}\binom{q}{s}\left(\frac{-3}{20}\right)^s\left(\frac{9}{2}\right)^i\left(\frac{-23}{9}\right)^k(-1)^p\left(\frac{-10}{23}\right)^q$$

*N'*

réseau de niveau *k*

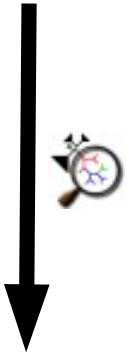
| Feuilles | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| niveau 1 non enr. | - | 2 | 15 | 192 | 3 450 | 79 740 |
| niveau 1 enr. | 3 | 36 | 723 | 20 280 | 730 755 | 32 171 580 |
| niveau 2 non enr. | - | 9 | 282 | 14 697 | 1 071 750 | 100 467 405 |

Berry, Bouvel, Gambette & Paul, 2012

# Reconstruction depuis les clades souples

{arbres}

{clades}



*N'*

réseau de niveau *k*

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

**Algorithme de complexité paramétrée en *k***
$O(f(k).poly(n))$

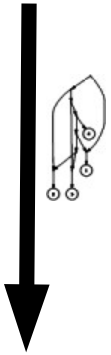
Kelk & Scornavacca, 2011

Approche théorique basée sur l'idée des générateurs

# Reconstruction depuis les triplets

{arbres}

{triplets}



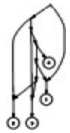
*N'*
réseau
de niveau *k*

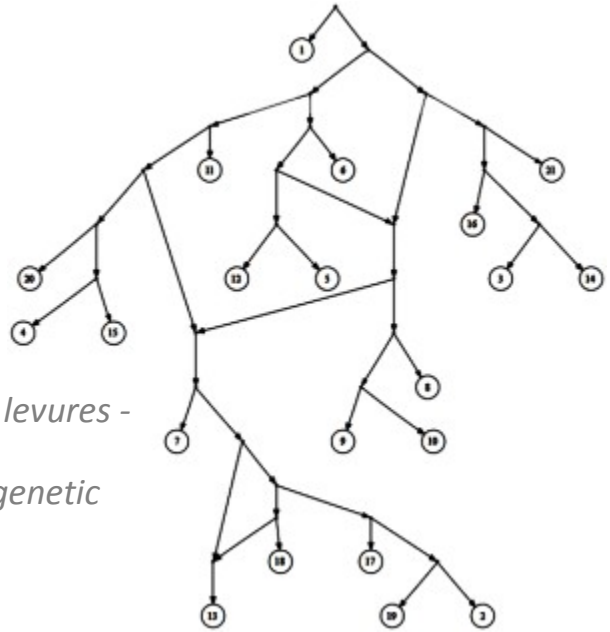
Méthodes exactes pour reconstruire un **réseau de niveau 1 et 2** (s'il en existe un) à partir d'un ensemble dense de **triplets**

Jansson, Nguyen & Sung, SODA'05 : $O(n^3)$ pour niveau 1,
van Iersel, Kelk & al, RECOMB'08 : $O(n^8)$ pour niveau 2,
To & Habib, CPM'09 : $O(n^{5k+4})$ pour niveau $k$

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

**Programme Simplistic**





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

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

# Reconstruction depuis les triplets

{arbres}

{triplets}

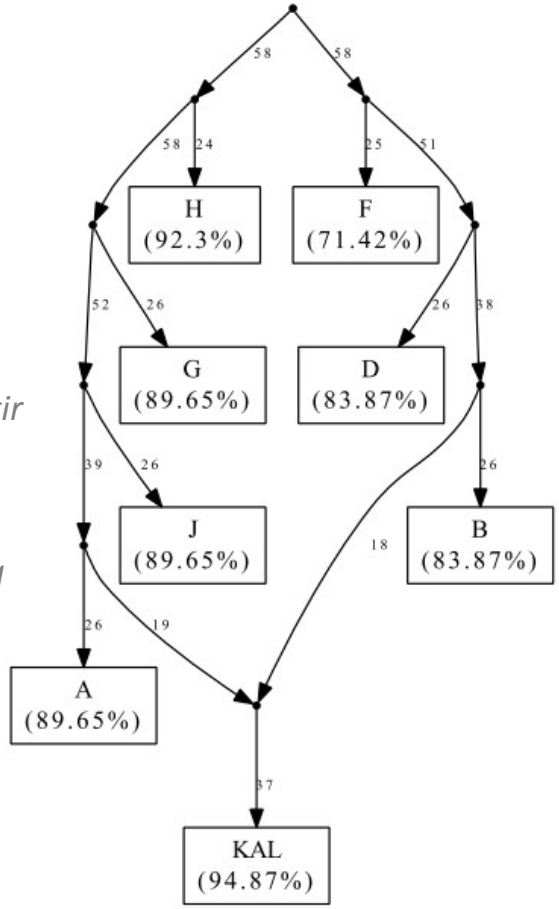
*N'*
réseau
de niveau 1

Méthode heuristique rapide pour reconstruire un **réseau de niveau 1** contenant un **maximum de triplets**

Huber, van Iersel, Kelk & Suchecki, TCBB, 2011

**Programme Lev1athan**

*Réseau phylogénétique construit à partir des triplets extraits de deux arbres de souches de VIH-1*
*Huber, van Iersel, Kelk & Suchecki*
*A practical algorithm for reconstructing level-1 phylogenetic networks*
*TCBB, 2011*




http://homepages.cwi.nl/~kelk/lev1athan/

# Reconstruction depuis les arbres

{arbres}

*N'*
réseau
optimal

Méthodes exactes pour reconstruire un **réseau optimal** contenant les 2 arbres en entrée

**Complexité théorique**
NP-difficile, APX-difficile

Bordewich & Semple, 2007

Complexité **paramétrée en *h*** :
$O((28h)^h + n^3)$ Bordewich & Semple, 2007
$O((6h)^h \text{ poly}(h))$ Kelk, 2011
$O(3.18^h n)$ Whidden, Beiko & Zeh, 2011

**Lien avec Feedback Vertex Set**

van Iersel, Kelk, Lekic & Scornavacca, 2012

→ Pas de 1.36-approximation sauf si P=NP
→ Pas de (2-$\varepsilon$)-approximation sauf si la *Conjecture des Jeux Uniques* est fausse
→ $O(\log h \log \log h)$-approximation

# Reconstruction depuis les arbres

{arbres}

*N'*
réseau
optimal

Méthodes pour reconstruire un **réseau optimal** contenant les 2 arbres en entrée

**Méthodes exactes**

Programmes **Dendroscope** et **HybridNet**

Albrecht, Scornavacca, Cenci & Huson, *Bioinformatics*, 2012
Chen & Wang, *TCBB*, 2012

Un réseau (ou plusieurs) avec le moins d'hybridations

*En pratique* : limite à 30-40 hybridations

# Reconstruction depuis les arbres

{arbres}

*N'*
réseau
optimal

Méthodes pour reconstruire un **réseau optimal** contenant les 2 arbres en entrée

**Méthode approchée**

Programme **CycleKiller**

van Iersel, Kelk, Lekic & Scornavacca, 2012

Algorithme **exponentiel**
(rapide en pratique : jusqu'à 97 hybridations)
de **2-approximation**
(performant en pratique : optimal trouvé dans >95% des cas)

Version plus rapide pour une 4-approximation
Réseaux de 10000 hybridations pour 10000 espèces en 10 minutes

# Reconstruction depuis les arbres

{arbres}

*N'*
réseau
optimal

Méthode exacte pour reconstruire un **réseau optimal** contenant les arbres en entrée

**Programme HybridNet**
Un ou tous les réseaux avec le moins d'hybridations
$O(k^2 n 3^{d''} + kn 6^{d} 2^{-d''} + (d\text{-}1)^{d-d'+2} 6^{d'} 2^{-d''})$ ...

Chen & Wang, *TCBB*, 2012

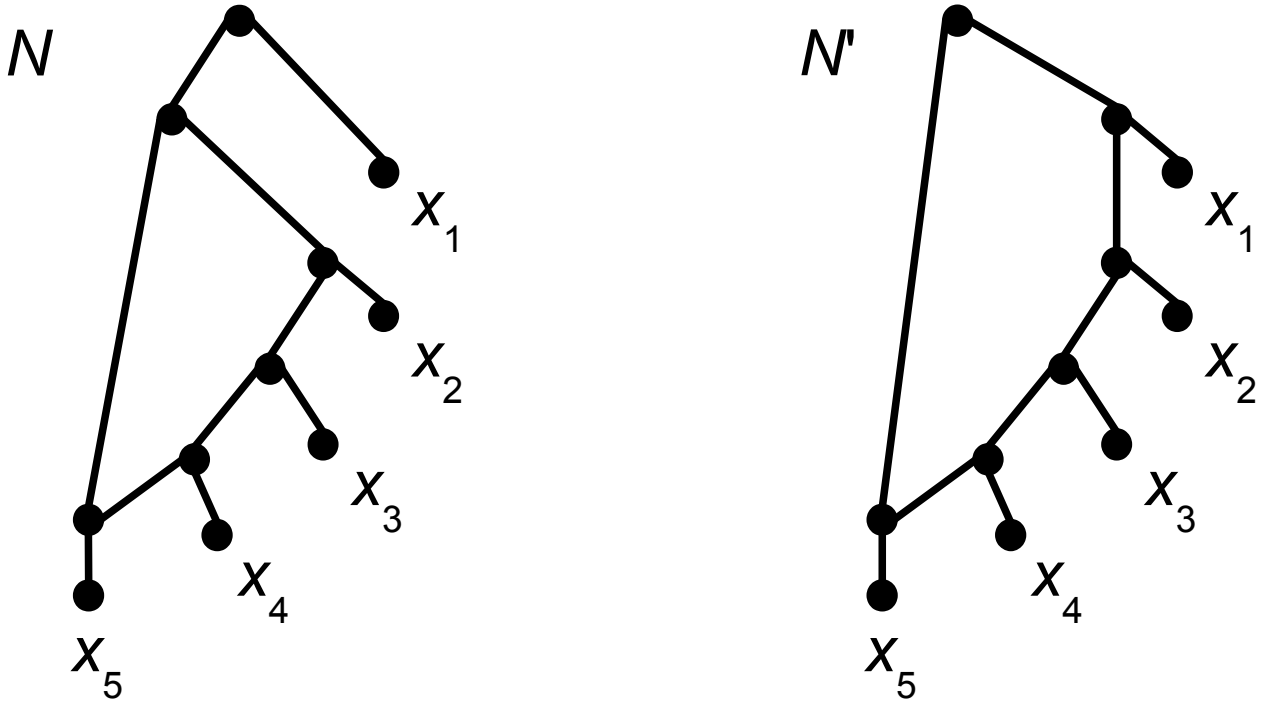
# Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Méthodes combinatoires de reconstruction
- Utilisation pratique
- Illustrations
- Perspectives

# Ambiguïté des solutions

- **Ambiguïté** de la reconstruction

Tous les triplets de *N* sont dans *N*'
Tous les clades souples de *N* sont dans *N*'



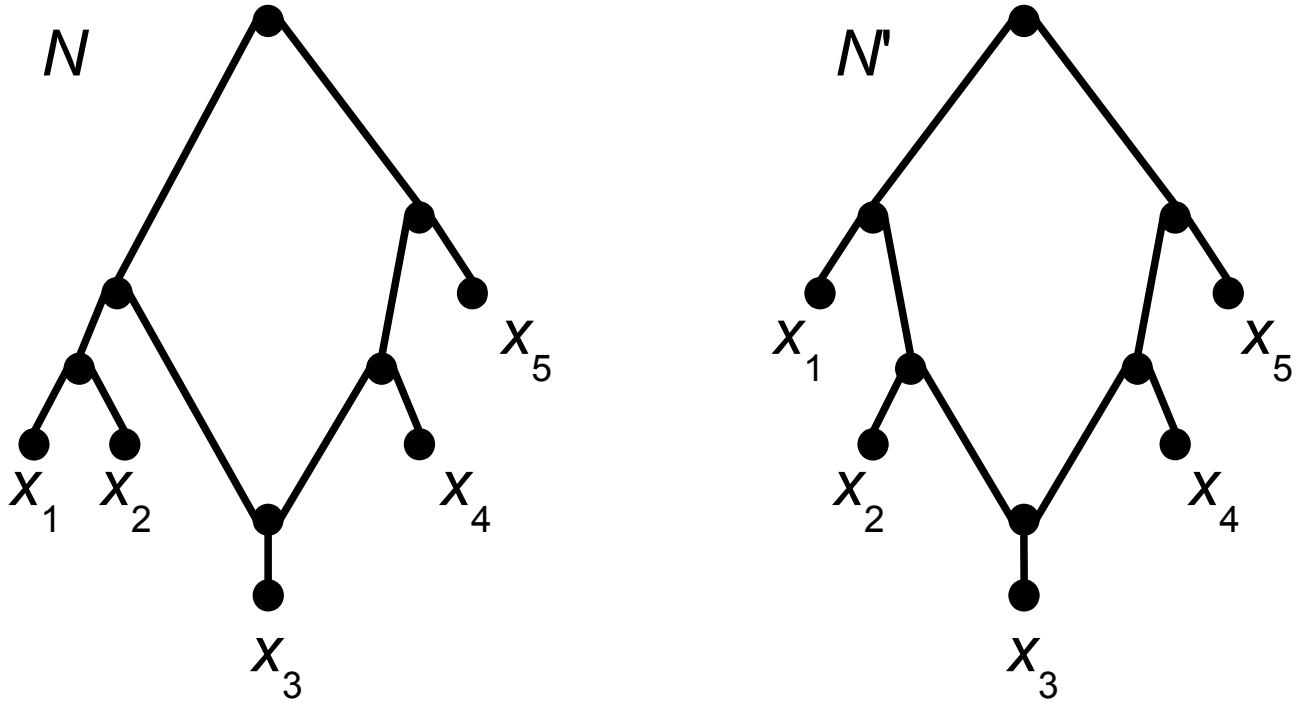

Si on a en entrée les triplets de *N*,
on ne sait pas si le vrai réseau est *N* ou *N*'

Gambette, Huber & Kelk, 2012

# Ambiguïté des solutions

- **Ambiguïté** de la reconstruction

Tous les triplets de *N* sont dans *N*'
Tous les clades souples de *N* sont dans *N*'



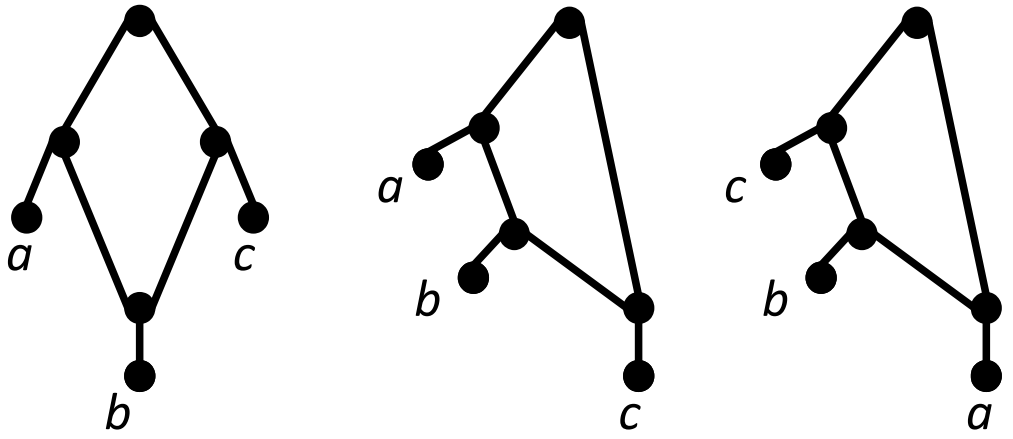

Si on a en entrée les triplets de *N*,
on ne sait pas si le vrai réseau est *N* ou *N*'

Gambette, Huber & Kelk, 2012

# Ambiguïté des solutions

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

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




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

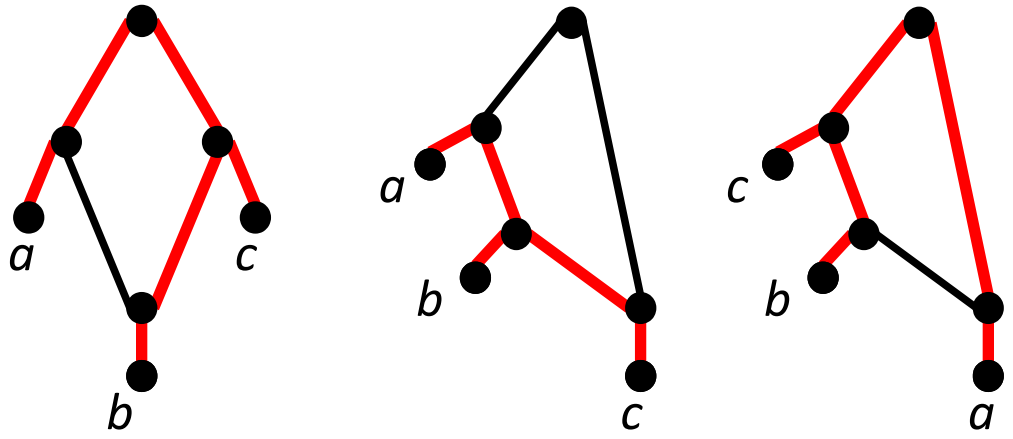
Gambette & Huber, *JMB*, 2012

# Ambiguïté des solutions

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

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

a|bc




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

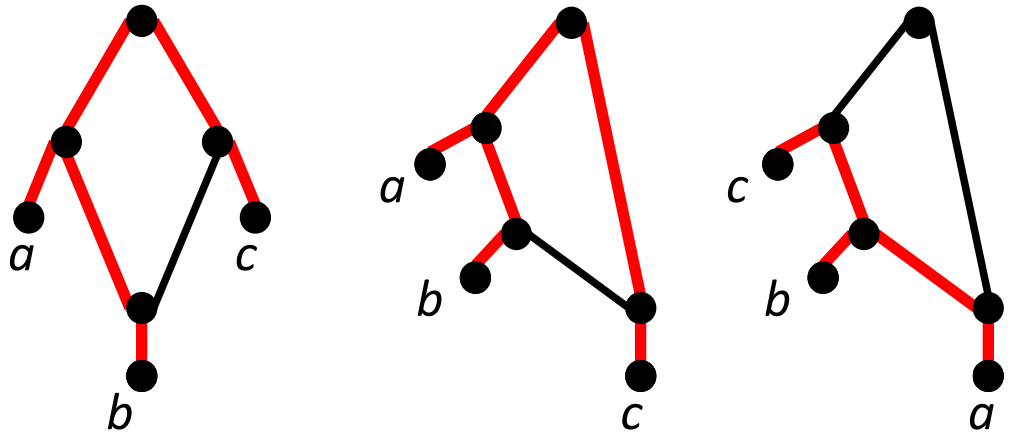
Gambette & Huber, *JMB*, 2012

# Ambiguïté des solutions

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

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

c|ab



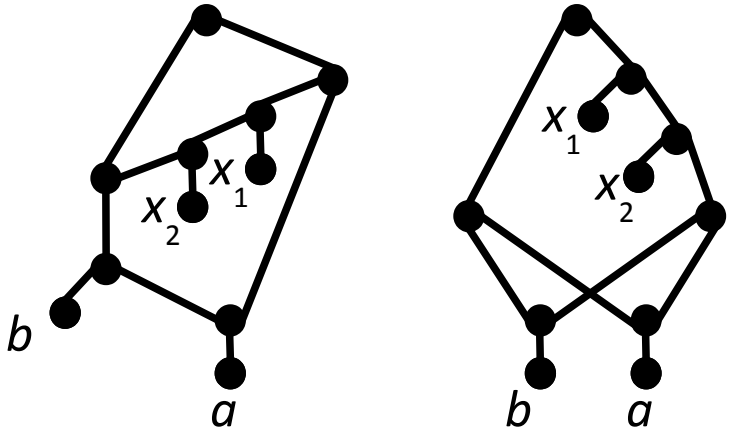

Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 sommets)

Gambette & Huber, *JMB*, 2012

# Ambiguïté des solutions

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

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



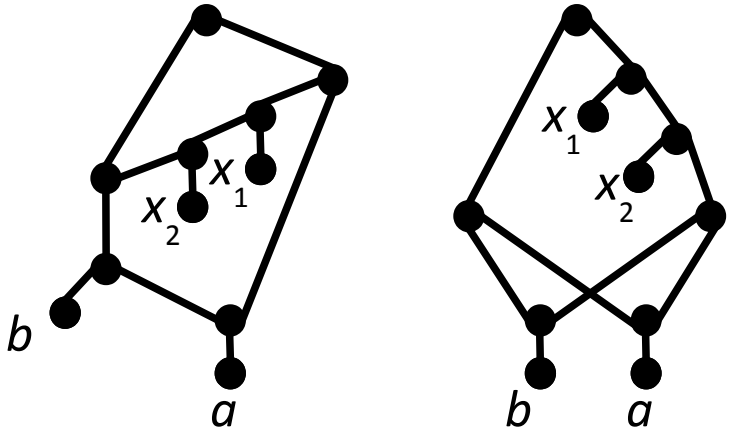

2 réseaux de niveau 2 avec le même ensemble de triplets

Gambette & Huber, *JMB*, 2012

# Ambiguïté des solutions

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

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




2 réseaux de niveau 2 avec le même ensemble de triplets

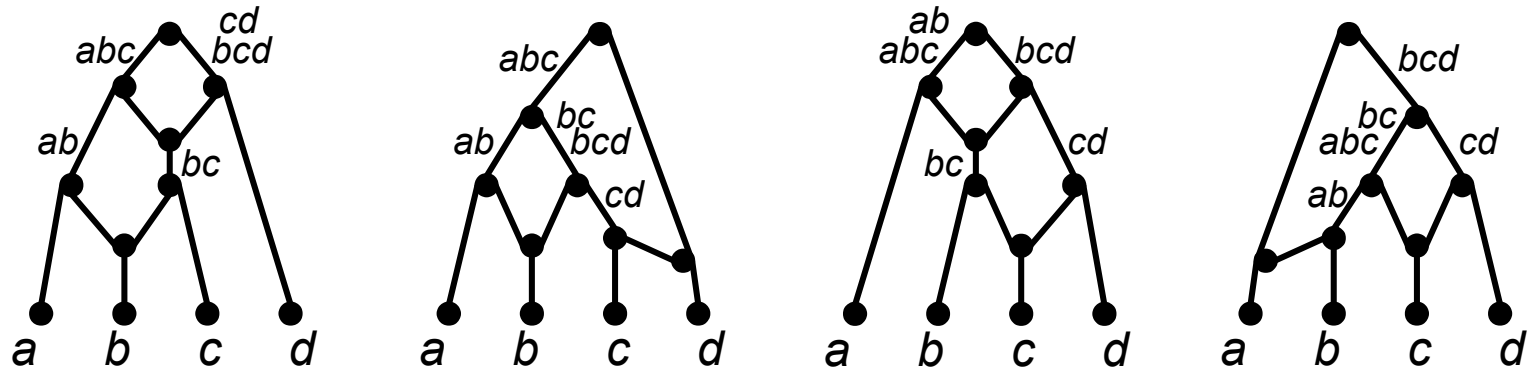
Même avec des données de départ **complètes** et **correctes**, impossible de choisir entre les formes ambiguës !

Gambette & Huber, *JMB*, 2012

# Ambiguïté des solutions

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

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




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

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

Gambette & Huber, *JMB*, 2012

# Utilisation pratique

existant
*en cours d'étude*

| Conditions d'utilisation | Données disponibles | Traitements possibles |
|---|---|---|
| arbres enracinés | arbres non enracinés | *enracinement avec un arbre des espèces de référence, ou contraintes topologiques* |
| arbres sans taxon répété | arbres avec gènes issus de duplications | traitement des MUL-trees<br>Scornavacca, Berry & Ranwez, 2009 |
| clades et triplets corrects | données bruitées | nettoyage des arbres<br>PhySIC_IST, 2008<br>filtre des données (clades avec bonne valeur de bootstrap, présents dans >x% des arbres)<br>*édition des données : solution compatible avec le plus de données en entrée* |
| données complètes (ensembles denses de triplets, clades complets) | données partielles, ou gènes ayant été supprimés | *sélection d'un grand nombre d'arbres couvrant un grand nombre d'espèces*<br>*sélection du nombre maximal de taxons avec densité de triplets*<br>problèmes NP-complets |

# Plan

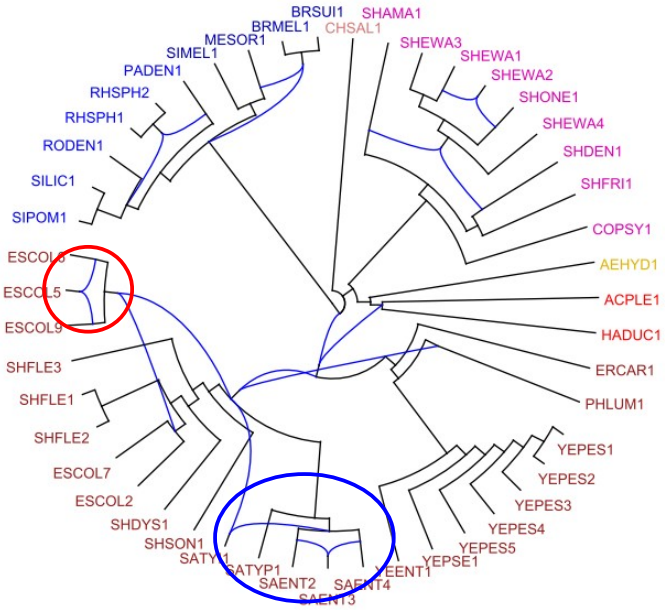
- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Méthodes combinatoires de reconstruction
- Utilisation pratique
- Illustrations
- Perspectives

# Exemples de résultats

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

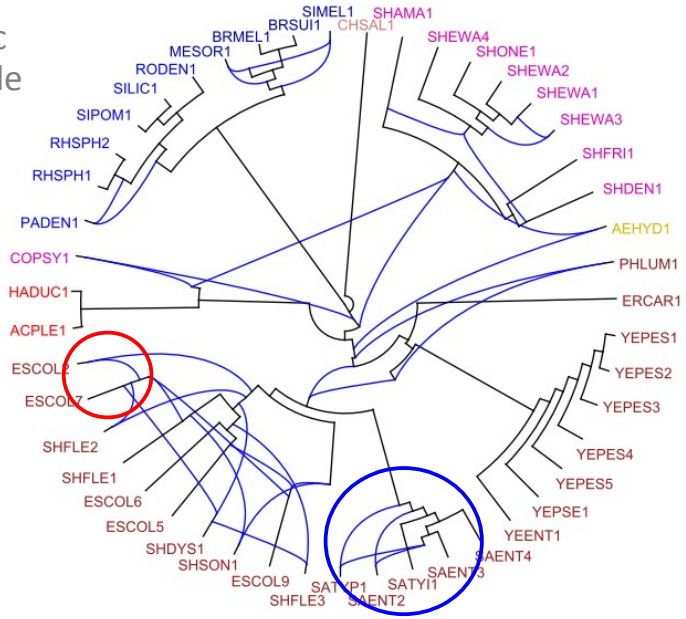
24 Enterobacteriales
2 Pasteurellales
1 Aeromonadales
9 Alteromonadales
1 Oceanospirillales
6 Rhodobacterales
4 Rhizobiales

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

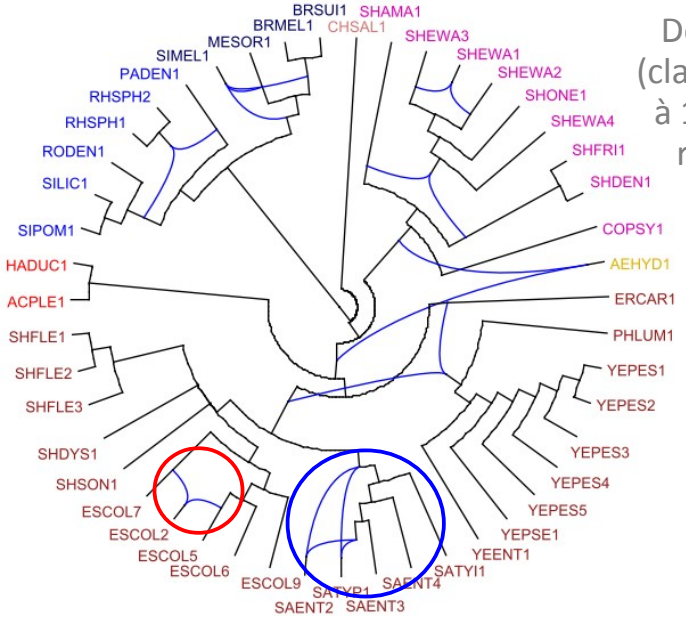



Lev1athan (heuristique triplets, niveau 1) 24 sec.

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




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

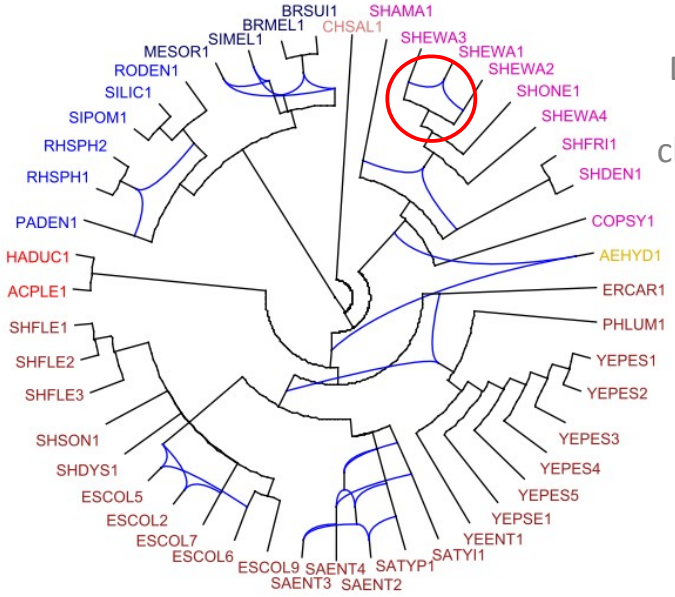


# Exemples de résultats

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

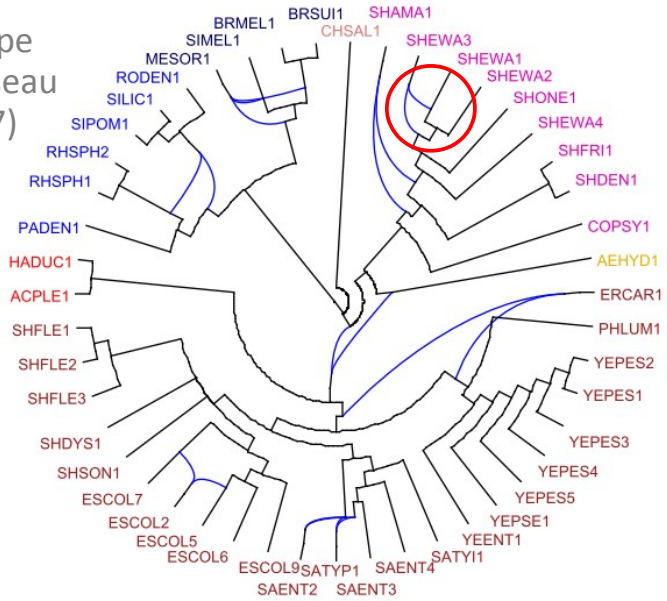
24 Enterobacteriales
2 Pasteurellales
1 Aeromonadales
9 Alteromonadales
1 Oceanospirillales
6 Rhodobacterales
4 Rhizobiales

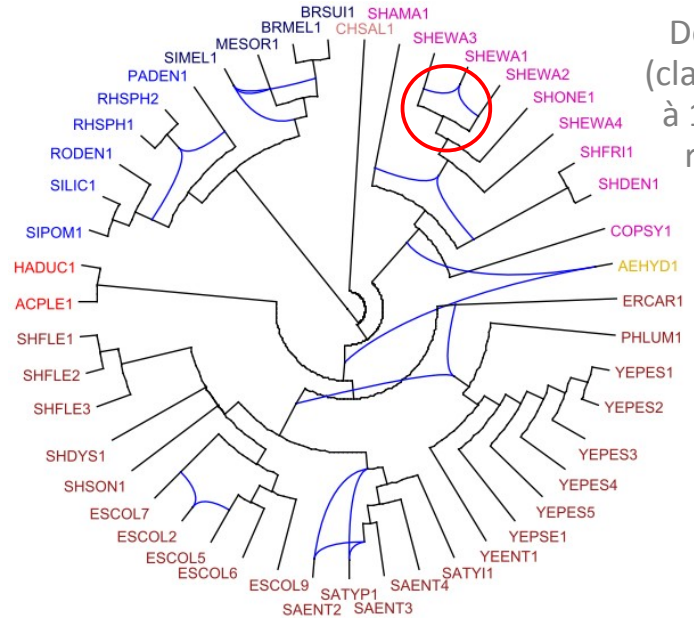
Réseaux contenant les clades souples présents dans au moins 20% des arbres




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

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



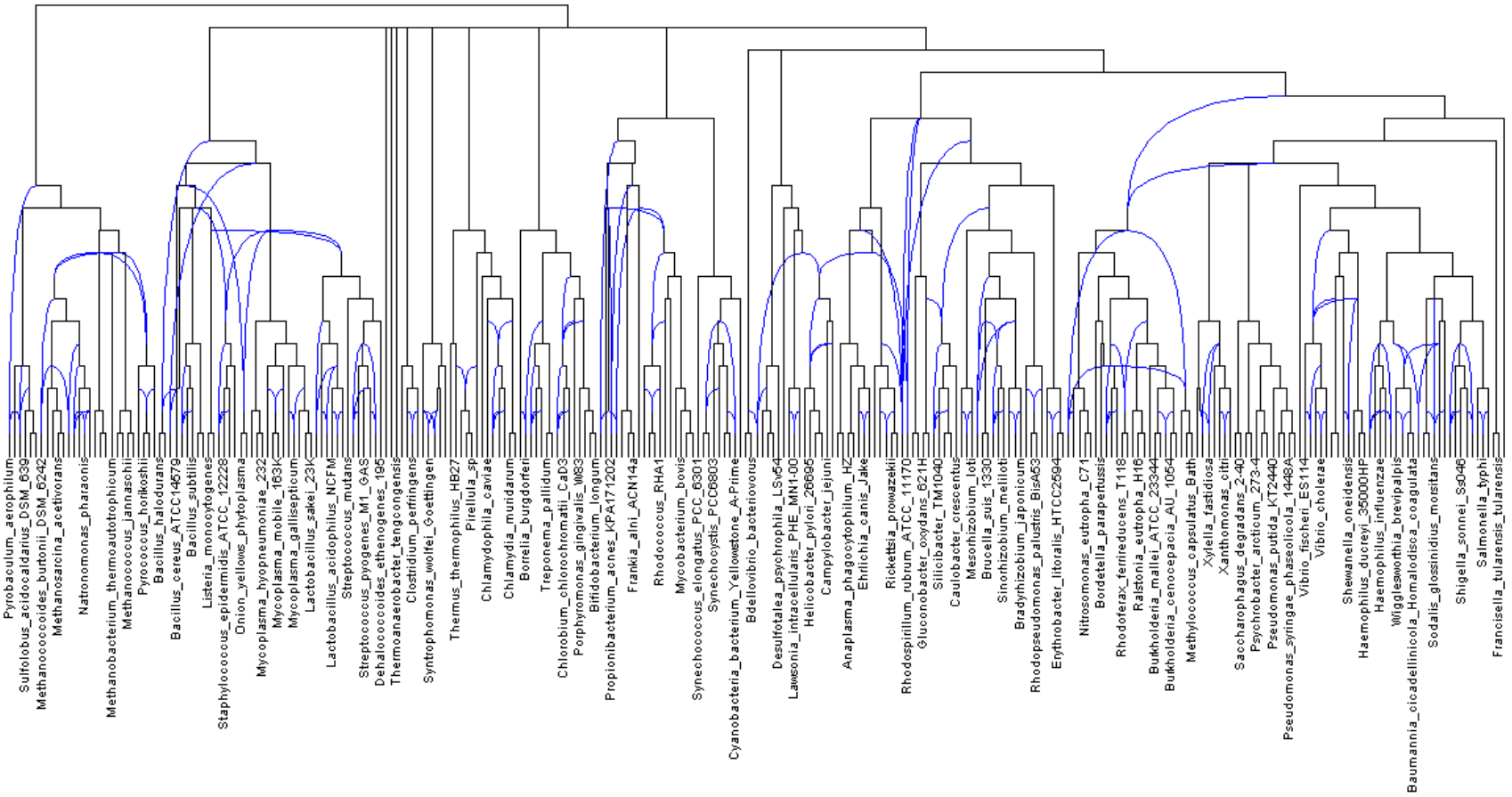




Dendroscope (clades, réseau à 1 couche de réticulation)
<1 sec.

# Exemples de résultats

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

Auch, Steigele, Huson & Henz, 2009




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

# Illustrations

**23 trees, 45 species** from the 3 domains of life clusters with 80% bootstrap confidence present in at least 2 trees

Dendroscope (galled network) <1 sec.


Data from Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: Universal trees based on large combined protein sequence data sets. Nat Genet 2001, 28:281--285

# Illustrations

**23 trees, 45 species** from the 3 domains of life clusters with 80% bootstrap confidence present in at least 2 trees

Dendroscope (level-3 network) <1 sec.


Data from Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: Universal trees based on large combined protein sequence data sets. Nat Genet 2001, 28:281--285

# Plan

- Les réseaux phylogénétiques
- Motivations de l'approche combinatoire
- Méthodes combinatoires de reconstruction
- Application pratique
- Illustrations
- Perspectives

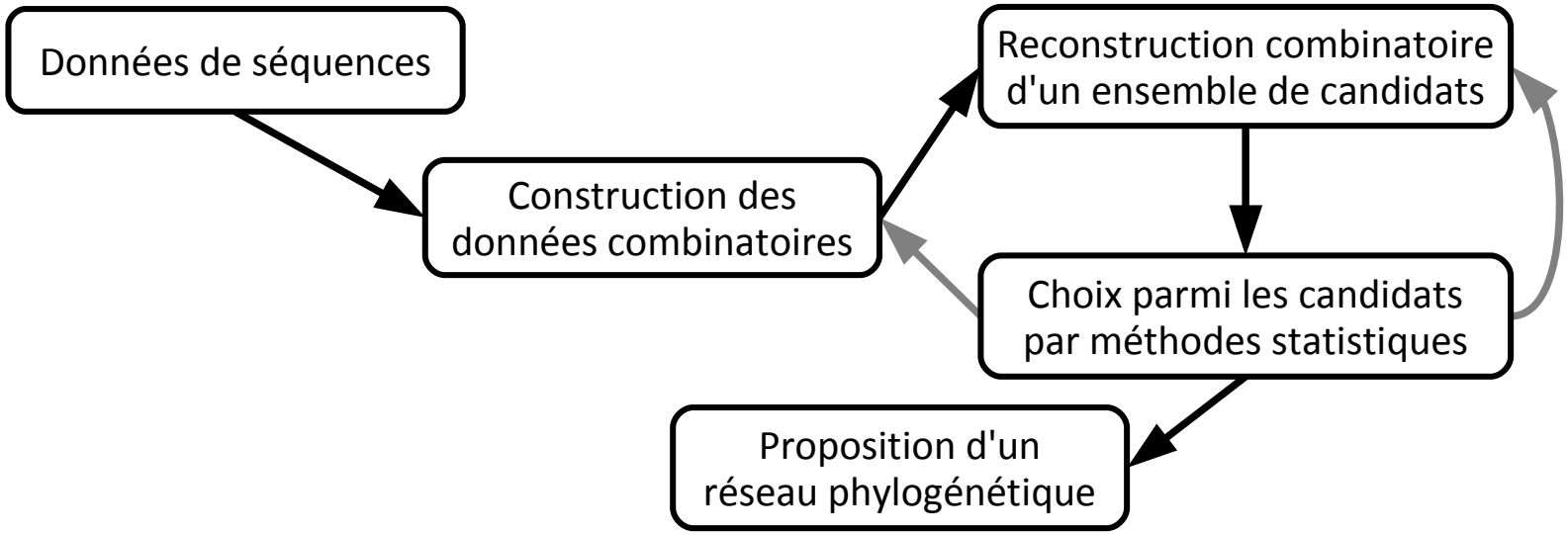
# Perspectives de recherche

**Combinatoire :**

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

**Bioinformatique :**

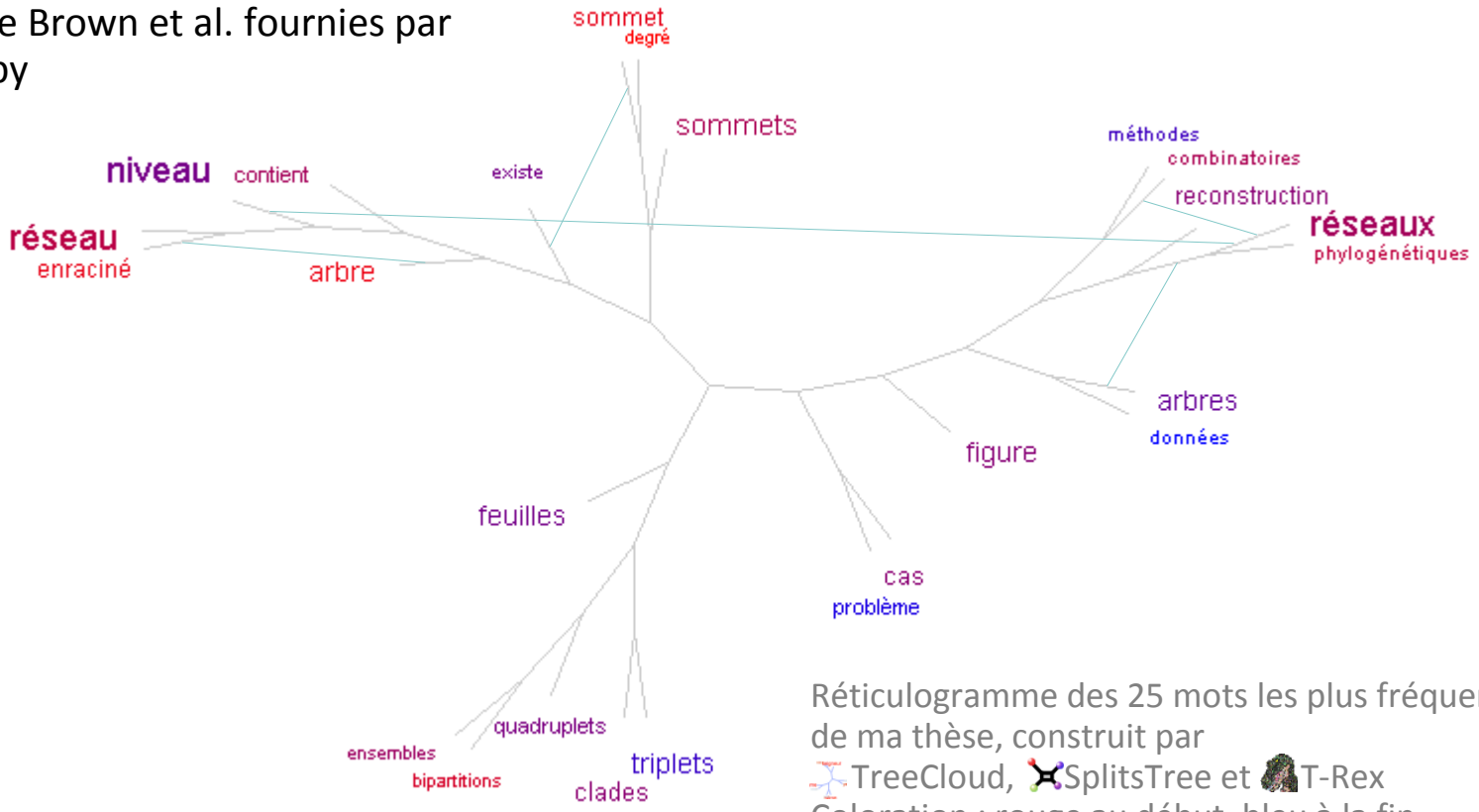
- Fonction des gènes transférés (“autoroutes de transfert”)
- Intégration des méthodes combinatoires dans une approche statistique



# Merci !

Coauteurs des travaux présentés :
- Vincent Berry, Christophe Paul (LIRMM)
- Mathilde Bouvel (Bordeaux)
- Katharina Huber (East Anglia)
- Daniel Huson, Regula Rupp (Tübingen)

Données de Brown et al. fournies par Sophie Abby




Réticulogramme des 25 mots les plus fréquents de ma thèse, construit par TreeCloud, SplitsTree et T-Rex
Coloration : rouge au début, bleu à la fin

http://www.treecloud.org