

09/02/2011 - Rencontres Alphy, Lyon

Practical use of combinatorial methods for phylogenetic network reconstruction

Philippe Gambette



Outline

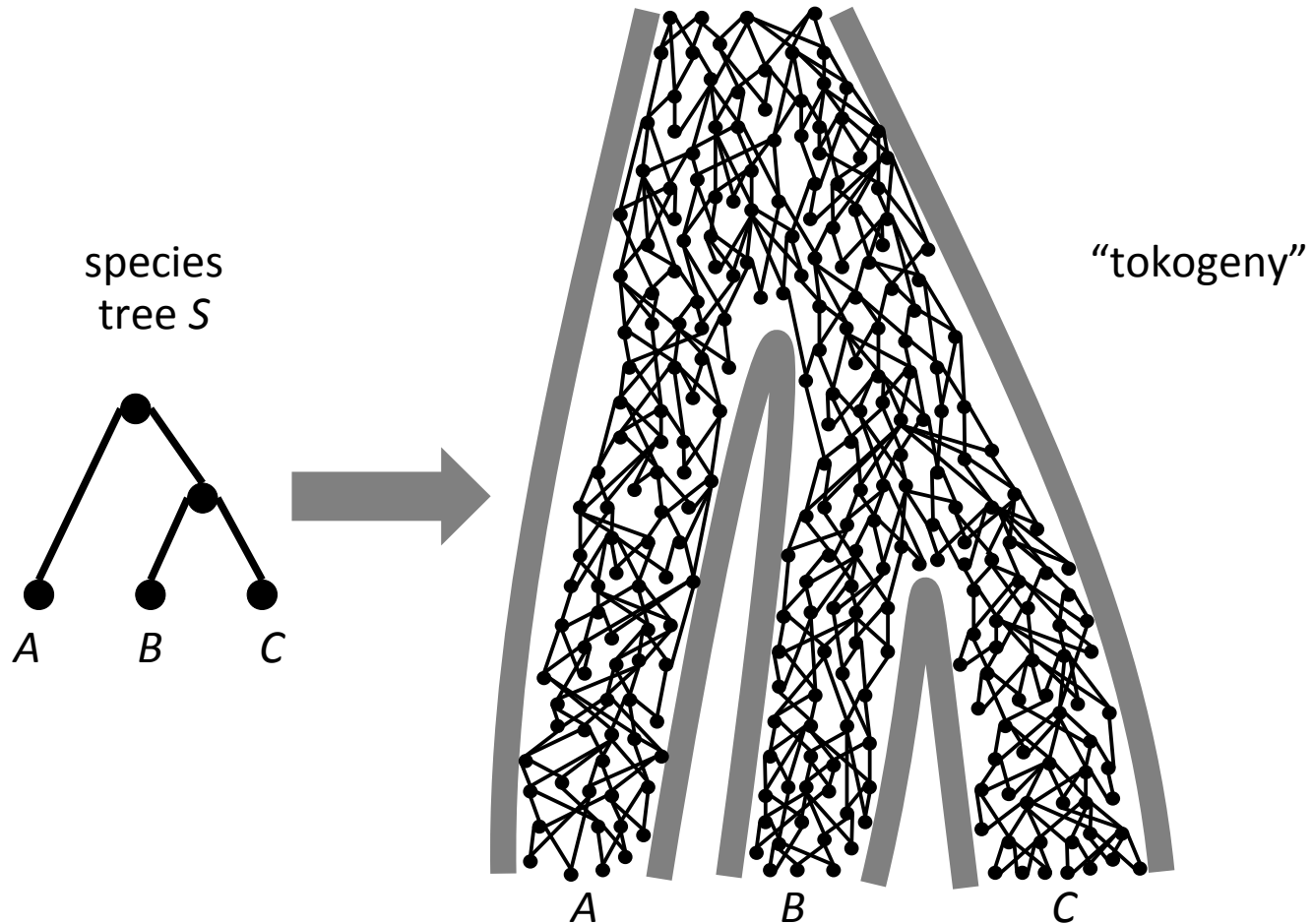
- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- Combinatorial reconstruction methods
- Practical use
- Illustrations
- Perspectives

Outline

- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- Combinatorial reconstruction methods
- Practical use
- Illustrations
- Perspectives

Phylogenetic trees

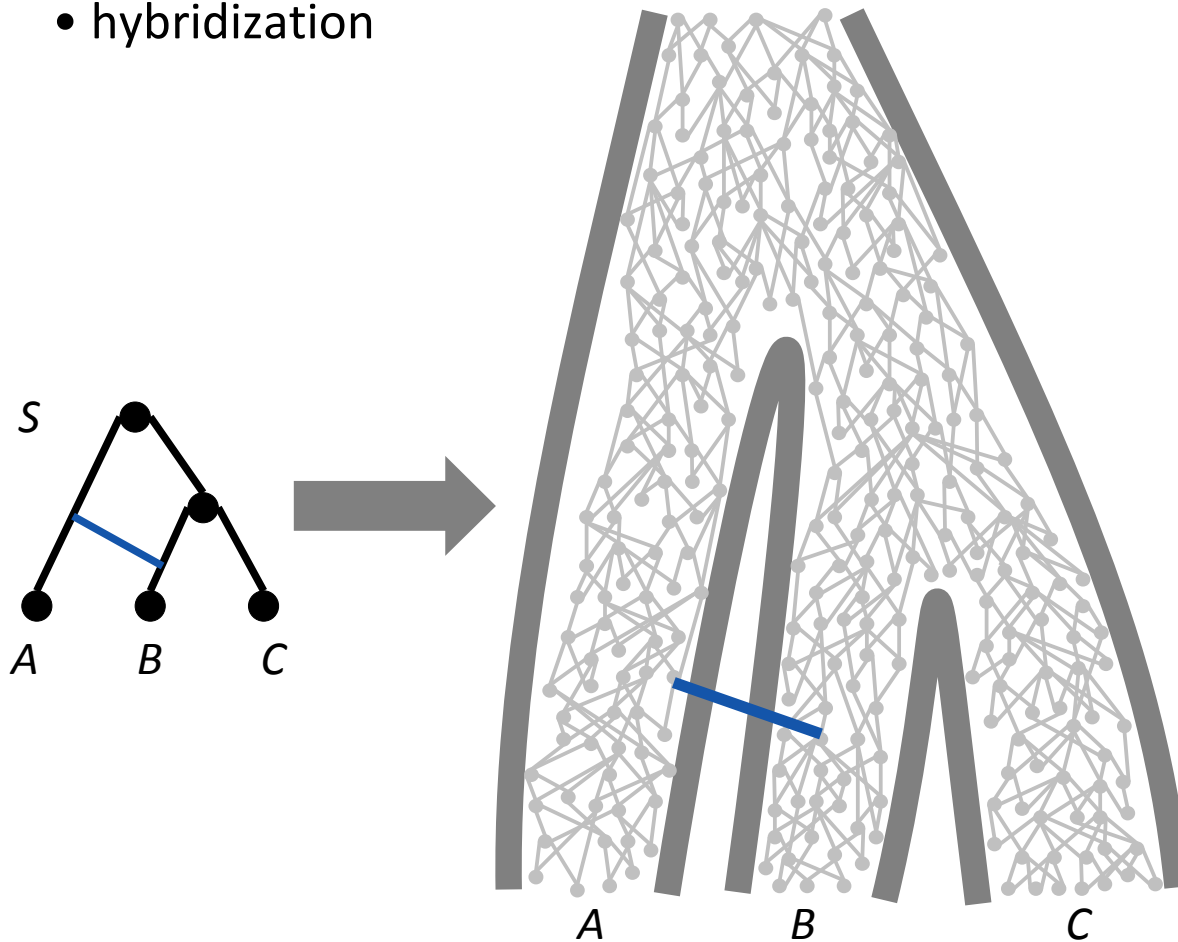
Phylogenetic tree of a species set



Genetic material transfer

Genetic material transfers between coexisting species:

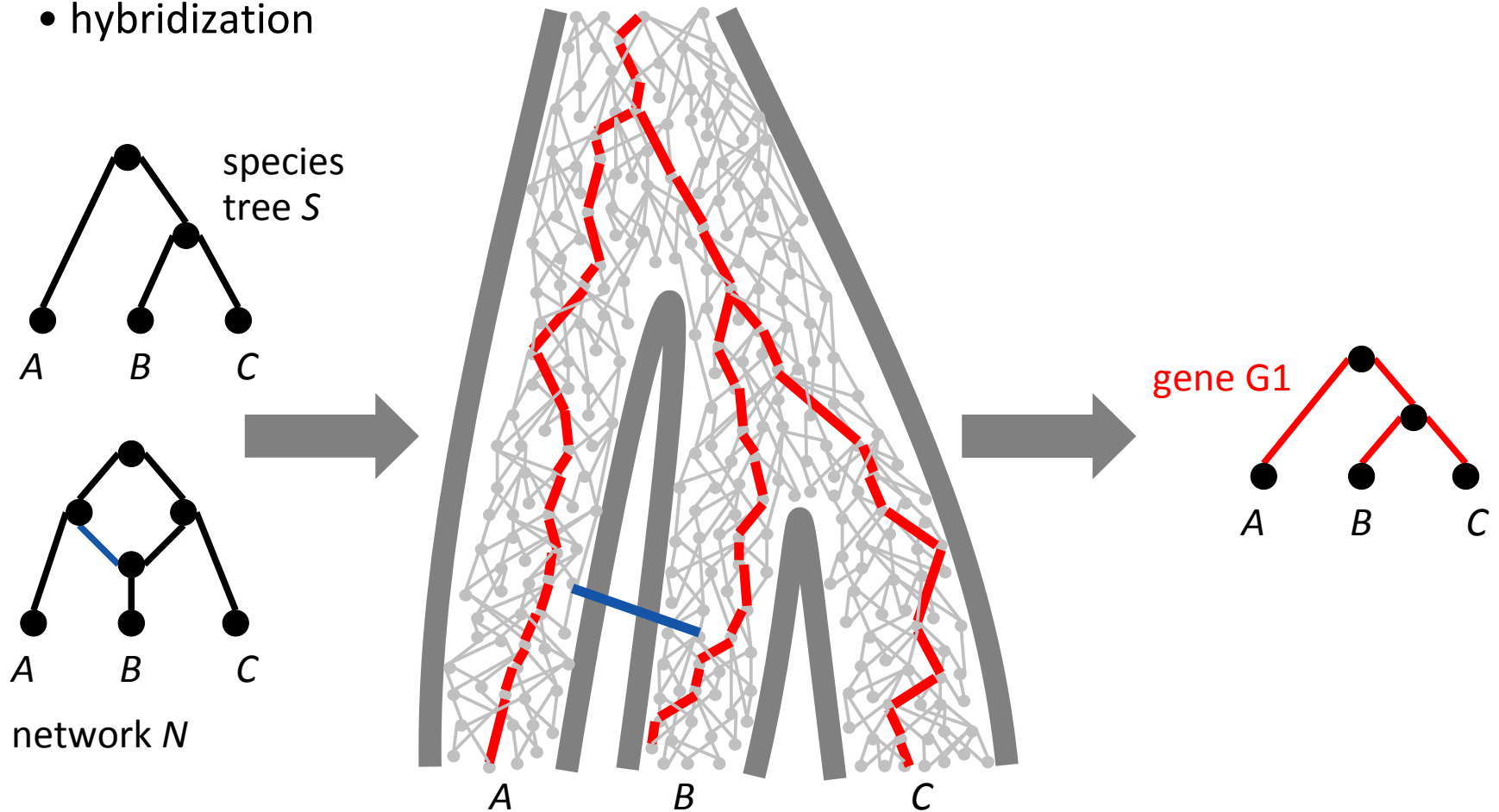
- horizontal gene transfer
- hybridization



Genetic material transfer

Genetic material transfers between coexisting species:

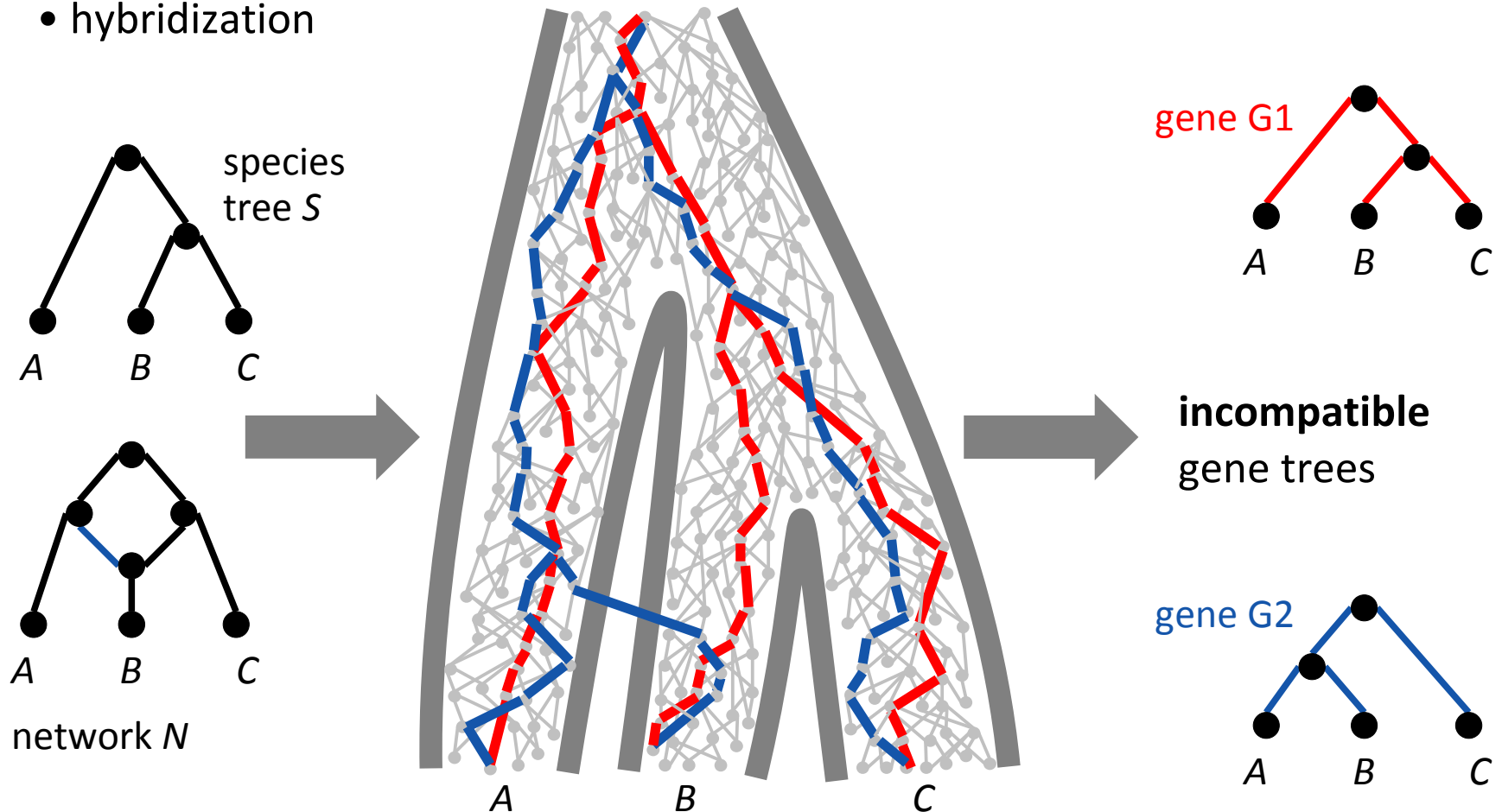
- horizontal gene transfer
- hybridization



Genetic material transfer

Genetic material transfers between coexisting species:

- horizontal gene transfer
- hybridization

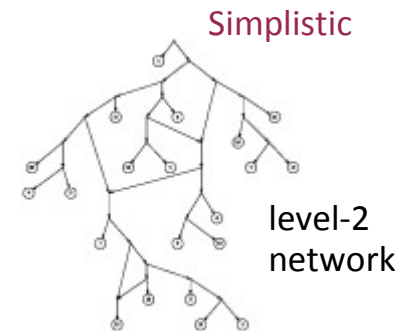
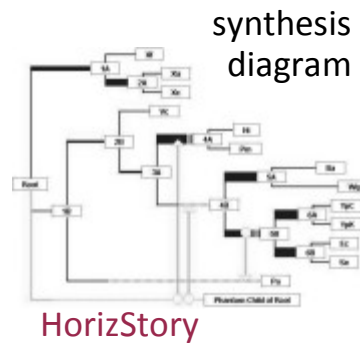
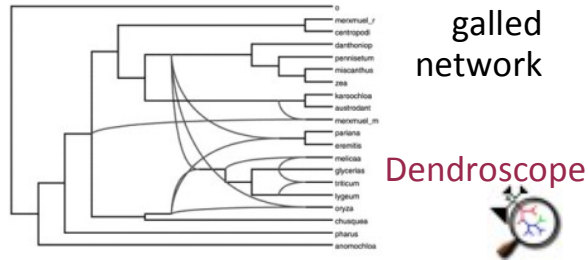


Phylogenetic networks

Phylogenetic network: network representing evolution data

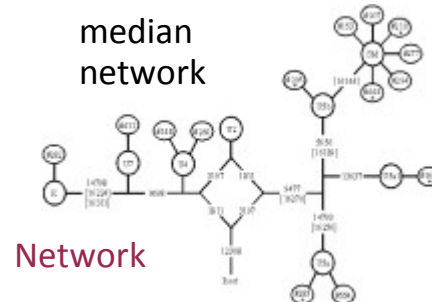
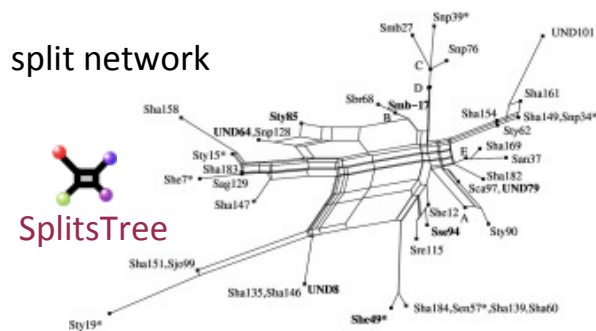
- **explicit** phylogenetic networks

to **model** evolution

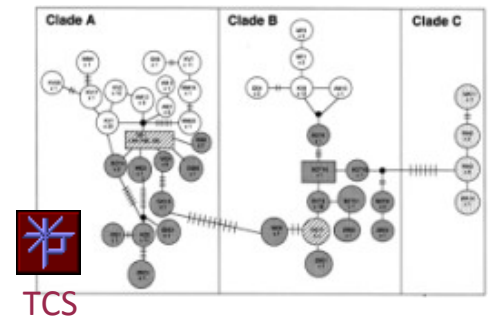


- **abstract** phylogenetic network

to **classify, visualize** data



minimum covering network



Phylogenetic network software

Who is Who in Phylogenetic Networks - Articles, Authors & Programs [RSS](#)

Index | Browse Contribute! | My selection

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

Publications - Index (All 376 publications) Selection by: Year | Category | Keyword | Author

Selection by Year

Number of publications per year on phylogenetic networks
Click on a year to display the publications

Year	Number of Publications
1975	1
1980	1
1985	3
1990	2
1995	6
2000	17
2001	9
2002	12
2003	28
2004	39
2005	34
2006	44
2007	44
2008	40
2009	44
2010	38

See how the community working on phylogenetic networks evolved in the last 10 years with the [coauthor graphs!](#)

Selection by Category

Article (Journal) (222)	InProceedings (94)	InBook (21)
Book (1)	PhdThesis (20)	MastersThesis (1)
Misc (17)	Programs (52)	

Selection by Keyword

abstract-network(48) approximation(8) APX-hard(2) ARG(5) bayesian(2) block-realization(1) branch-and-bound(1) cactus-graph(1) characterization(8) circular-split-system(8) clustering(3) consistency(2) cophylogeny(1) distance-between-networks(21) diversity(1) duplication(11) explicit-network(97) exponential-algorithm(2) FPT(17) from-clusters(11) from-distance(6) from-network(12) from-quartets(7) **from-rooted-trees(65)** from-sequencing(26) from-splits(10) from-trees(6) from-triplets(18) from-unrooted-trees(12) galled-network(8) haplotype-network(2) haplotyping(1) heuristic(12) HMM(2) hybridization(39) linear-programming(2) labeling(4) lateral-gene-transfer(35) level-k-phylogenetic-network(4) sorting(6) MASN(4) median-network(16) MedianJoining(2) minimum-number(19) minimum-selection(2) mu-distance(2) NeighborNet(11) nested-network(2) netting(3) normal-network(4) realization(2) parsimony(32) perfect(5) **phylogenetic-network(233)** polynomial(47) Program-Arlequin(5) Program-Beagle(3) Program-Bio-PhyloNetwork(4) Program-constNJ(1) Program-Dendroscope(8) Program-EEEP(3) Program-GalledTree(1) Program-HapBound(1) Program-

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

Based on BibAdmin by Sergiu Chelcea + tag clouds, date histogram, journal lists, co-author graphs, keyword definitions.

Who is Who in Phylogenetic Networks - Articles, Authors & Programs [RSS](#)

Index | Browse Contribute! | My selection

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

Programs to compute, evaluate, compare, visualize... **phylogenetic networks**
This page is automatically built from all publications tagged by Program* in the [database](#).

Program Arlequin

The goal of *Arlequin* is to provide the average user in population genetics with quite a large set of basic methods and statistical tests, in order to extract information on genetic and demographic features of a collection of population samples. In particular, Arlequin implements a Minimum Spanning Network algorithm to embed the set of all minimum spanning trees computed from a distance matrix of haplotypes (<http://cmpg.unibe.ch/software/arlequin3/>).

[5 publications in the database mention Program Arlequin](#)

Program Beagle

Beagle is a small collection of related programs for analysing the minimum number of recombinations required for a SNP data set under the infinite sites model. Available at <http://www.stats.ox.ac.uk/~lyngsoe/beagle/>.

[3 publications in the database mention Program Beagle](#)

Program Bio PhyloNetwork

Bio-PhyloNetwork is a Perl package that relies on the BioPerl bundle and implements many algorithms on phylogenetic networks (<http://dmi.uib.es/~gcardona/BioInfo/Bio-PhyloNetwork.tgz>). It is used in a Java Applet which can compare and draw two phylogenetic networks entered in eNewick format with the same set of leaves (<http://dmi.uib.es/~gcardona/BioInfo/alignment.php>)

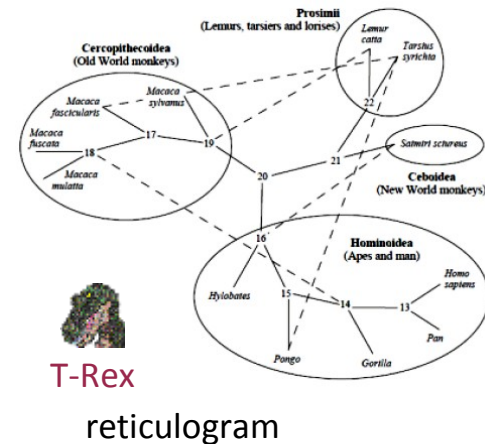
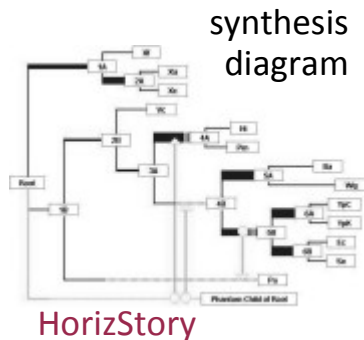
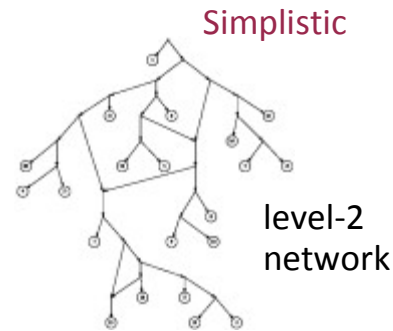
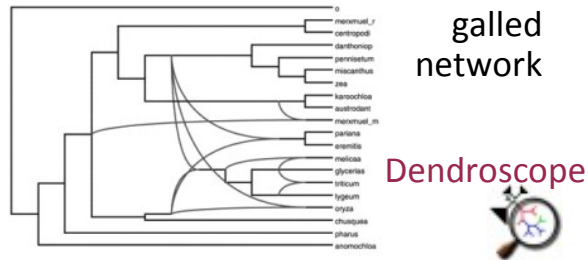
[4 publications in the database mention Program Bio PhyloNetwork](#)

Explicit phylogenetic networks

Phylogenetic network: network representing evolution data

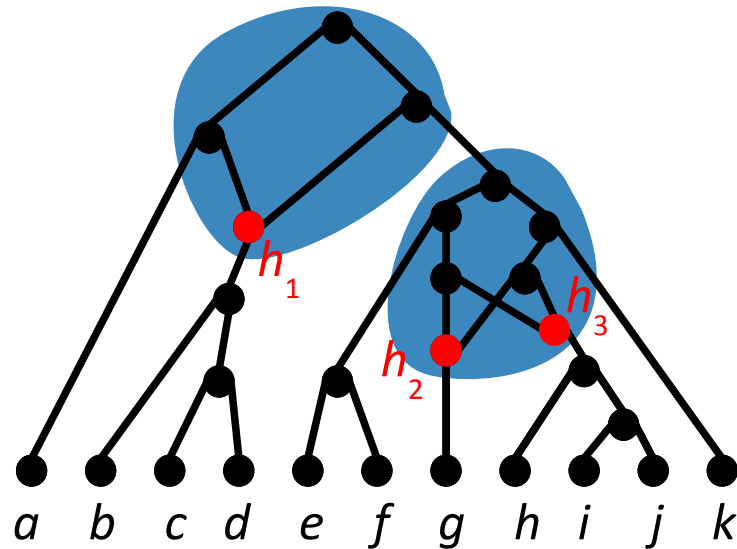
- **explicit** phylogenetic networks

evolution model



Explicit phylogenetic networks

Rooted explicit phylogenetic network: tree-like parts + *blobs*.



vertices with more than one parent:
reticulations

Plan

- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- Combinatorial reconstruction methods
- Practical use
- Illustrations
- Perspectives

Phylogenetic network reconstruction

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

{gene sequences}

distance methods

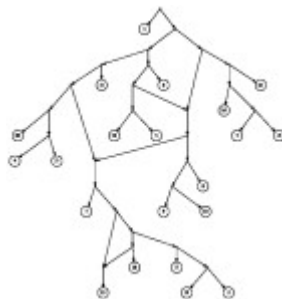
*Bandelt & Dress 1992 - Legendre & Makarenkov
2000 - Bryant & Moulton 2002*

parsimony methods

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

likelihood methods

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



network *N*

Phylogenetic network reconstruction

**Problem: usually slow,
lots of sequences available.**

espèce 1 : AATTGCAG TAGCCCCAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 : TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAT
espèce 7 : ACCTGGTG TAAAAT

G1 **G2**

{gene sequences}

distance methods

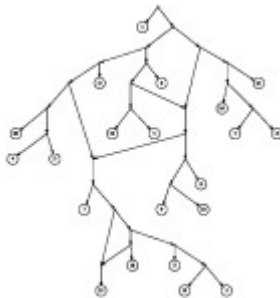
*Bandelt & Dress 1992 - Legendre & Makarenkov
2000 - Bryant & Moulton 2002*

parsimony methods

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

likelihood methods

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

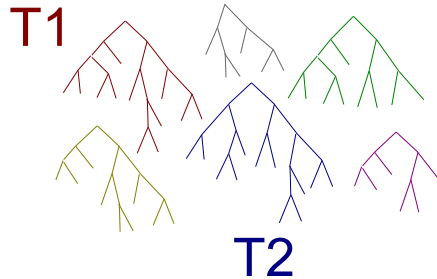


network *N*

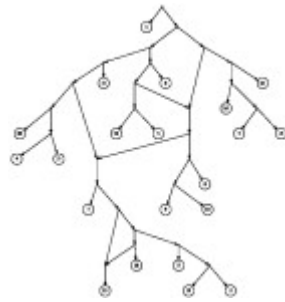
Phylogenetic network reconstruction

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



rooted
explicit
network



{gene sequences}

Reconstruction of one tree for each gene present in several species

Guindon & Gascuel, SB, 2003

{trees}

HOGENOM database

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



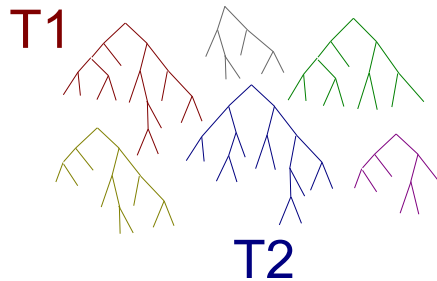
Tree consensus or reconciliation

optimal network N

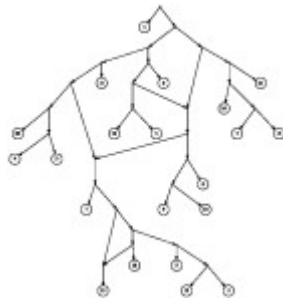
Phylogenetic network reconstruction

espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAAT
espèce 5 : TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAAT

G1 G2



rooted
explicit
network



{gene sequences}

Reconstruction of one tree for each gene present in several species

Guindon & Gascuel, SB, 2003

{trees}

HOGENOM database



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

> 500 species, >70 000 trees

Tree consensus or reconciliation

optimal network N

Problem: tree reconciliation is difficult even for 2 trees

(NP-complete for 2 trees with minimum reticulation number)

Bordewich & Semple, DAM, 2007

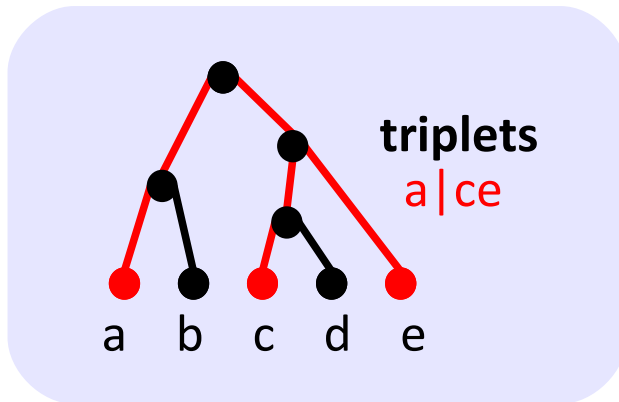
Triplets and clusters

Problem:

Reconstructing the **supernetwork** of a set of trees is
hard.

Idea:

reconstruct a network containing all:



of the input trees ?

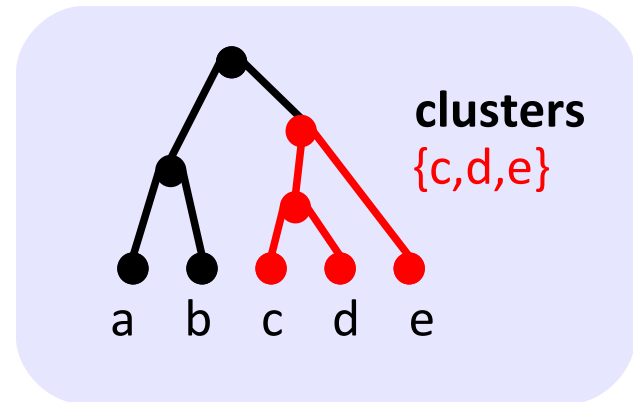
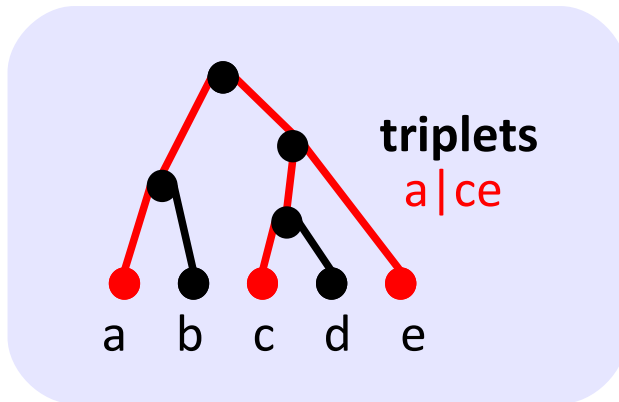
Triplets and clusters

Problem:

Reconstructing the **supernetwork** of a set of trees is
hard.

Idea:

reconstruct a network containing all:

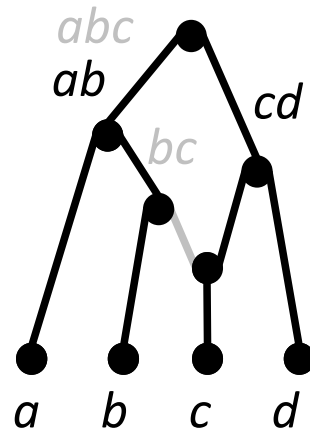


of the input trees ?

Softwired clusters

“softwired” cluster : cluster of a tree contained in the network

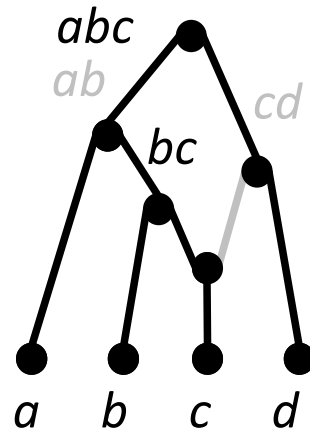
Tree-like model of gene transmission:
each gene comes from a single parent



Softwired clusters

“softwired” cluster : cluster of a tree contained in the network

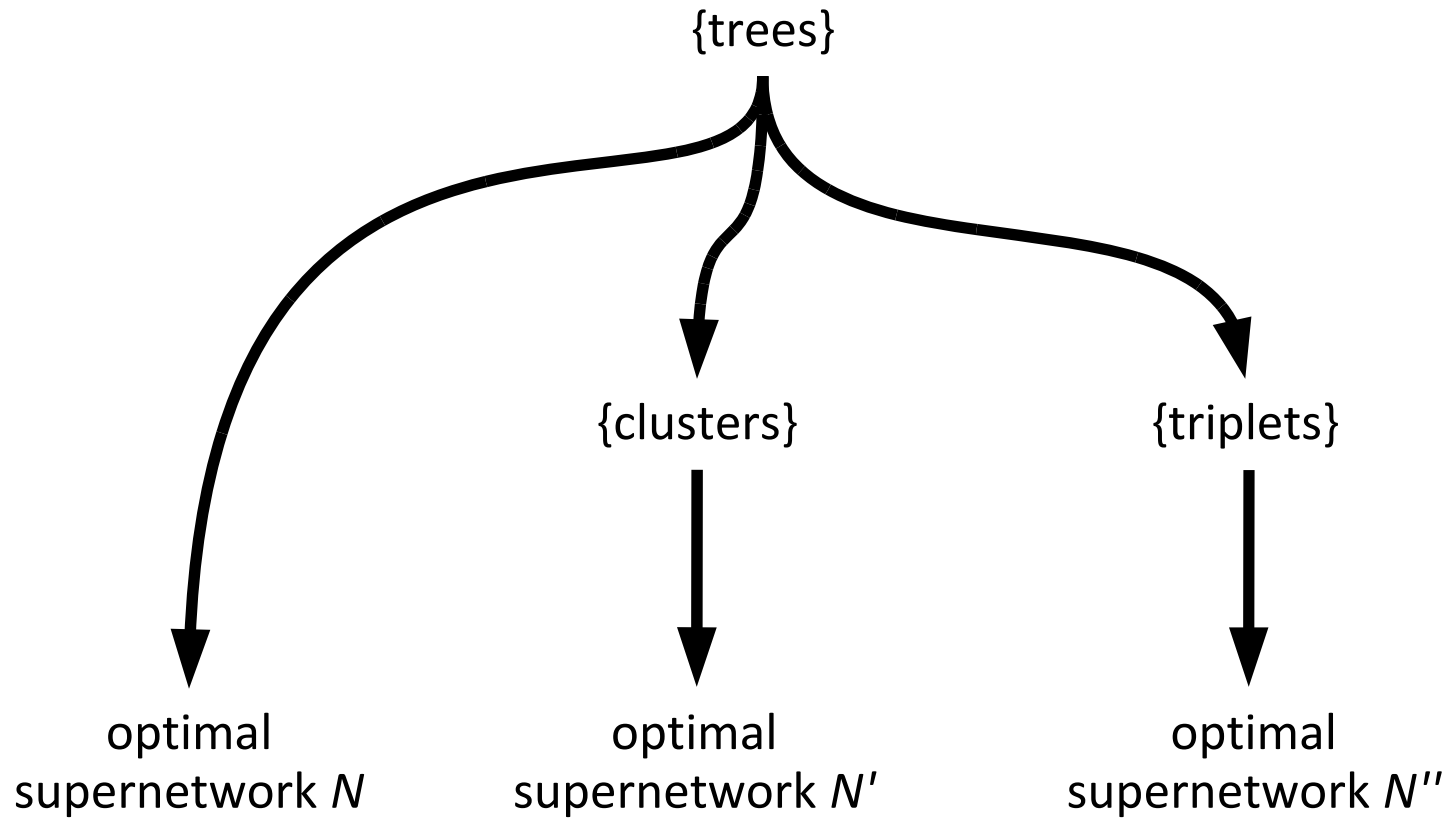
Tree-like model of gene transmission:
each gene comes from a single parent



Combinatorial phylogenetic network reconstruction

Idea:

change the type of data to process

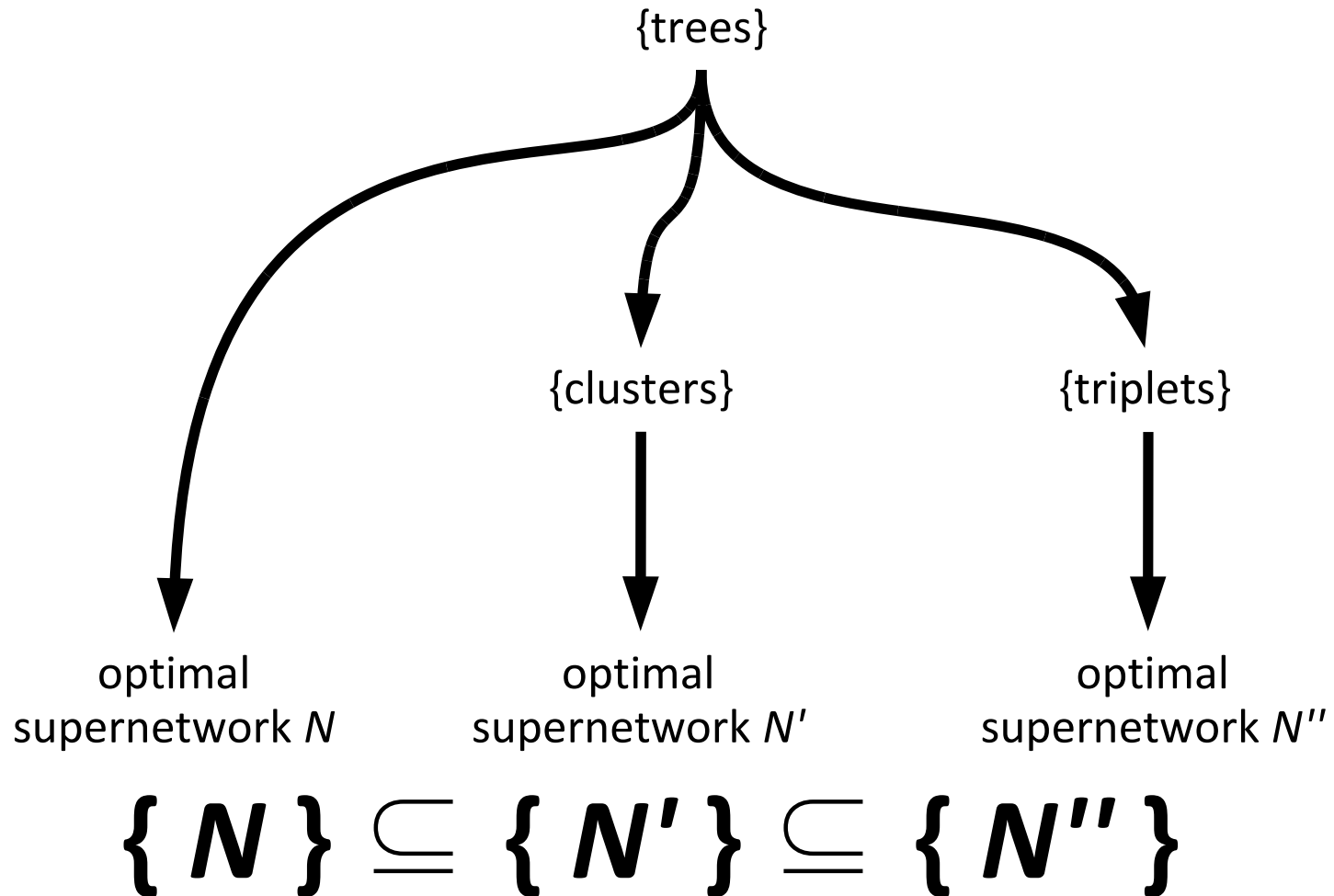


$$N = N' = N''?$$

Combinatorial phylogenetic network reconstruction

Idea:

change the type of data to process



Plan

- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- **Combinatorial reconstruction methods**
- Practical use
- Illustrations
- Perspectives

Reconstruction from softwired clusters

{trees}



{clusters}



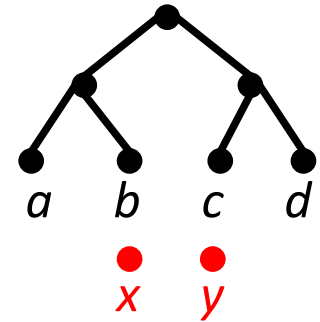
N'

galled network

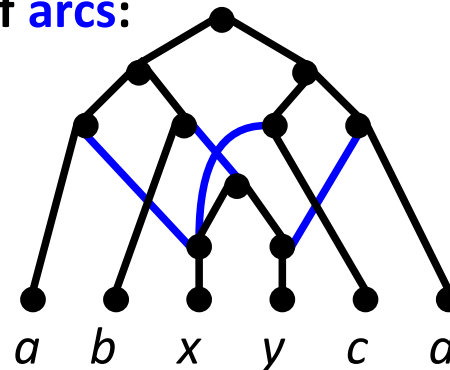
Fast exact **galled network** reconstruction method from **softwired clusters**

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

Step 1- Solve cluster conflicts by **deleting taxa**,
reconstruct tree on remaining taxa
MAXIMUM COMPATIBLE SUBSET



Step 2- Attach taxa involved in conflicts to the tree
with the **minimum number of arcs**:
MINIMUM ATTACHMENT



Reconstruction from softwired clusters

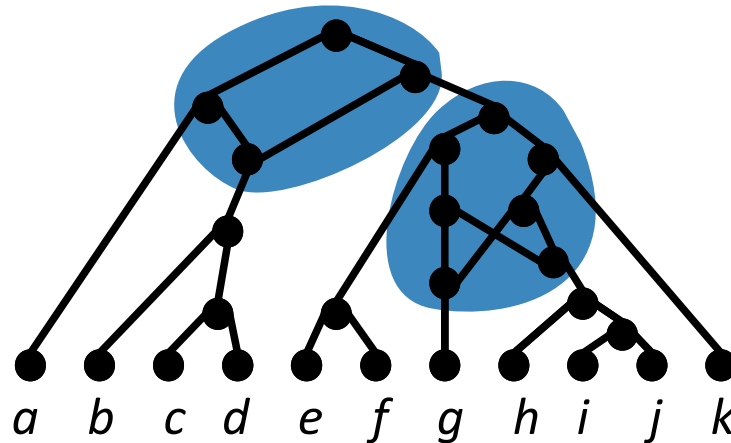
{arbres}

Exact method for **level k network** reconstruction from **softwired clusters**

Iersel, Kelk, Rupp & Huson, ISMB 2010

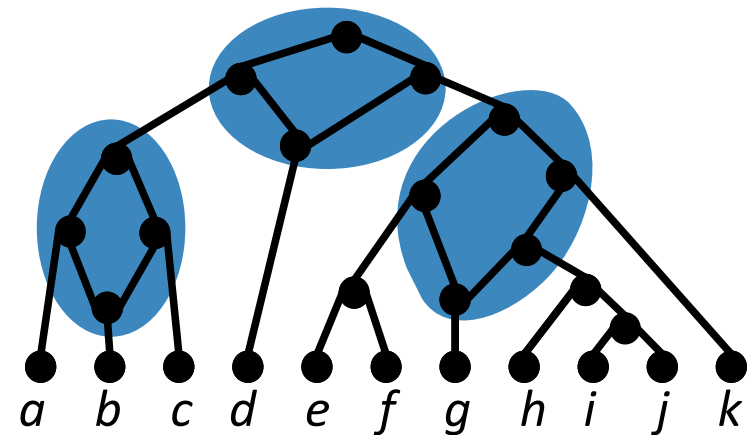
{clusters}

↳ less reticulations, but slower for level > 2 .



level-2 network

level =
maximum number of reticulations
per blob.



level-1 network
(*"galled tree"*)

N'
level- k
network

Reconstruction from triplets

{trees}



{triplets}



N'
level- k
network

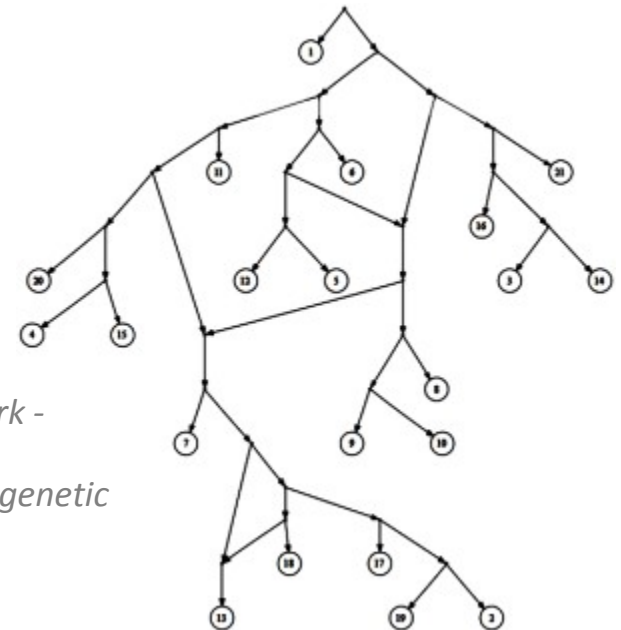
Exact methods to reconstruct **level-1 and 2 networks** (if there exist any) from a dense **triplet set**

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

T **dense** triplet set =

On any subset of 3 leaves, T contains at least one triplet

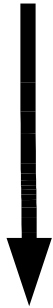
Program Simplistic



*Yeast phylogenetic network -
Van Iersel et al. :
Constructing level-2 phylogenetic
networks from triplets.
RECOMB 2008*

Reconstruction from triplets

{trees}



{triplets}



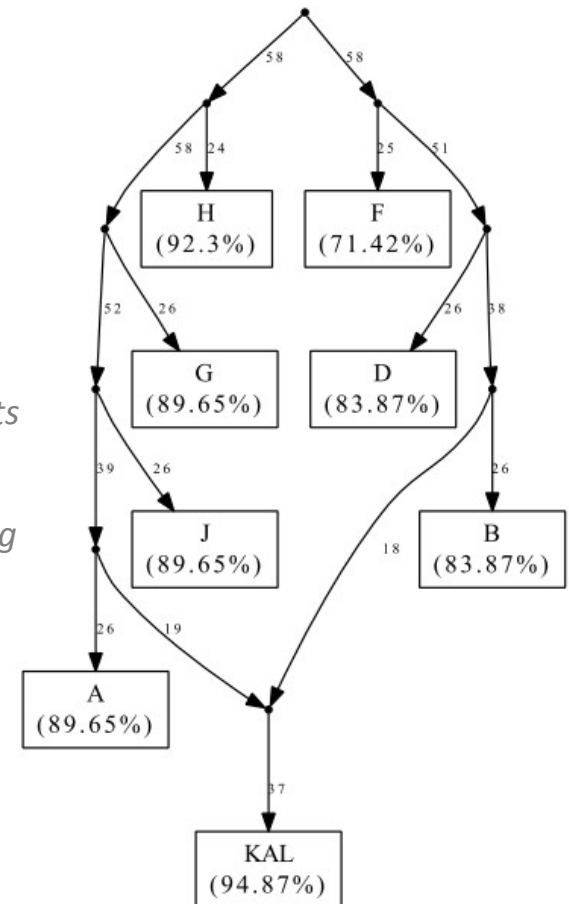
N'
level-1
network

Fast heuristic method to reconstruct a **level-1 network** containing **most of the input triplets**

Program Lev1athan

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

*Phylogenetic network built from triplets
extracted from 2 trees of HIV-1 strains
Huber, van Iersel, Kelk & Sucheccki
A practical algorithm for reconstructing
level-1 phylogenetic networks
TCBB, 2011*



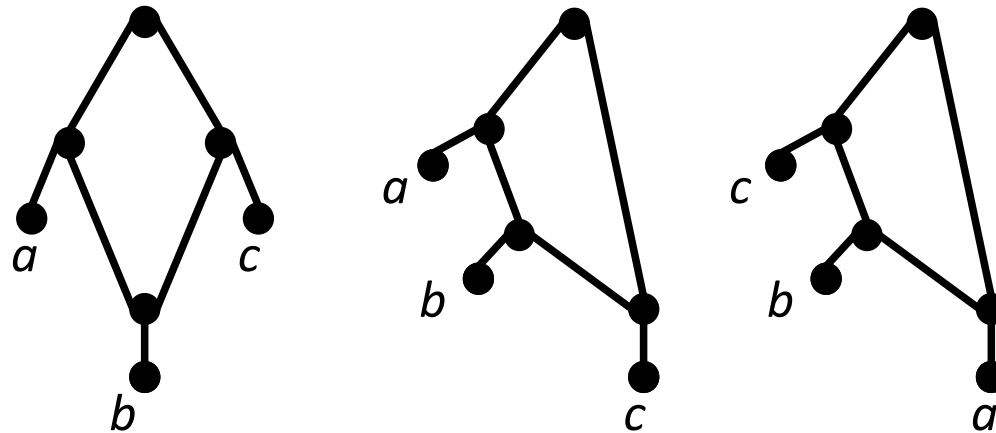
Outline

- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- Combinatorial reconstruction methods
- **Practical use**
- Illustrations
- Perspectives

Solution ambiguity

Ambiguity of the results even with complete and correct data

Many **distinct** minimal networks have exactly the **same** set of contained trees, triplets, and clusters.

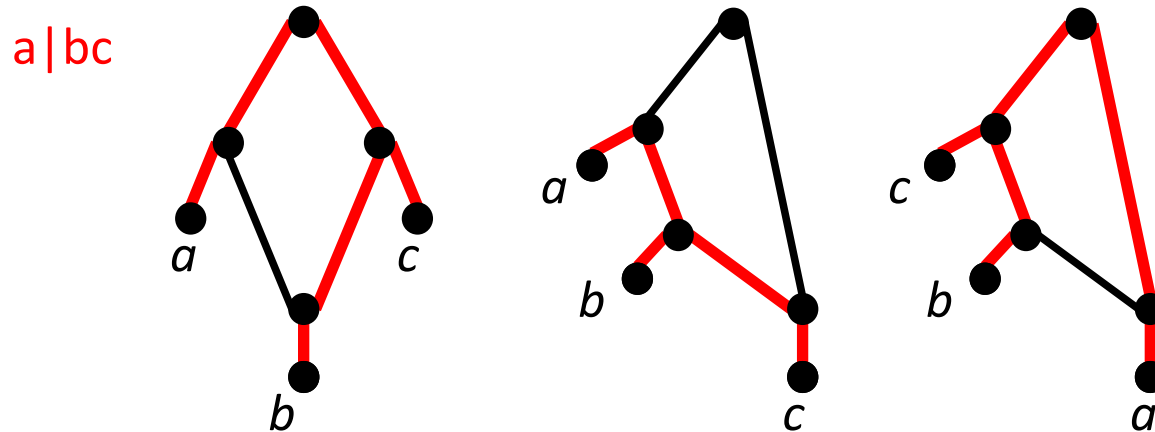


Characterization for level-1 networks :
The only ambiguous cases have such blobs (< 5 vertices)

Solution ambiguity

Ambiguity of the results even with complete and correct data

Many **distinct** minimal networks have exactly the **same** set of contained trees, triplets, and clusters.

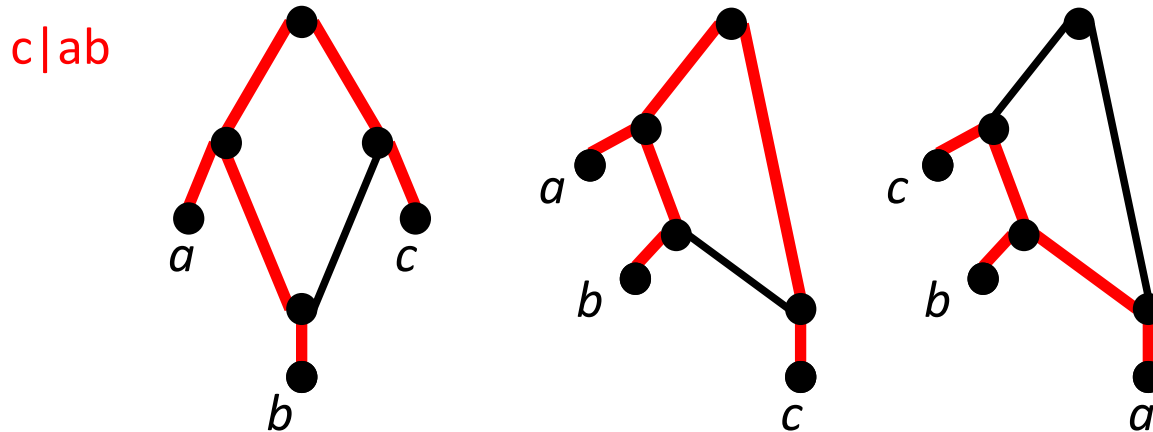


Characterization for level-1 networks :
The only ambiguous cases have such blobs (< 5 vertices)

Solution ambiguity

Ambiguity of the results even with complete and correct data

Many **distinct** minimal networks have exactly the **same** set of contained trees, triplets, and clusters.

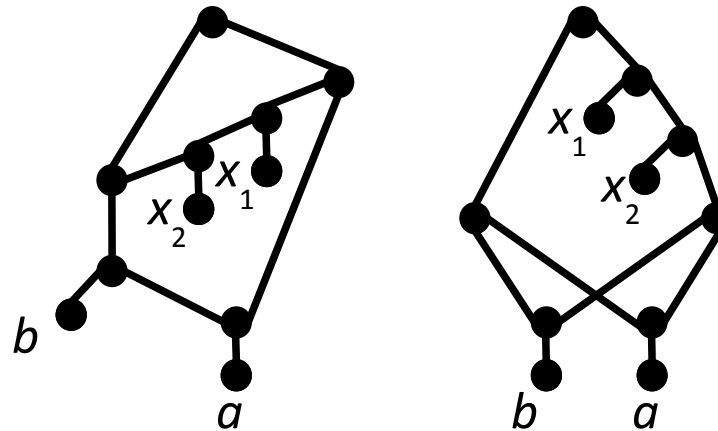


Characterization for level-1 networks :
The only ambiguous cases have such blobs (< 5 vertices)

Solution ambiguity

Ambiguity of the results even with complete and correct data

Many **distinct** minimal networks have exactly the **same** set of contained trees, triplets, and clusters.

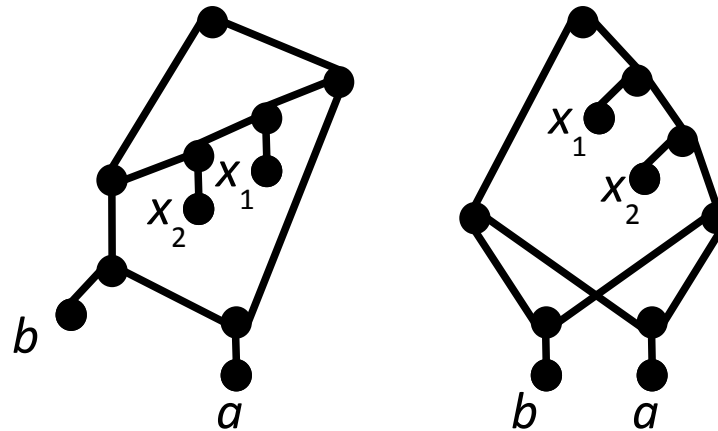


2 level-2 networks with exactly the same triplet set

Solution ambiguity

Ambiguity of the results even with complete and correct data

Many **distinct** minimal networks have exactly the **same** set of contained trees, triplets, and clusters.



2 level-2 networks with exactly the same triplet set

Even with **complete** and **correct** data,
impossible to choose among the ambiguous configurations!

Practical use

existing methods
still work to do!

Conditions for use	Available data	Possible processings
rooted trees	unrooted trees	<i>Rooting with a reference species tree or topological constraints</i>
Single-copy gene trees	MUL-trees (with duplicated genes)	MUL-tree processing Scornavacca, Berry & Ranwez, 2009
Correct clusters and triplets	noisy data	Tree cleaning PhySIC_IST, 2008 Data filtering (clusters with high bootstrap value, present in >x% of the trees) <i>Data editing : solution containing most of the input data</i>
Complete data (density for triplet sets, complete clusters)	partial data, deleted genes	<i>Selection of a large number of trees with a large number of common species</i> <i>Selection of the maximal number of taxa with triplet density</i> NP-complete problems

Outline

- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- Combinatorial reconstruction methods
- Practical use
- **Illustrations**
- Perspectives

Illustrations

16 trees on 47 taxa from the HOGENOM database

(proteobacteria)

24 Enterobacteriales

2 Pasteurellales

1 Aeromonadales

9 Alteromonadales

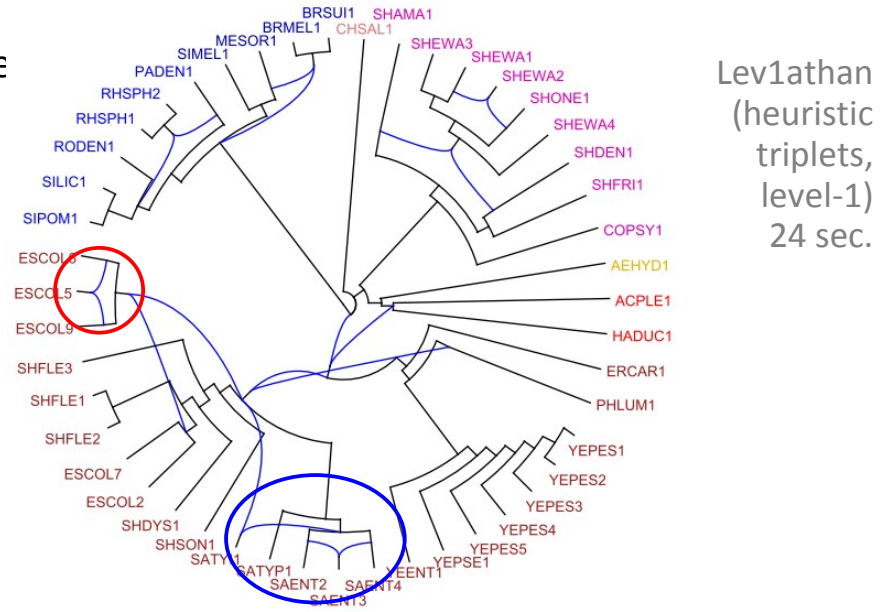
1 Oceanospirillales

6 Rhodobacterales

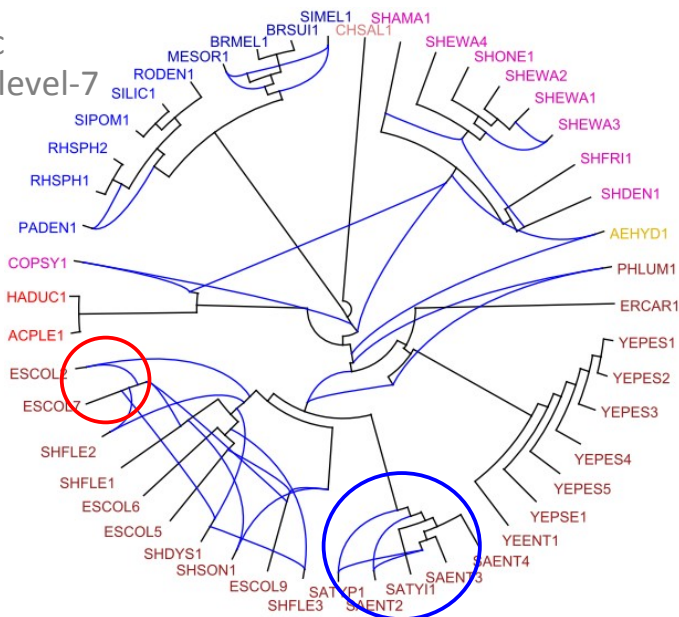
4 Rhizobiales



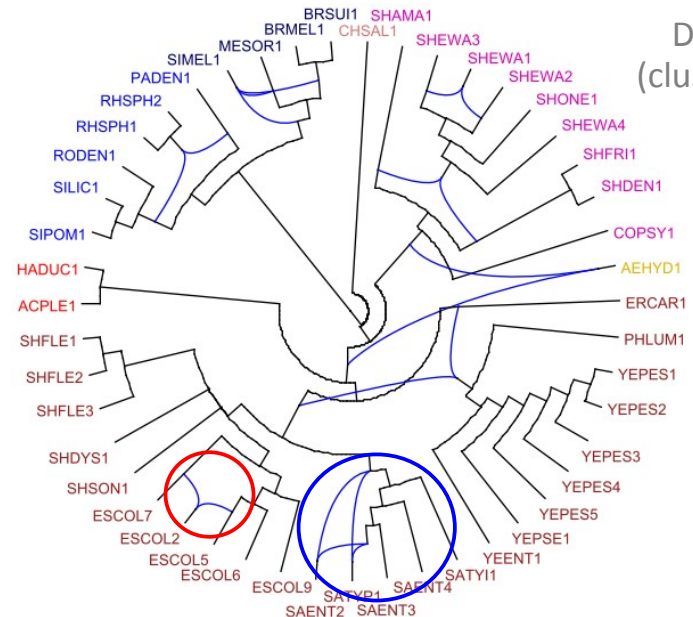
Networks containing triplets, softwired clusters, present in at least 20% of the trees



Simplistic
(triplets, level-7 network)
63 sec.



Dendroscope
(clusters, galled network)
<1 sec.



Illustrations

16 trees on 47 taxa from the HOGENOM database

(proteobacteria)

24 Enterobacteriales

2 Pasteurellales

1 Aeromonadales

9 Alteromonadales

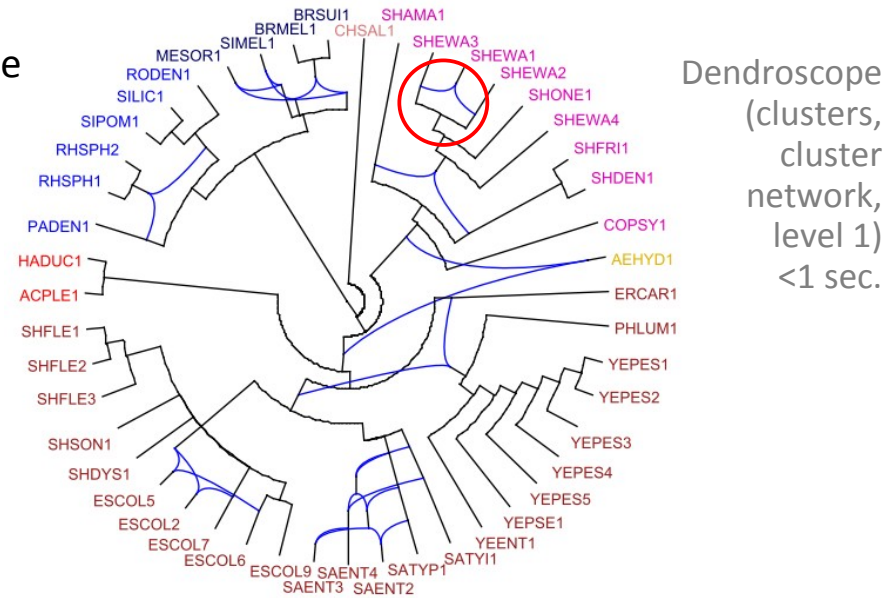
1 Oceanospirillales

6 Rhodobacterales

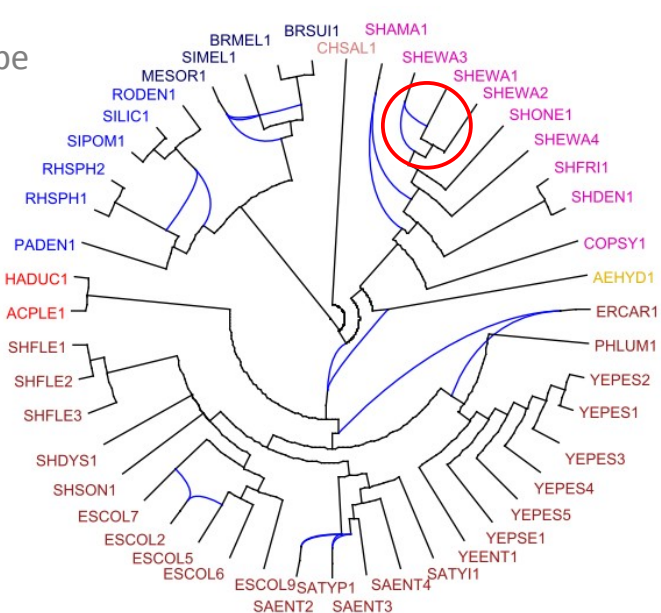
4 Rhizobiales



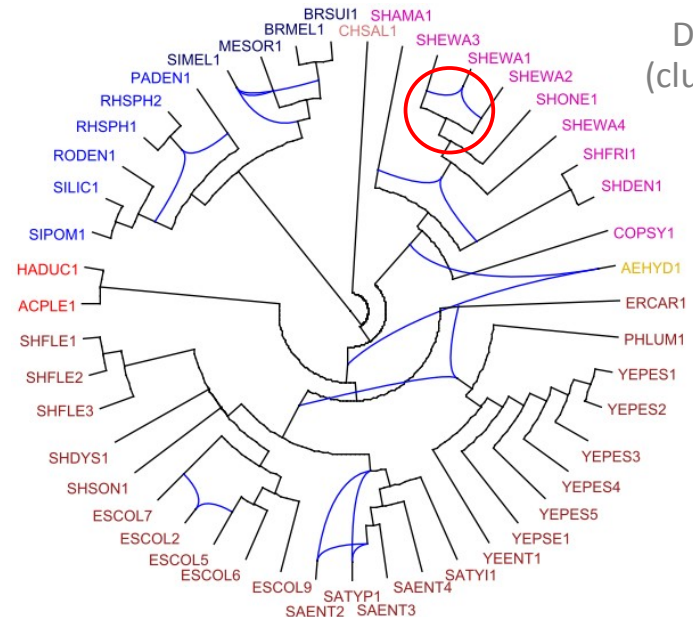
Networks containing triplets, softwired clusters, present in at least 20% of the trees



Dendroscope
(clusters,
level-7
network)
2 sec.



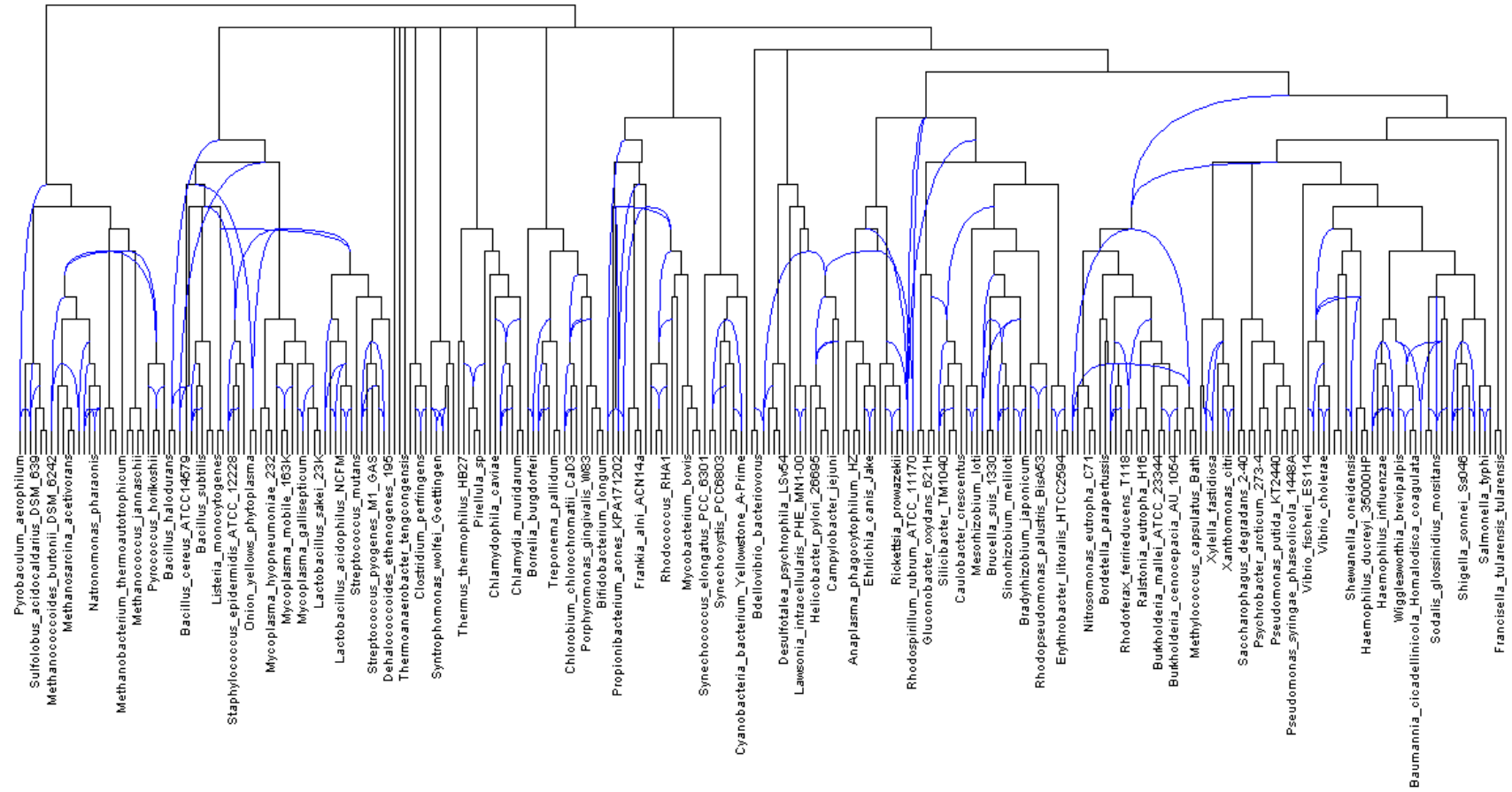
Dendroscope
(clusters, galled
network)
<1 sec.



Illustrations

9 trees on 279 procaryote species Clusters in at least 2 trees

Auch, Steigle, Huson & Henz, 2009

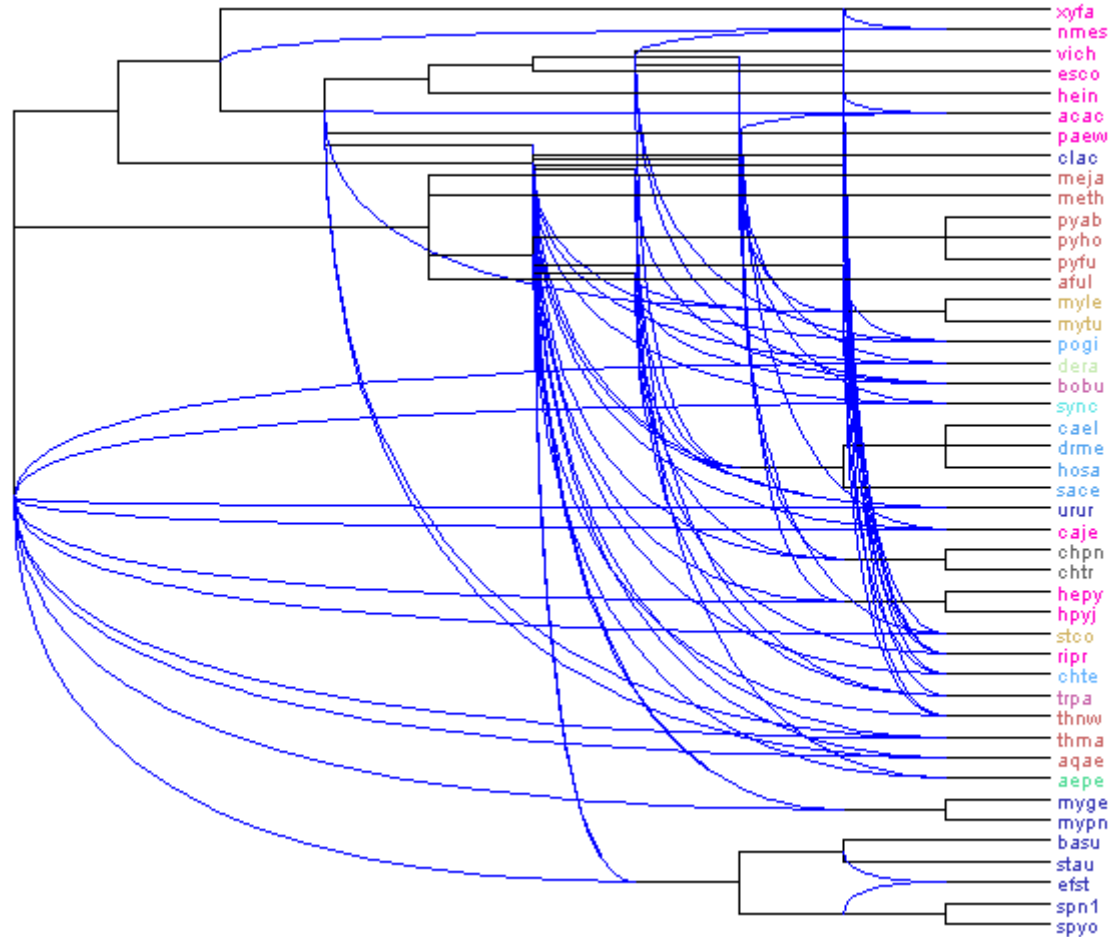


Dendroscope
(clusters, galled network)
2 sec.

Illustrations

23 trees, 45 species from the 3 domains of life
clusters with 99% bootstrap confidence

Dendroscope
(galled
network)
4 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
(galled
network)
<1 sec.

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.

Outline

- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- Combinatorial reconstruction methods
- Practical use
- Illustrations
- **Perspectives**

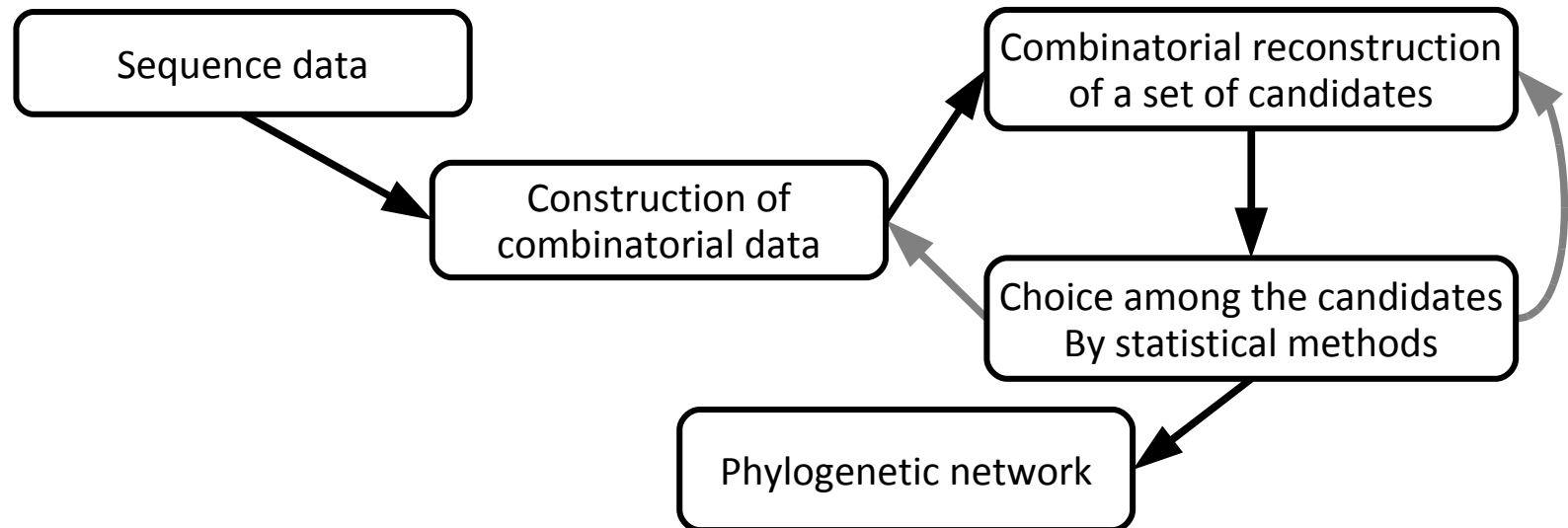
Perspectives

Combinatorics :

- Better knowledge of small level networks
- Update of a network with new data
- Unrooted explicit phylogenetic network reconstruction

Bioinformatics :

- Function of transferred genes (“transfer highways”)
- Integration of combinatorial methods in a statistical framework

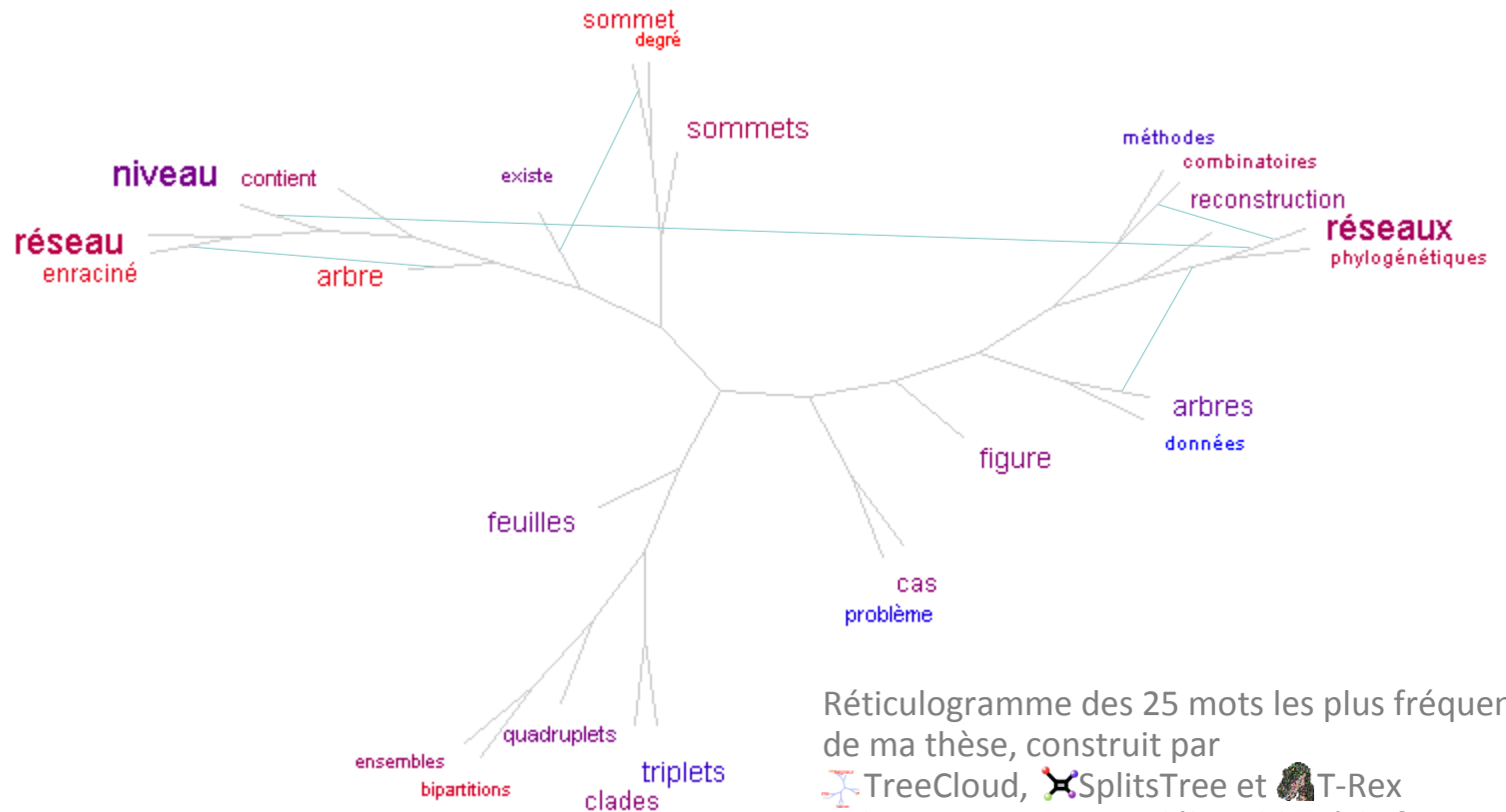





Thank you !

Coauthors:

- Vincent Berry, Christophe Paul (LIRMM)
- Daniel Huson, Regula Rupp (Tübingen)
- Katharina Huber (East Anglia)

Brown et al.'s data provided by 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