07/04/2011

Séminaire MaBioVis - LABRI

Réseaux phylogénétiques pour la visualisation de données biologiques et textuelles

Philippe Gambette

Plan

- Les réseaux phylogénétiques
- Méthodes de reconstruction
- Limites des méthodes combinatoires
- Illustration sur des données biologiques
- Utilisation sur des données textuelles
- Perspectives

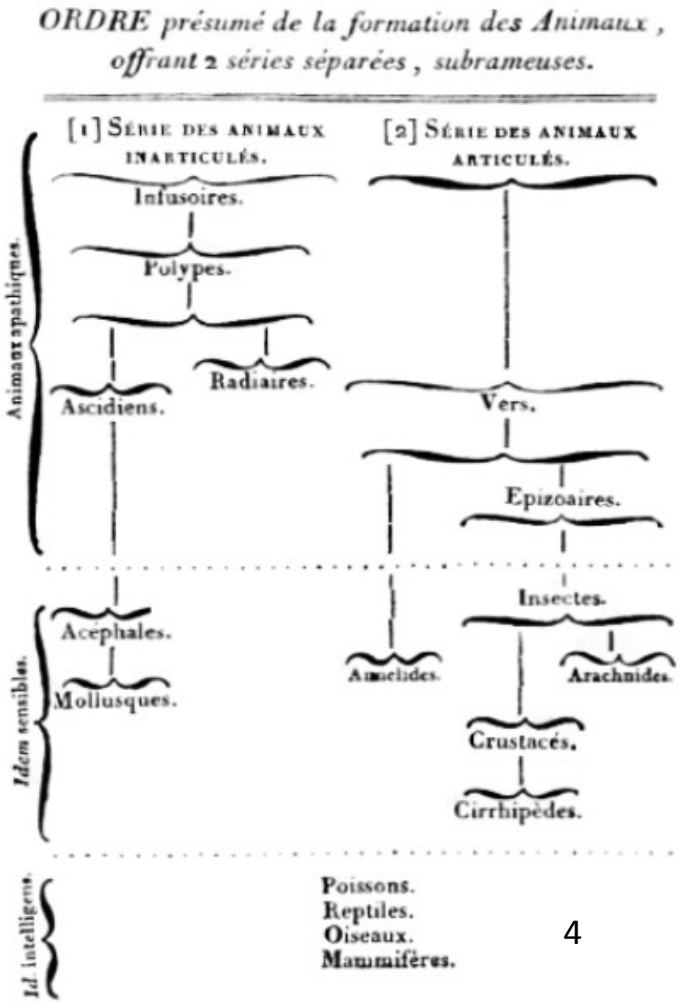
Plan

- Les réseaux phylogénétiques
- Méthodes de reconstruction
- Limites des méthodes combinatoires
- Illustration sur des données biologiques
- Utilisation sur des données textuelles
- Perspectives

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

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



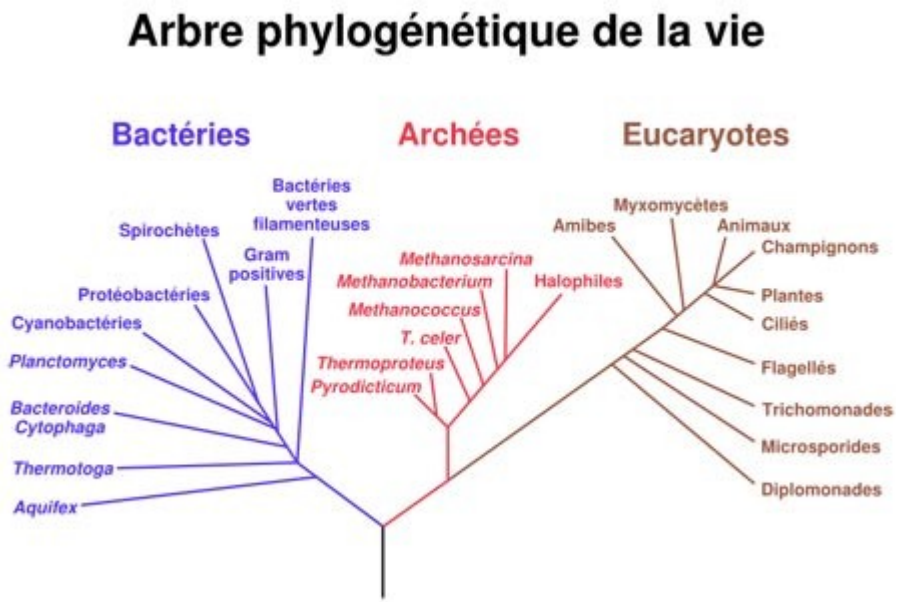
D'après Lamarck : Histoire naturelle des animaux sans vertèbres (1815)

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

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

modélisation



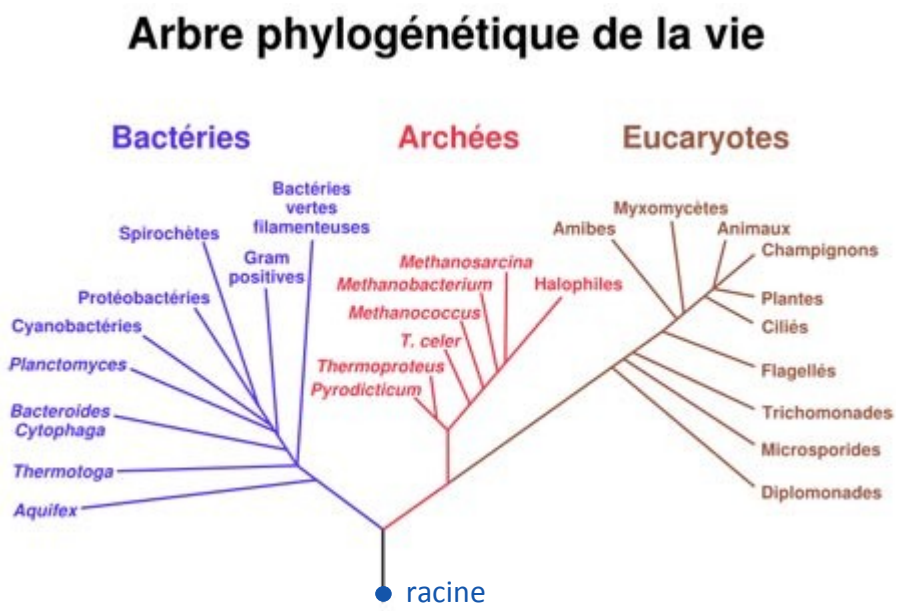
D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

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

modélisation



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

Les arbres phylogénétiques

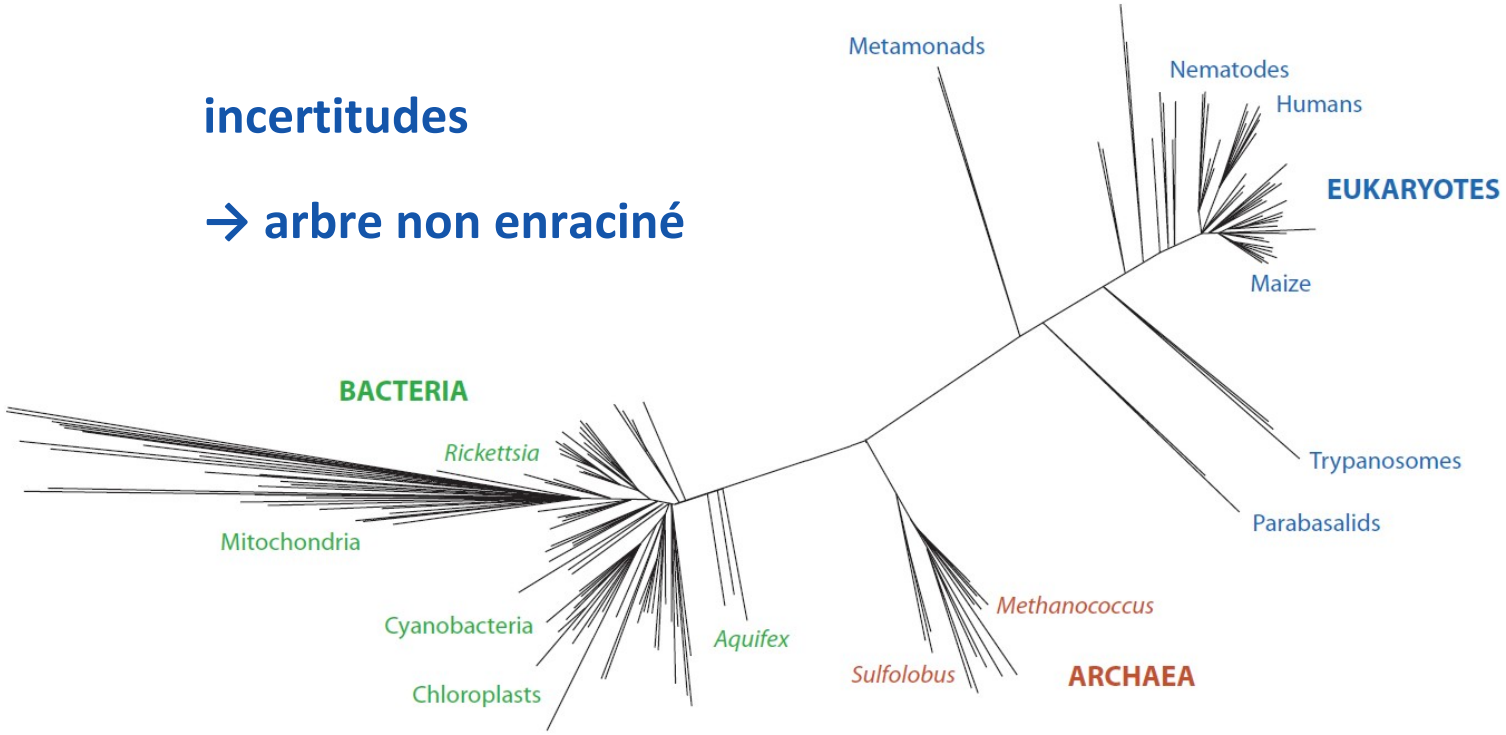
Arbre phylogénétique d'un ensemble d'espèces :

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

modélisation

incertitudes

→ arbre non enraciné

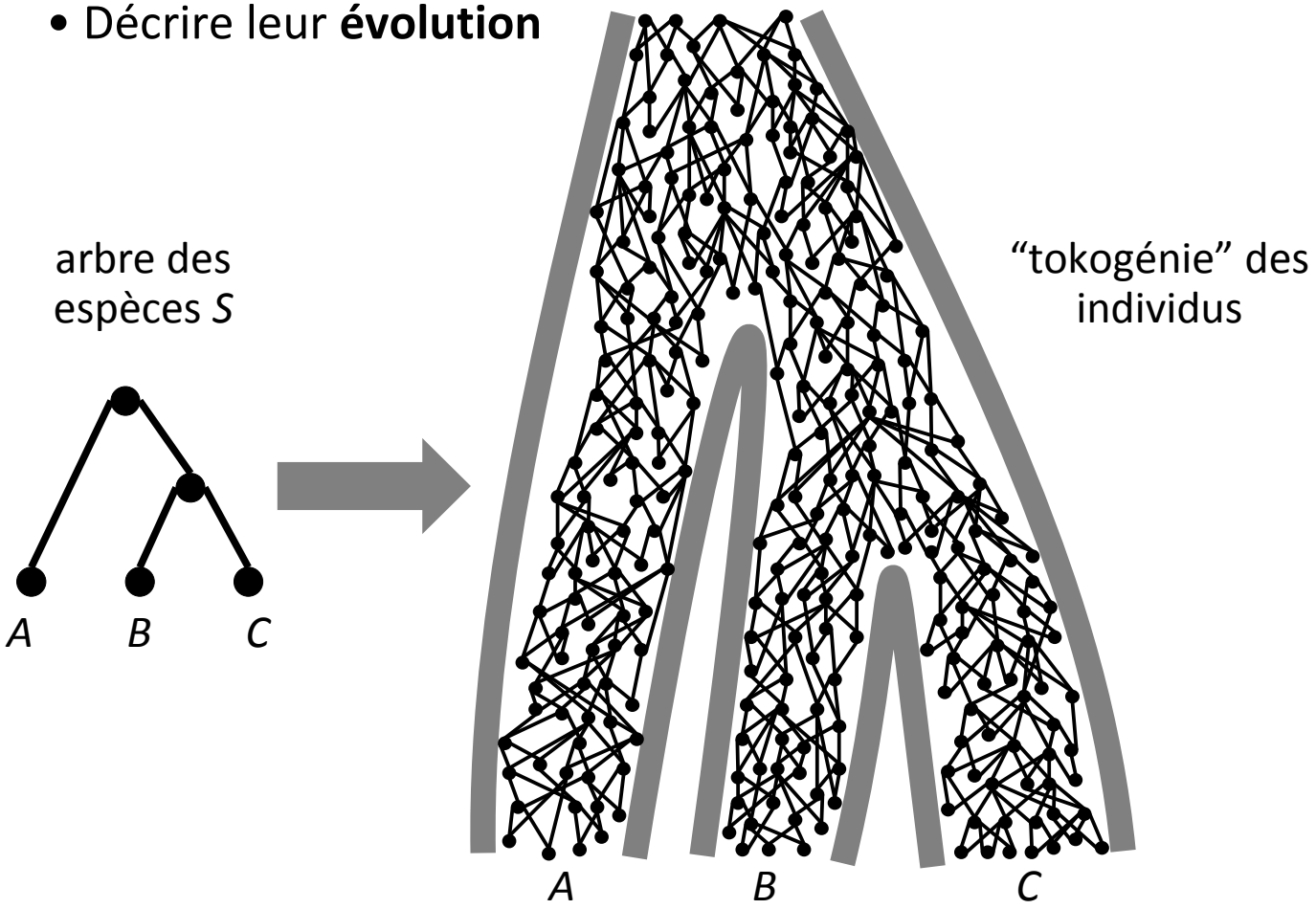


D'après Christophe Blumrich, David S. Spencer,
cité dans Doolittle : Uprooting the Tree of Life, Scientific American (Fév. 2000)

Les arbres phylogénétiques

Arbre phylogénétique d'un ensemble d'espèces :

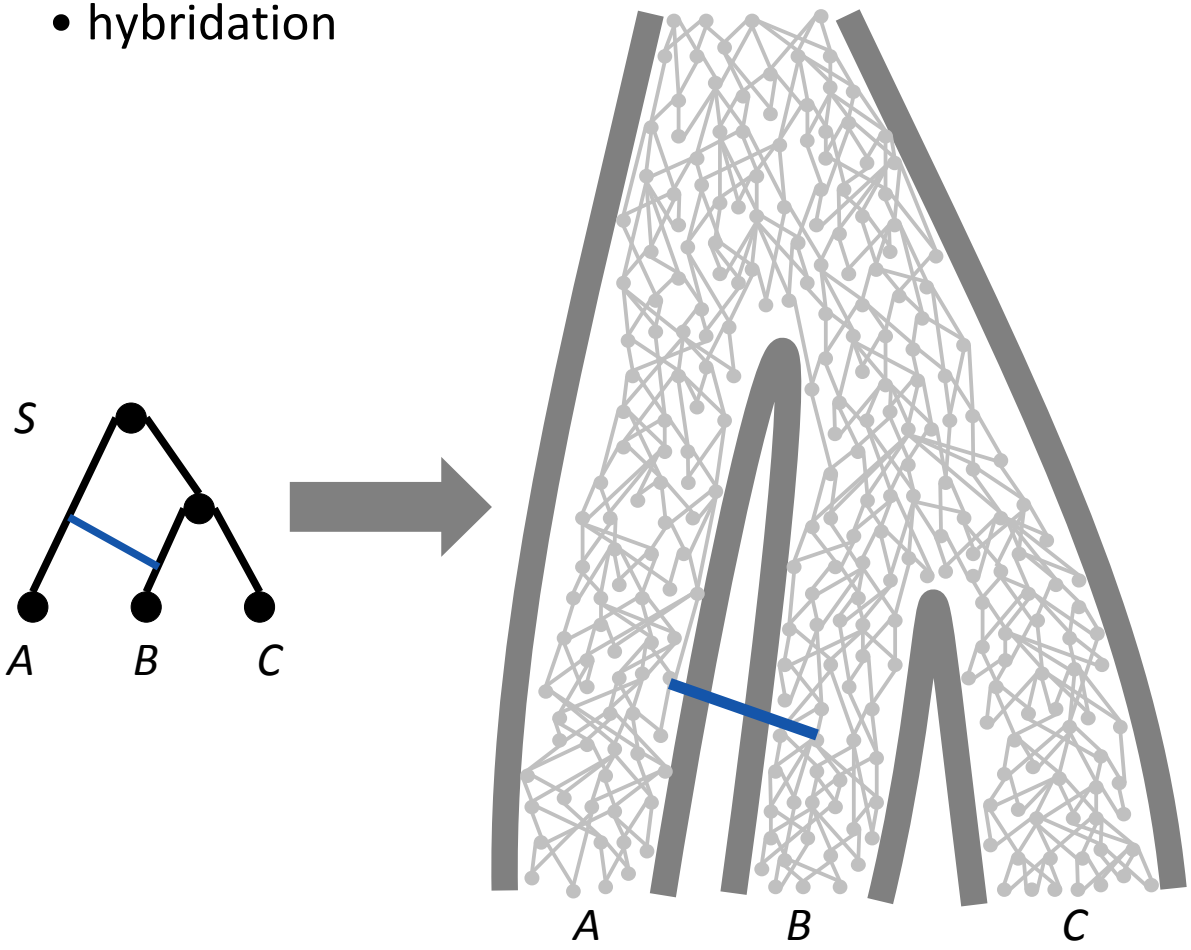
- Les organiser en fonction de caractères communs
- Décrire leur **évolution**



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

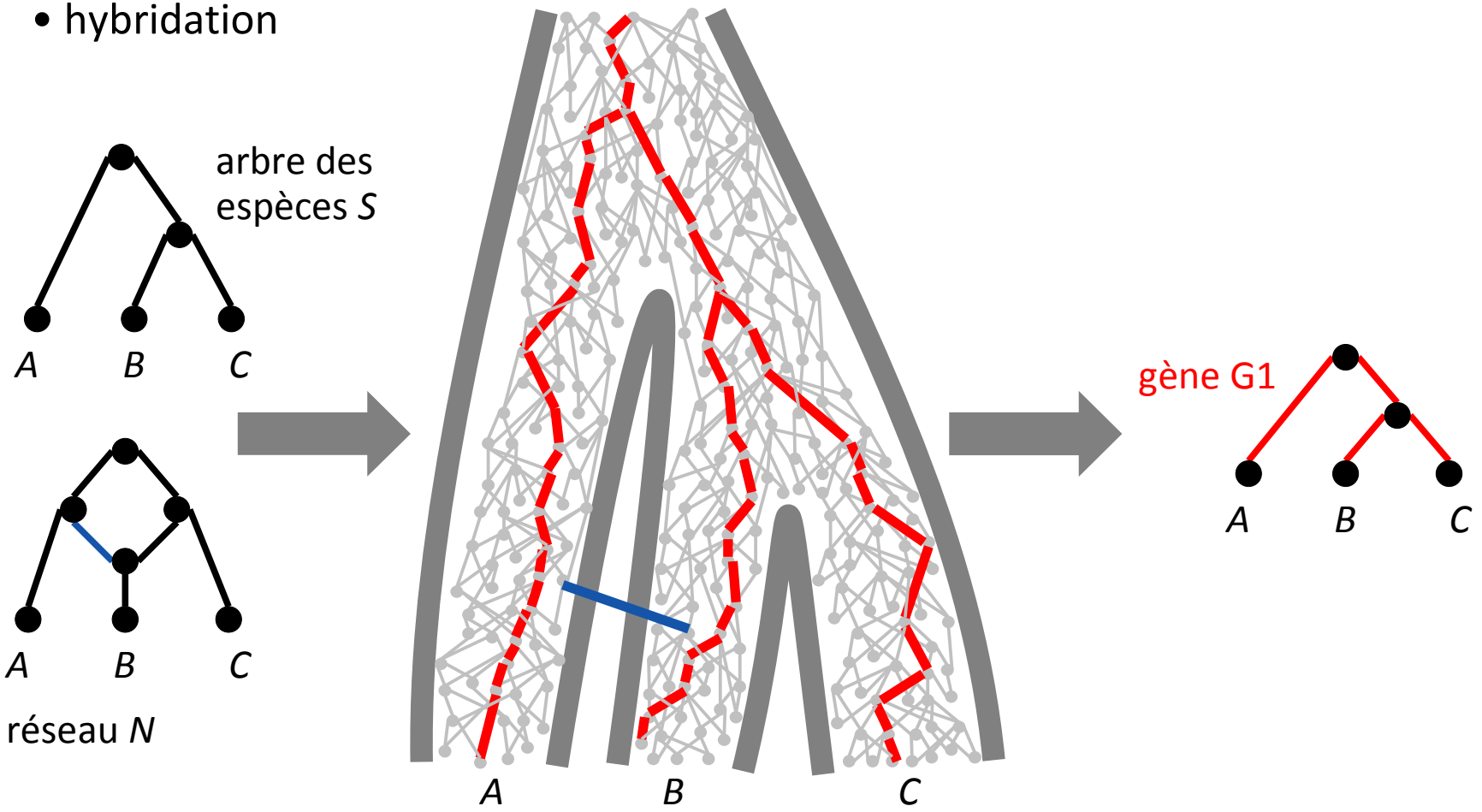
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

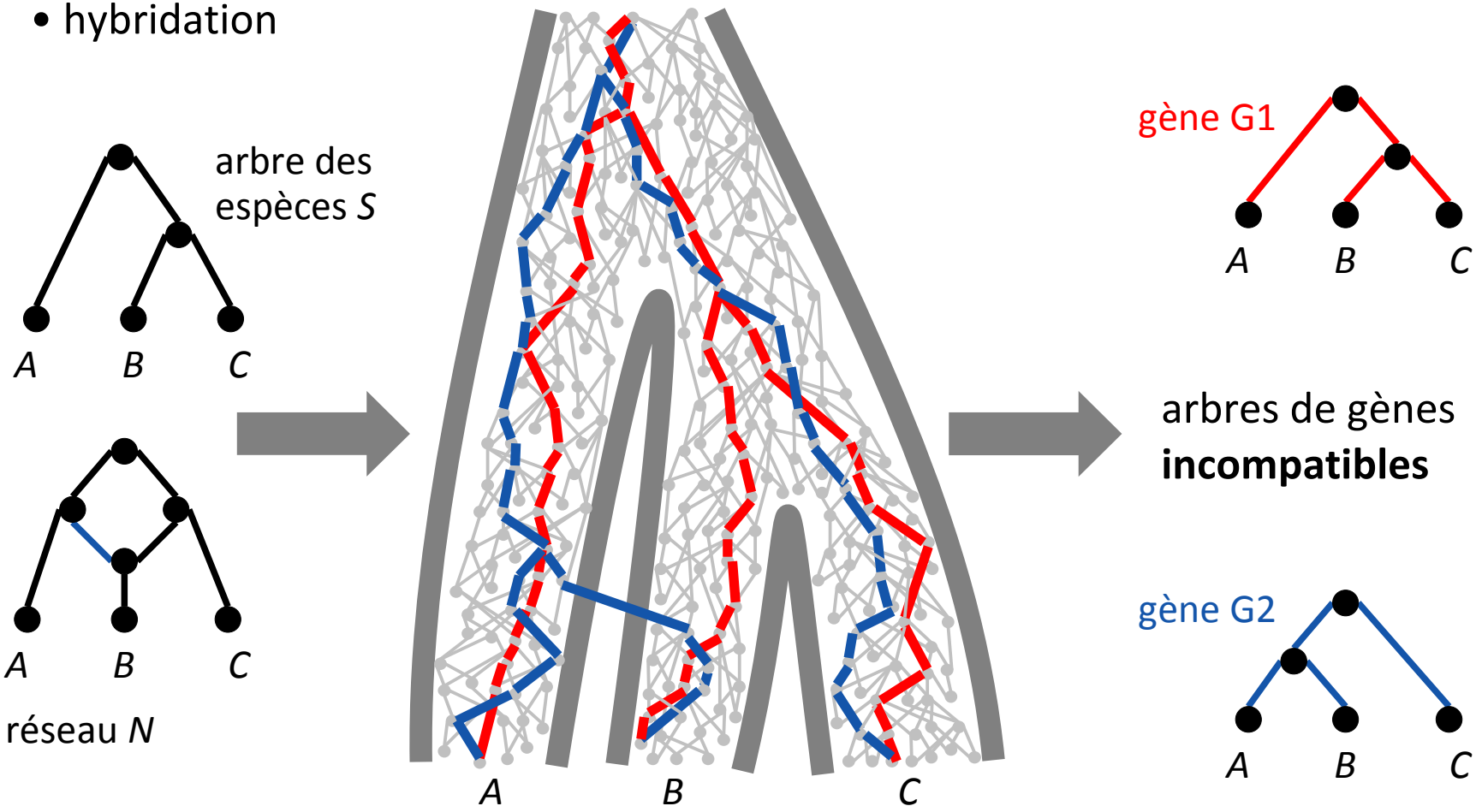
- transfert horizontal
- hybridation



Transferts de matériel génétique

Transferts de matériel génétique entre espèces coexistantes :

- transfert horizontal
- hybridation



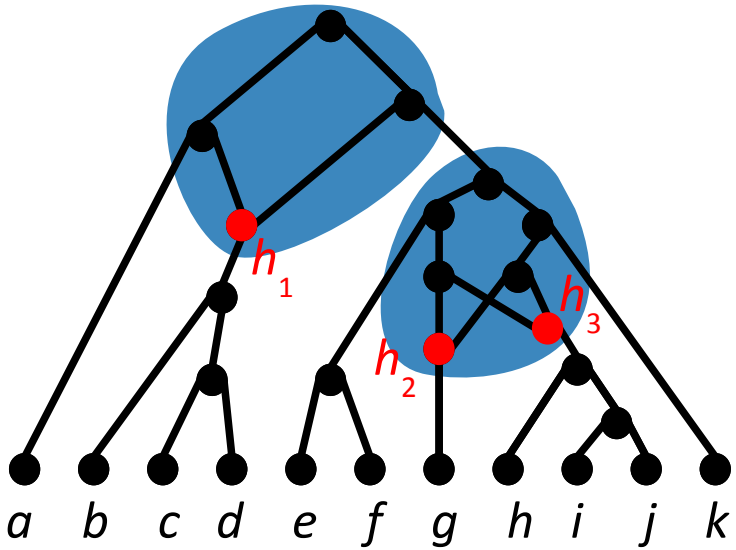
Réseaux phylogénétiques explicites

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **explicites**

modélisation de l'évolution

Réseau phylogénétique explicite enraciné :



Sommets à plus d'un parent :
réticulations

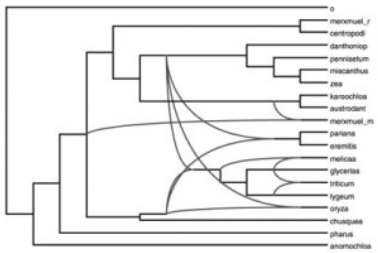
Partie **non arborée** : ***blob***.

Réseaux phylogénétiques explicites

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **explicites**

modélisation de l'évolution



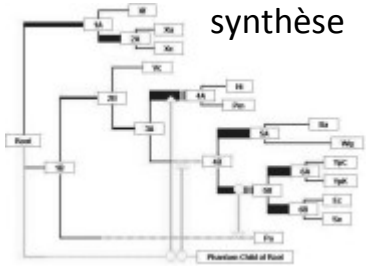
réseau à une couche de réticulation

Dendroscope

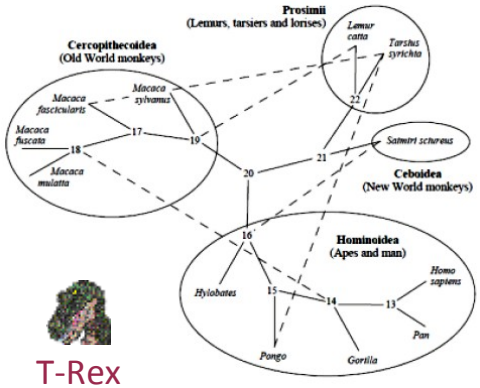
Simplistic

réseau de niveau 2

diagramme de synthèse



HorizStory



T-Rex

réticulogramme

Réseaux phylogénétiques abstraits

Réseau phylogénétique : réseau représentant des données d'évolution

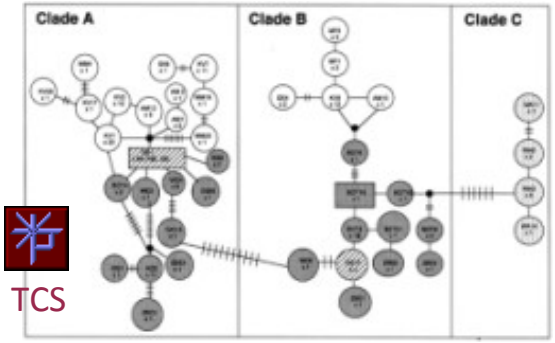
- réseaux phylogénétiques **abstraits**

classification, **visualisation** **de données**

réseau de bipartitions

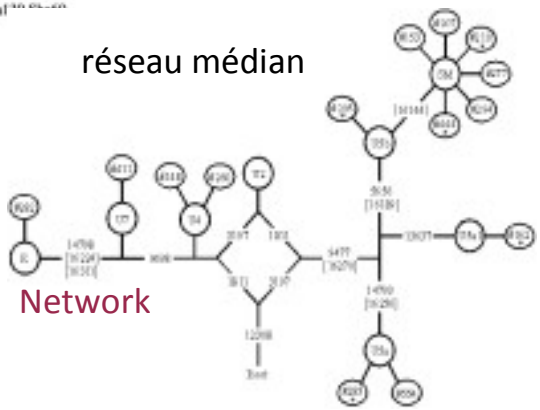
SplitsTree

réseau couvrant minimum



TCS

réseau médian



Network

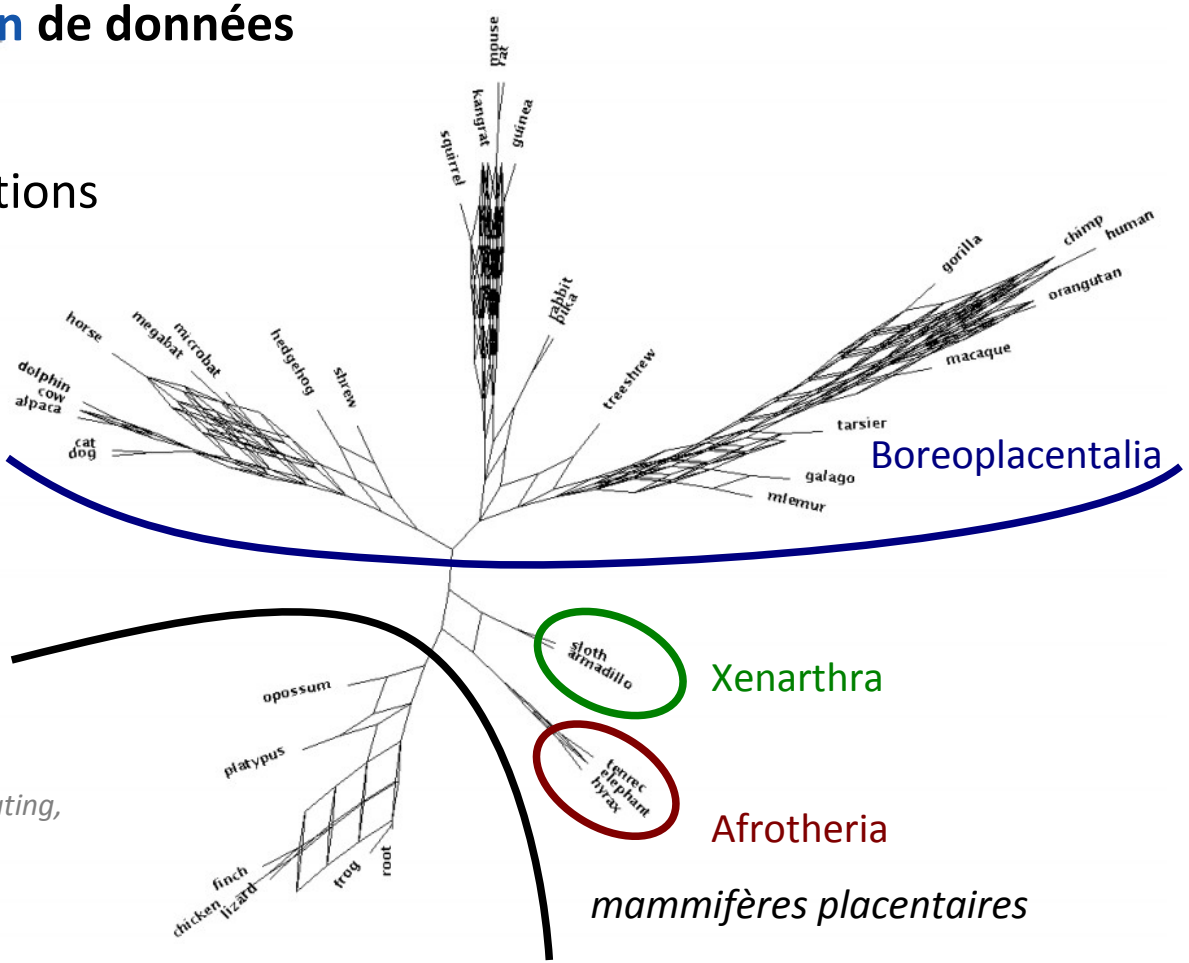
Réseaux phylogénétiques abstraits

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **abstraits**

classification, visualisation de données

Réseau abstrait de bipartitions
(“*consensus network*”) :



D'après Björn M. Hallström, Axel Janke - Mammalian evolution may not be strictly bifurcating, MBE, 2010

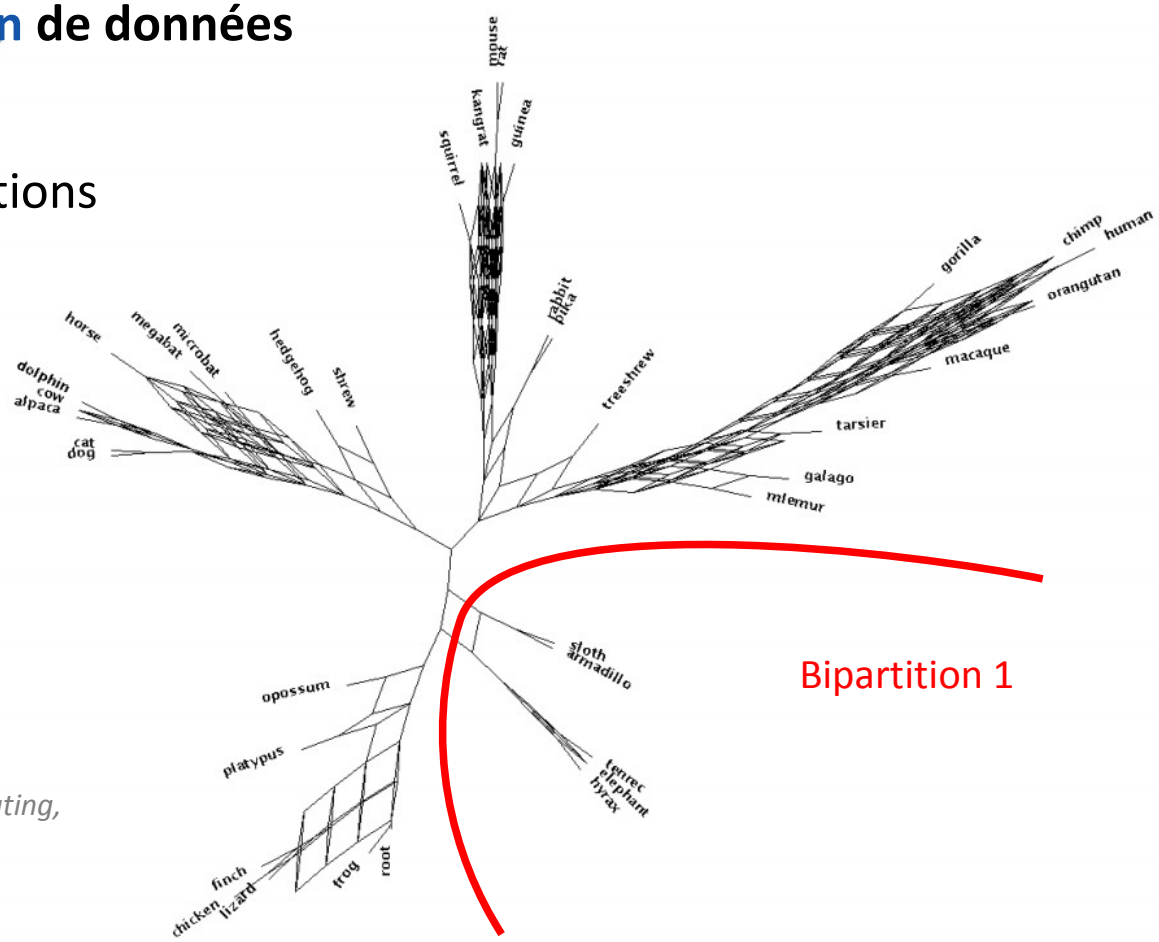
Réseaux phylogénétiques abstraits

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **abstraits**

classification, visualisation de données

Réseau abstrait de bipartitions
(“*consensus network*”) :



D'après Björn M. Hallström, Axel Janke - Mammalian evolution may not be strictly bifurcating, MBE, 2010

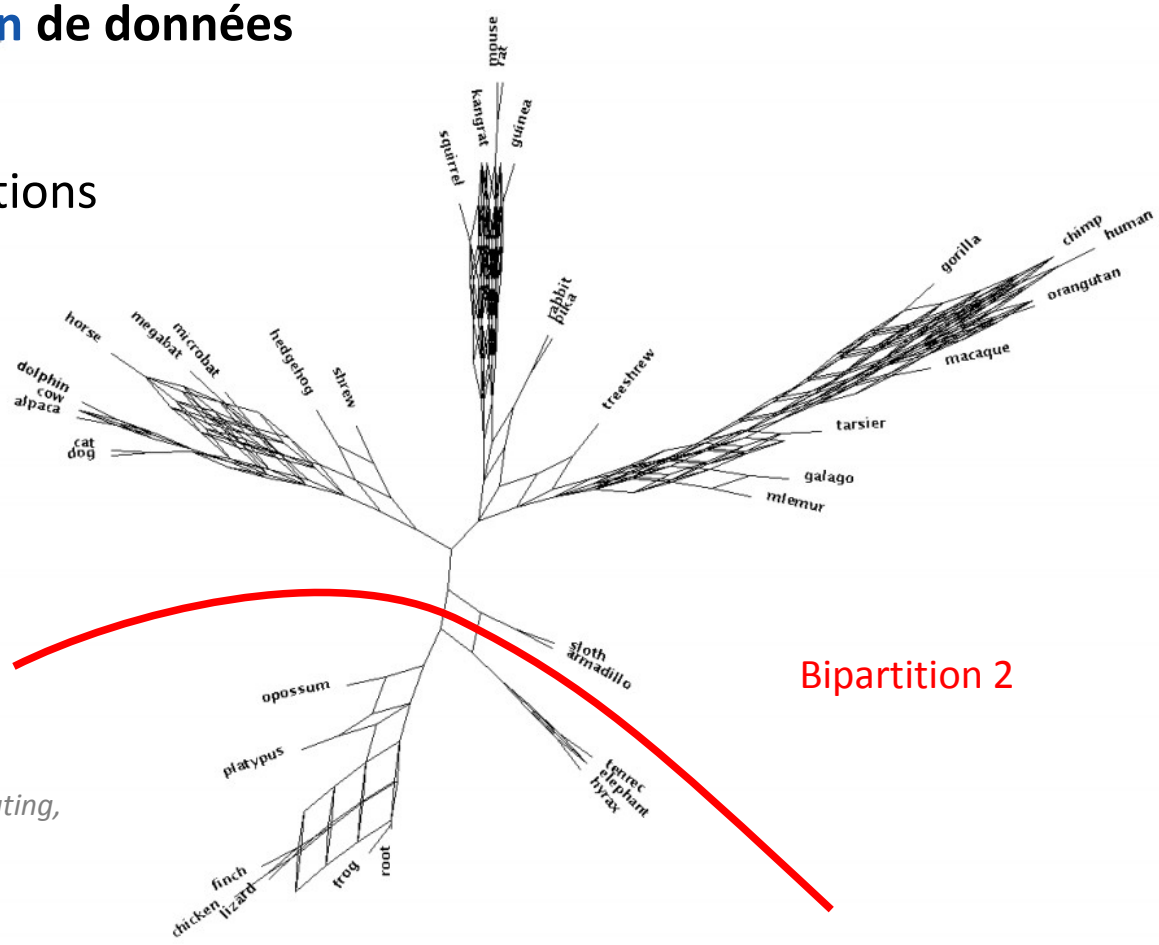
Réseaux phylogénétiques abstraits

Réseau phylogénétique : réseau représentant des données d'évolution

- réseaux phylogénétiques **abstraits**

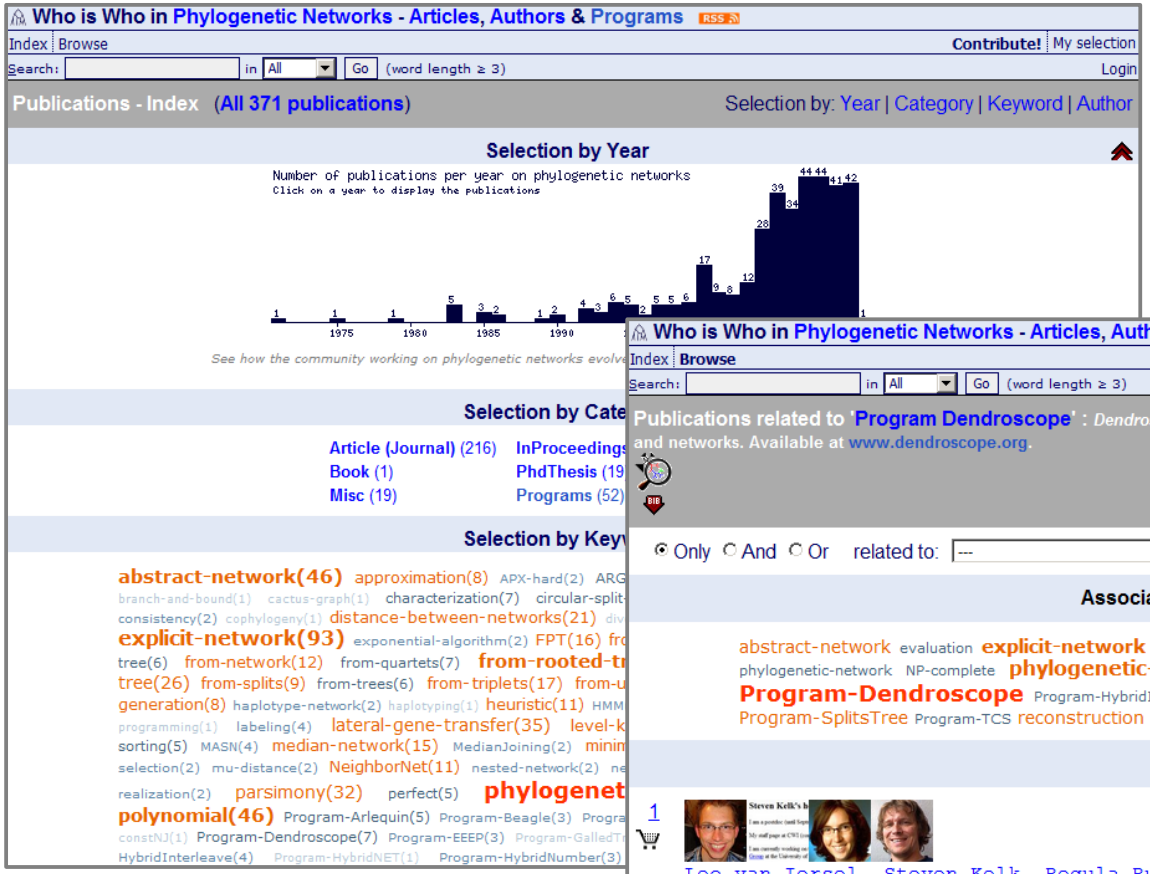
classification, visualisation de données

Réseau abstrait de bipartitions
(“*consensus network*”) :



D'après Björn M. Hallström, Axel Janke - Mammalian evolution may not be strictly bifurcating, MBE, 2010

Plate-forme bibliographique



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

Plan

- Les réseaux phylogénétiques
- Méthodes de reconstruction
- Limites des méthodes combinatoires
- Illustration sur des données biologiques
- Utilisation sur des données textuelles
- Perspectives

Reconstruction de réseaux phylogénétiques

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

G1 G2

{séquences de gènes}

méthodes de distance

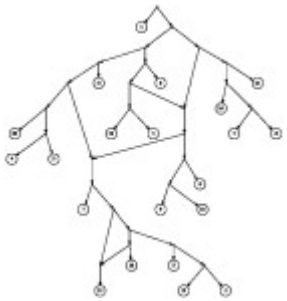
Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002

méthodes de parcimonie

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

méthodes de vraisemblance

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



réseau *N*

Reconstruction de réseaux phylogénétiques

```
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

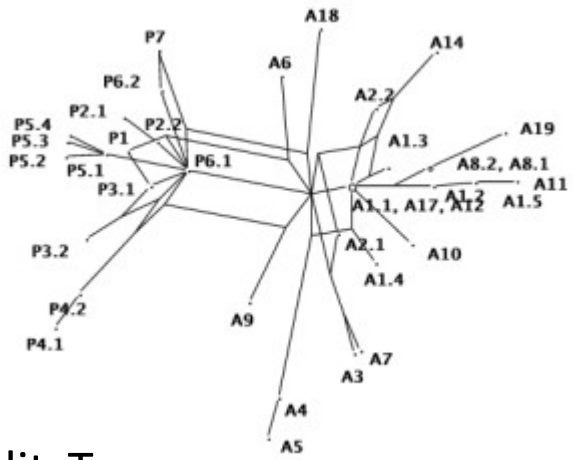
{séquences de gènes}

méthodes de distance

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

réseau *N*

"Split decomposition"



SplitsTree

D'après David Bryant & Daniel H. Huson, Application of Phylogenetic Networks in Evolutionary Studies, MBE, 2006

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

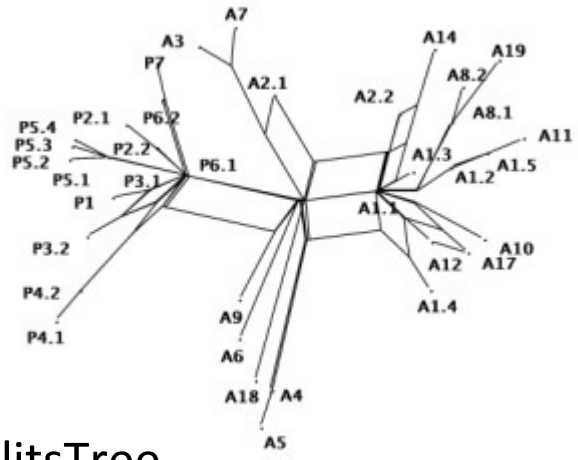
{séquences de gènes}

méthodes de distance

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

réseau *N*

"Neighbor Net"



SplitsTree

D'après David Bryant & Daniel H. Huson, Application of Phylogenetic Networks in Evolutionary Studies, MBE, 2006

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

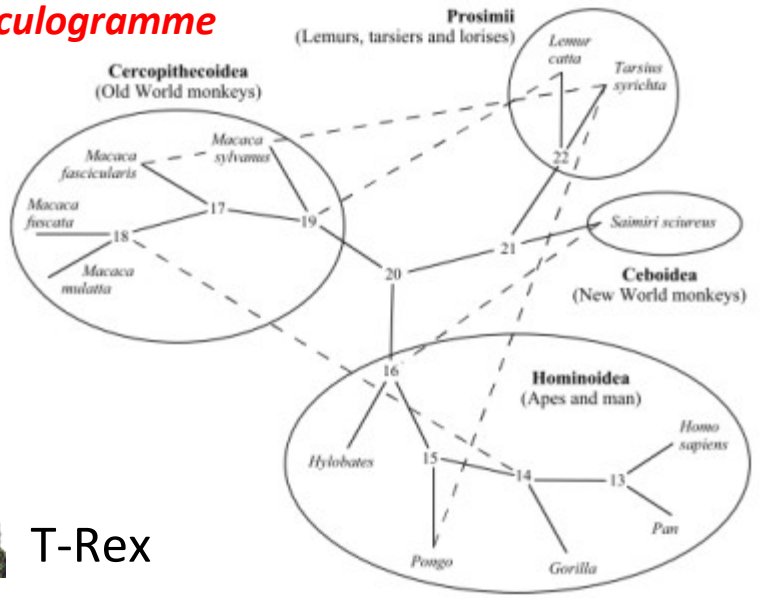
{séquences de gènes}

méthodes de distance

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

réseau *N*

Réticulogramme



T-Rex

D'après Vladimir Makarenkov & Pierre Legendre, Improving the additive tree representation of a dissimilarity matrix using reticulations, DACRM, 2000

Reconstruction de réseaux phylogénétiques

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

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

G1 G2

{séquences de gènes}

méthodes de distance

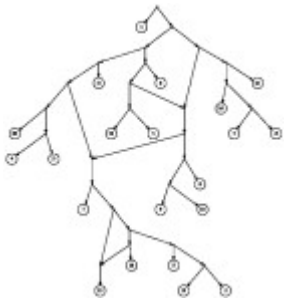
Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002

méthodes de parcimonie

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

méthodes de vraisemblance

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



réseau *N*

Reconstruction de réseaux phylogénétiques

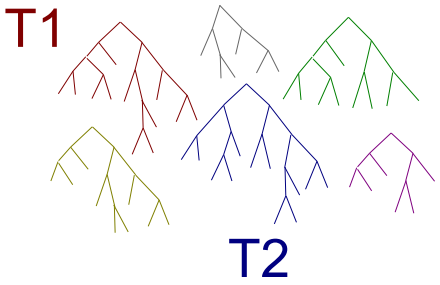
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

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

Guindon & Gascuel, SB, 2003



{arbres}

Base HOGENOM

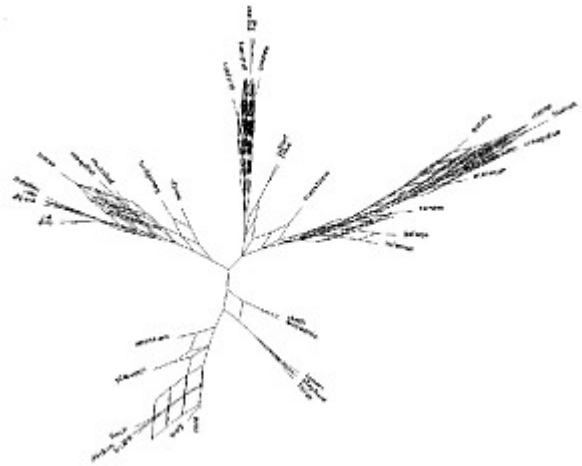
Phyl-ARIANE

ANR

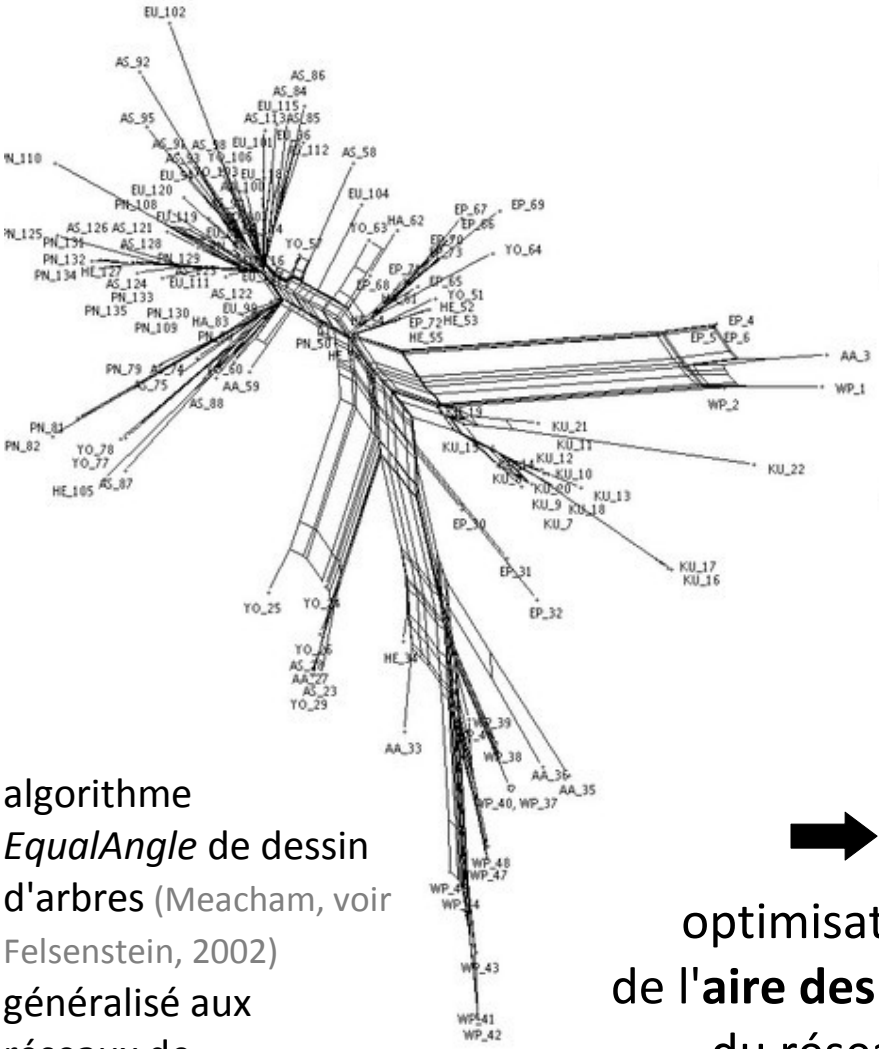
Dufayard, Duret, Penel, Gouy, Rechenmann & Perrière, BioInf, 2005

*Réconciliation ou **consensus** d'arbres*

super-réseau de consensus *N*

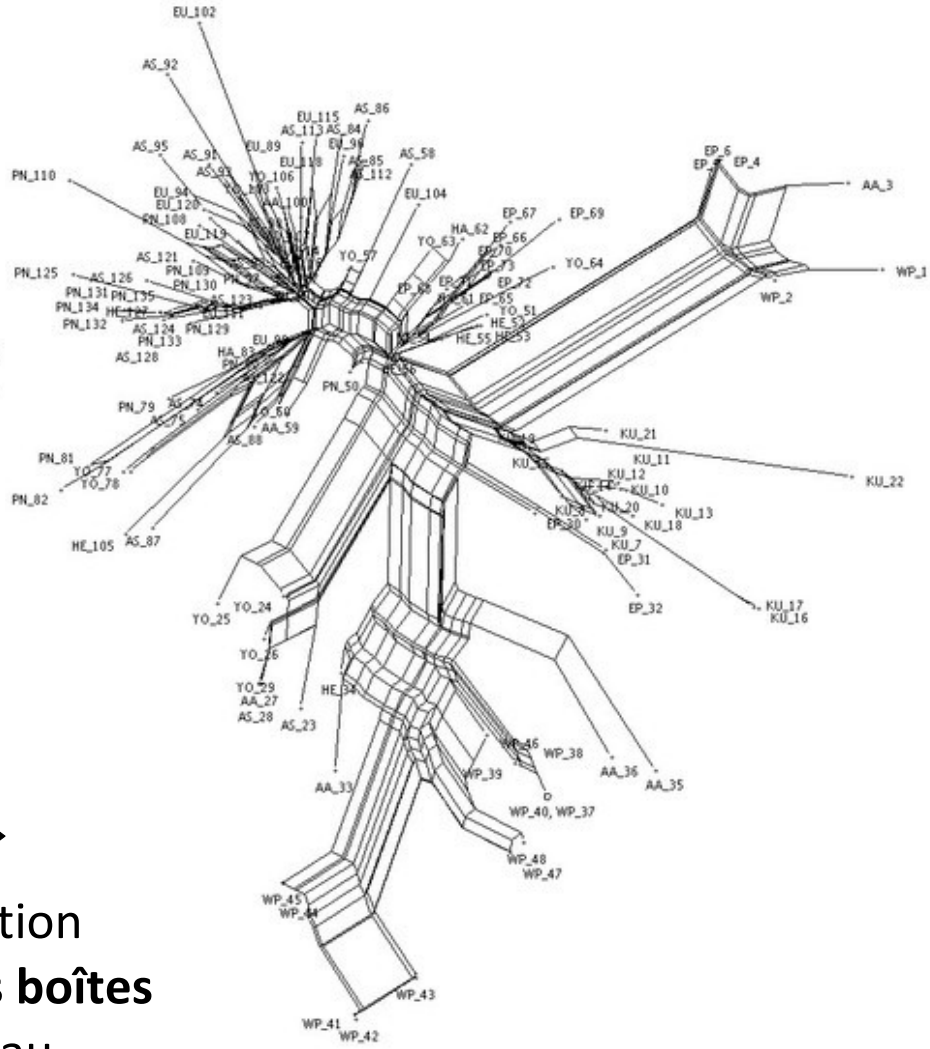


Visualisation des réseaux de bipartitions



algorithme *EqualAngle* de dessin d'arbres (Meacham, voir Felsenstein, 2002) généralisé aux réseaux de bipartitions.

optimisation de l'**aire des boîtes** du réseau

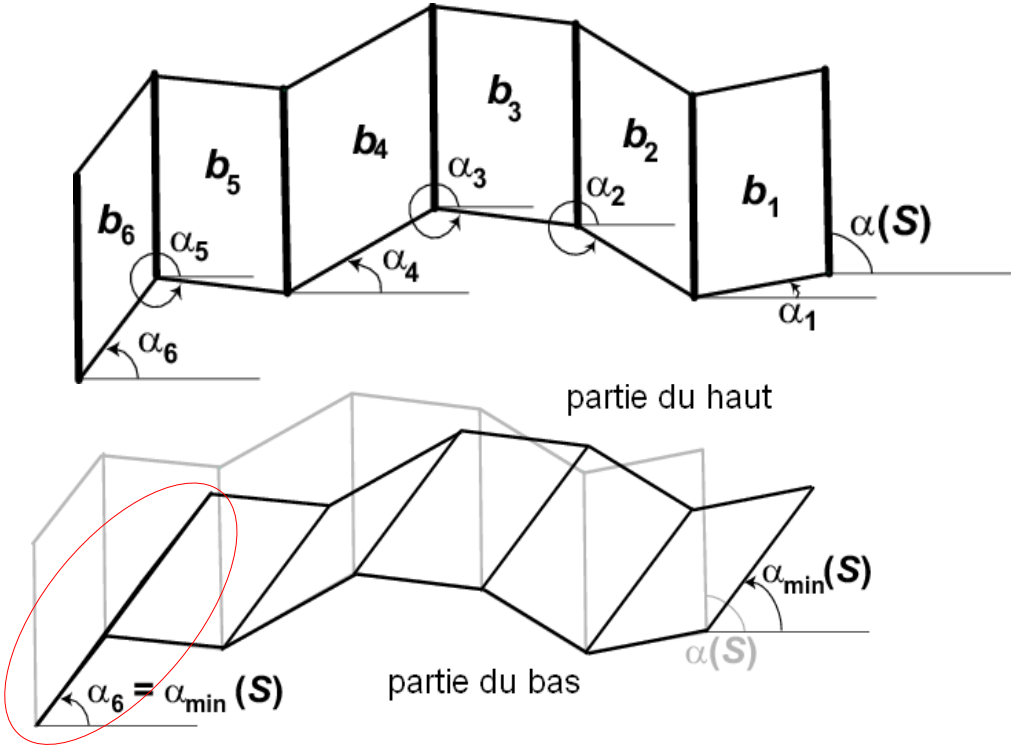


Gambette & Huson, TCBB, 2008

Algorithme Box-opening

Collisions locales :

deux **angles critiques** $\alpha_{min}(S)$ et $\alpha_{max}(S)$ pour l'ange de la bipartition $\alpha(S)$



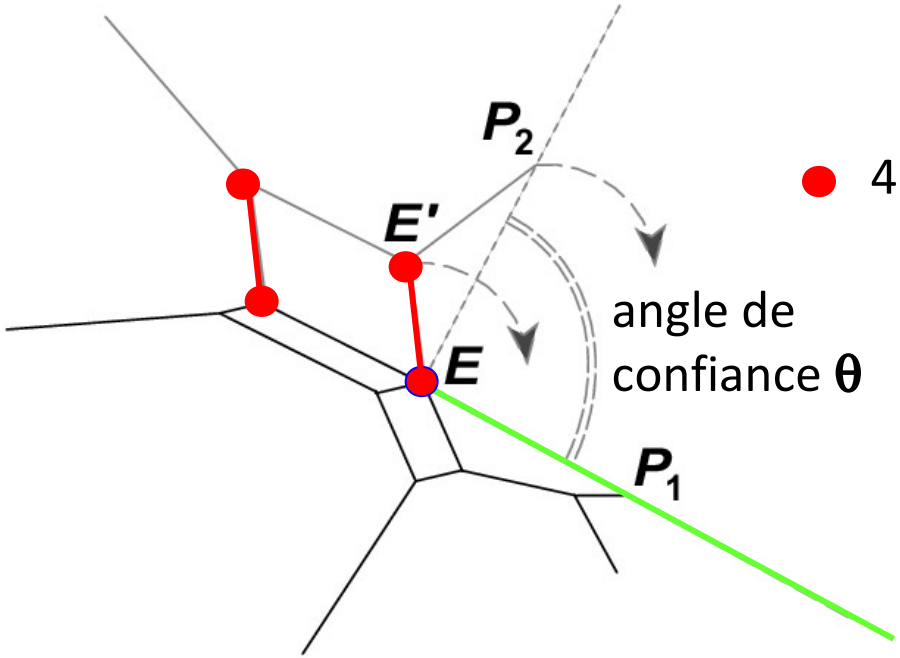
$$\alpha_{max}(S) = \alpha(S) + \min_{\text{boîte } b_i}\{(\alpha_i - \alpha(S) - \pi) \bmod 2\pi\}$$

$$\alpha_{min}(S) = \alpha(S) - \min_{\text{boîte } b_i}\{(\alpha(S) - \alpha_i) \bmod 2\pi\}$$

Gambette & Huson, TCBB, 2008

Algorithme Box-opening

Choix de l'angle $\alpha(S)$: **collisions globales**

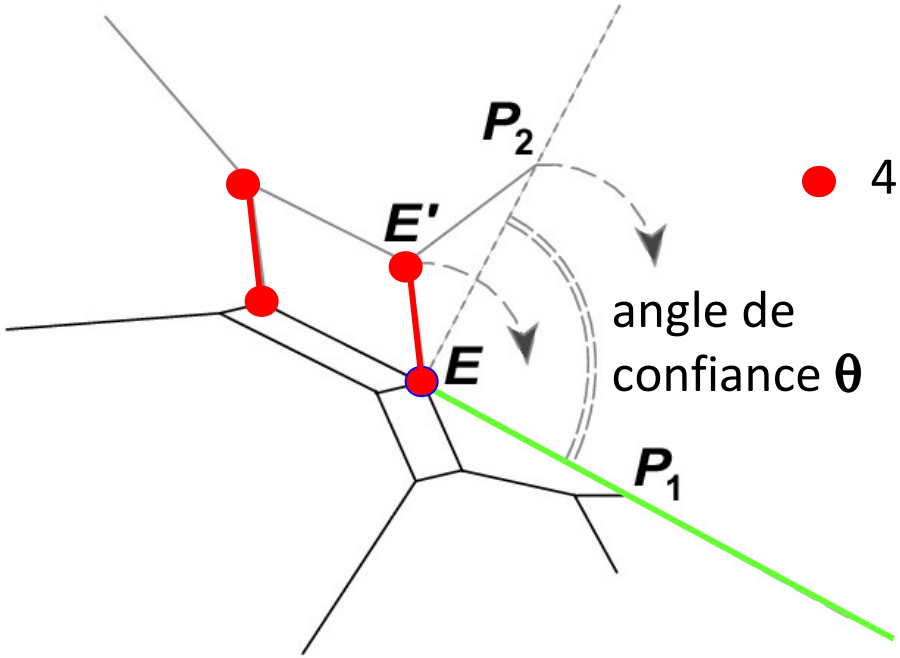


- 4 nœuds extrêmes

Gambette & Huson, TCBB, 2008

Algorithme Box-opening

Choix de l'angle $\alpha(S)$: **collisions globales**



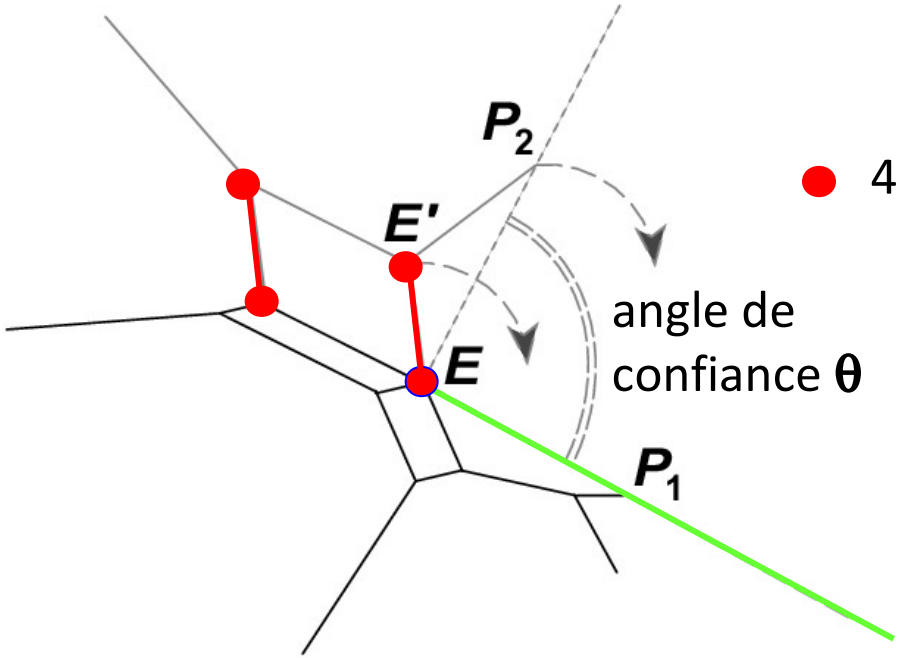
- 4 nœuds extrêmes

➡ **optimisations locales** de l'aire

Gambette & Huson, TCBB, 2008

Algorithme Box-opening

Choix de l'angle $\alpha(S)$: **collisions globales**



- 4 nœuds extrêmes

➡ **optimisations locales** de l'aire

➡ **optimisations globale** du réseau
métaheuristique (recuit simulé)

Gambette & Huson, TCBB, 2008

Reconstruction de réseaux phylogénétiques

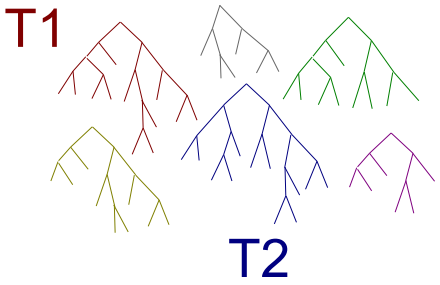
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{séquences de gènes}

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

Guindon & Gascuel, SB, 2003



{arbres}

Base HOGENOM

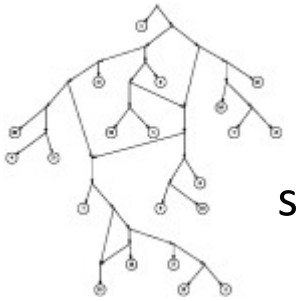
Phyl-ARIANE

ANR

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

Réconciliation ou consensus d'arbres

réseau explicite



super-réseau optimal *N*

Reconstruction de réseaux phylogénétiques

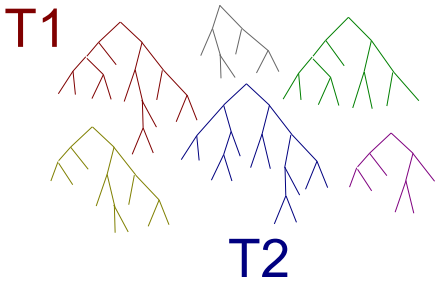
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

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

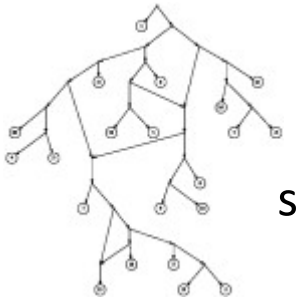
ANR

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

Triplets et quadruplets, clades et bipartitions

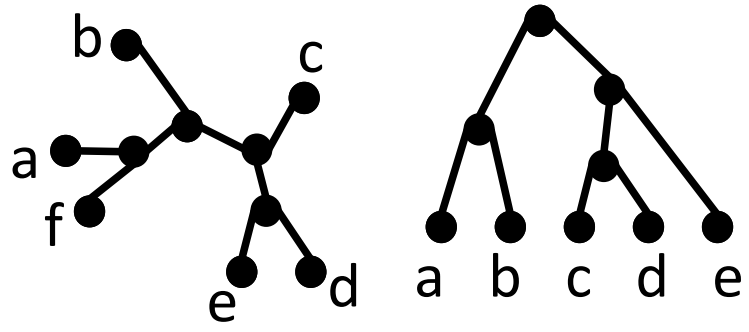
Problème :
Reconstruire le **super-réseau** d'un ensemble d'arbres est

difficile.

Idée :
reconstuire un réseau contenant tous les :

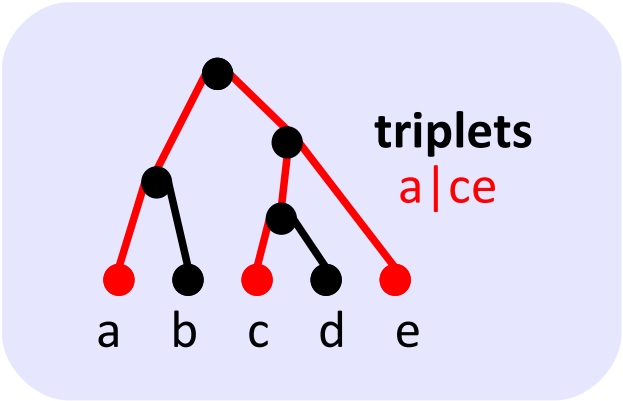
triplets
quadruplets
clades
bipartitions

des arbres en entrée ?



Triplets et quadruplets, clades et bipartitions

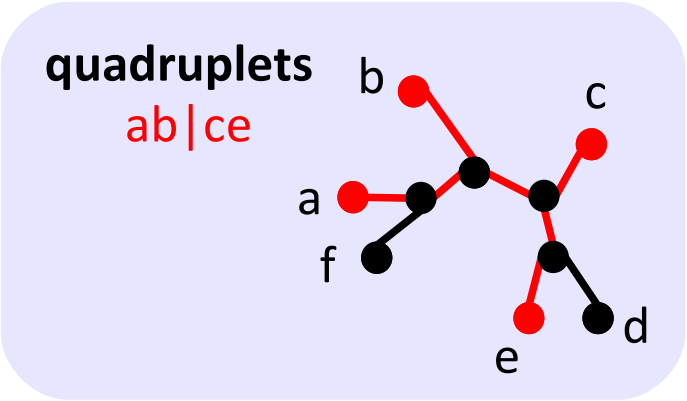
Idée :
reconstuire un réseau contenant tous les :

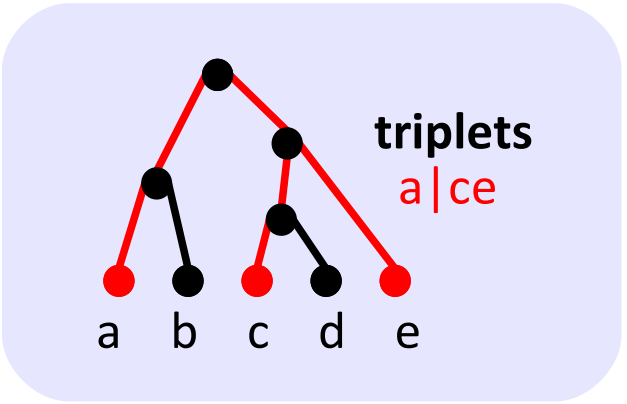


des arbres en entrée ?

Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :



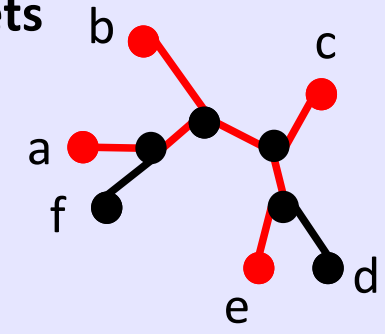


des arbres en entrée ?

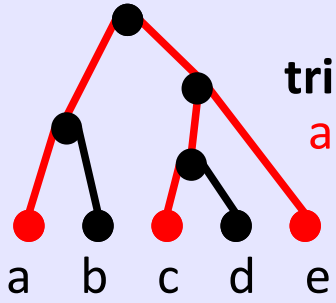
Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :

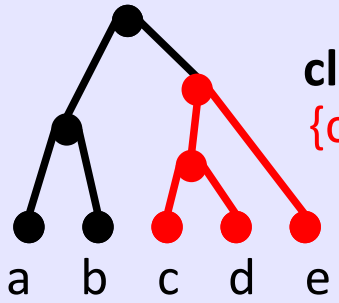
quadruplets
ab|ce



triplets
a|ce



clades
{c,d,e}

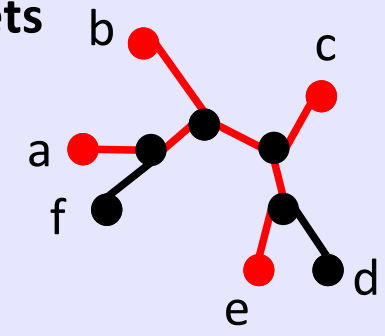


des arbres en entrée ?

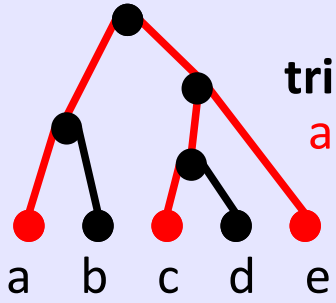
Triplets et quadruplets, clades et bipartitions

Idée :
reconstuire un réseau contenant tous les :

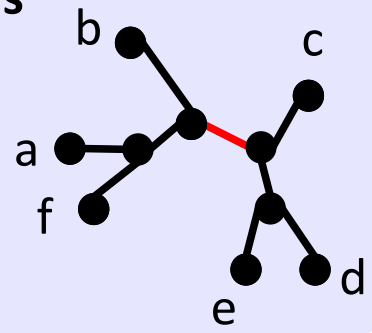
quadruplets
ab|ce



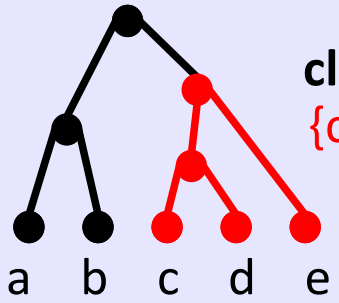
triplets
a|ce



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



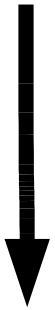
clades
{c,d,e}



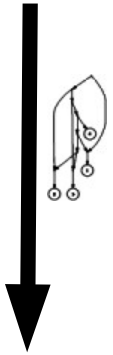
des arbres en entrée ?

Reconstruction depuis les triplets

{arbres}



{triplets}



N'
réseau
de niveau *k*

Reconstruction d'un réseau de **niveau** *k* à partir d'un ensemble de **triplets**

Jansson, Nguyen & Sung, JOC, 2006 : NP-complet pour niveau 1,
Van Iersel, Kelk & Mnich, JBCB, 2009 : NP-complet pour niveau *k*

niveau =
mesure de “complexité”, d'éloignement par rapport à une structure d'arbre.

Choy, Jansson, Sadakane & Sung, TCS, 2005

Reconstruction depuis les triplets

{arbres}

{triplets}

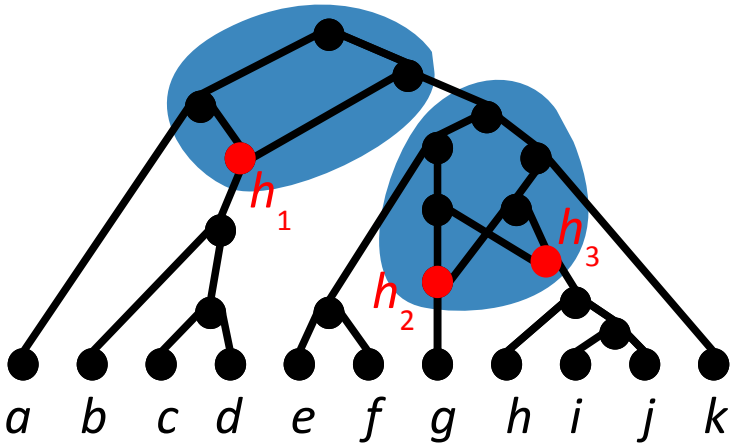
N'
réseau
de niveau *k*

Reconstruction d'un réseau de **niveau** *k* à partir d'un ensemble de **triplets**

Jansson, Nguyen & Sung, JOC, 2006 : NP-complet pour niveau 1,
Van Iersel, Kelk & Mnich, JBCB, 2009 : NP-complet pour niveau *k*

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

Choy, Jansson, Sadakane & Sung, TCS, 2005



réseau de niveau 2

Reconstruction depuis les triplets

{arbres}

{triplets}

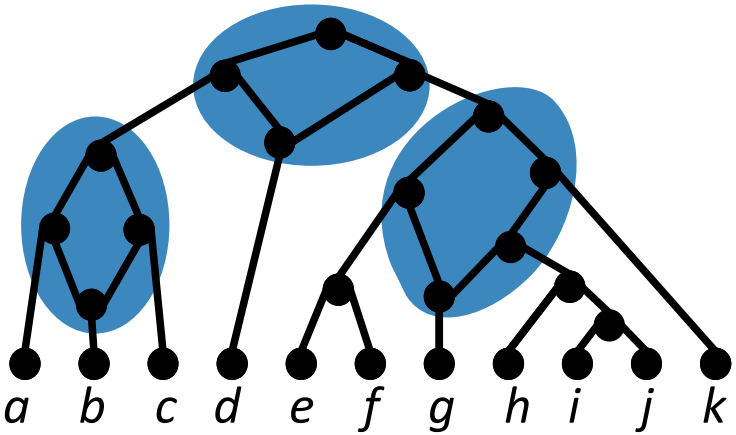
N'
réseau
de niveau *k*

Reconstruction d'un réseau de **niveau** *k* à partir d'un ensemble de **triplets**

Jansson, Nguyen & Sung, JOC, 2006 : NP-complet pour niveau 1,
Van Iersel, Kelk & Mnich, JBCB, 2009 : NP-complet pour niveau *k*

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

Choy, Jansson, Sadakane & Sung, TCS, 2005

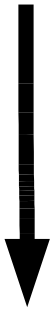


réseau de niveau 1 “galled tree”

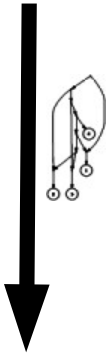
Gusfield, Eddhu & Langley, IJOC, 2004

Reconstruction depuis les triplets

{arbres}



{triplets}

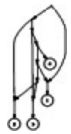


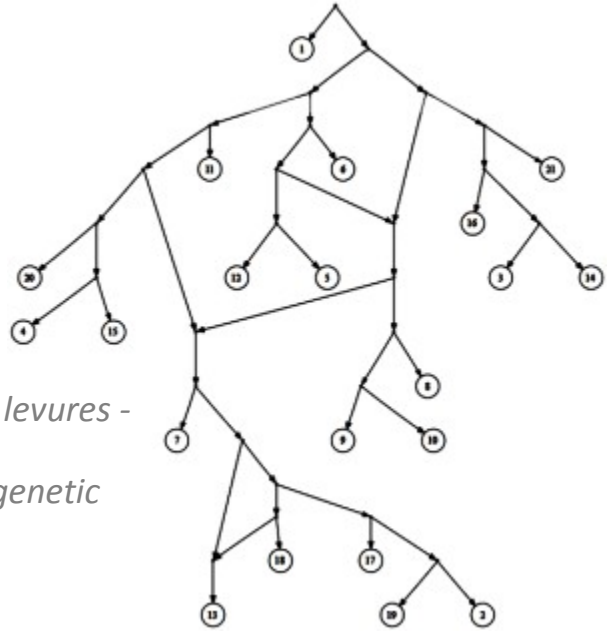
N'
réseau
de niveau *k*

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

Jansson, Nguyen & Sung, SODA'05 : O(n^3) pour niveau 1,
van Iersel, Kelk & al, RECOMB'08 : O(n^8) pour niveau 2,
To & Habib, CPM'09 : $O(n^{5k+4})$ pour niveau k
Van Iersel & Kelk, J. Theor. Biol., 2011 : NP-complet de trouver le niveau minimal

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

Programme Simplistic 

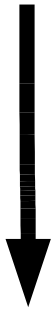


Réseau phylogénétique de levures - Van Iersel et al. : Constructing level-2 phylogenetic networks from triplets. RECOMB 2008

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

Reconstruction depuis les clades souples

{arbres}



{clades}

N'

réseau à 1 couche de réticulation

Consensus de clades souples :
Dendroscope

Huson et al., BMCB, 2007

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

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

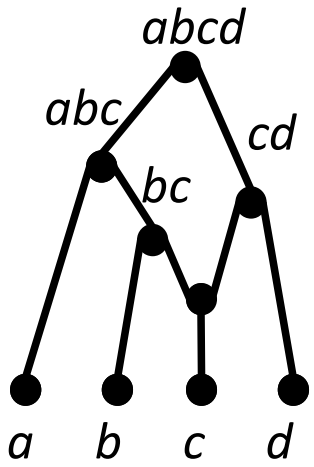
Méthode exacte de reconstruction de **réseaux de niveau *k*** à partir de **clades souples**

Iersel, Kelk, Rupp & Huson, ISMB 2010

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

Clades stricts et souples

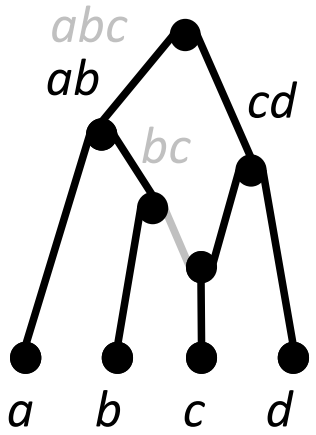
Clade “strict” : ensemble des feuilles sous un noeud du réseau



Clades stricts et souples

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

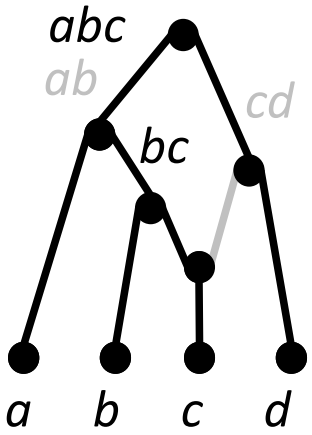
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



Clades stricts et souples

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

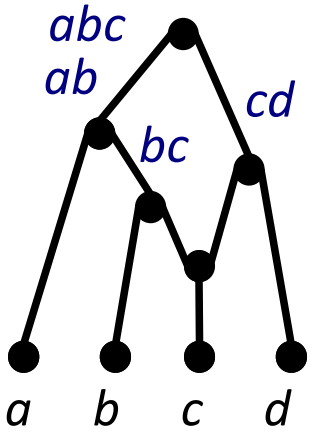
Modèle de **transmission arborée** des gènes
(gène transmis intégralement)



Clades stricts et souples

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

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



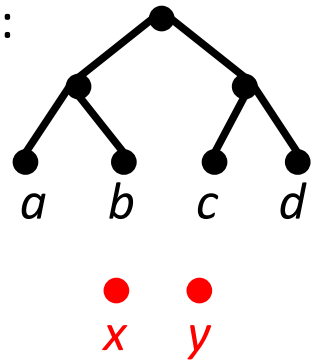
L'ensemble *S*(*N*) de **tous les clades souplement compatibles** avec *N* peut être de taille **exponentielle.**
Tester si un **clade souple** appartient à un réseau : **NP-complet.**

Kanj, Nakhleh, Than & Xia, TCS, 2008

Une approche de reconstruction en deux étapes

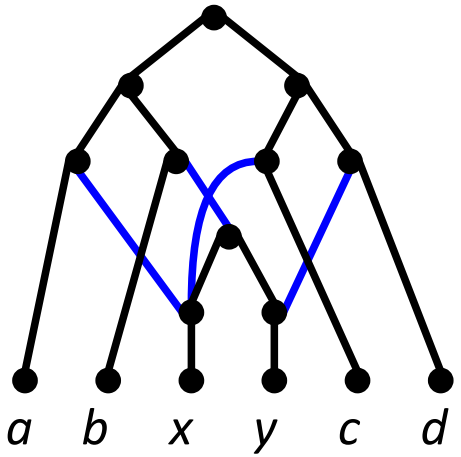
1- Trouver un **ensemble minimum de conflits** parmi les clades :
- partie sans conflits ➡ arbre,
- taxons impliqués dans des conflits ➡ sous les réticulations.

MAXIMUM COMPATIBLE SUBSET



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

MINIMUM ATTACHMENT PROBLEM



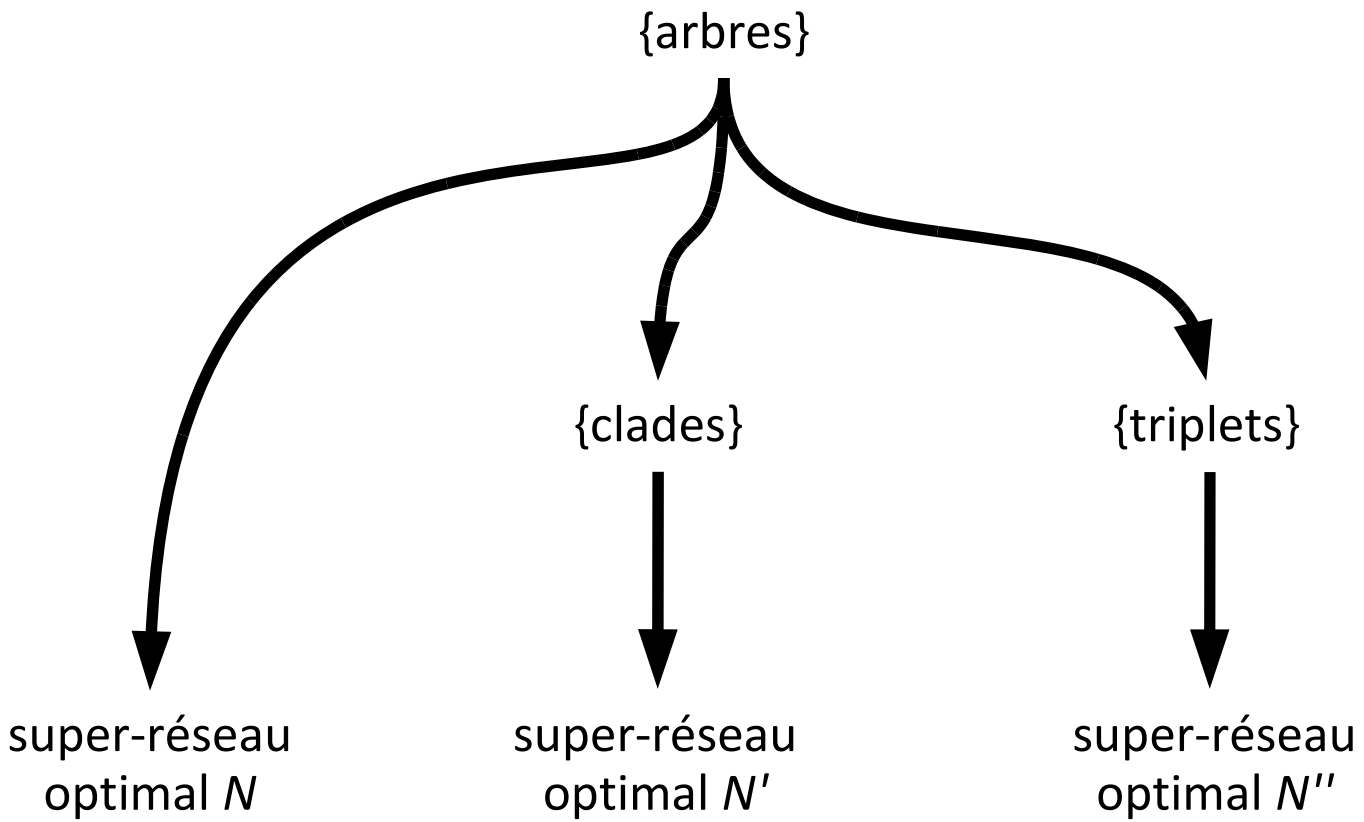
Huson, Rupp, Berry, Gambette & Paul, IMSB 2009

Plan

- Les réseaux phylogénétiques
- Méthodes de reconstruction
- Limites des méthodes combinatoires
- Illustration sur des données biologiques
- Utilisation sur des données textuelles
- Perspectives

Reconstruction combinatoire de réseaux phylogénétiques

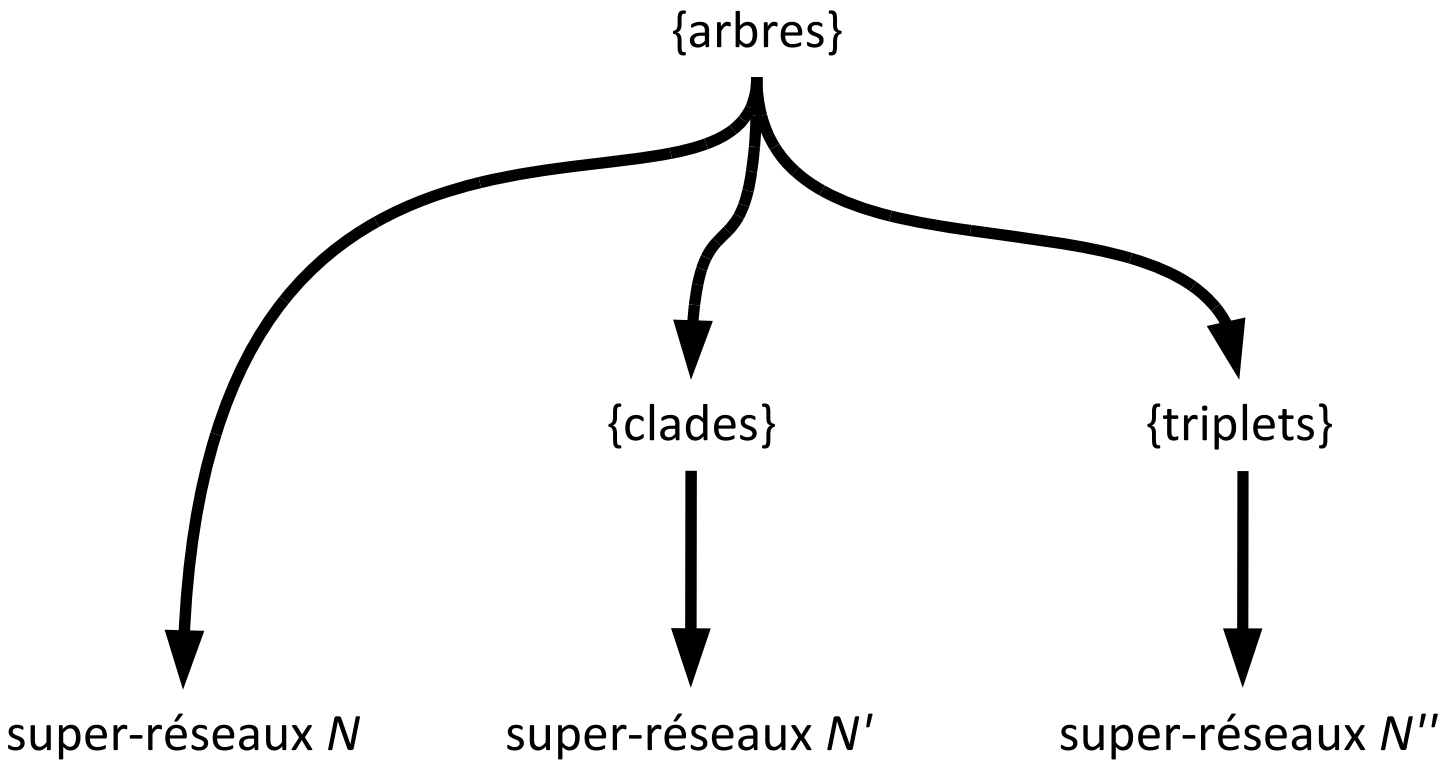
Idée :
modifier le type de données à traiter



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

Reconstruction combinatoire de réseaux phylogénétiques

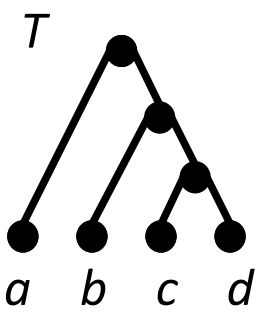
Idée :
modifier le type de données à traiter

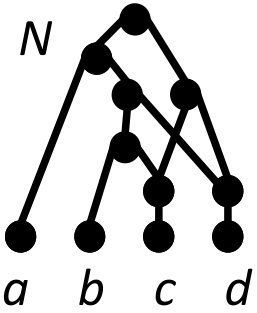


$$\{N\} \subseteq \{N'\} \subseteq \{N''\}$$

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***.



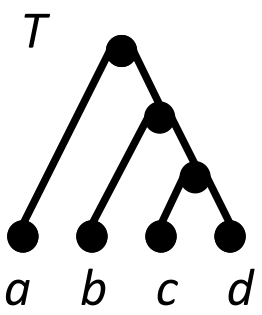


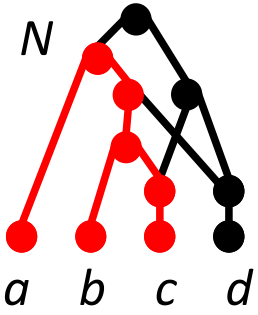
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



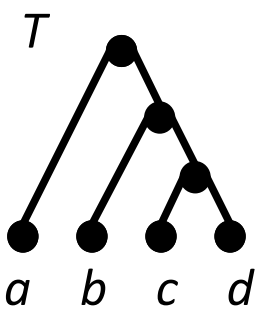


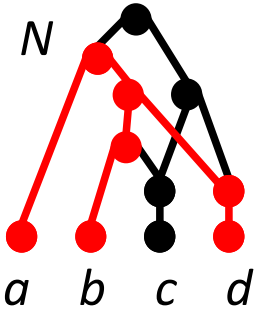
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



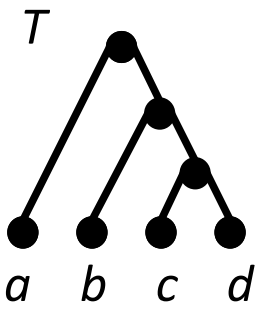


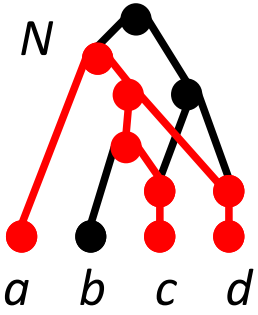
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



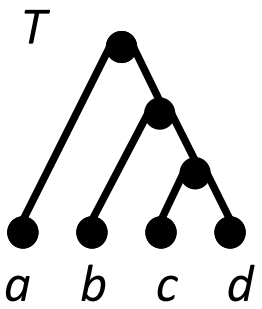


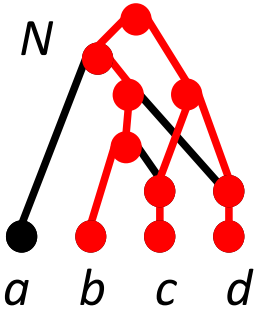
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



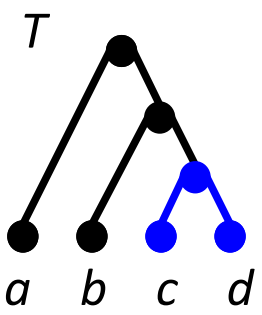


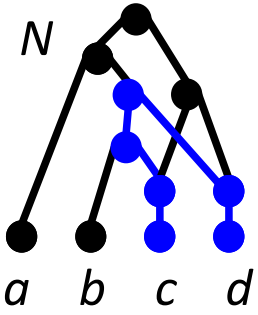
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



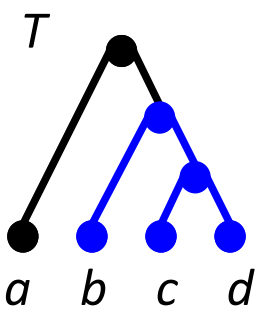


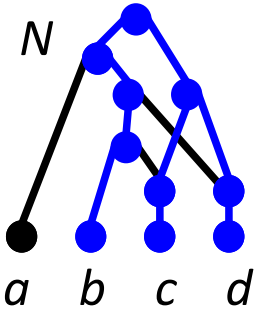
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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***.



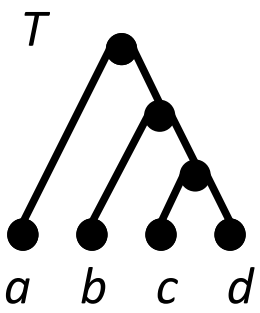


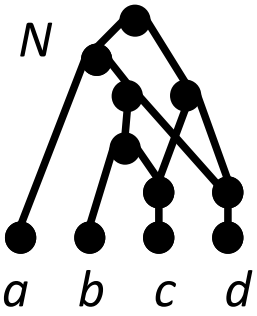
contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

Reconstruction combinatoire de réseaux phylogénétiques

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

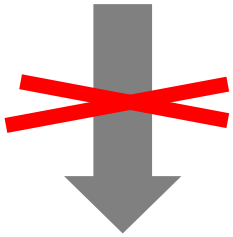




contient {*a*|*bc*,*a*|*bd*,*a*|*cd*,*b*|*cd*}
mais pas *T*

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

contient les clades / triplets d'un arbre *T*

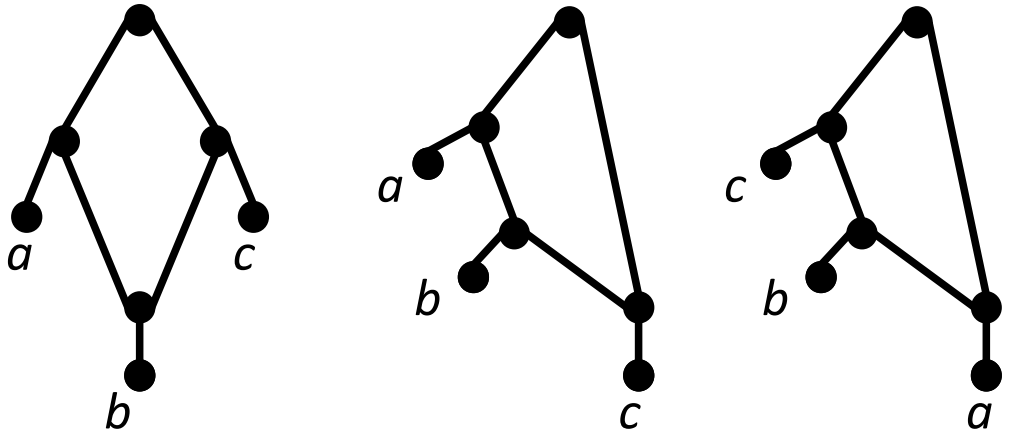


contient *T*.

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 somments)

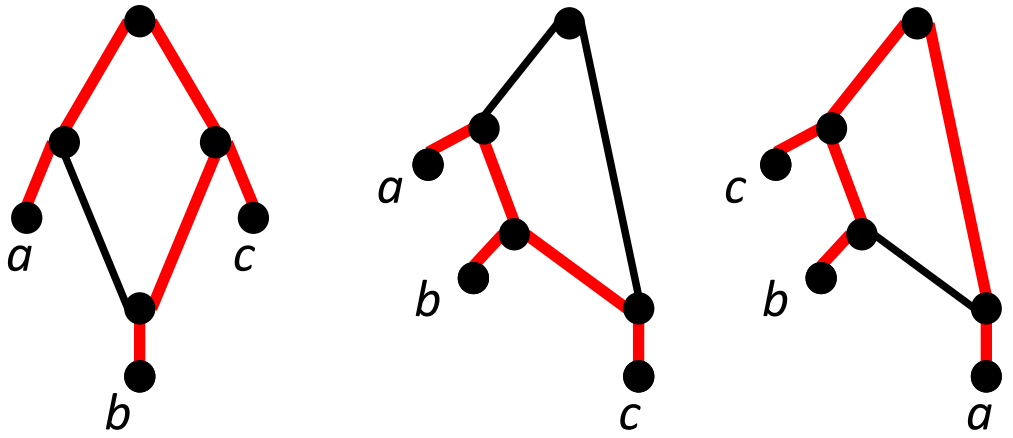
Gambette & Huber, 2011

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 somments)

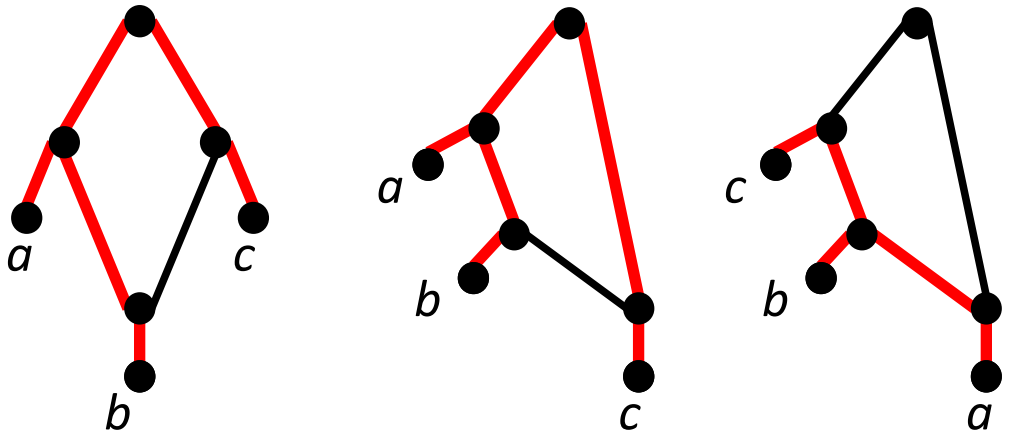
Gambette & Huber, 2011

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



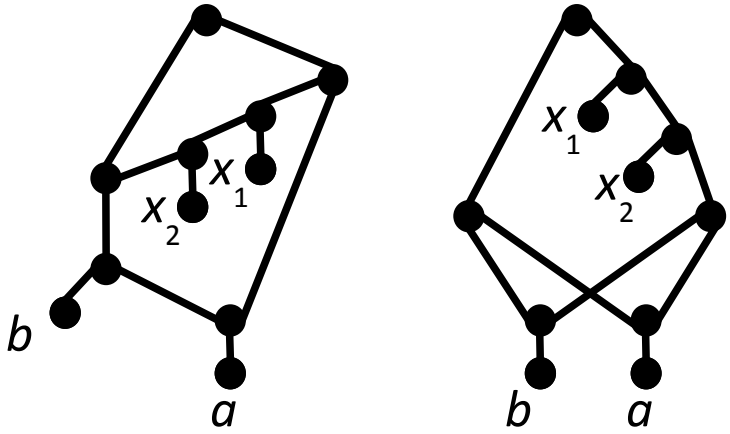
Caractérisation pour les réseaux de niveau 1 :
les seuls cas ambigus sont les blobs ci-dessus (< 5 somments)

Gambette & Huber, 2011

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.



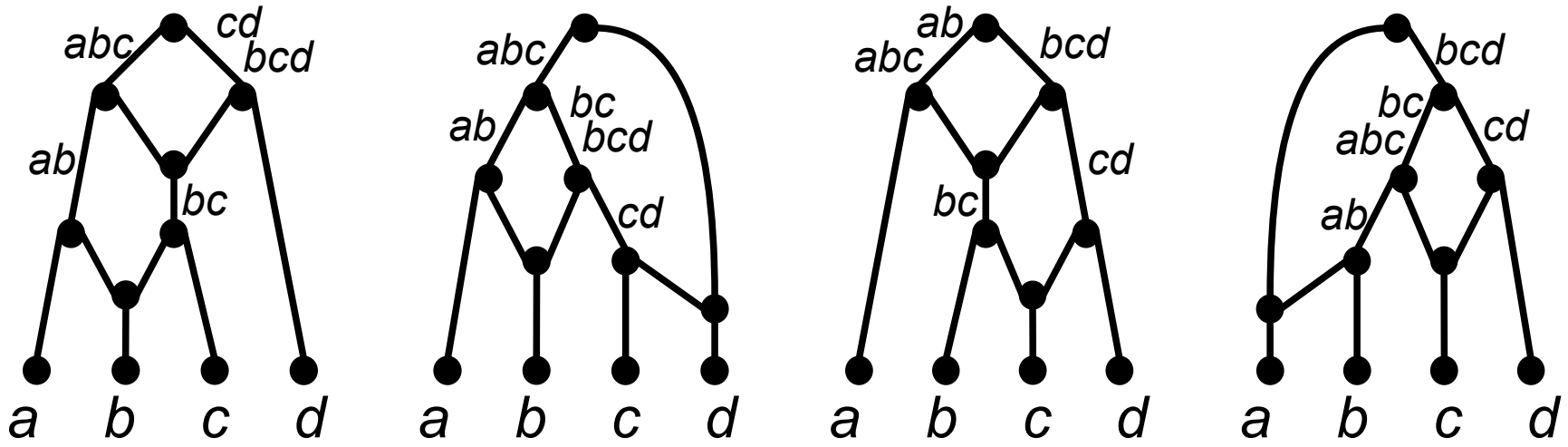
2 réseaux de niveau 2 avec le même ensemble de triplets

Gambette & Huber, 2011

Ambiguïté des solutions

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

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



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

Gambette & Huber, 2011

Utilisation pratique

méthodes existantes
à faire ou améliorer

Conditions d'utilisation	Données disponibles	Traitements possibles
arbres enracinés	arbres non enracinés	*Enracinement à partir d'un arbre d'espèces de référence ou des contraintes topologiques*
arbres de gènes simple copie	"MUL-trees" (avec gènes dupliqués)	Traitement des MUL-trees Scornavacca, Berry & Ranwez, 2009
Clades et triplets corrects	données bruitées	Nettoyage d'arbre PhySIC_IST, 2008 Filtrage des données (clades avec support élevé, présent dans plus de x% des arbres) *Edition de données : solution contenant le maximum des données d'entrée*
Données complètes (ensembles denses de triplets, clades complets)	données partielles, gènes supprimés	*Sélection d'un grand nombre d'arbres avec un grand nombre d'espèces en commun* *Sélection du nombre maxium de taxons avec densité des triplets* problèmes NP-complets

Plan

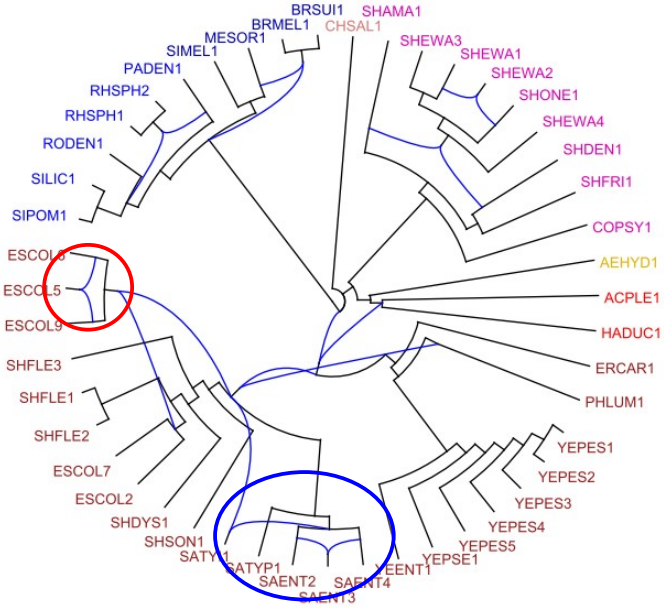
- Les réseaux phylogénétiques
- Méthodes de reconstruction
- Limites des méthodes combinatoires
- Illustration sur des données biologiques
- Utilisation sur des données textuelles
- Perspectives

Illustrations

16 arbres sur 47 taxons de la base HOGENOM (proteobacteria)

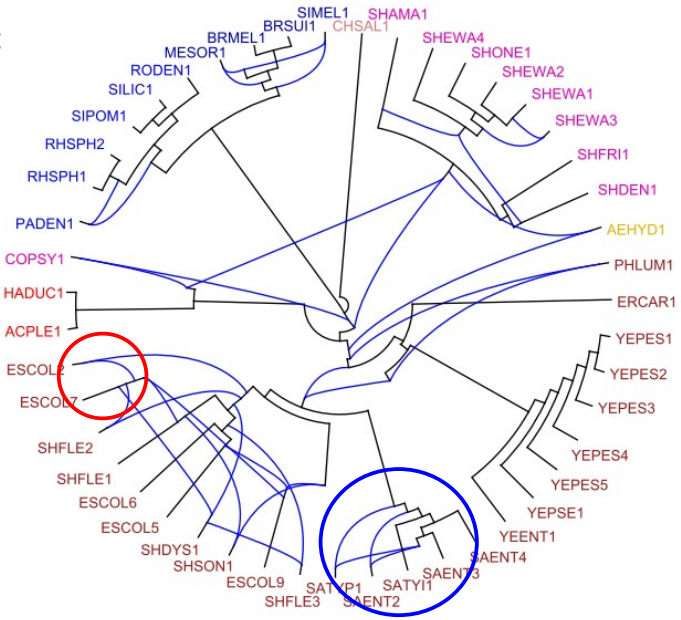
24 Enterobacteriales
2 Pasteurellales
1 Aeromonadales
9 Alteromonadales
1 Oceanospirillales
6 Rhodobacterales
4 Rhizobiales

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

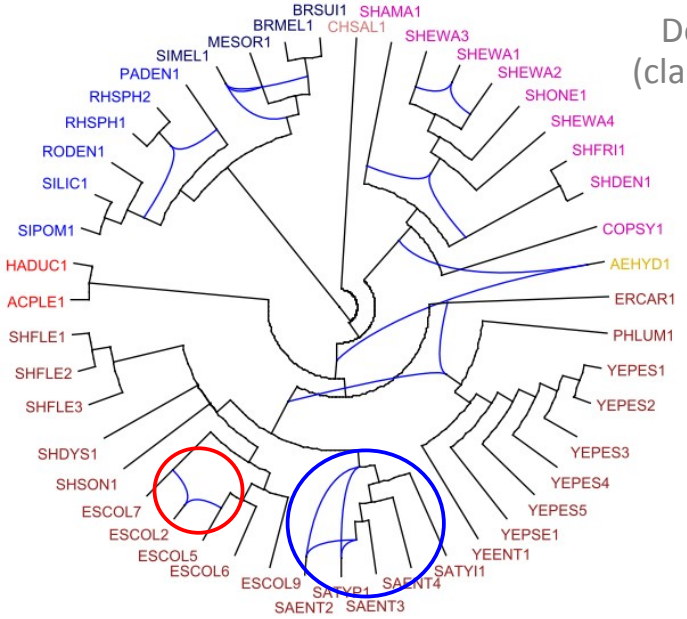


Lev1athan (heuristique triplets, niveau 1) 24 sec.

Simplistic (triplets, niveau 7) 63 sec.



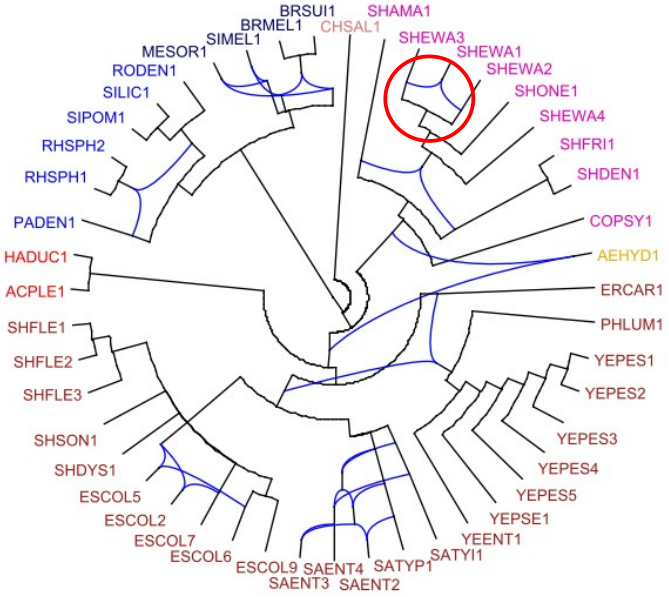
Dendroscope (clades, “galled network”) <1 sec.



Illustrations

16 arbres sur 47 taxons de la base HOGENOM (proteobacteria)

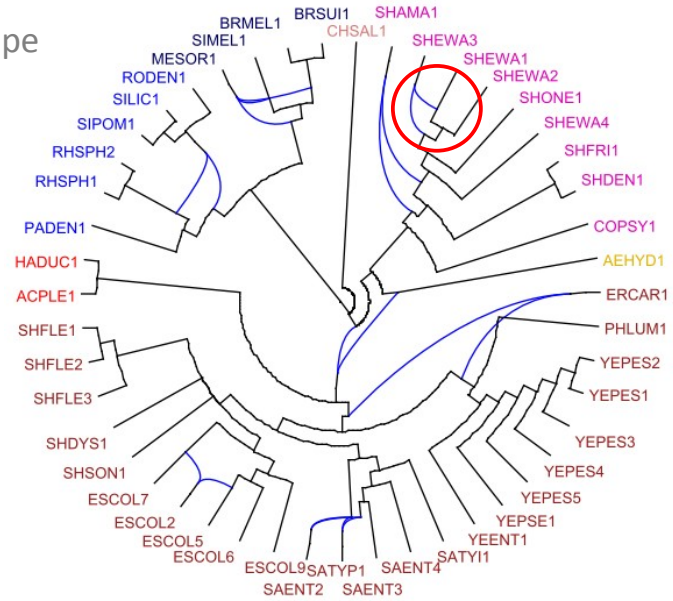
24 Enterobacteriales
2 Pasteurellales
1 Aeromonadales
9 Alteromonadales
1 Oceanospirillales
6 Rhodobacterales
4 Rhizobiales

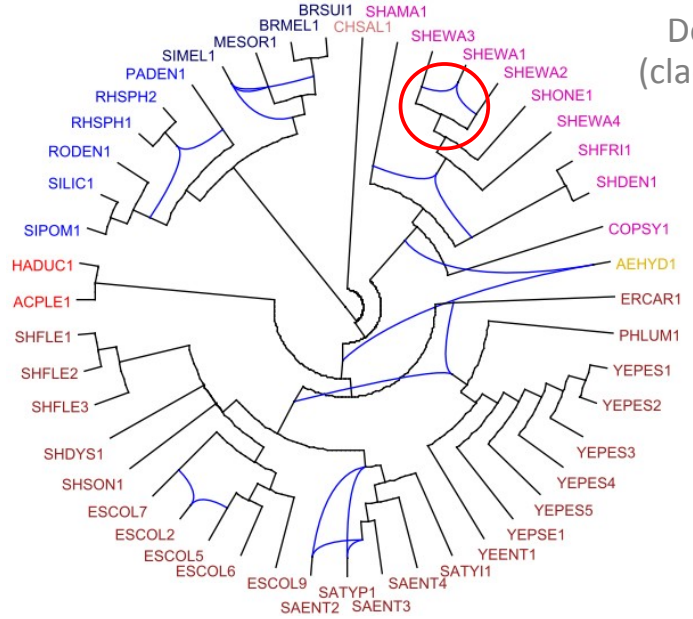


Dendroscope (clades stricts, “cluster network”, level 1) <1 sec.

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

Dendroscope (clades, niveau 2) 2 sec.



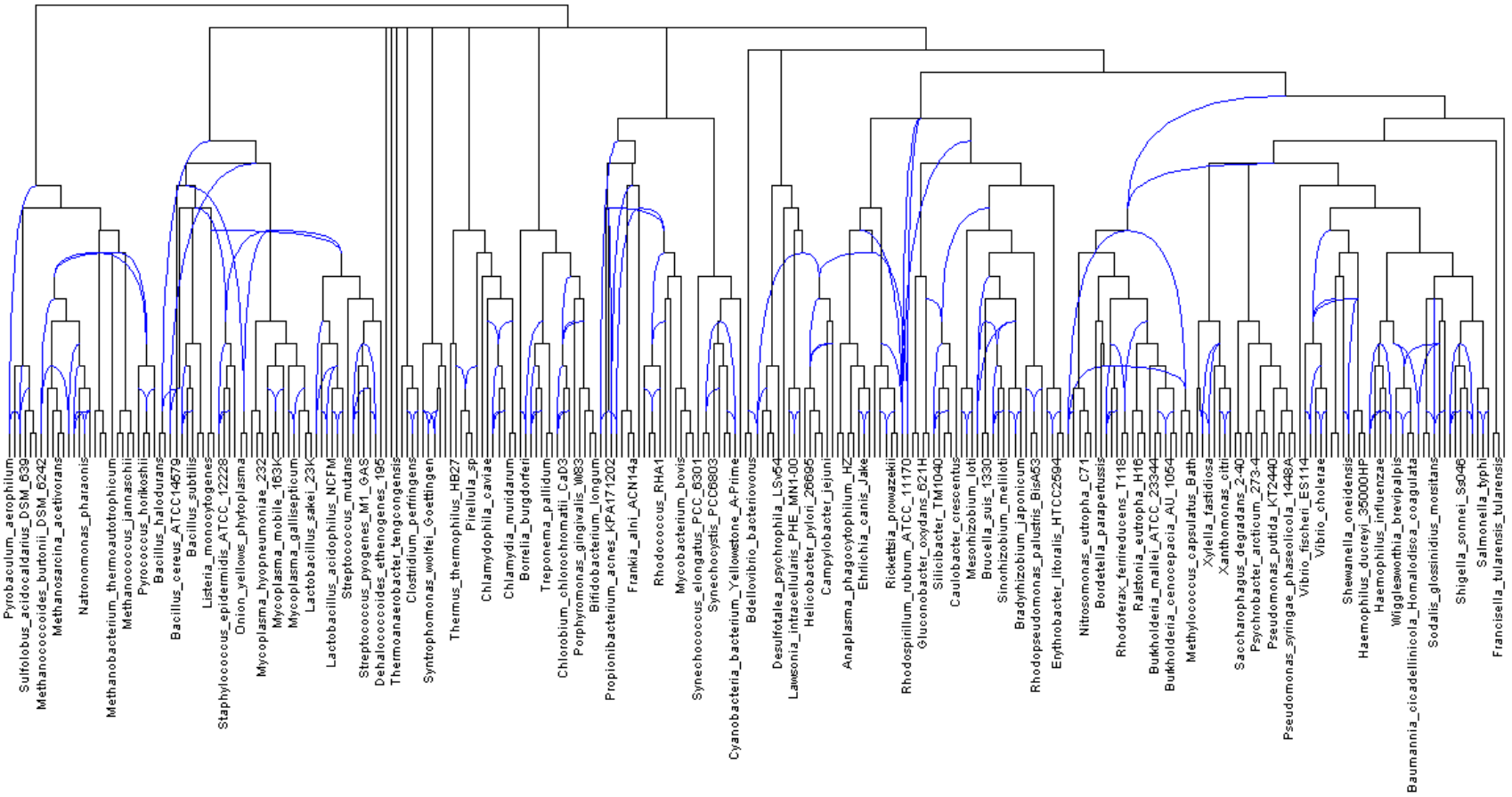


Dendroscope (clades, “galled network”) <1 sec.

Illustrations

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

Auch, Steigele, Huson & Henz, 2009

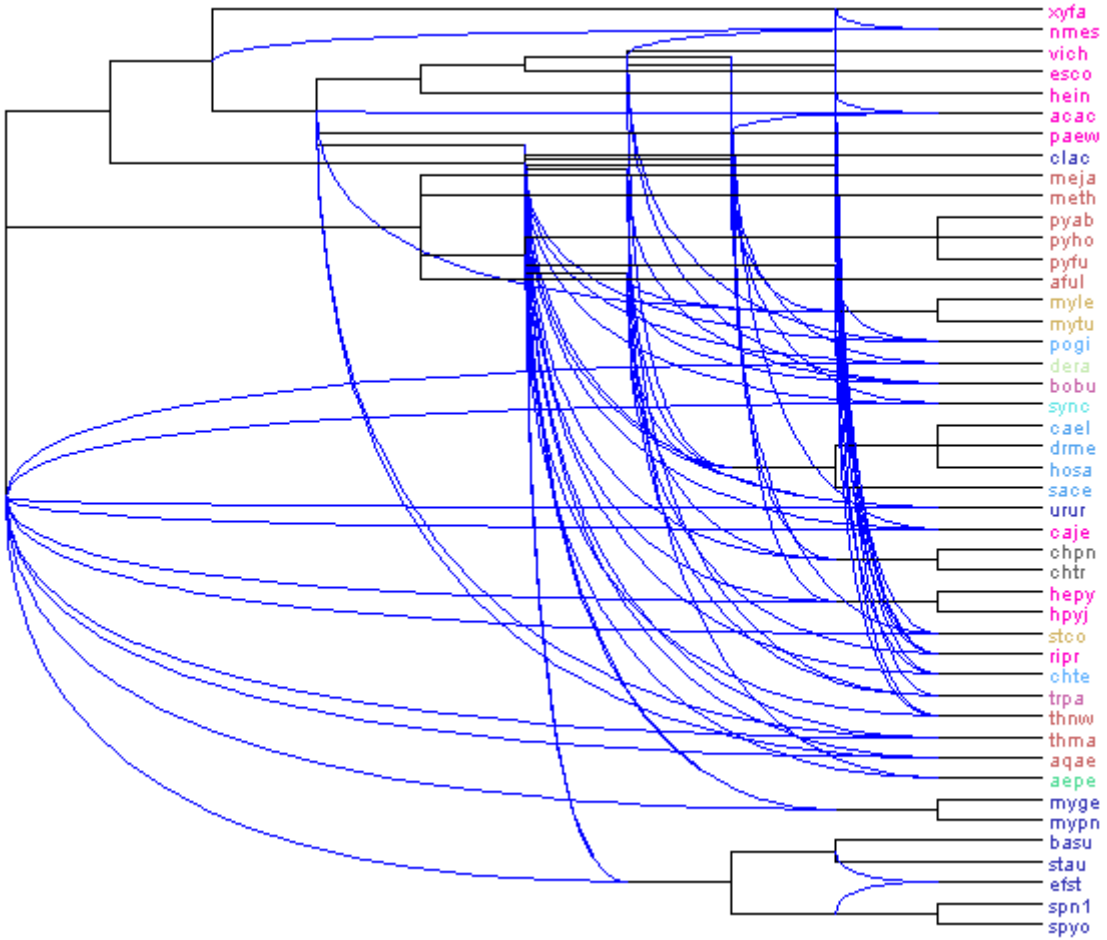


Dendroscope
(clusters, galled network)
2 sec.

Illustrations

23 arbres, 45 espèces des 3 domaines du vivant
clades avec 99% de confiance (bootstrap)

Dendroscope
(“galled
network”)
4 sec.

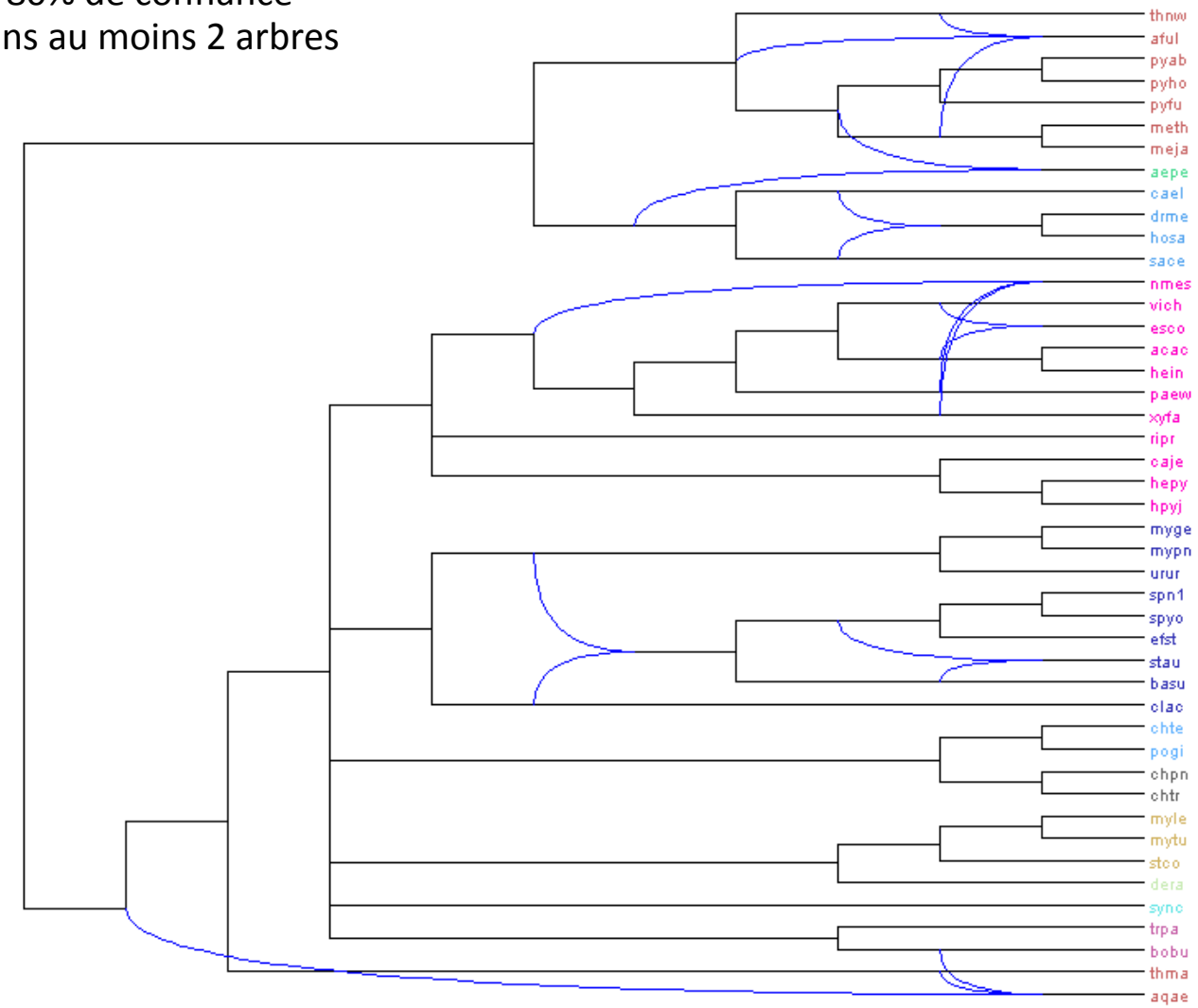


Données de Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ:
Universal trees based on large combined protein sequence data sets. Nat Genet 2001, 28:281--285

Illustrations

23 arbres, 45 espèces des 3 domaines du vivant
clades avec 80% de confiance
présents dans au moins 2 arbres

Dendroscope
(“galled
network”)
<1 sec.

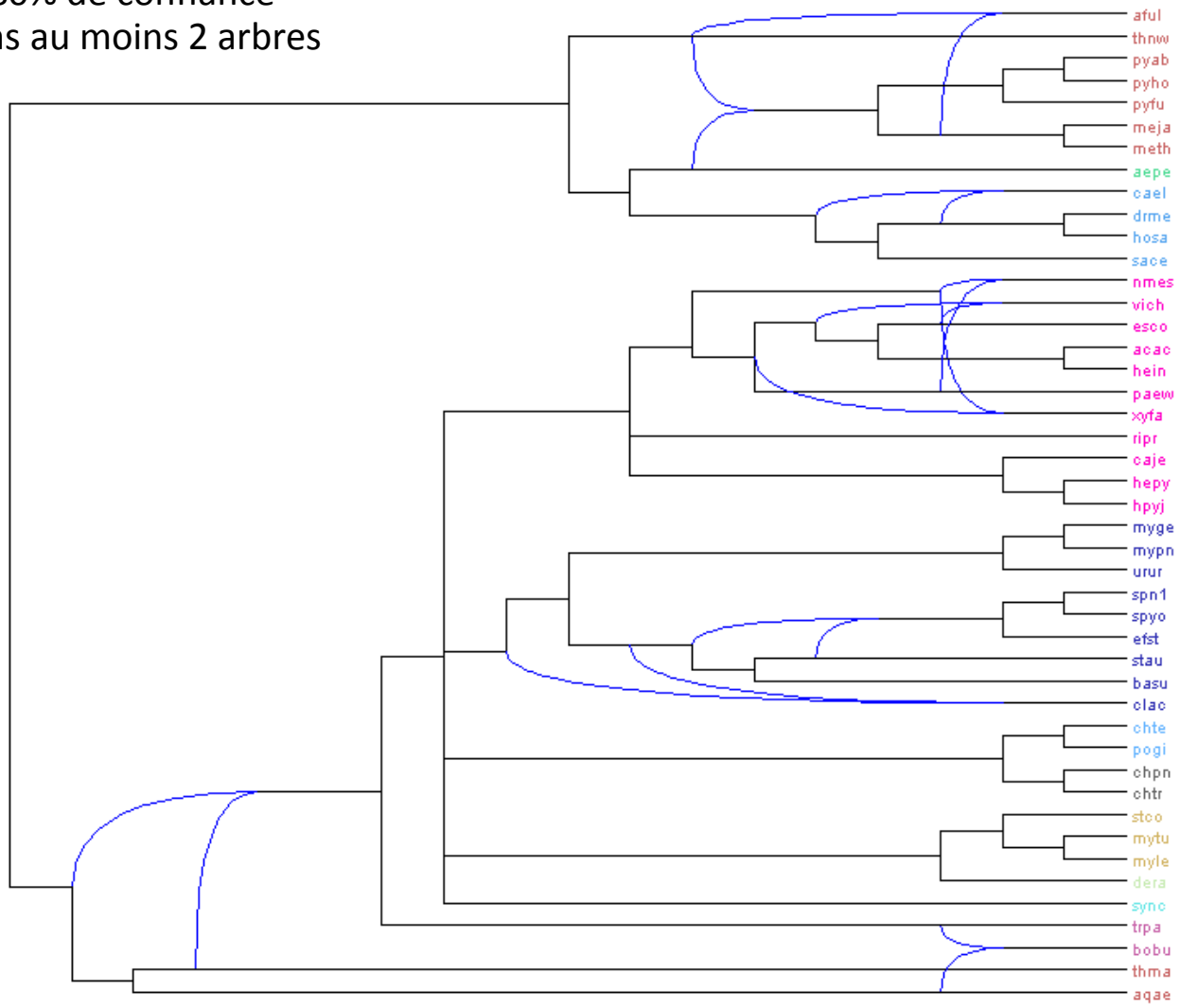


Données de Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ:
Universal trees based on large combined protein sequence data sets. Nat Genet 2001, 28:281--285

Illustrations

23 arbres, 45 espèces des 3 domaines du vivant
clades avec 80% de confiance
présents dans au moins 2 arbres

Dendroscope
(réseau de
niveau 3)
<1 sec.

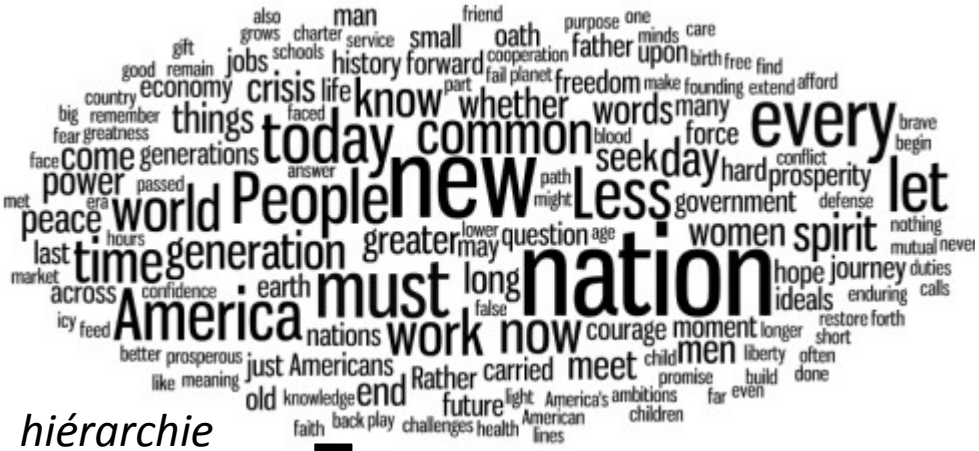


Données de Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ:
Universal trees based on large combined protein sequence data sets. Nat Genet 2001, 28:281--285

Plan

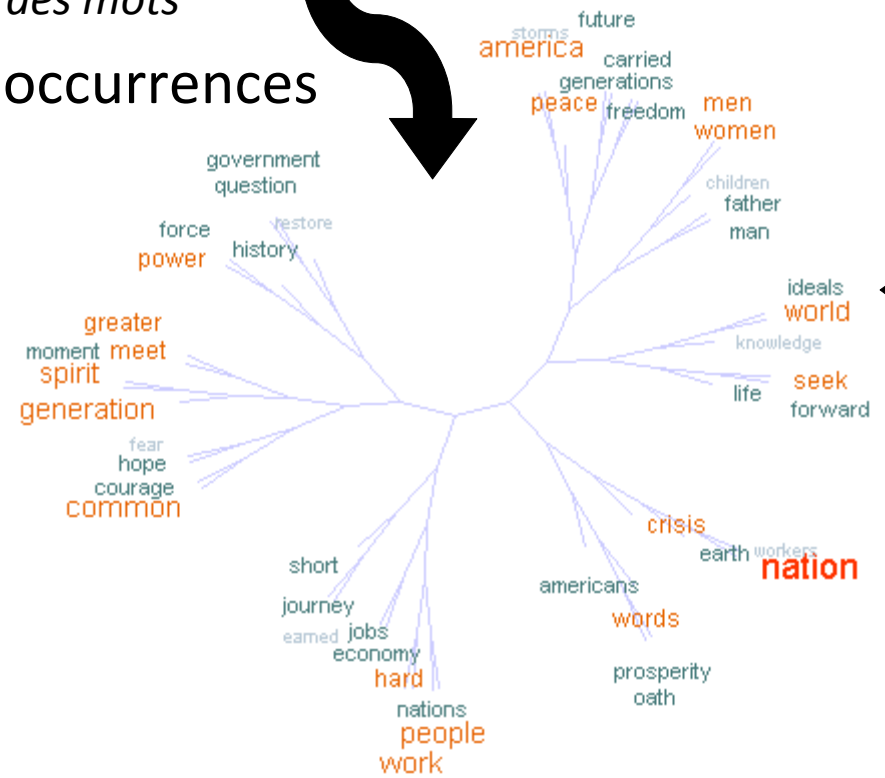
- Les réseaux phylogénétiques
- Méthodes de reconstruction
- Limites des méthodes combinatoires
- Illustration sur des données biologiques
- Utilisation sur des données textuelles
- Perspectives

Nuage arboré, une information double

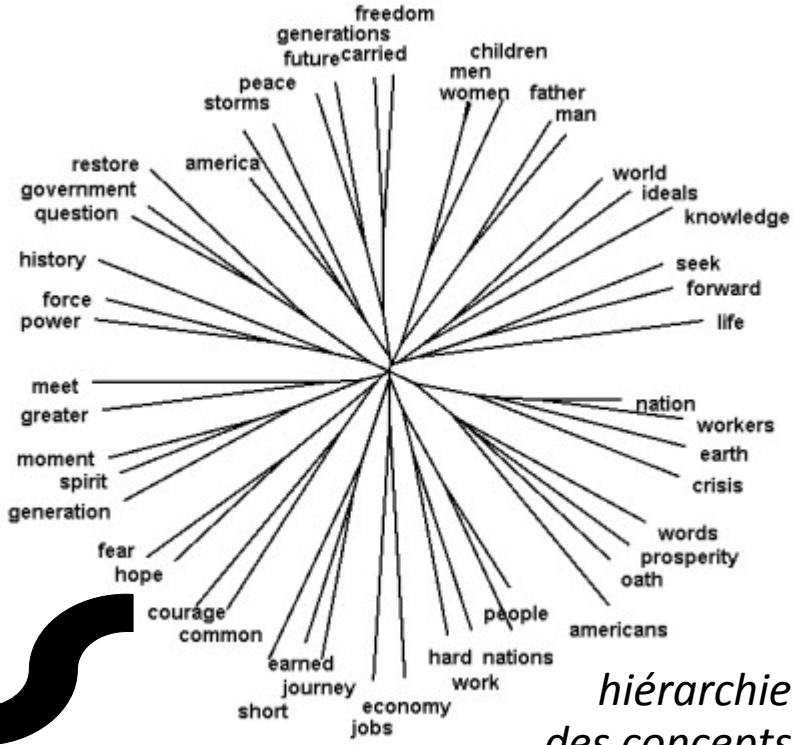


hiérarchie des mots

occurrences



cooccurrences



hiérarchie des concepts

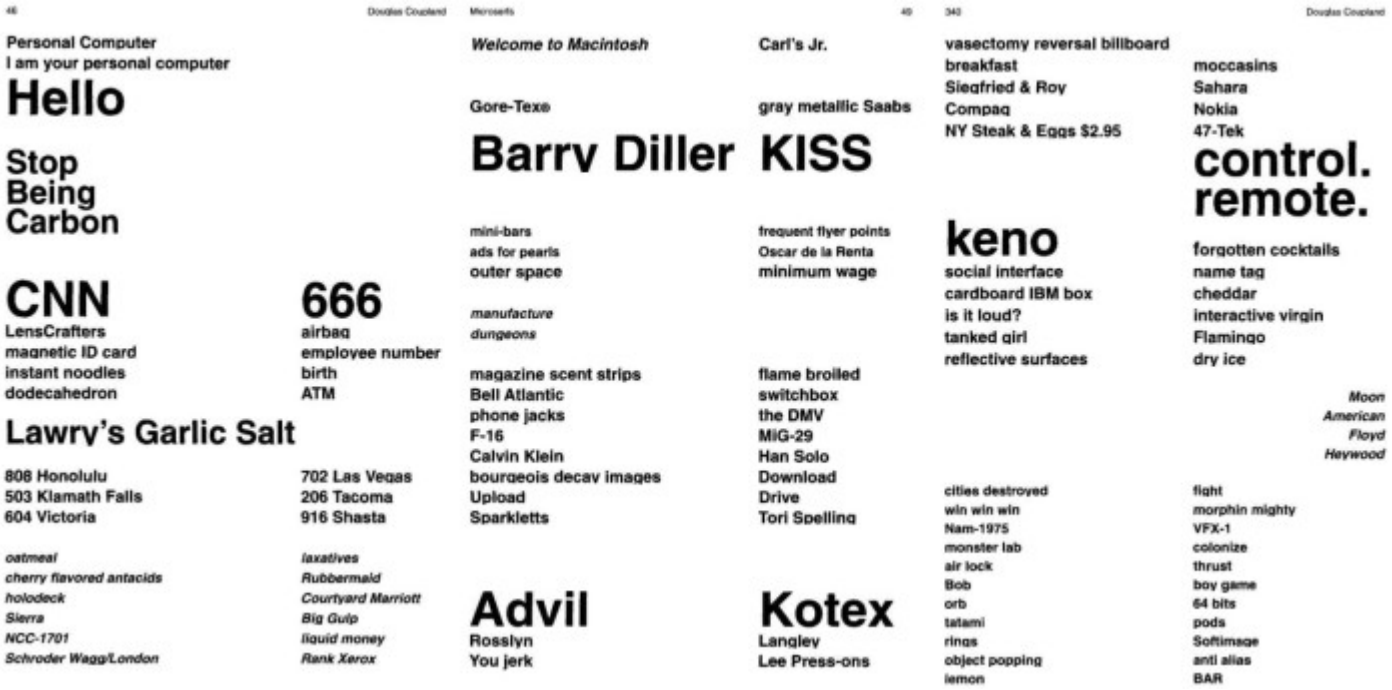
construit avec

TreeCloud.org

SplitsTree4

Nuages de tags

- Construits depuis un ensemble de tags
- Taille de police liée à la fréquence



Ce qui est habituellement cité comme le premier nuage de tags, dans *Microserfs* de D. Coupland, HarperCollins, Toronto, 1995

Nuages de tags

- Construits depuis l'ensemble **des mots d'un texte**
- Taille de police liée à la fréquence
- Se sont popularisés avec Wordle
- Donnent un bon aperçu d'un texte

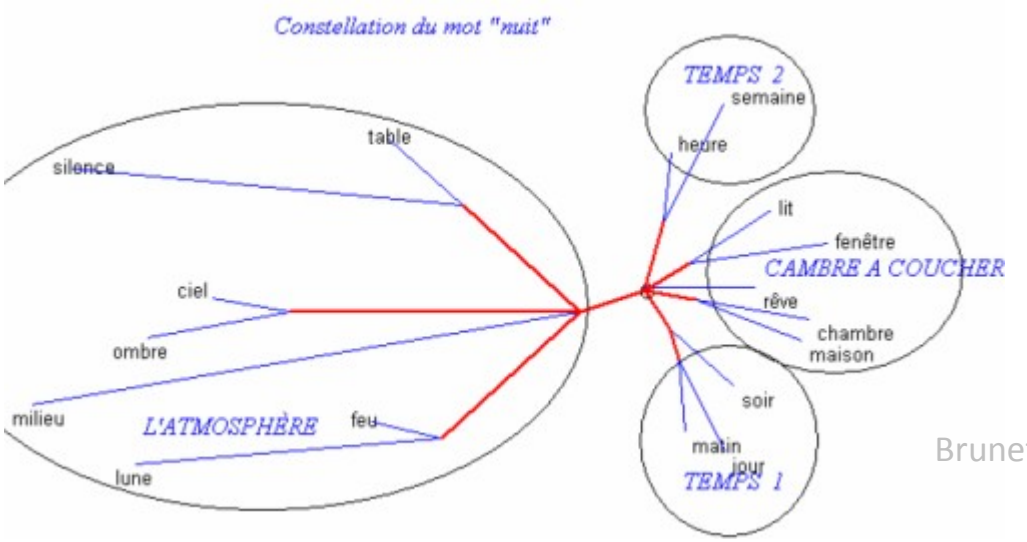
Nuages Wordle des mots les plus utilisés début 2009 dans les blogs des Top 100 politique et high-tech de Wikio
http://aixtal.blogspot.com/2009/04/web-de-quoi-parlent-les-blogs.html

Extraire l'information sémantique d'un texte

- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)

Extraire l'information sémantique d'un texte

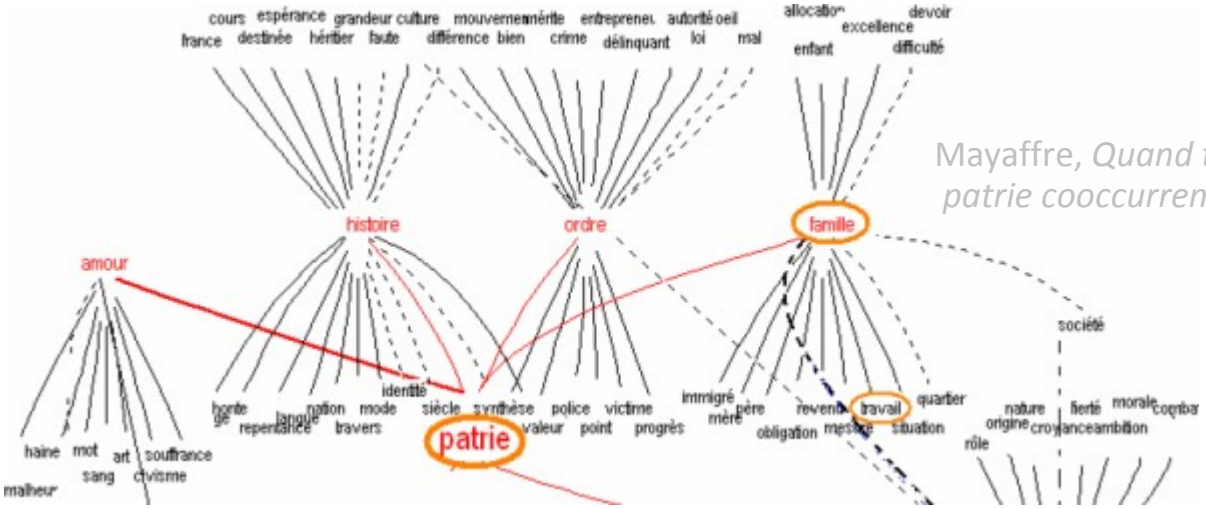
- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)



Brunet, *Les séquences (suite)*, JADT'08

Extraire l'information sémantique d'un texte

- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)

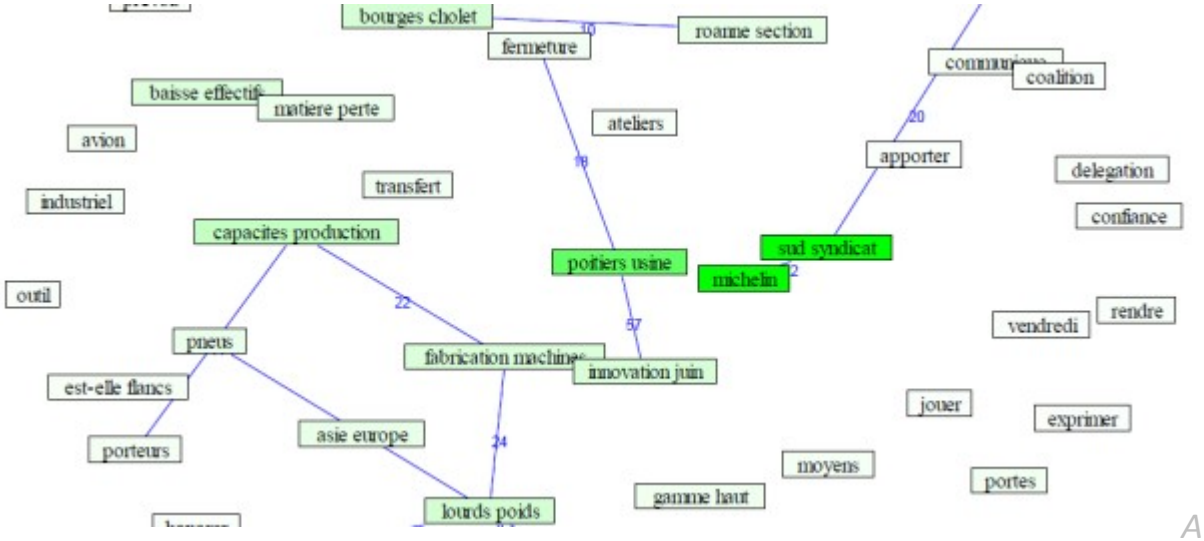


Mayaffre, *Quand travail, famille, et patrie cooccurrent dans le discours de Nicolas Sarkozy*, JADT'08

Extraire l'information sémantique d'un texte

- analyse arborée
- graphe de coocurrence
- graphe sémantique
- lexicogramme récursif
- désambiguïsation lexicale
- réseau Phrasenet
- projection géodésique

Brunet (Hyperbase)
Brunet (Hyperbase)
Grimmer (Wordmapper)
Martinez (Coocs)
Véronis (Hyperlex)
Viegas et al. (IBM Many Eyes)
Viprey (Astartex)

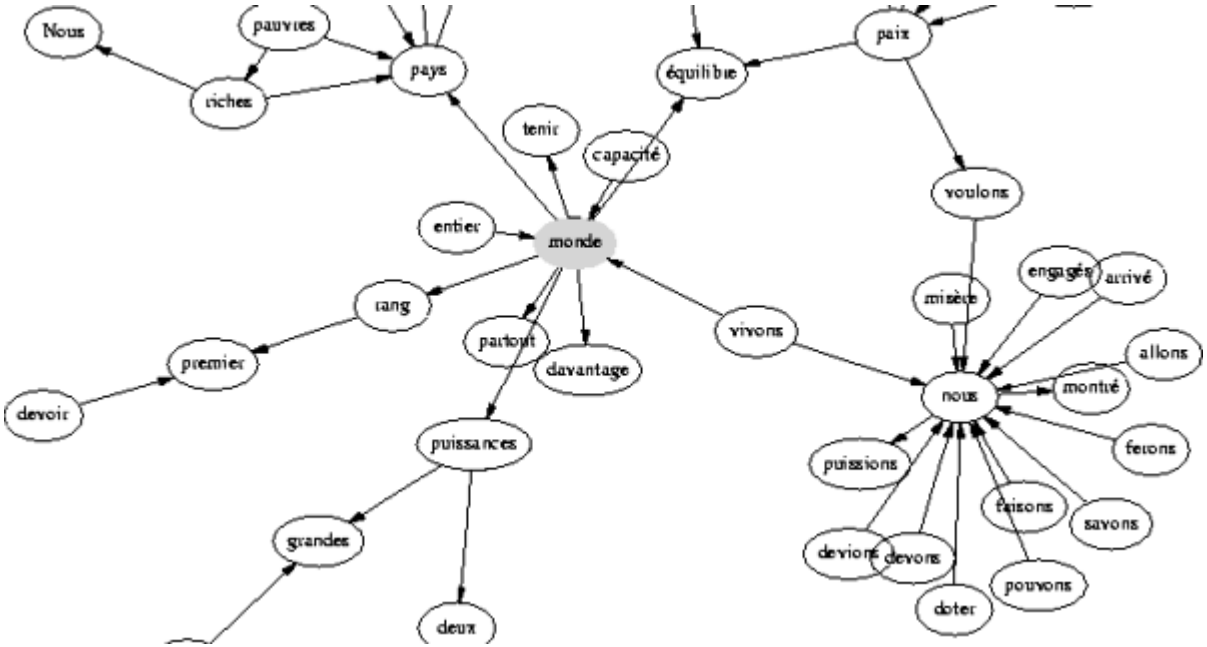


Peyrat-Guillard, *Analyse du discours syndical sur l'entreprise*, JADT'08

Extraire l'information sémantique d'un texte

- analyse arborée
- graphe de coocurrence
- graphe sémantique
- lexicogramme récursif
- désambiguïsation lexicale
- réseau Phrasenet
- projection géodésique

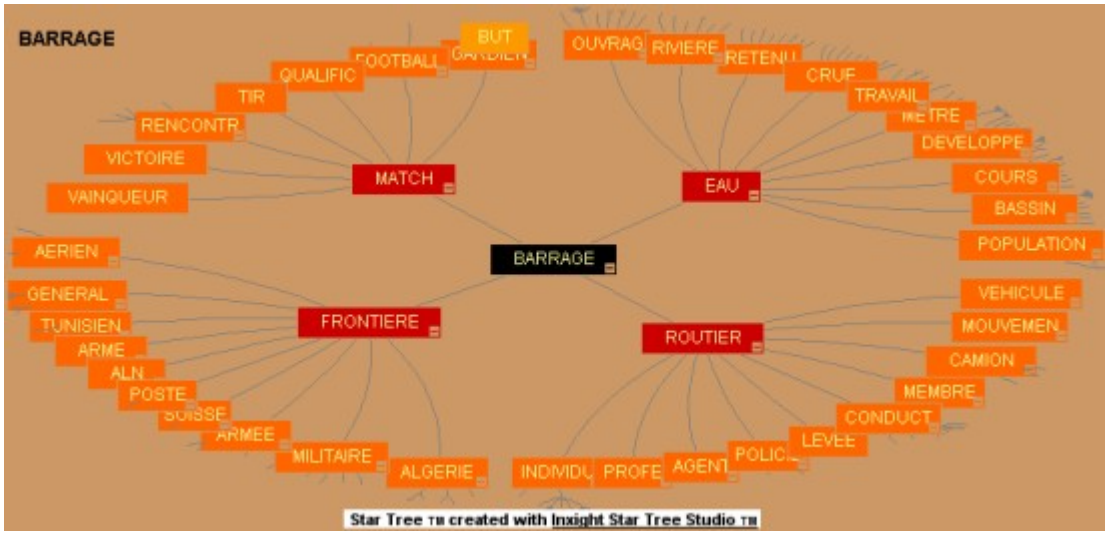
Brunet (Hyperbase)
Brunet (Hyperbase)
Grimmer (Wordmapper)
Martinez (Coocs)
Véronis (Hyperlex)
Viegas et al. (IBM Many Eyes)
Viprey (Astartex)



Leblanc, Martinez
L'analyse contrastive des réseaux de cooccurrence
JADT 2006.

Extraire l'information sémantique d'un texte

- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)



Désambiguïsation du mot “barrage”.

Véronis, *HyperLex: Lexical Cartography for Information Retrieval*, 2004

Extraire l'information sémantique d'un texte

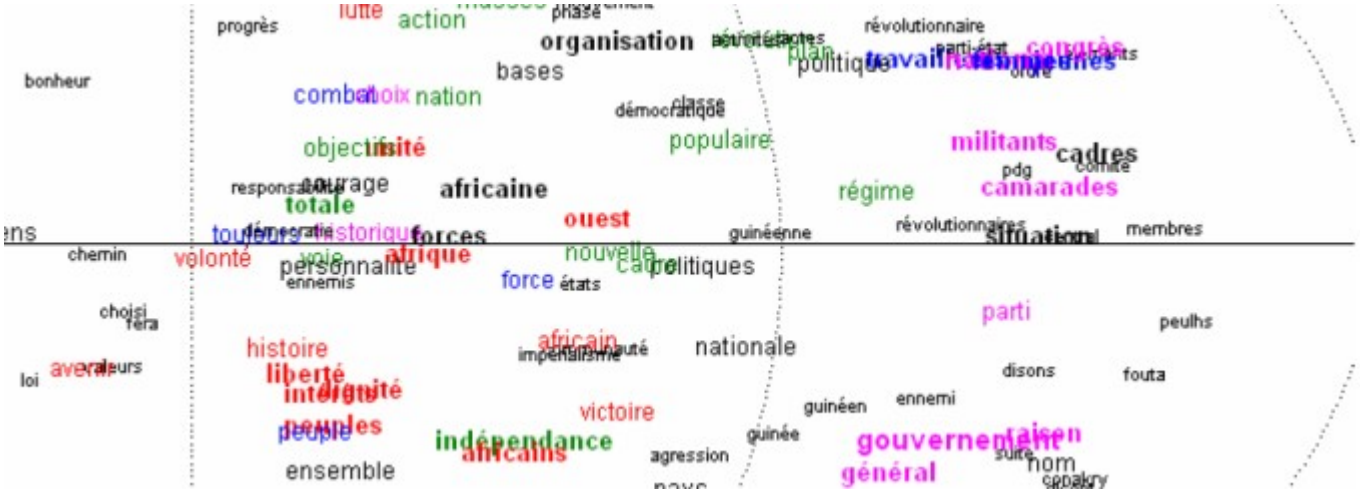
- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)

Visualisation PhraseNet
de paroles des Beatles
créé avec Many Eyes (IBM)
http://many-eyes.com

http://visualthinkmap.ning.com/photo/phrasenet_beatles-many-eyes

Extraire l'information sémantique d'un texte

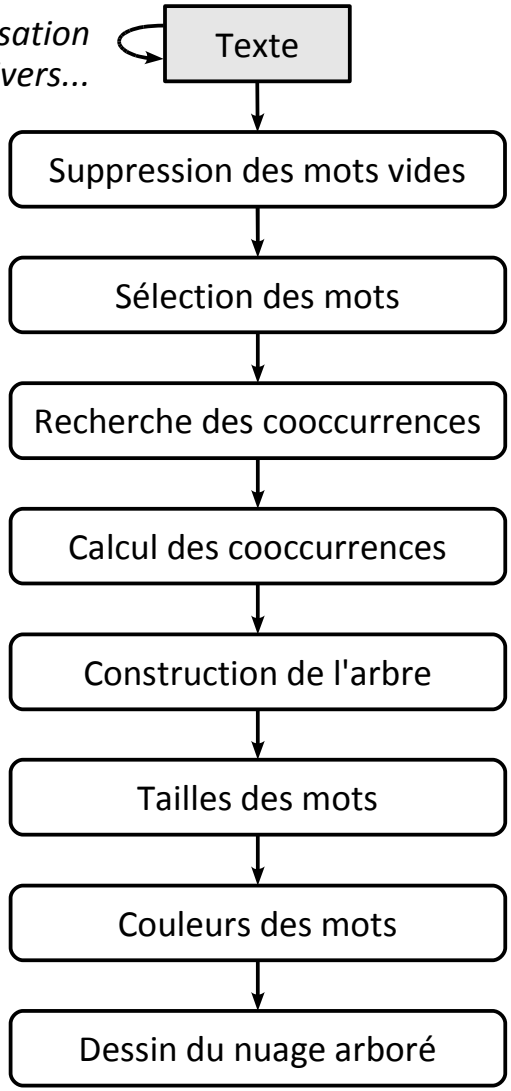
- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)

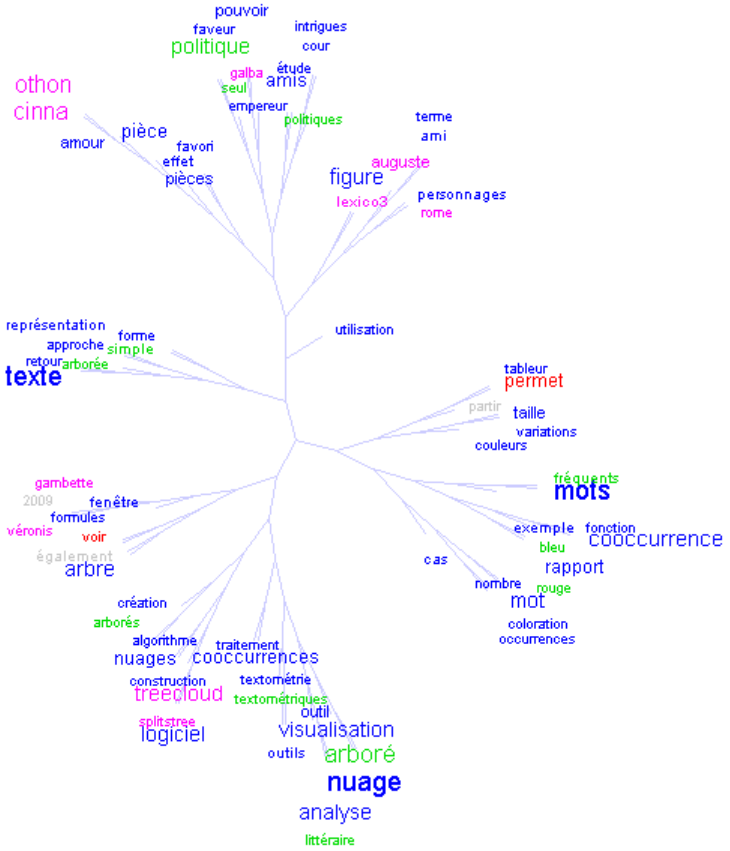


Barry, Viprey, *Approche comparative des résultats d'exploration textuelle des discours de deux leaders africains Keita et Touré*, JADT'08

Construction des nuages arborés

Concordance d'un mot, lemmatisation ou remplacements divers...

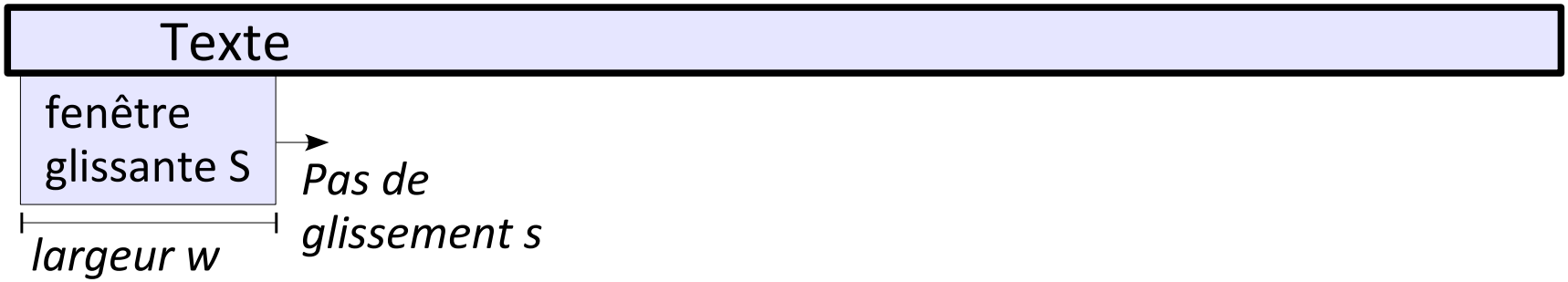




Gambette & Véronis, IFCS 2009
Amstutz & Gambette, JADT 2010

Construction de la matrice de cooccurrences

Formules de distance sémantique à base de cooccurrence



matrices de cooccurrence
O_{11}, O_{12}, O_{21}, O_{22}

	$v \in S$	$v \notin S$
$u \in S$	O_{11}	O_{12}
$u \notin S$	O_{21}	O_{22}

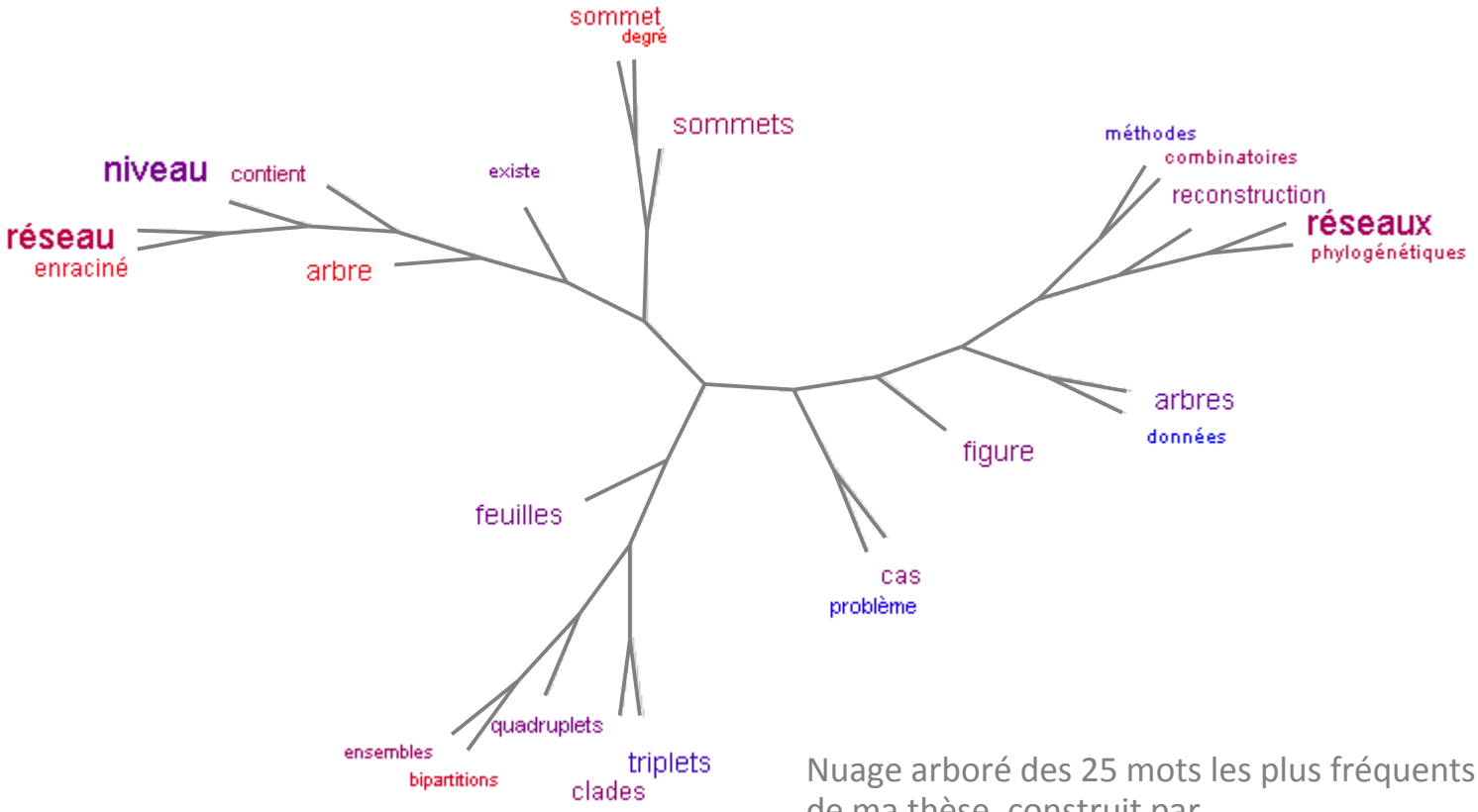
→ matrice de dissimilarité sémantique

chi squared, mutual information, liddel, dice, jaccard, gmean, hyperlex, minimum sensitivity, odds ratio, zscore, log likelihood, poisson-stirling...

Evert, *Statistics of words cooccurrences*, thèse, 2005
Gambette, *User manual for TreeCloud*, 2009

Construction du réseau phylogénétique

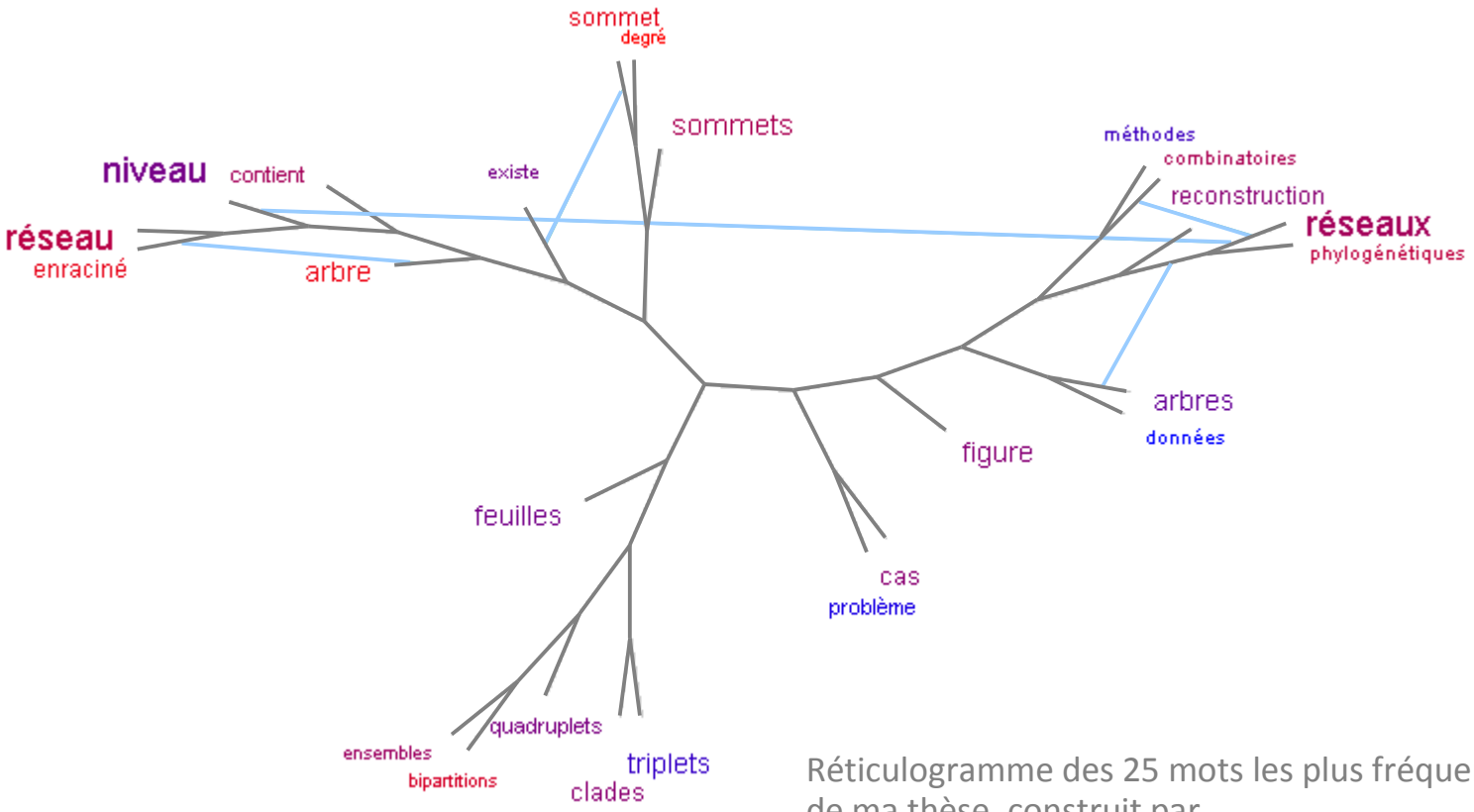
Méthode T-Rex bien adaptée



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

Construction du réseau phylogénétique

Méthode T-Rex bien adaptée



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

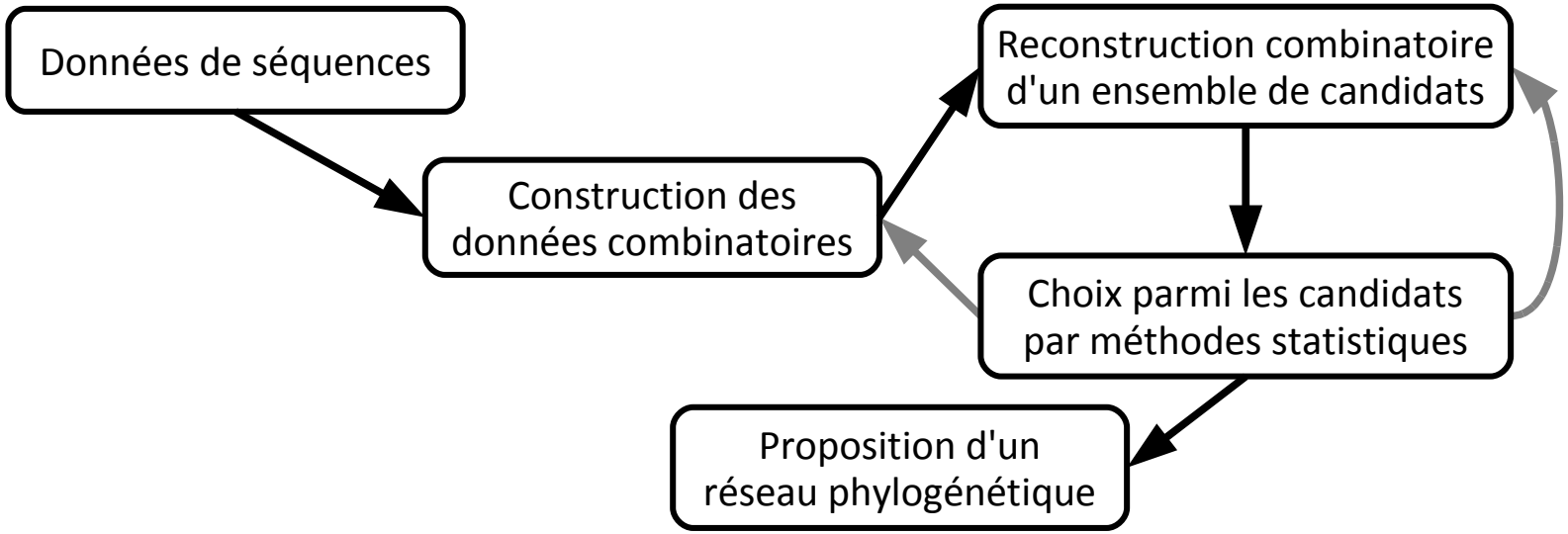
Plan

- Les réseaux phylogénétiques
- Méthodes de reconstruction
- Limites des méthodes combinatoires
- Illustration sur des données biologiques
- Utilisation sur des données textuelles
- Perspectives

Perspectives de recherche

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

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



Autres applications des réseaux phylogénétiques :
- visualiser la polysémie dans les nuages arborés

Merci !

Who is who in Phylogenetic Networks

http://atgc.lirmm.fr/phylnet

Dendroscope

http://www.dendroscope.org

TreeCloud

http://www.treecloud.org

SplitsTree

http://www.splitstree.org