Réunion Projet PhylAriane
Montpellier – 08/12/2009

Réseaux désorientés et arbres dans les nuages

Philippe Gambette

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

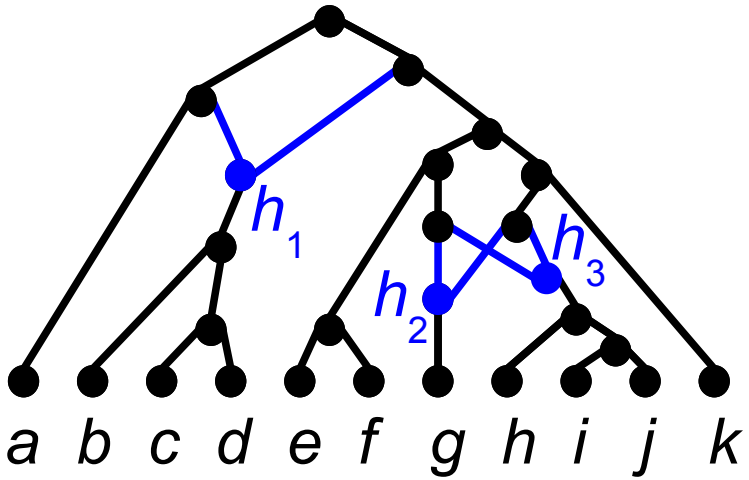
LIRMM

cnrs
dépasser les frontières

Agence Nationale de la Recherche
ANR

Réseau à une couche de réticulation

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

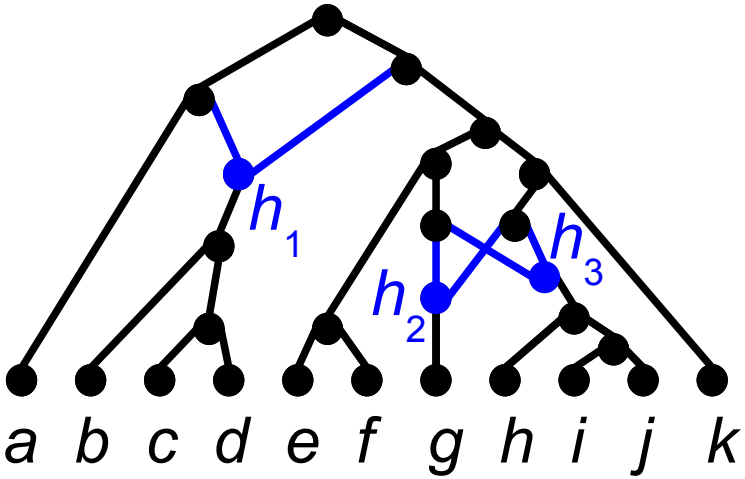


réseau à une couche de réticulation.

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

Réseau à une couche de réticulation

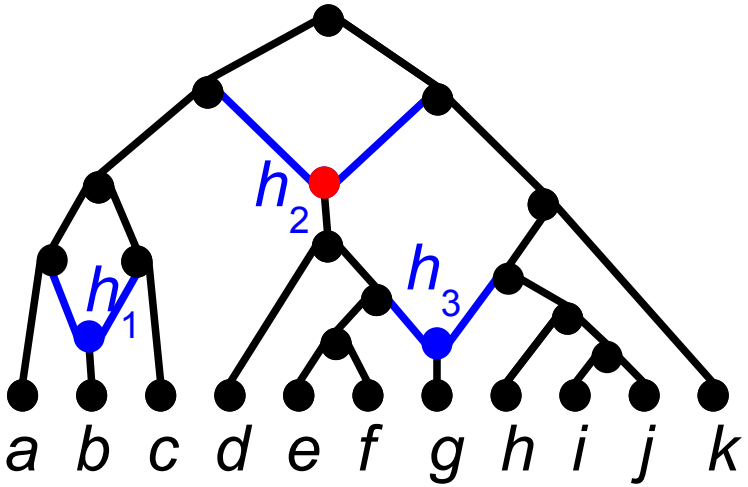
Algorithmes rapides pour des réseaux à **structure proche d'un arbre**.



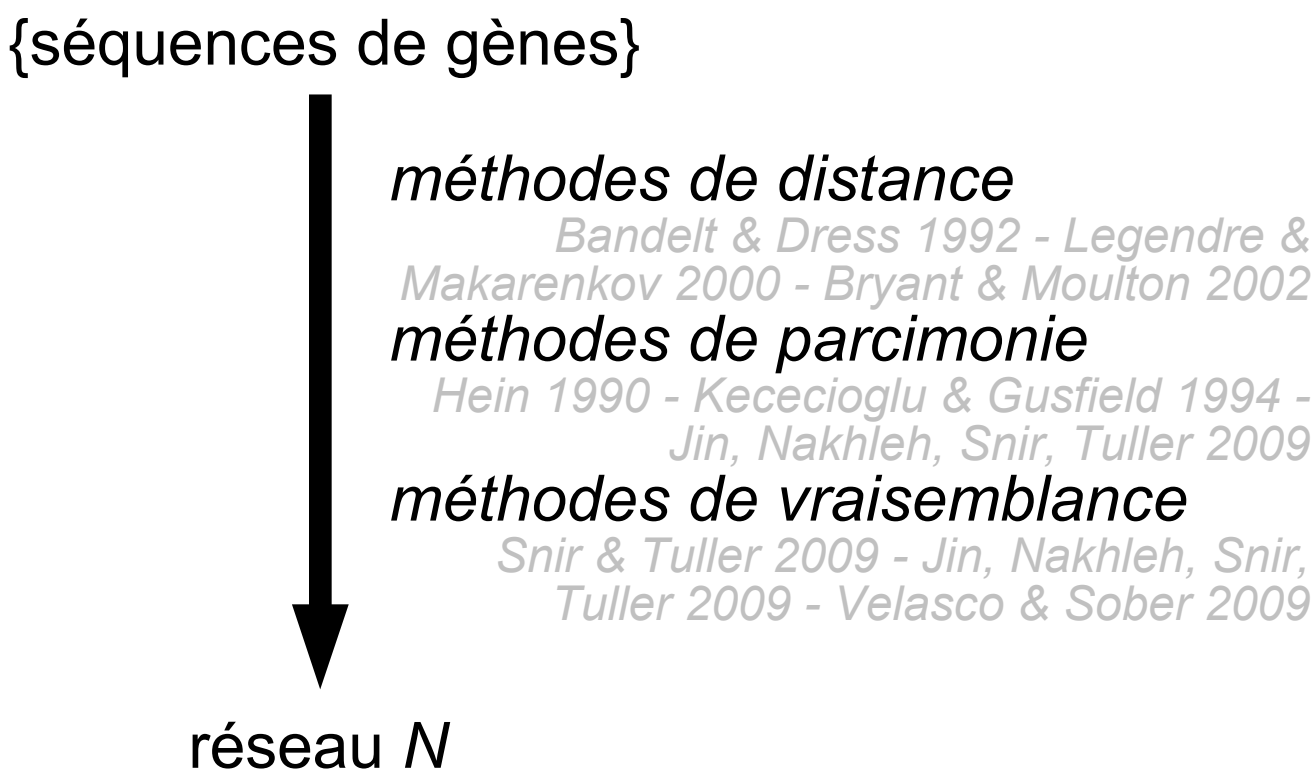
réseau à une couche de réticulation.

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

réseau à deux couches de réticulation.



Reconstruction de réseaux



Reconstruction de réseaux

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

{séquences de gènes}

méthodes de distance

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

méthodes de parcimonie

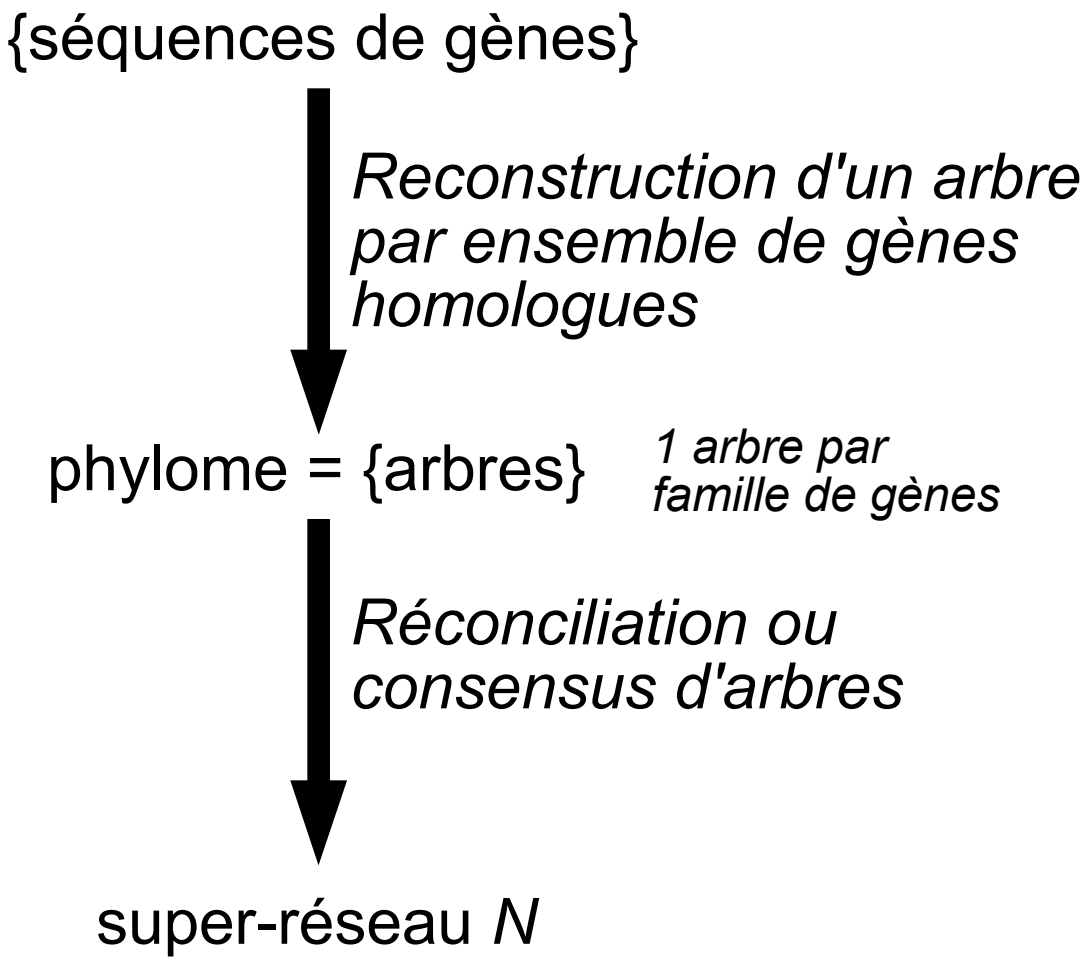
Hein 1990 - Kececioglu & Gusfield 1994 - Jin, Nakhleh, Snir, Tuller 2009

méthodes de vraisemblance

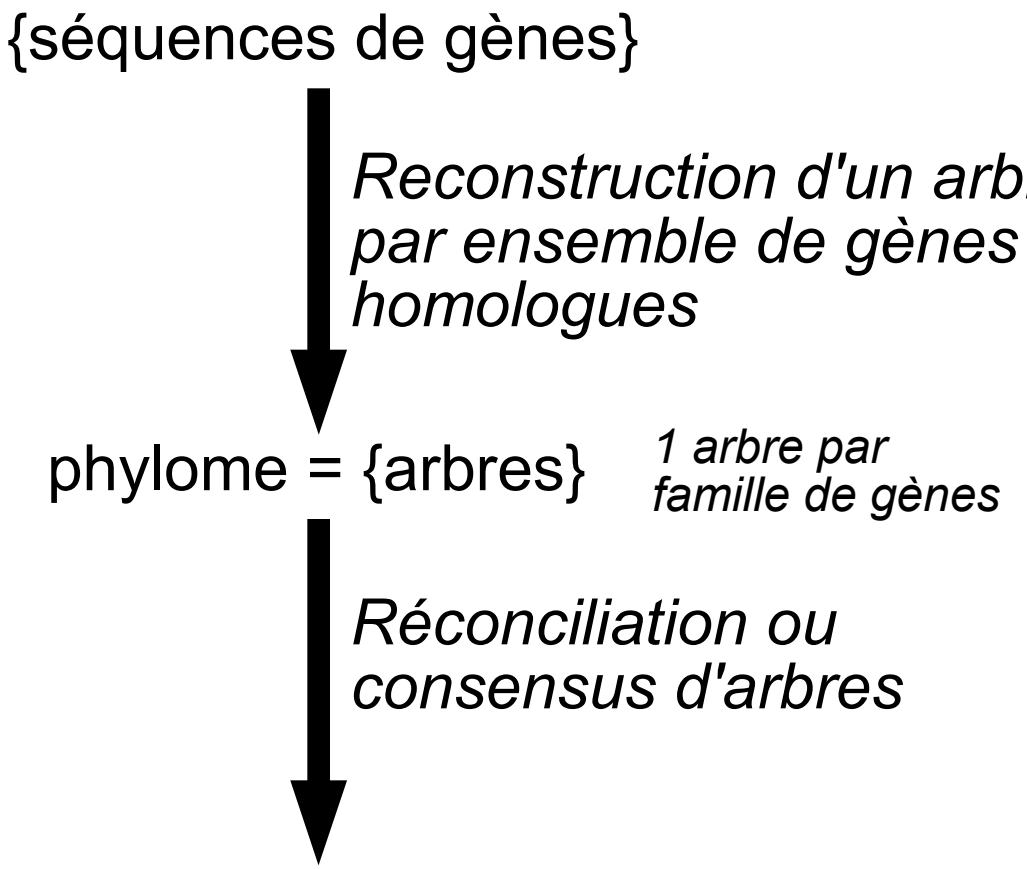
Snir & Tuller 2009 - Jin, Nakhleh, Snir, Tuller 2009 - Velasco & Sober 2009

réseau *N*

Reconstruction combinatoire de réseaux



Reconstruction combinatoire de réseaux



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

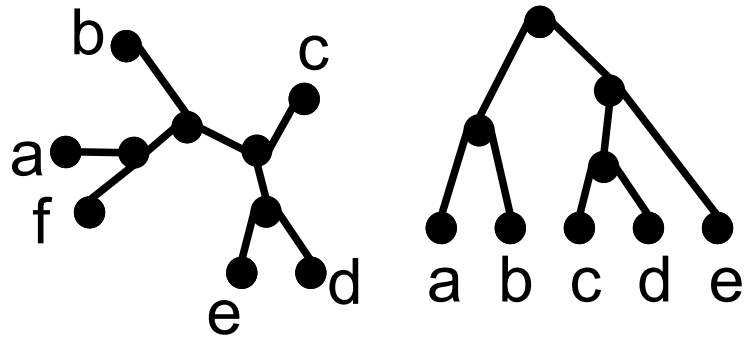
Triplets/quadruplets, splits/clades

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

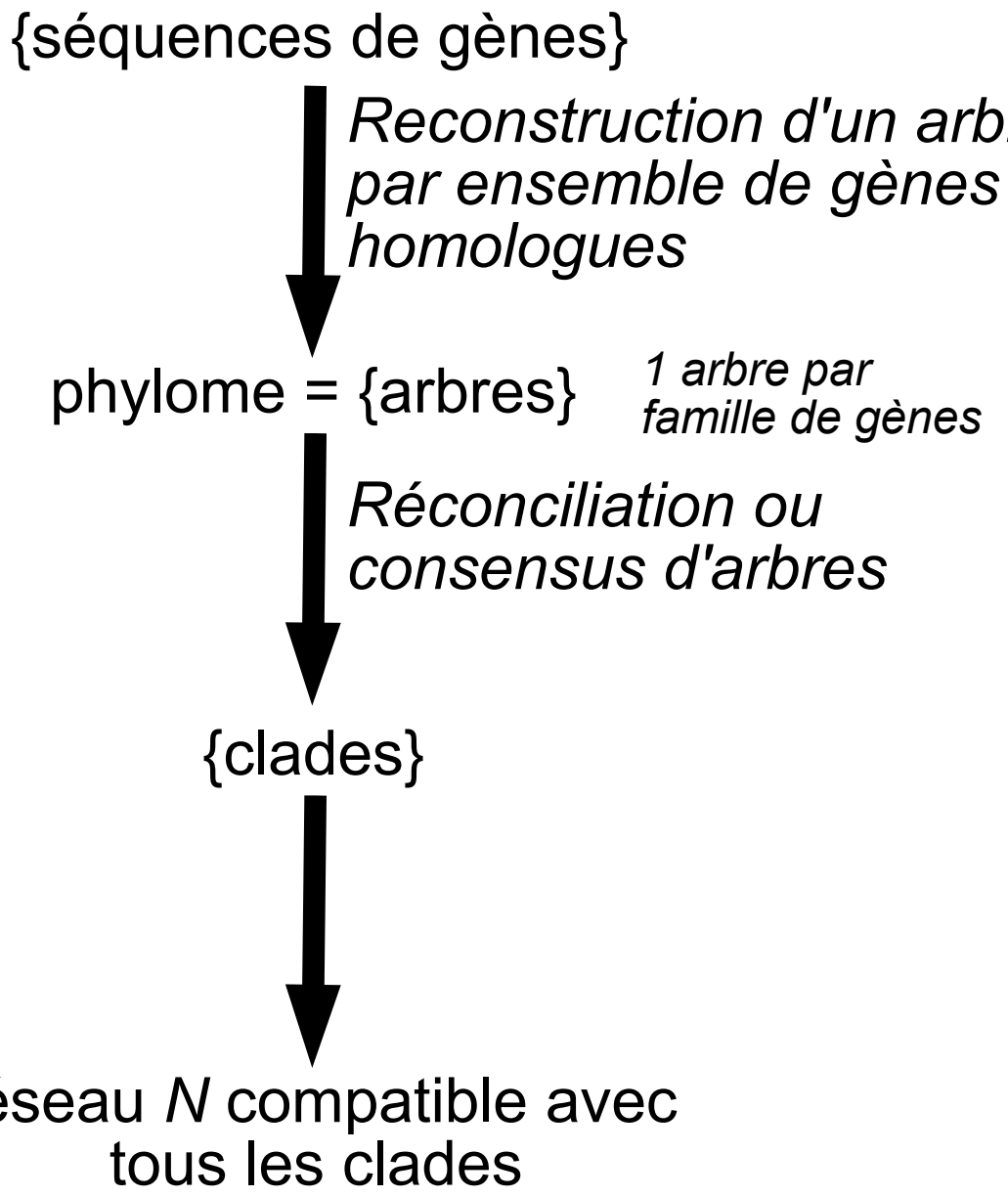
Idée :
reconstuire un réseau contenant les :

triplets
quadruplets
clades
splits

des arbres en entrée ?



Reconstruction combinatoire de réseaux



Reconstruction depuis les clades souples

{arbres}

{clades}

N réseau "galled network"

Consensus de clades souples :
Dendroscope

(Huson, Dezulian, Franz, Rausch, Richter & Rupp, 2007)

Méthode exacte rapide de reconstruction de **réseaux à 1 couche de réticulation** à partir de **clades souples** ("softwired clusters")

(Huson, Rupp, Berry, Gambette & Paul, 2009)

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

Reconstruction depuis les clades souples

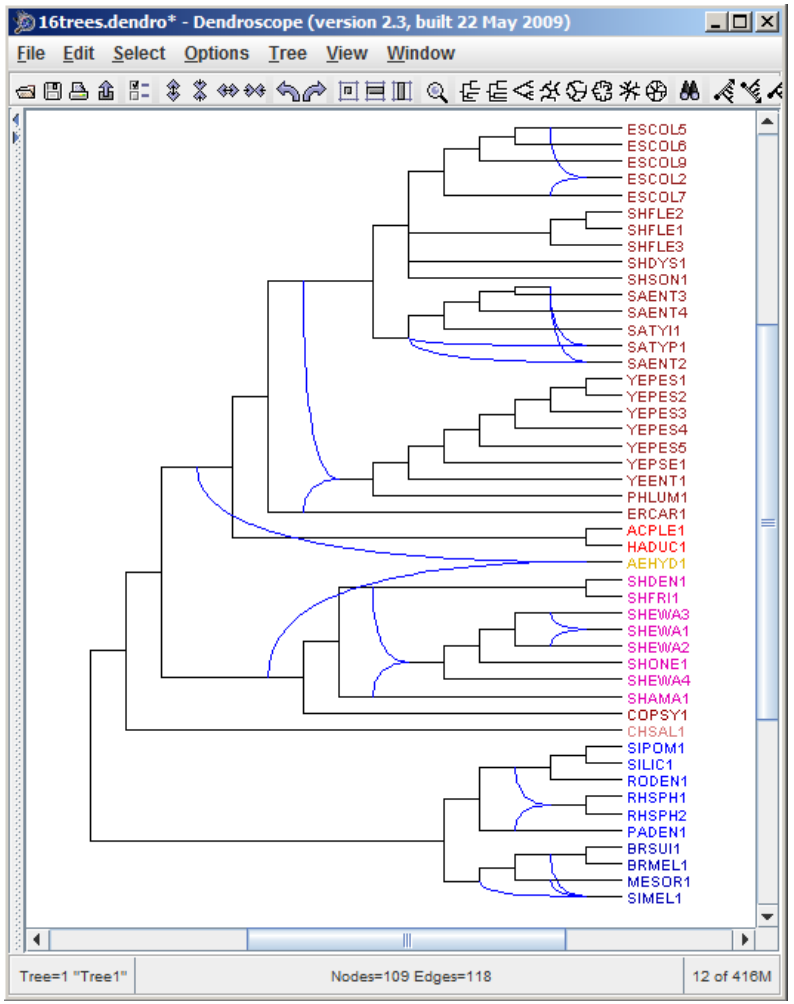
{arbres}

{clades}

N réseau "galled network"

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

Dendroscope



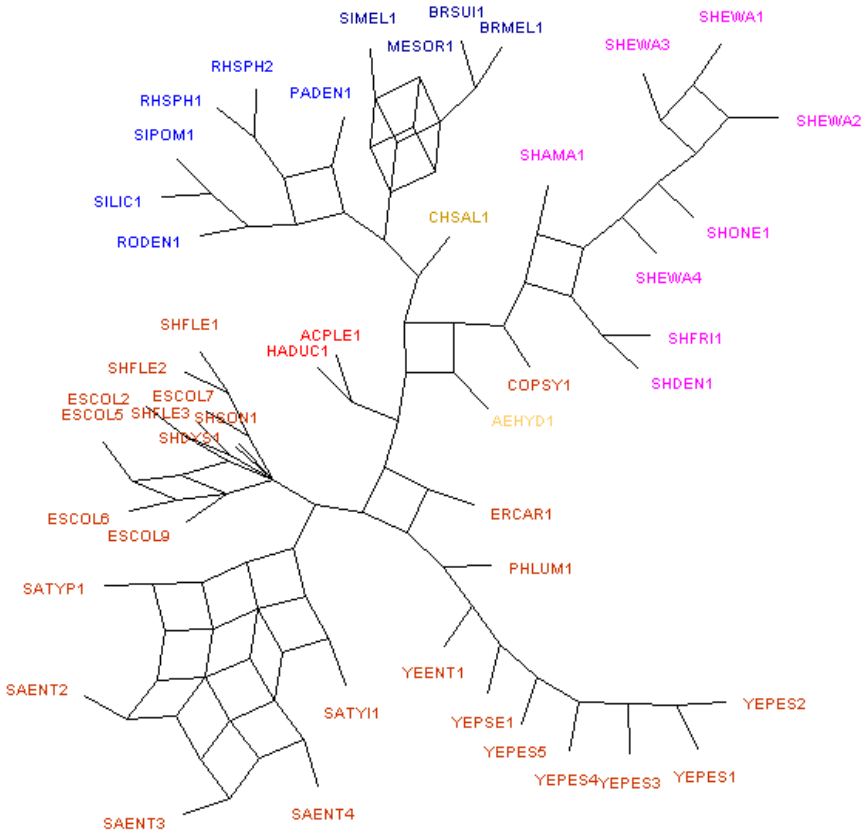
Enterobacteriales
Pasteurellales
Aeromonadales
Alteromonadales
Oceanospirillales
Rhodobacterales
Rhizobiales

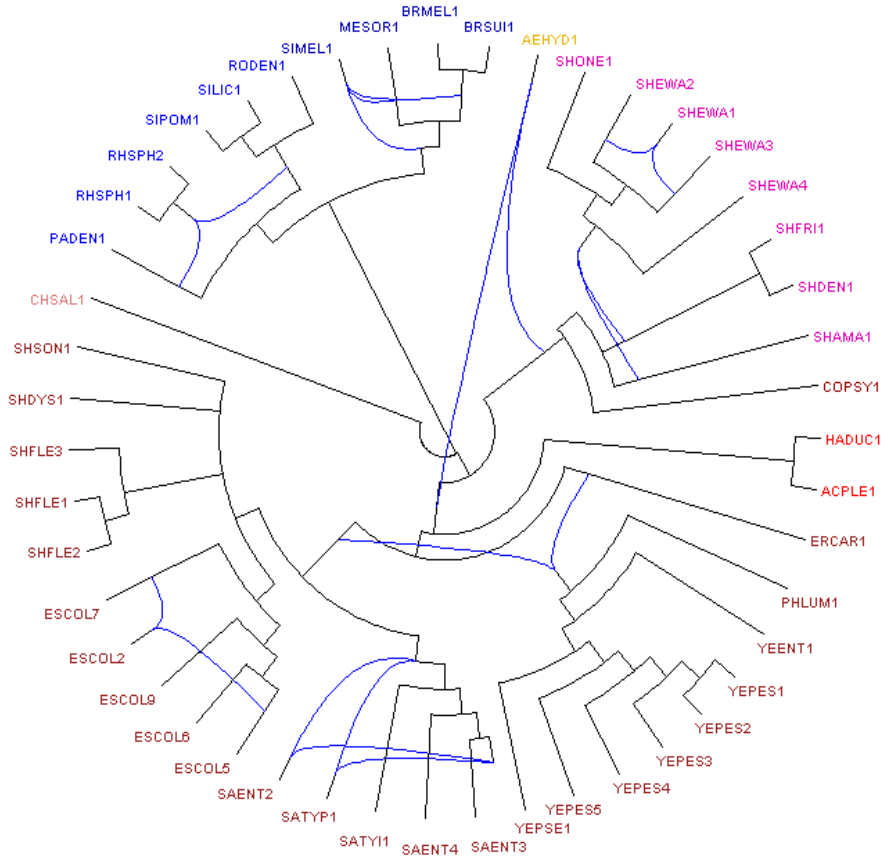
Reconstruction depuis les clades souples

{arbres}

SplitsTree **Dendroscope**

{splits} Filtre : présence dans 20% des arbres {clades}

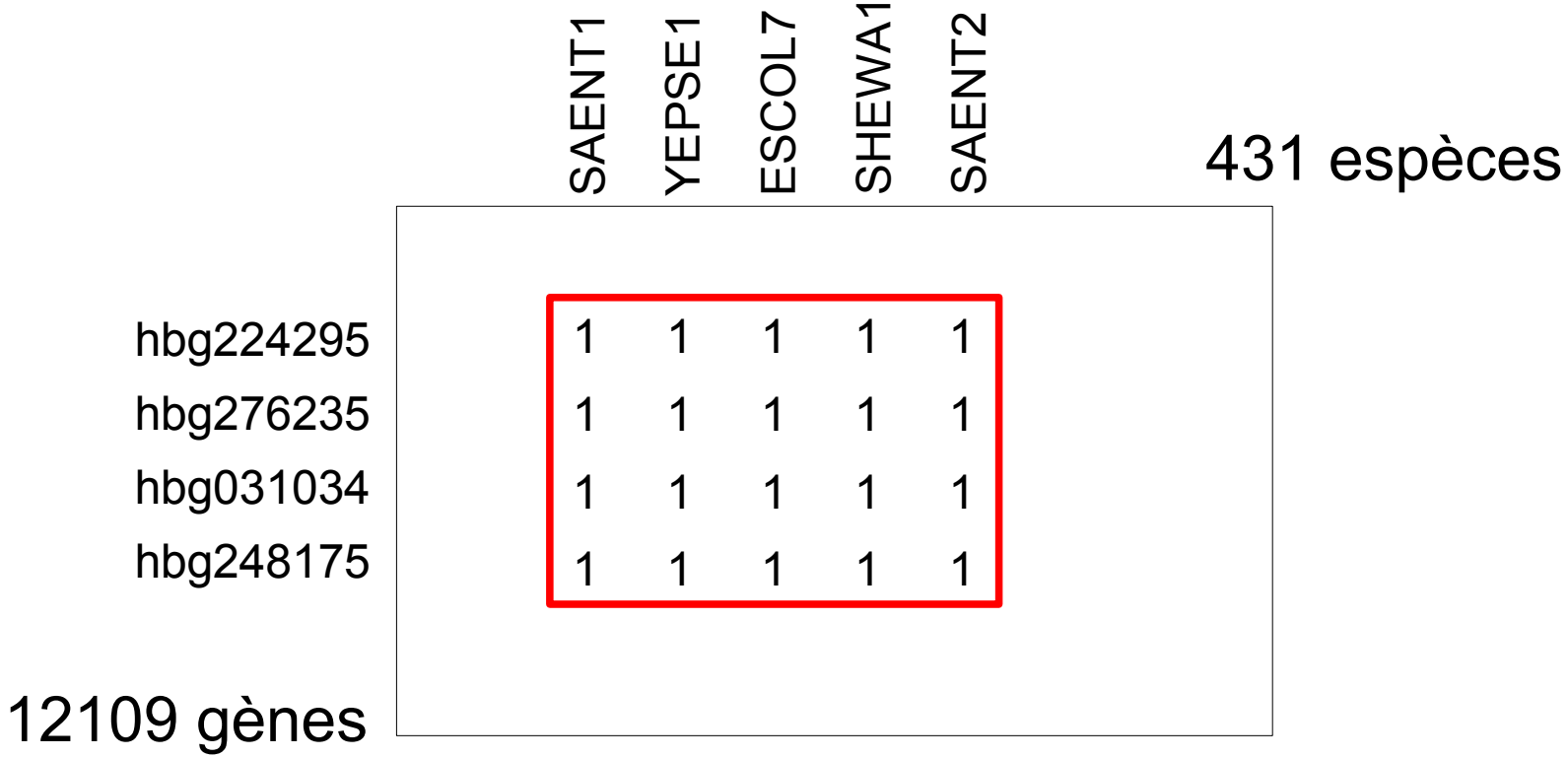




Données de la base HOGENOM, projet ANR PhylAriane

Quelles données en entrée ?

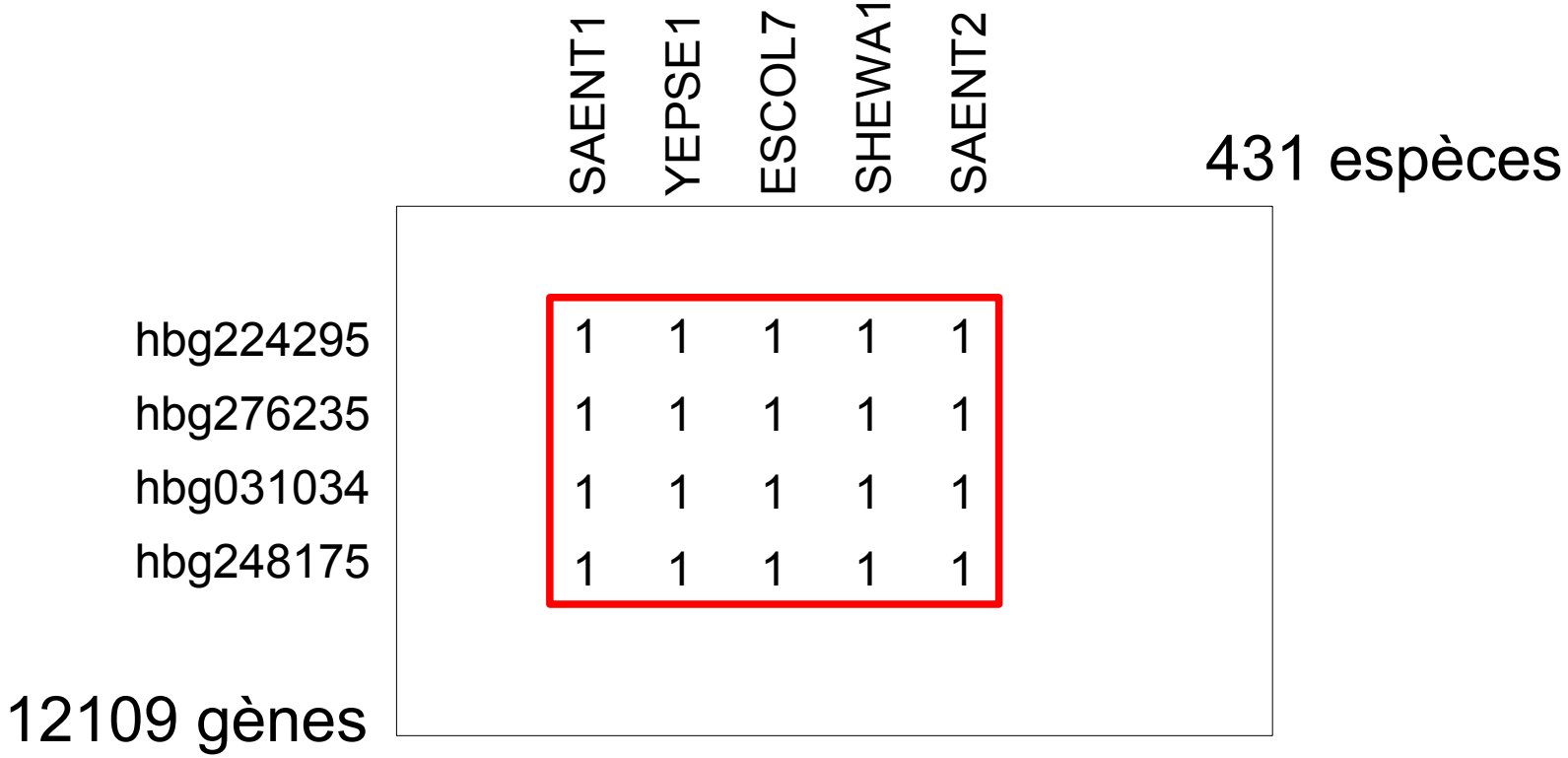
Contrainte : les méthodes de réseaux fonctionnent mieux
- pour des arbres de gènes **sans paralogues** (une seule copie du gène par espèce)
- pour des arbres de gènes portant **sur le même ensemble de taxons**



Approche “désorientée”

Trouver des données qui respectent ces contraintes pour tester la méthode.

➡ Trouver de gros blocs de 1

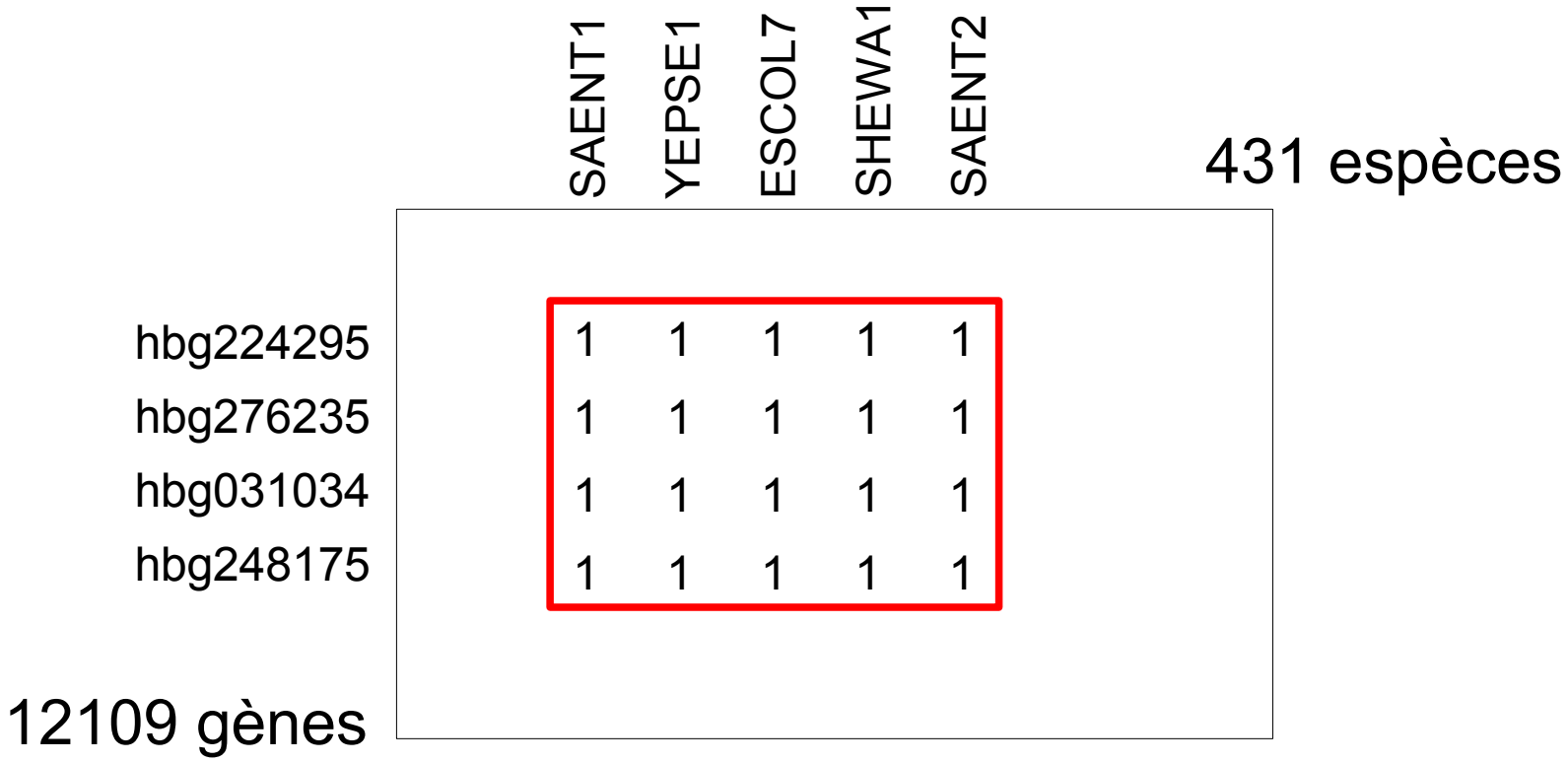


Approche “désorientée”

Trouver des données qui respectent ces contraintes pour tester la méthode.

➡ Trouver de gros blocs de 1

➡ Problème de la recherche du “core”
Méthode classique ? Heuristiques ?

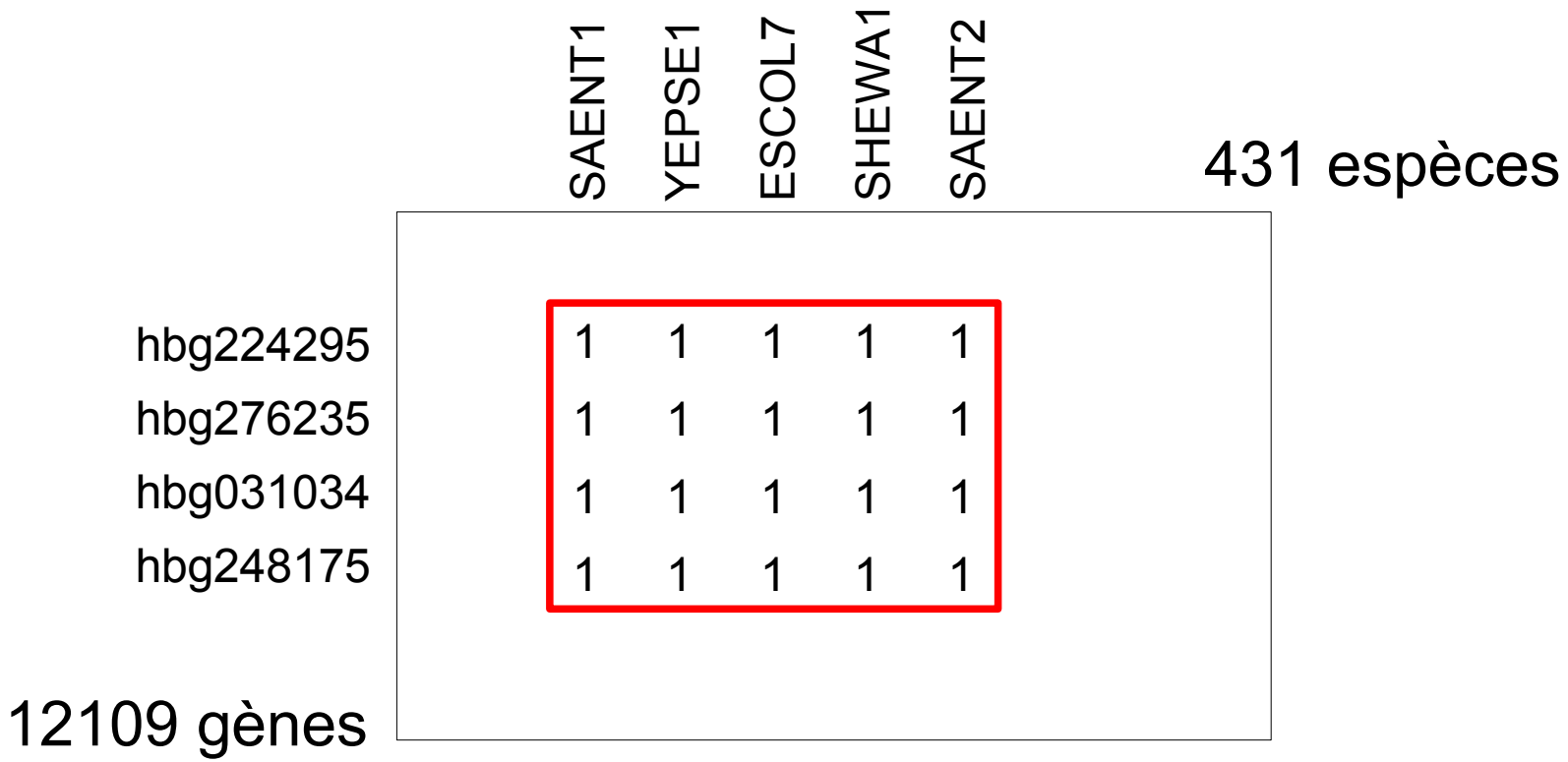


Approche ciblée

Trouver des données qui respectent ces contraintes pour tester la méthode, qui contiennent certains **gènes**, ou certaines **espèces d'intérêt**

➡ Trouver de gros blocs de 1 qui contiennent telles lignes ou telles colonnes

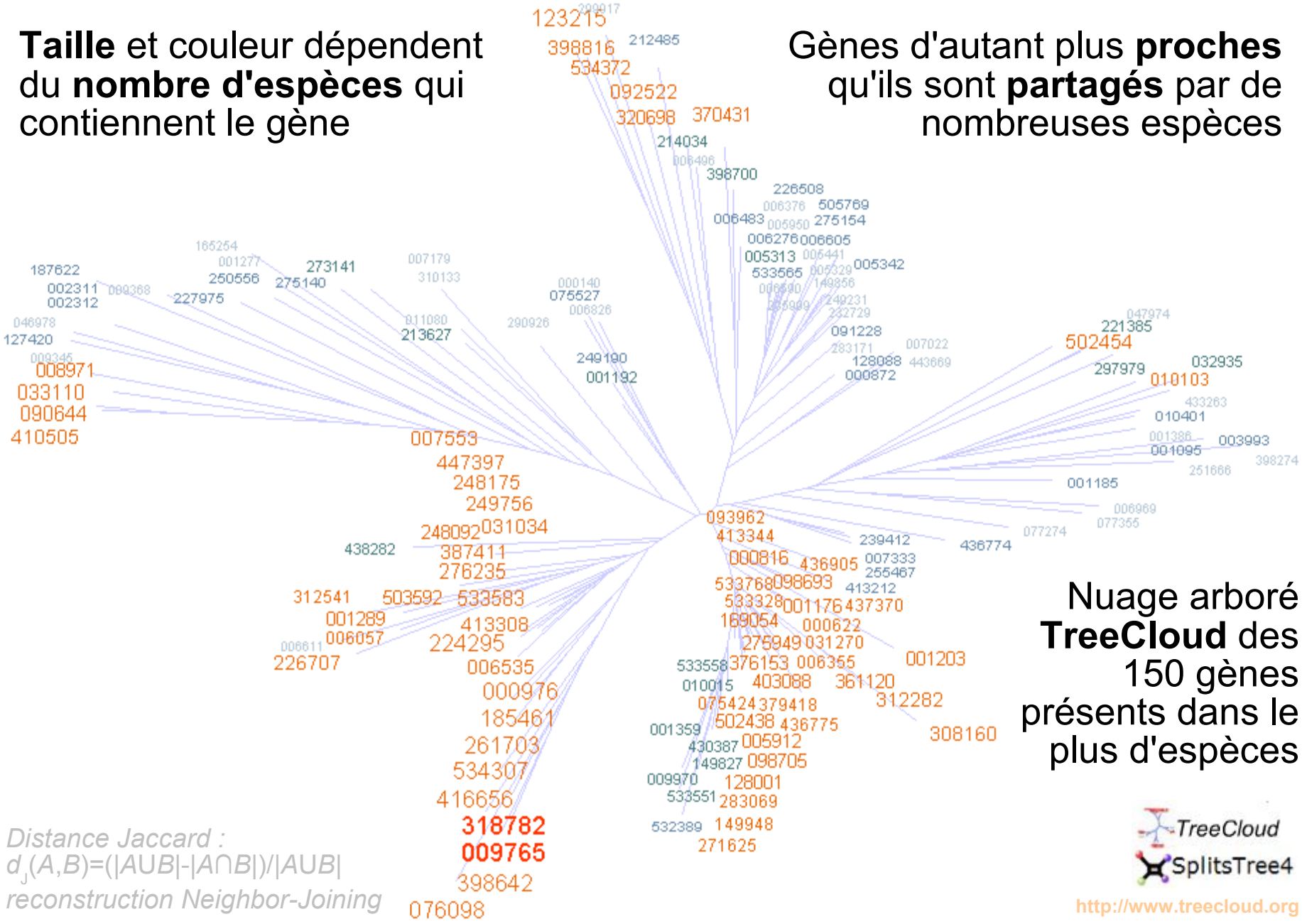
➡ Outil d'exploration de la base HOGENOM ?



Le nuage arboré pour trouver les “blocs”

Taille et couleur dépendent du **nombre d'espèces** qui contiennent le gène

Gènes d'autant plus **proches** qu'ils sont **partagés** par de nombreuses espèces



Nuage arboré **TreeCloud** des 150 gènes présents dans le plus d'espèces

Distance Jaccard :
$d_J(A,B)=(|AUB|-|A\cap B|)/|AUB|$
reconstruction Neighbor-Joining

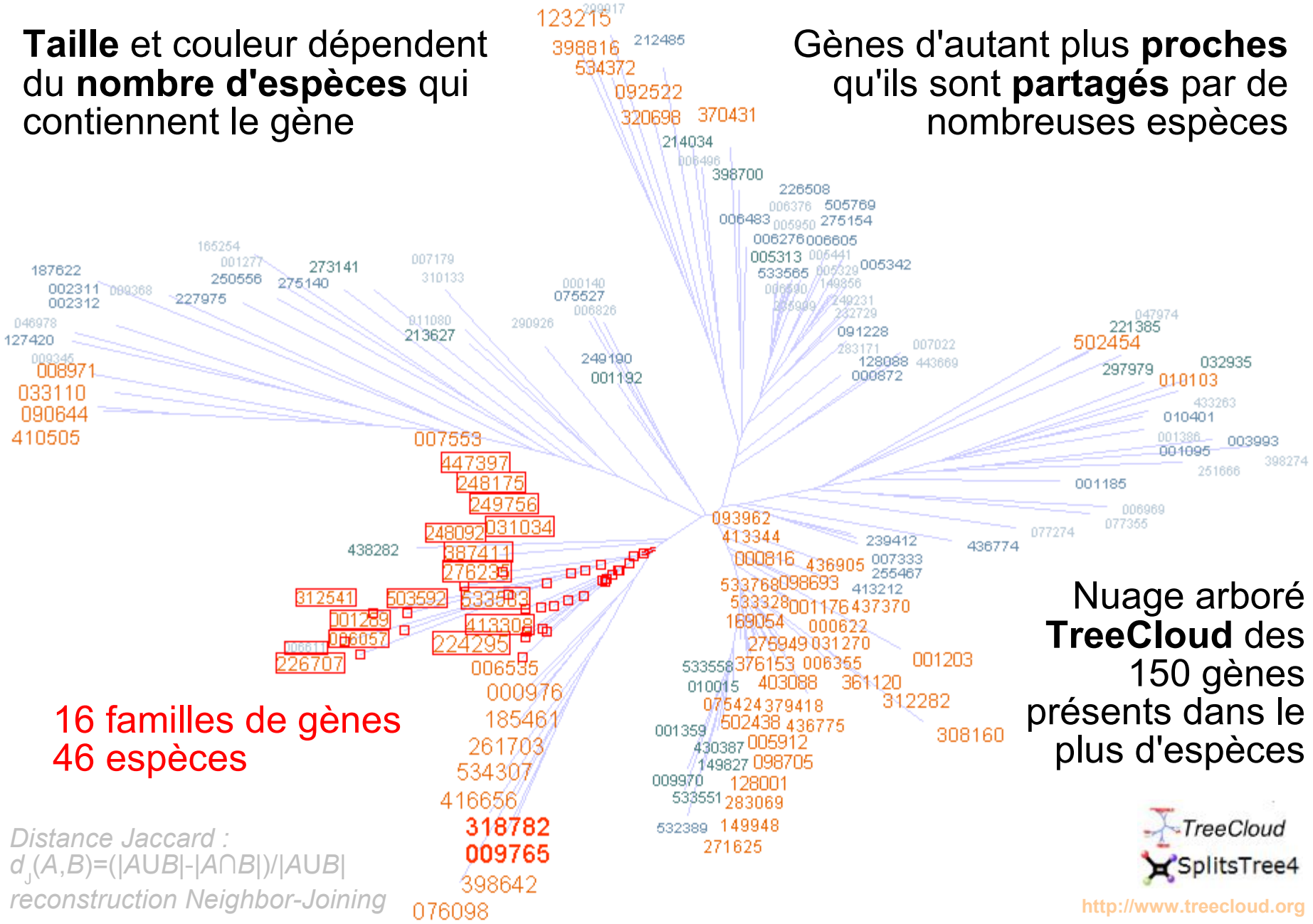
TreeCloud
SplitsTree4

http://www.treecloud.org

Le nuage arboré pour trouver les “blocs”

Taille et couleur dépendent du **nombre d'espèces** qui contiennent le gène

Gènes d'autant plus **proches** qu'ils sont **partagés** par de nombreuses espèces



16 familles de gènes
46 espèces

Nuage arboré **TreeCloud** des 150 gènes présents dans le plus d'espèces

Distance Jaccard :
$d_J(A,B)=(|AUB|-|A\cap B|)/|AUB|$
reconstruction Neighbor-Joining

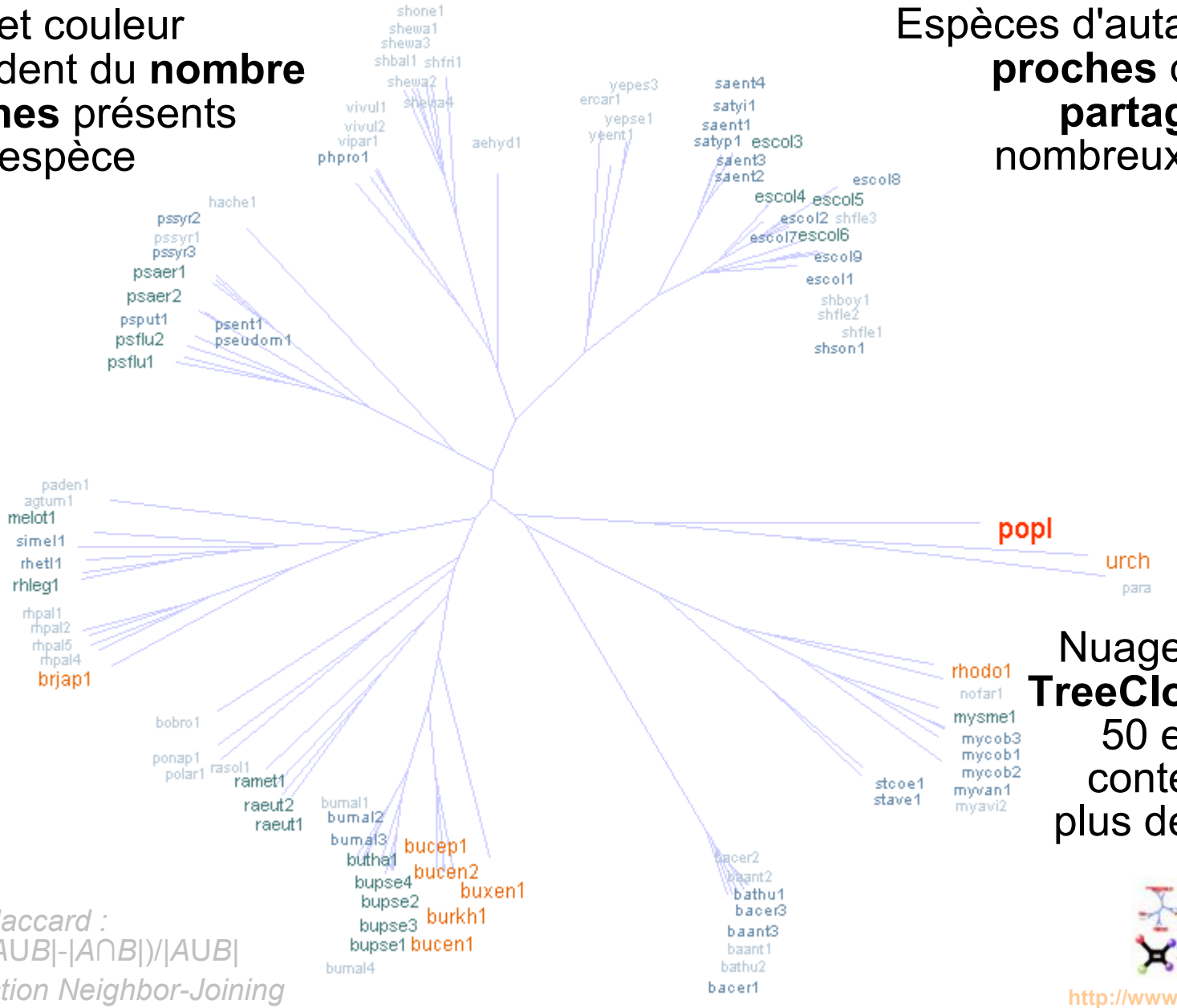
TreeCloud
SplitsTree4

http://www.treecloud.org

Le nuage arboré... inversé

Taille et couleur dépendent du **nombre de gènes** présents dans l'espèce

Espèces d'autant plus **proches** qu'elles **partagent** de nombreux gènes



Nuage arboré **TreeCloud** des 50 espèces contenant le plus de gènes

Distance Jaccard :
$d_J(A,B)=(|A \cup B|-|A \cap B|)/|A \cup B|$
reconstruction Neighbor-Joining

TreeCloud
SplitsTree4

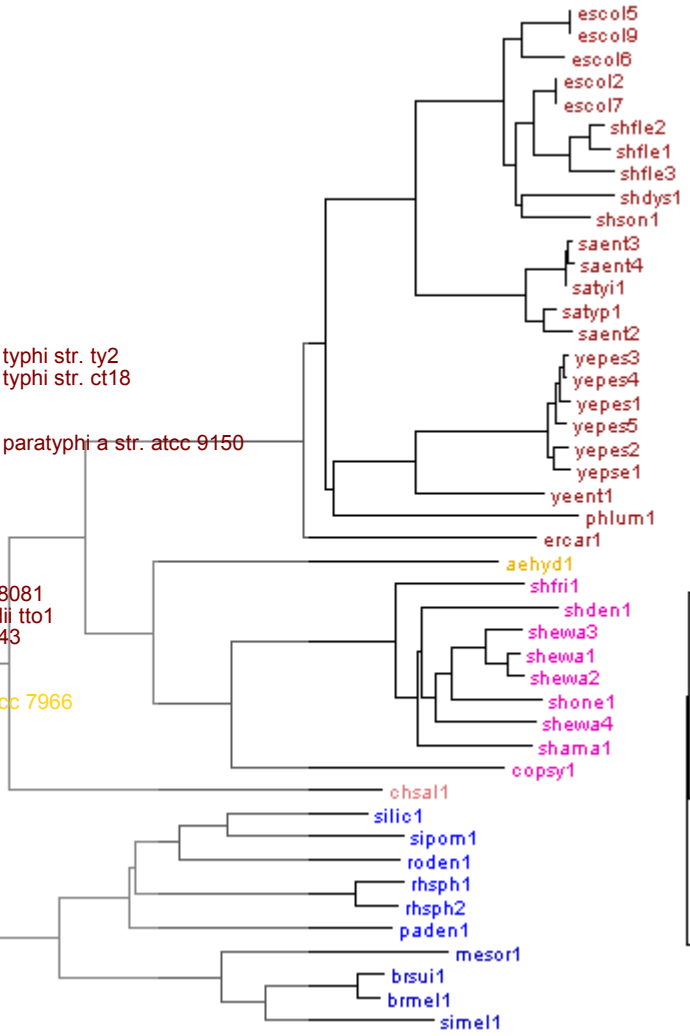
http://www.treecloud.org

Présence/absence de gènes

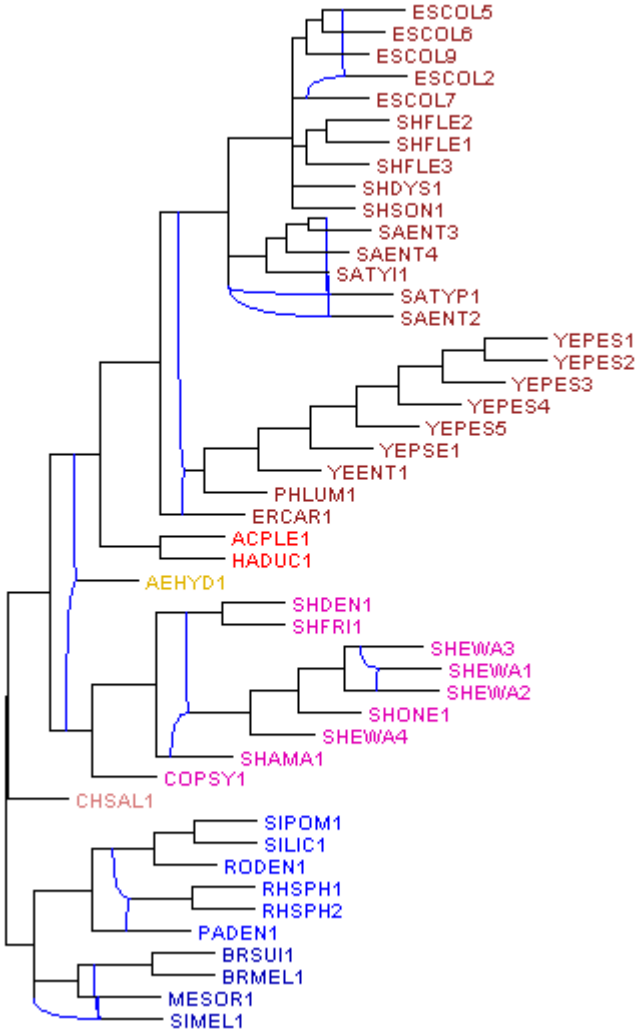
Enterobacteriales
Pasteurellales
Aeromonadales
Alteromonadales
Oceanospirillales
Rhodobacterales
Rhizobiales

escol5 escherichia coli o6
escol6 escherichia coli uti89
escol9 escherichia coli cft073
escol2 escherichia coli k12 k12
escol7 escherichia coli w3110
shfle2 shigella flexneri 2a str. 301
shfle1 shigella flexneri 2a str. 2457t
shfle3 shigella flexneri 5 str. 8401
shdys1 shigella dysenteriae sd197
shson1 shigella sonnei ss046
saent3 salmonella enterica subsp. enterica serovar typhi str. ty2
saent4 salmonella enterica subsp. enterica serovar typhi str. ct18
satyi1 salmonella typhi
satyp1 salmonella typhimurium lt2
saent2 salmonella enterica subsp. enterica serovar paratyphi a str. atcc 9150
yepes1 yersinia pestis antiqua
yepes2 yersinia pestis biovar microtus str. 91001
yepes3 yersinia pestis co92
yepes4 yersinia pestis kim
yepes5 yersinia pestis nepal516
yepse1 yersinia pseudotuberculosis ip 32953
yeent1 yersinia enterocolitica subsp. enterocolitica 8081
phlum1 photorhabdus luminescens subsp. laumondii tto1
ercar1 erwinia carotovora subsp. atroseptica scri1043
acple1 actinobacillus pleuropneumoniae l20
haduc1 haemophilus ducreyi 35000hp
aehyd1 aeromonas hydrophila subsp. hydrophila atcc 7966
shden1 shewanella denitrificans os217
shfri1 shewanella frigidimarina ncimb 400
shewa3 shewanella sp. ana-3
shewa1 shewanella sp. mr-4
shewa2 shewanella sp. mr-7
shone1 shewanella oneidensis mr-1
shewa4 shewanella sp. w3-18-1
shama1 shewanella amazonensis sb2b
copsy1 colwellia psychrerythraea 34h
chsal1 chromohalobacter salexigens dsm 3043
sipom1 silicibacter pomeroyi dss-3
silic1 silicibacter sp. tm1040
roden1 roseobacter denitrificans och 114
rhsph1 rhodobacter sphaeroides 2.4.1
rhsph2 rhodobacter sphaeroides atcc 17029
paden1 paracoccus denitrificans pd1222
brsui1 brucella suis 1330
brmel1 brucella melitensis 16m
mesor1 mesorhizobium sp. bnc1
simel1 sinorhizobium meliloti 1021

Extrait du nuage arboré



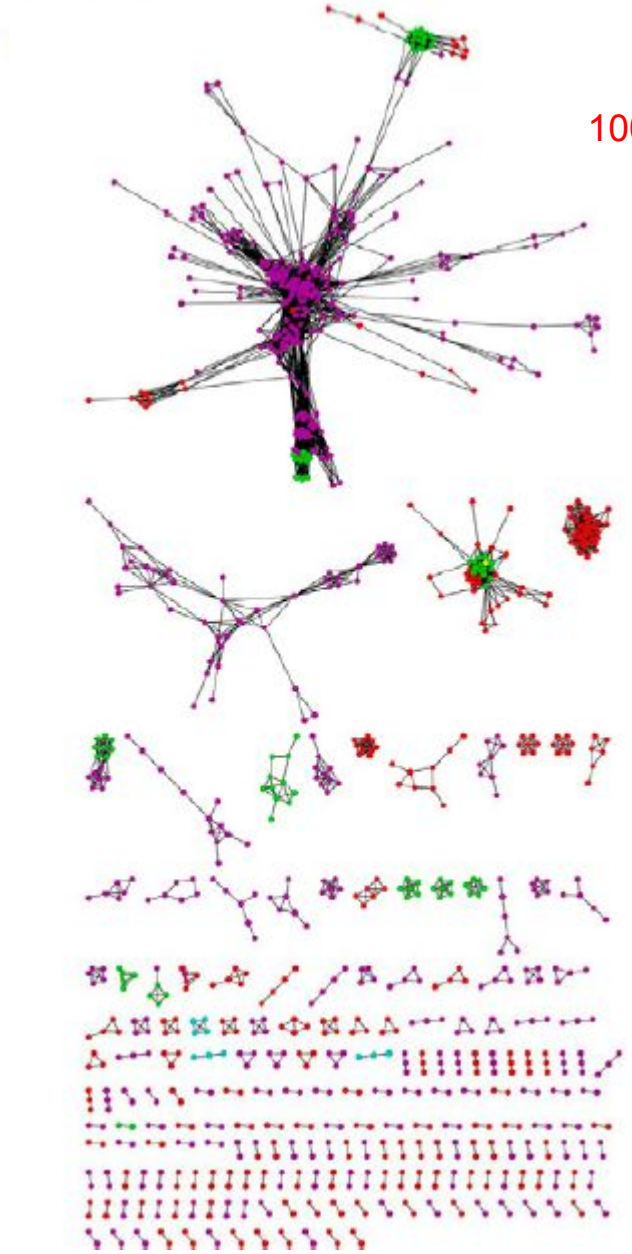
Réseau de clades



Les relations de **présence/absence de gènes** donnent une information phylogénétique !

Lima-Mendez, Van Helden, Toussaint, Leplae. 2008

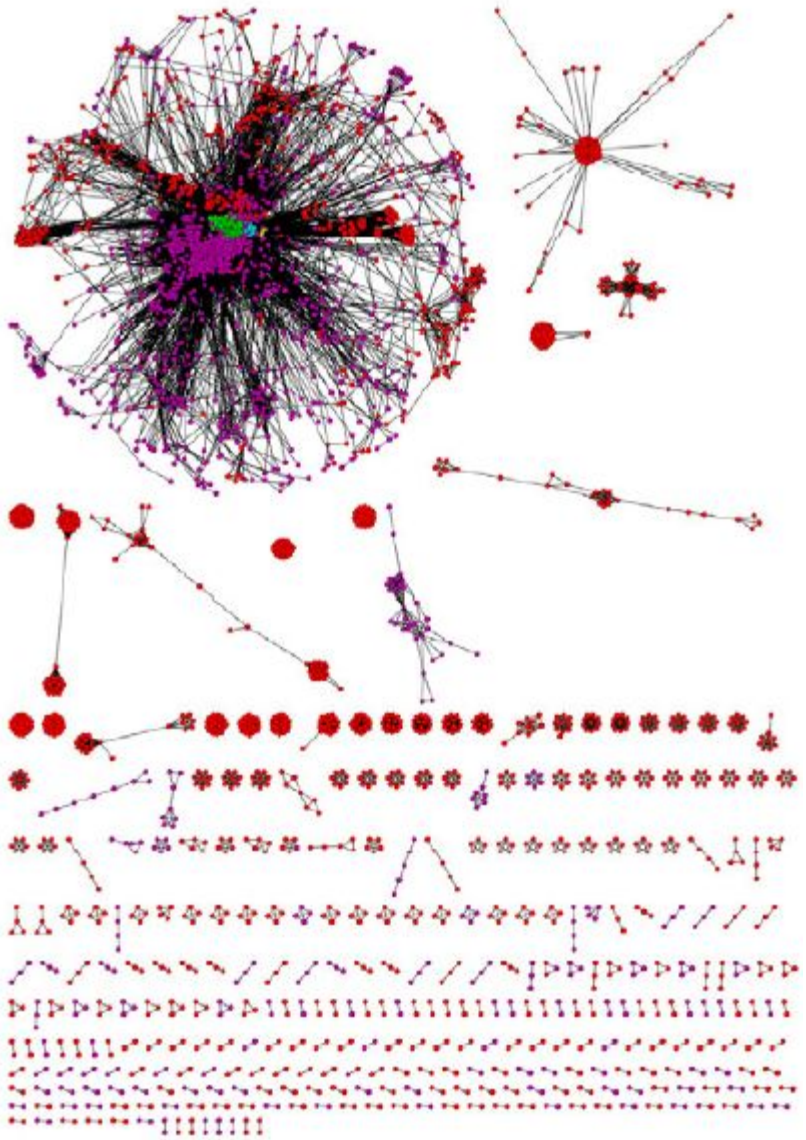
Réseaux de présence-absence



Arête : présence d'un gène avec

100% 20%

de similarité chez les deux espèces



Bacterial chromosomes

Archaeal chromosomes

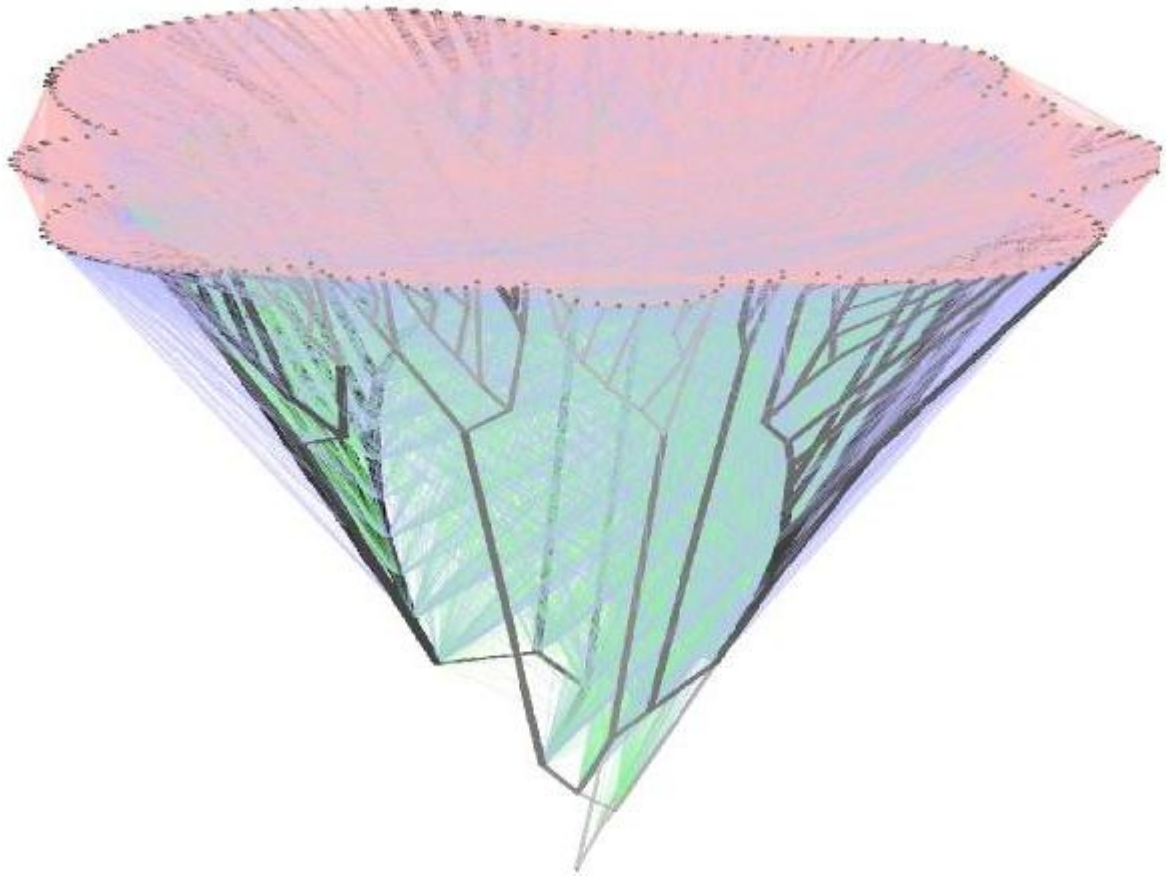
Eukaryotic chromosomes

Plasmids

Viruses

Halary S, Leigh J, Cheaib B, Lopez P, Bapteste E. *Network analyses structure genetic diversity in independent genetic worlds*. PNAS, à paraître

Visualisation des transferts



Dagan T, Artzy-Randrup Y, Martin W. *Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution*. PNAS 105(29), 2008

Approches ciblées

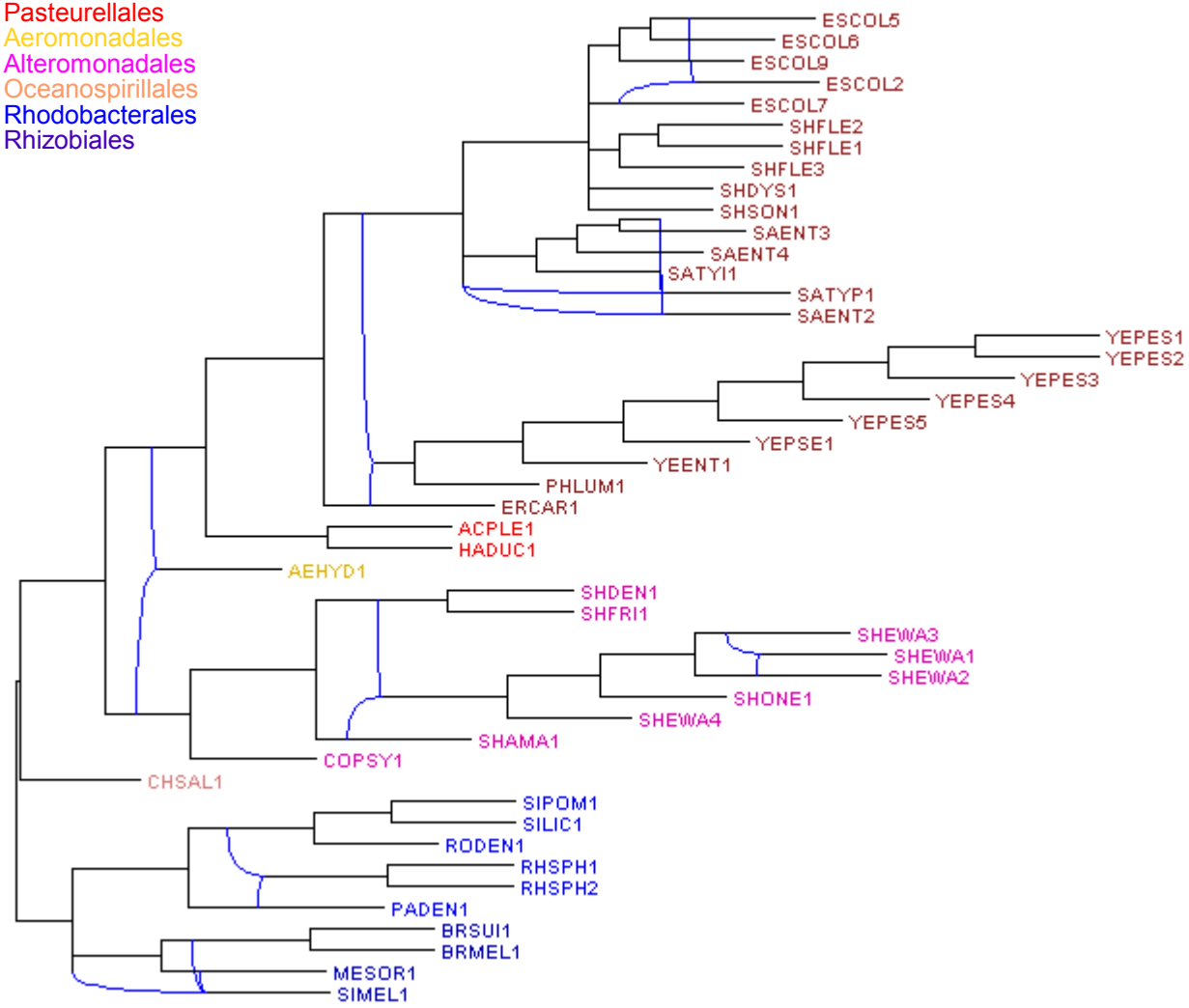
Quelles approches ciblées intéressantes ?

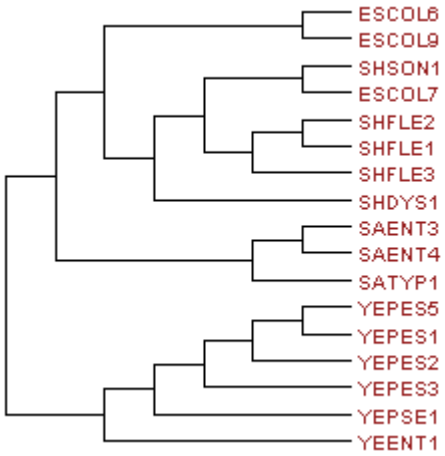
Choix des gènes ayant une fonction proche ?
➡ Récupération des données de fonction GeneOntology ?

Choix d'espèces d'intérêt ?
➡ Gamma-protéobactéries ?

Gamma-protéobactéries

Enterobacteriales
Pasteurellales
Aeromonadales
Alteromonadales
Oceanospirillales
Rhodobacterales
Rhizobiales





Superarbre de 1408 familles de gènes

Haggerty, Martin, Fitzpatrick, McInerney. *Gene and genome trees conflict at many levels*, PTRSB 364, 2009

Des questions ?

Merci pour votre attention !