09/02/2011 - Rencontres Alphy, Lyon

Practical use of combinatorial methods for phylogenetic network reconstruction

Philippe Gambette

Outline

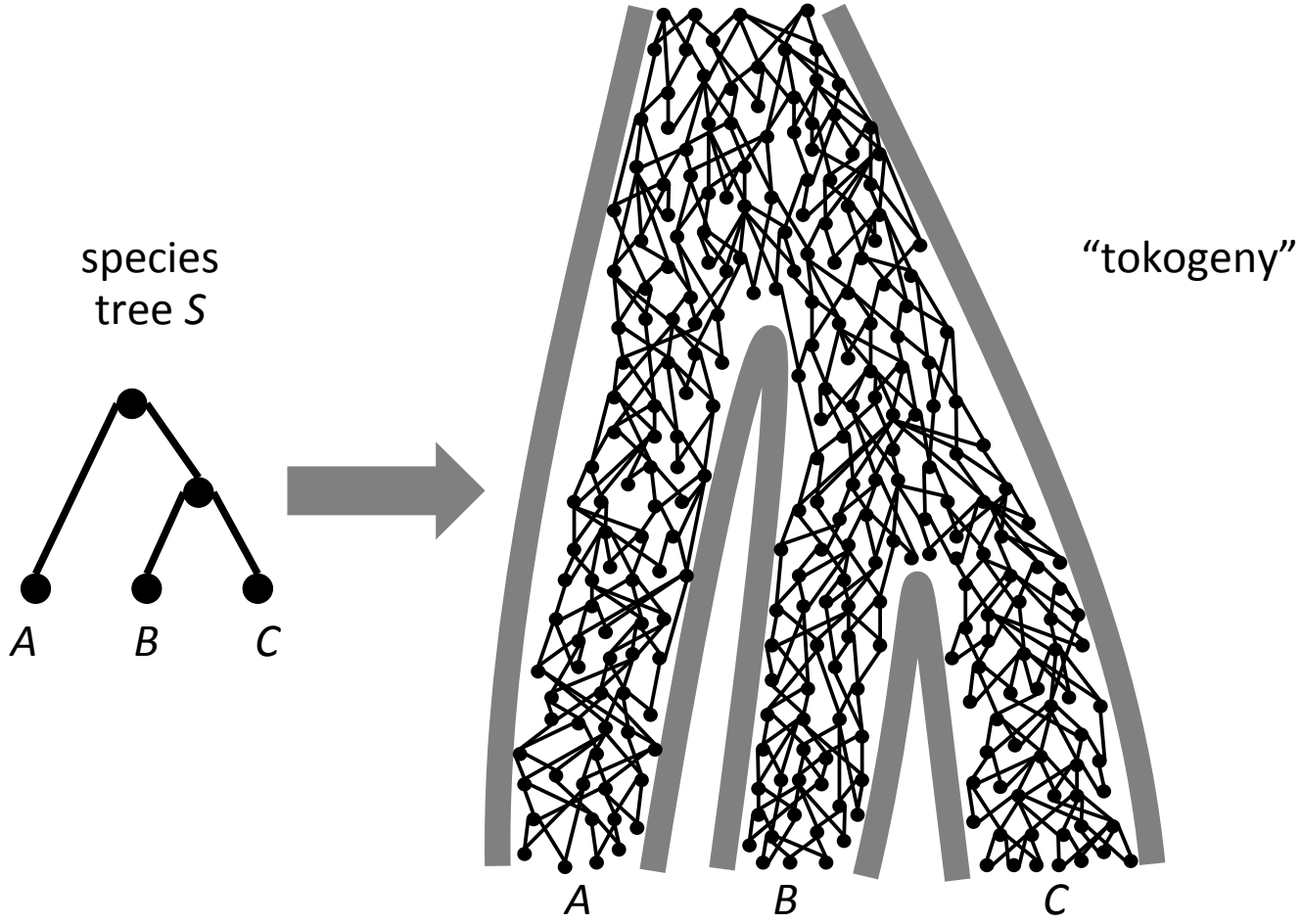
- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- Combinatorial reconstruction methods
- Practical use
- Illustrations
- Perspectives

Outline

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

Phylogenetic trees

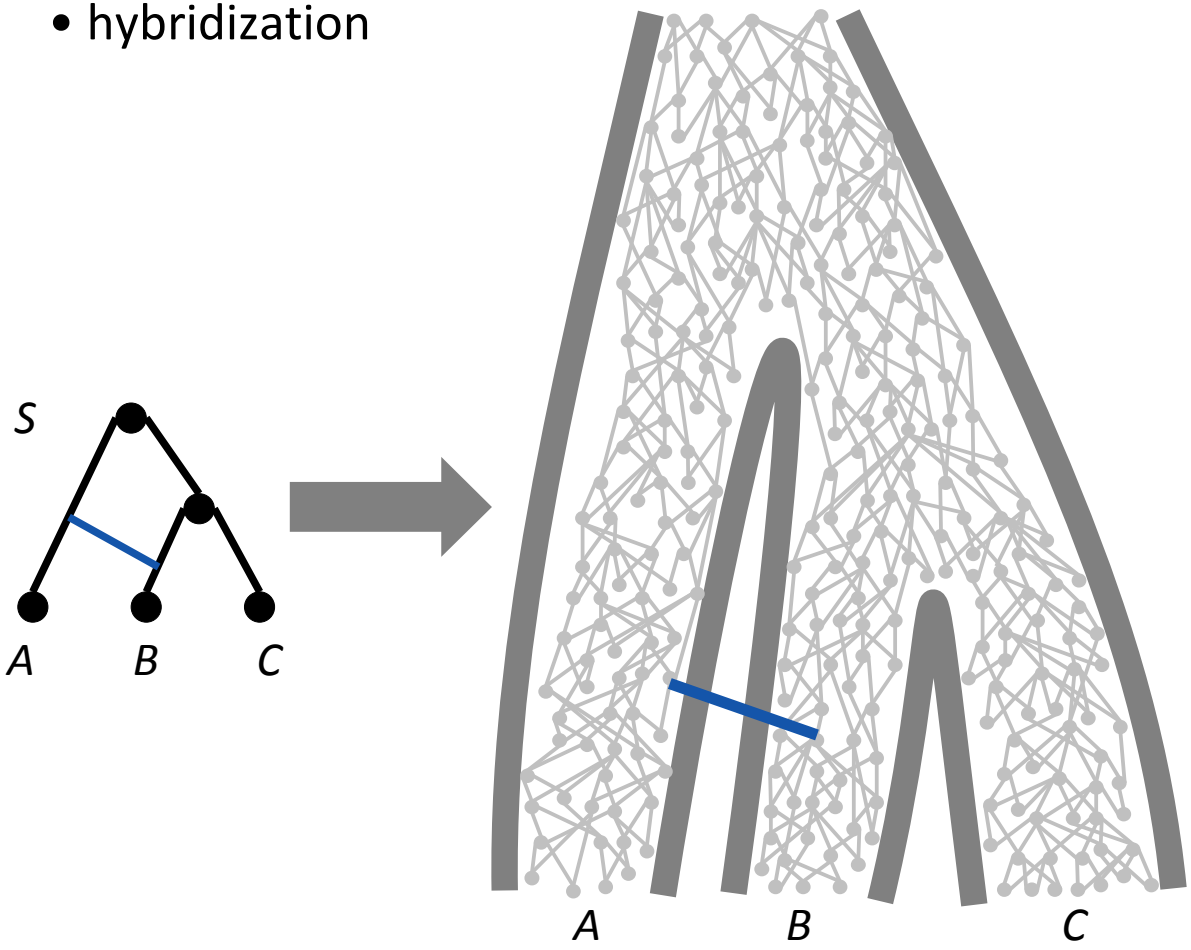
Phylogenetic tree of a species set



Genetic material transfer

Genetic material transfers between coexisting species:

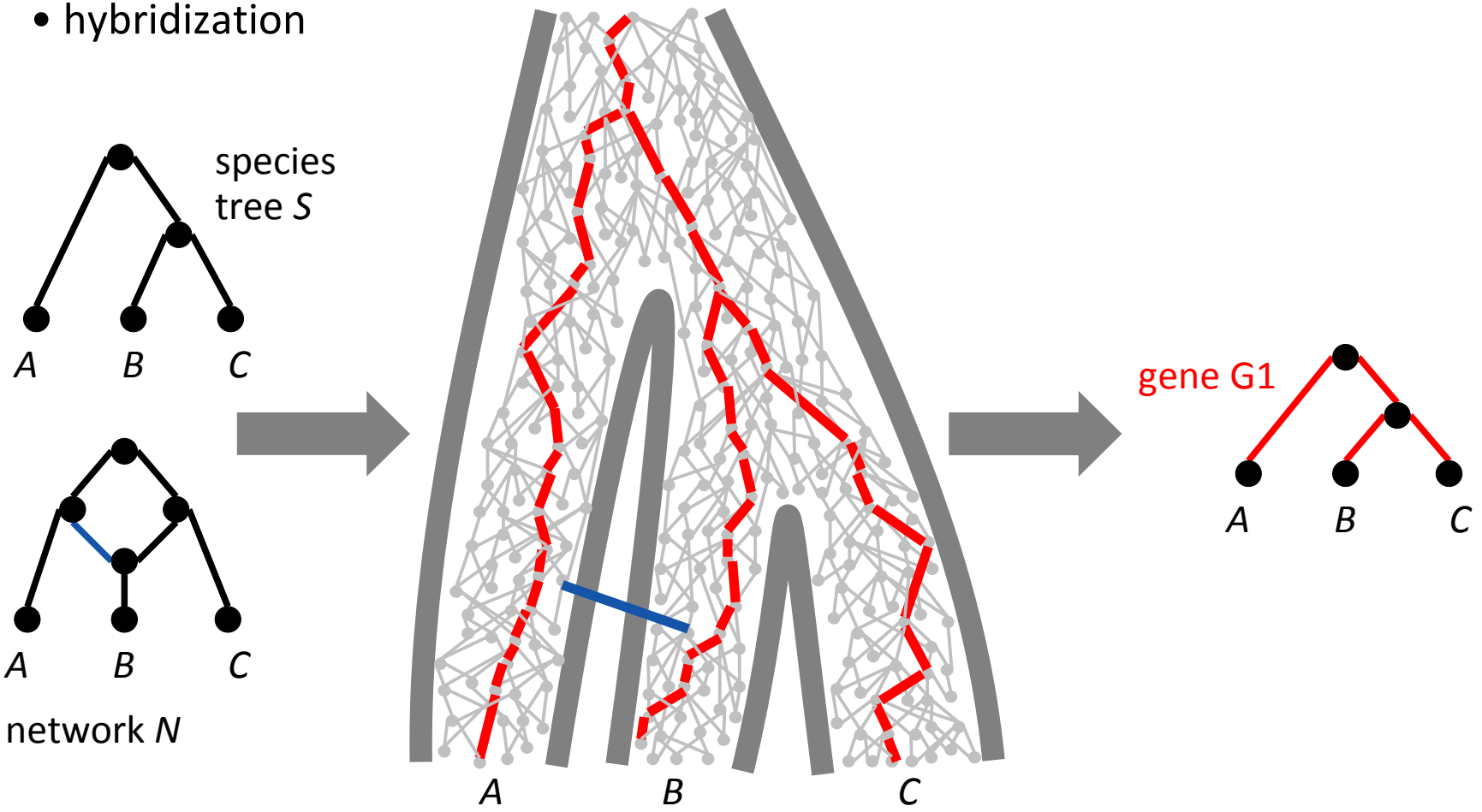
- horizontal gene transfer
- hybridization



Genetic material transfer

Genetic material transfers between coexisting species:

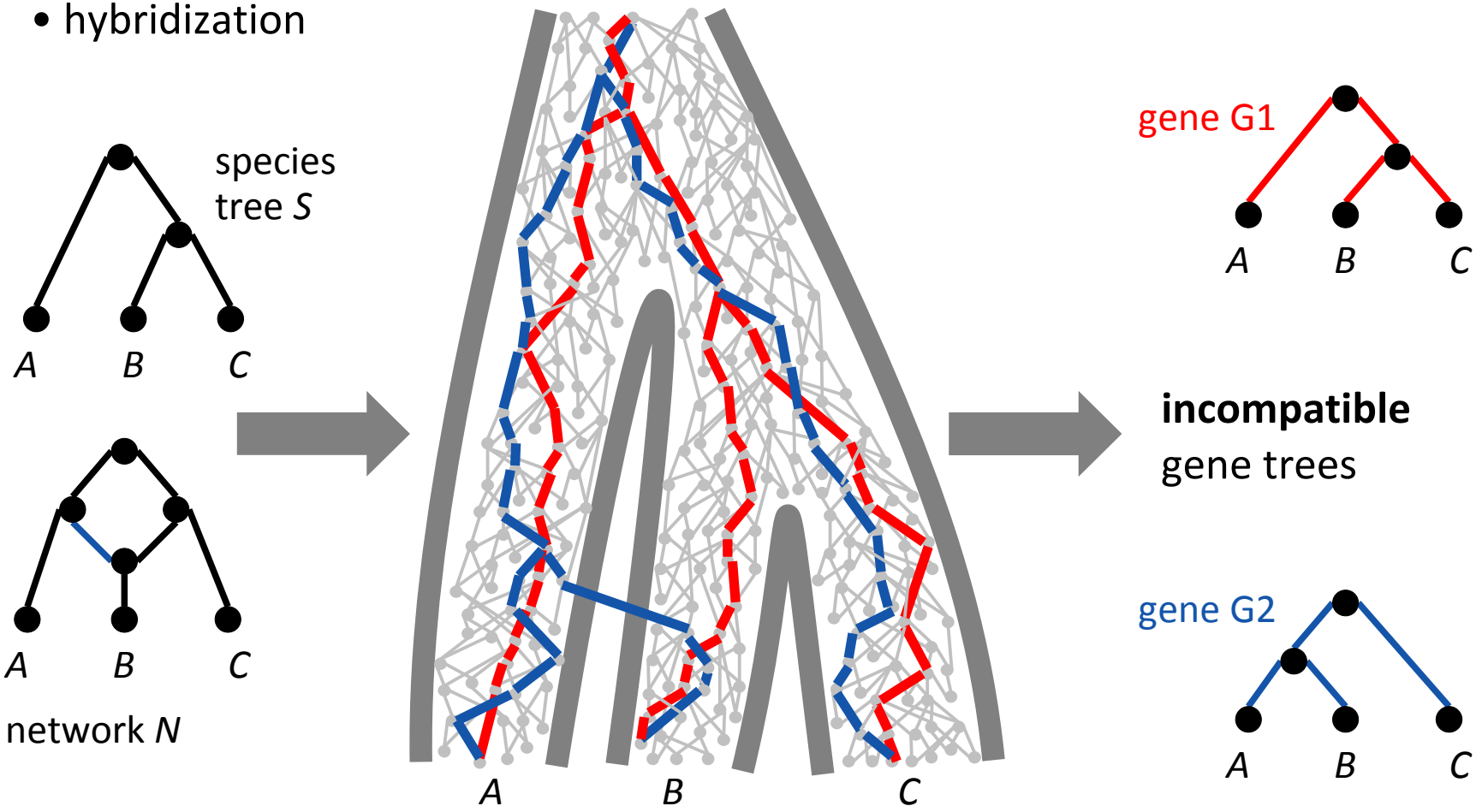
- horizontal gene transfer
- hybridization



Genetic material transfer

Genetic material transfers between coexisting species:

- horizontal gene transfer
- hybridization

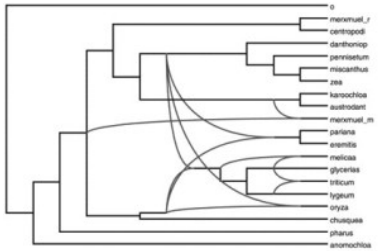


Phylogenetic networks

Phylogenetic network: network representing evolution data

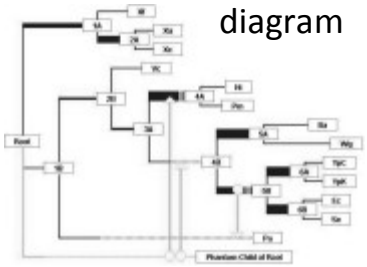
- **explicit** phylogenetic networks

to **model** evolution



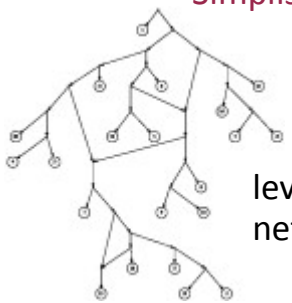
galled network

Dendroscope



synthesis diagram

HorizStory



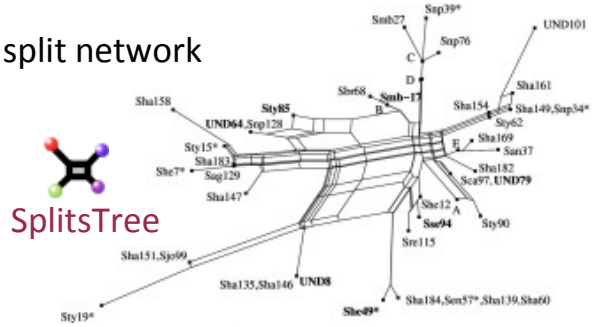
Simplistic

level-2 network

- **abstract** phylogenetic network

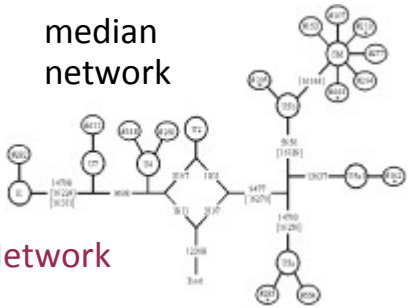
to **classify**, **visualize** data

split network



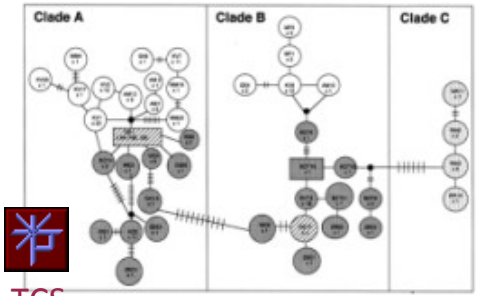
SplitsTree

median network



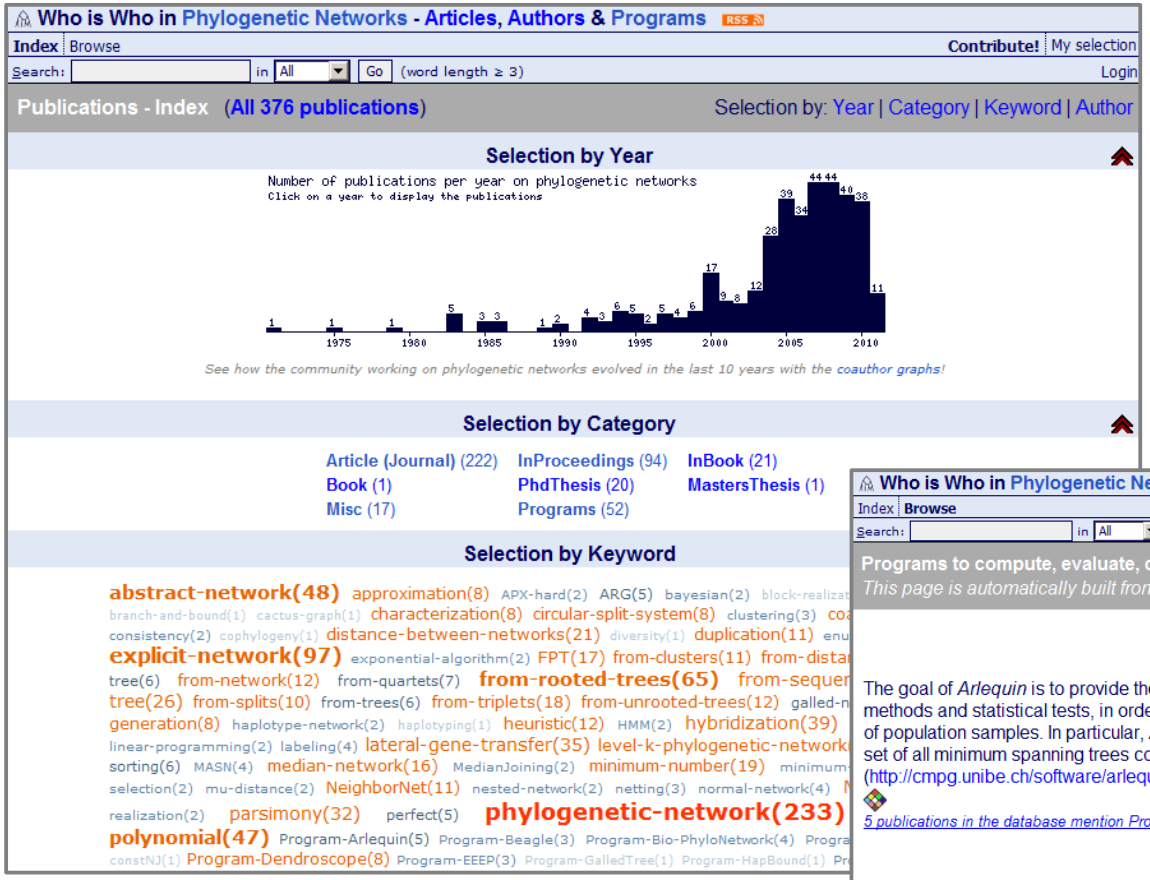
Network

minimum covering network



TCS

Phylogenetic network software



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.

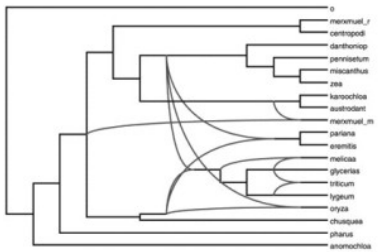
http://www.atgc-montpellier.fr/phylnet

Explicit phylogenetic networks

Phylogenetic network: network representing evolution data

- **explicit** phylogenetic networks

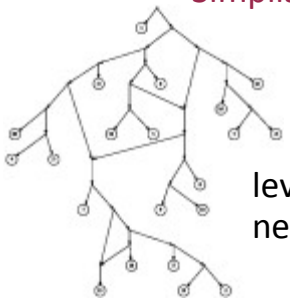
evolution model



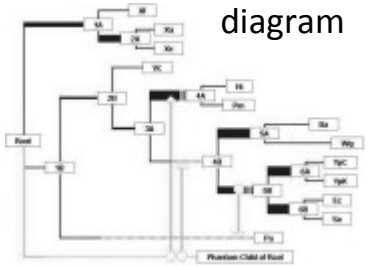
galled network

Dendroscope

Simplistic

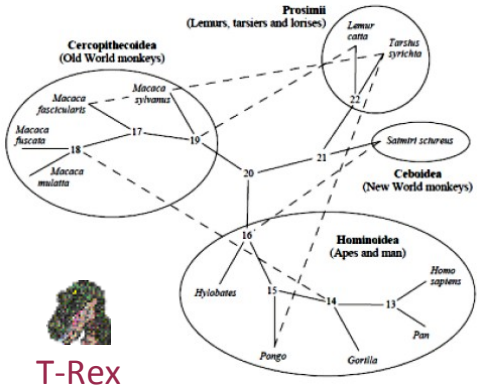


level-2 network



synthesis diagram

HorizStory

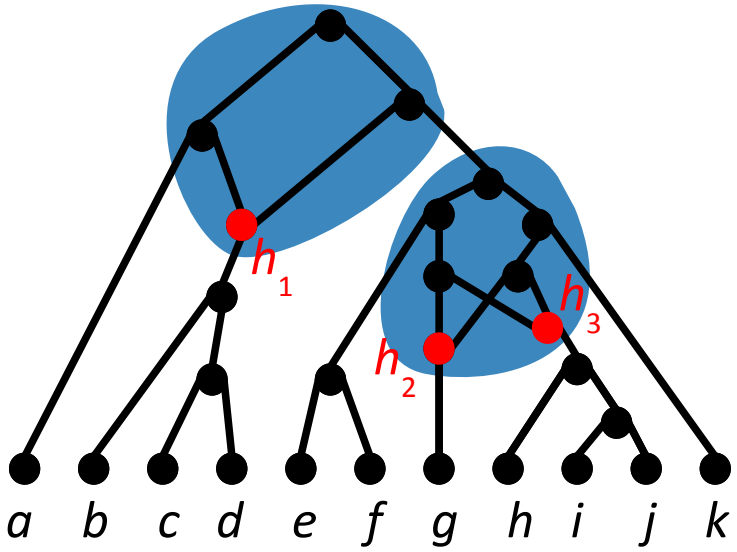


T-Rex

reticulogram

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

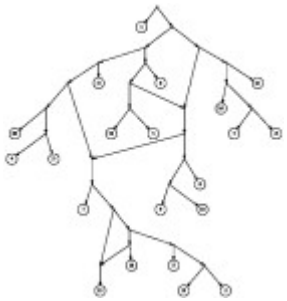
Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002

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 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

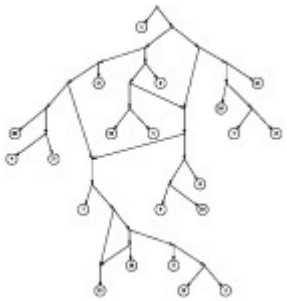
Bandelt & Dress 1992 - Legendre & Makarenkov 2000 - Bryant & Moulton 2002

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

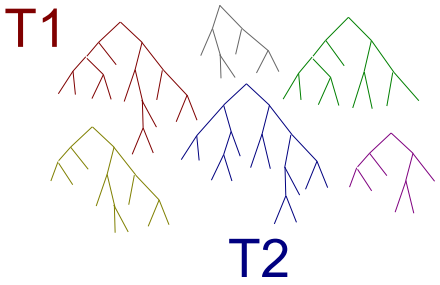
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{gene sequences}

Reconstruction of one tree for each gene present in several species

Guindon & Gascuel, SB, 2003



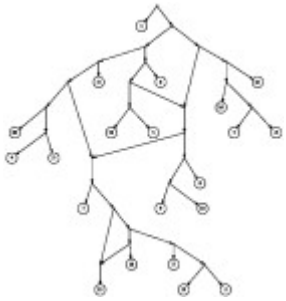
{trees}

HOGENOM database

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

Tree consensus or reconciliation

rooted explicit network



optimal network *N*

Phylogenetic network reconstruction

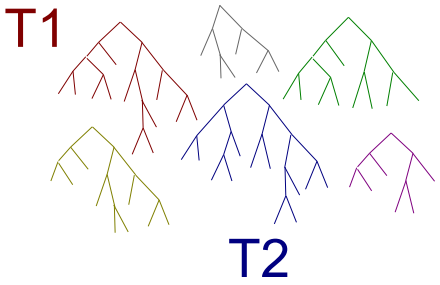
```
espèce 1 : AATTGCAG TAGCCCAAAAT
espèce 2 : ACCTGCAG TAGACCAAT
espèce 3 : GCTTGCCG TAGACAAGAAT
espèce 4 : ATTTGCAG AAGACCAAAT
espèce 5 :          TAGACAAGAAT
espèce 6 : ACTTGCAG TAGCACAAAAT
espèce 7 : ACCTGGTG TAAAAT
```

G1 G2

{gene sequences}

Reconstruction of one tree for each gene present in several species

Guindon & Gascuel, SB, 2003



{trees}

HOGENOM database

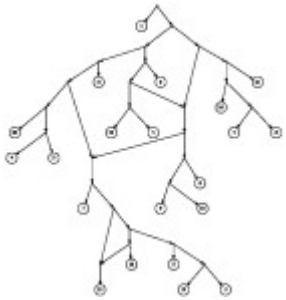
Phyl-ARIANE ANR

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

> 500 species, >70 000 trees

Tree consensus or reconciliation

rooted explicit network



optimal network *N*

Problem: tree reconciliation is difficult even for 2 trees

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

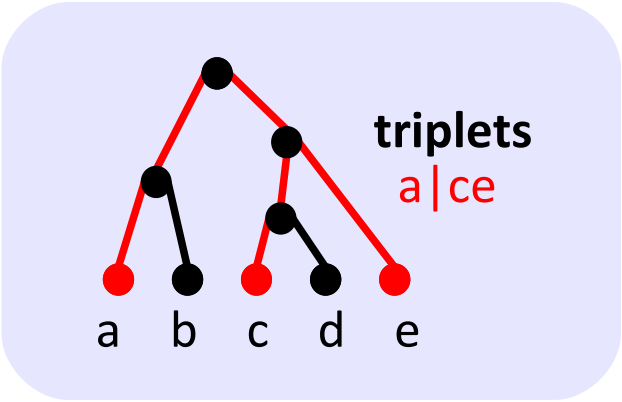
Bordewich & Semple, DAM, 2007

Triplets and clusters

Problem:
Reconstructing the **supernetwork** of a set of trees is

hard.

Idea:
reconstruct a network containing all:



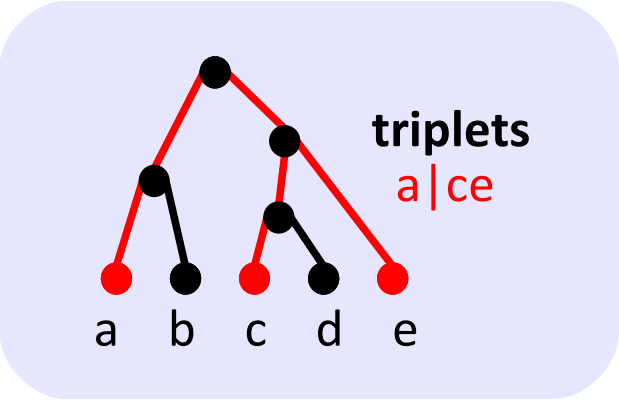
of the input trees ?

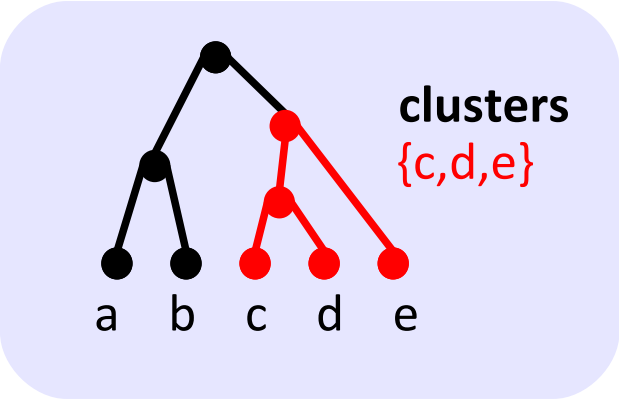
Triplets and clusters

Problem:
Reconstructing the **supernetwork** of a set of trees is

hard.

Idea:
reconstruct a network containing all:



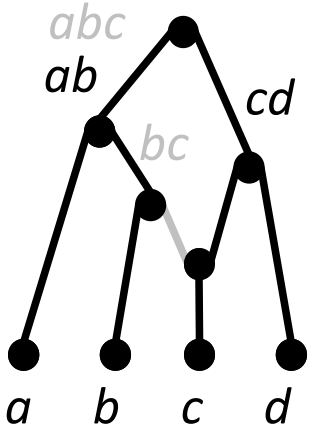


of the input trees ?

Softwired clusters

"softwired" cluster : cluster of a tree contained in the network

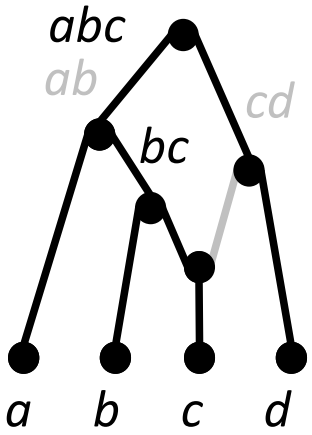
Tree-like model of gene transmission:
each gene comes from a single parent



Softwired clusters

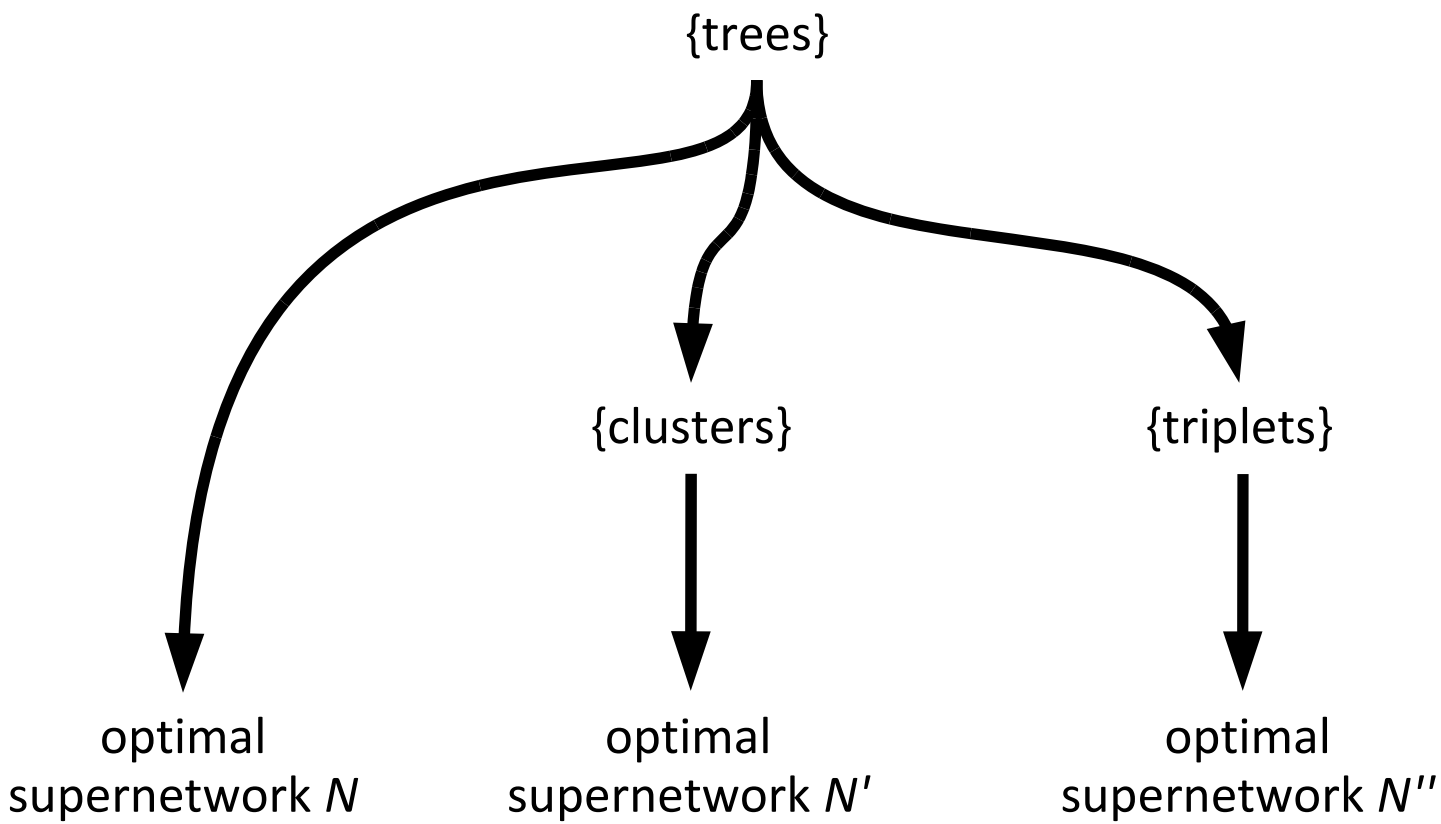
"softwired" cluster : cluster of a tree contained in the network

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



Combinatorial phylogenetic network reconstruction

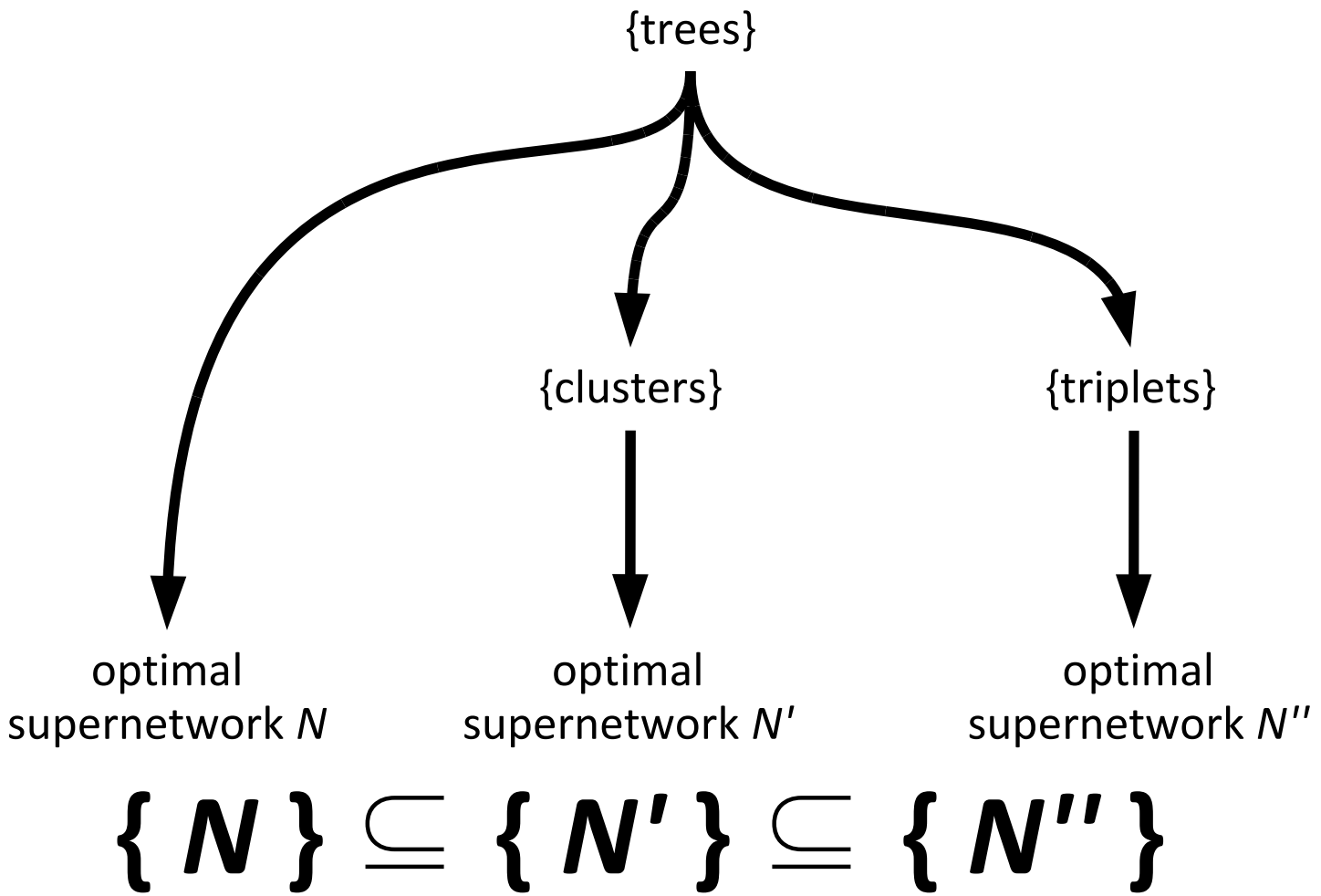
Idea:
change the type of data to process



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

Combinatorial phylogenetic network reconstruction

Idea:
change the type of data to process

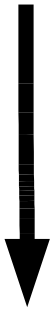


Plan

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

Reconstruction from softwired clusters

{trees}



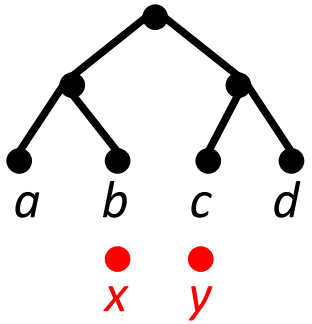
{clusters}

N'
galled network

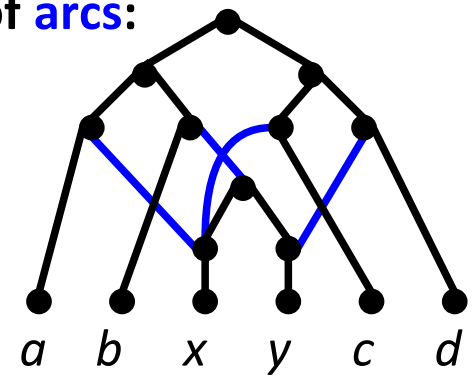
Fast exact **galled network** reconstruction method from **softwired clusters**

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}

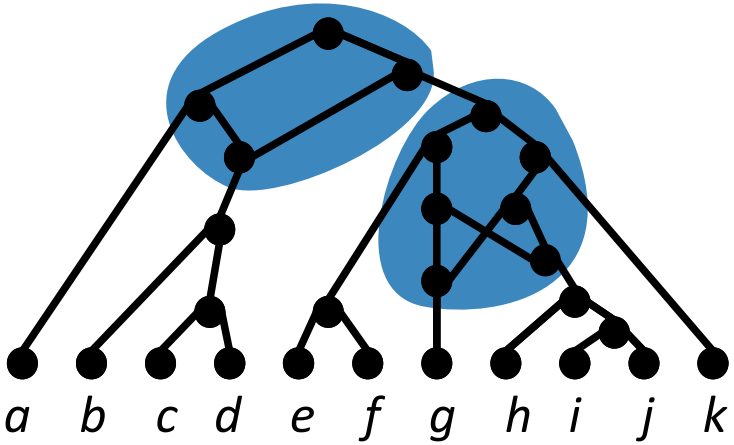
{clusters}

N'
level-*k*
network

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

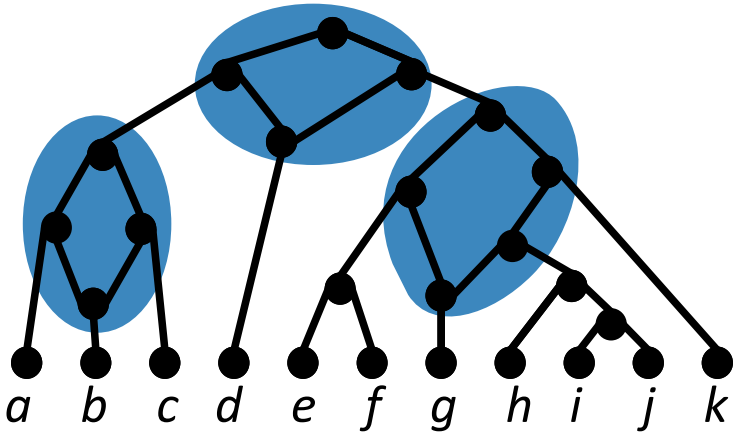
Iersel, Kelk, Rupp & Huson, ISMB 2010

less reticulations, but slower for level > 2.



level-2 network

level =
maximum number of reticulations per blob.

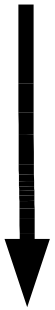


level-1 network
(“*galled tree*”)

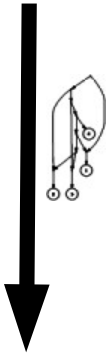
http://www.dendroscope.org

Reconstruction from triplets

{trees}



{triplets}



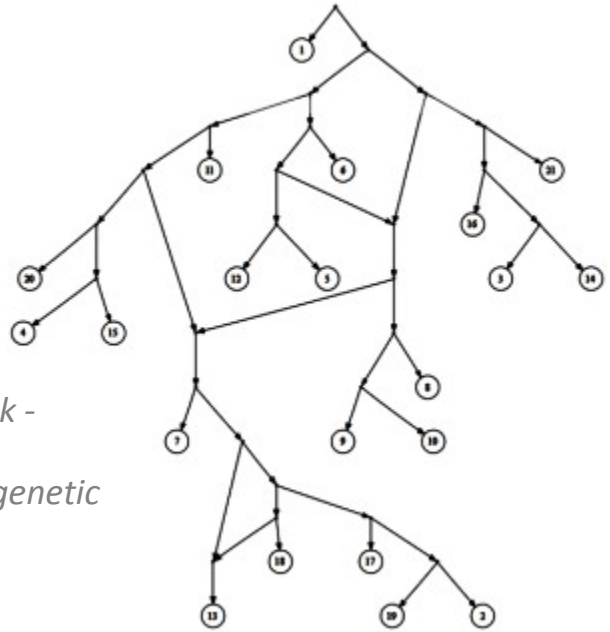
N'
level-*k*
network

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

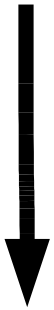


Yeast phylogenetic network -
Van Iersel et al. :
Constructing level-2 phylogenetic networks from triplets.
RECOMB 2008

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

Reconstruction from triplets

{trees}



{triplets}

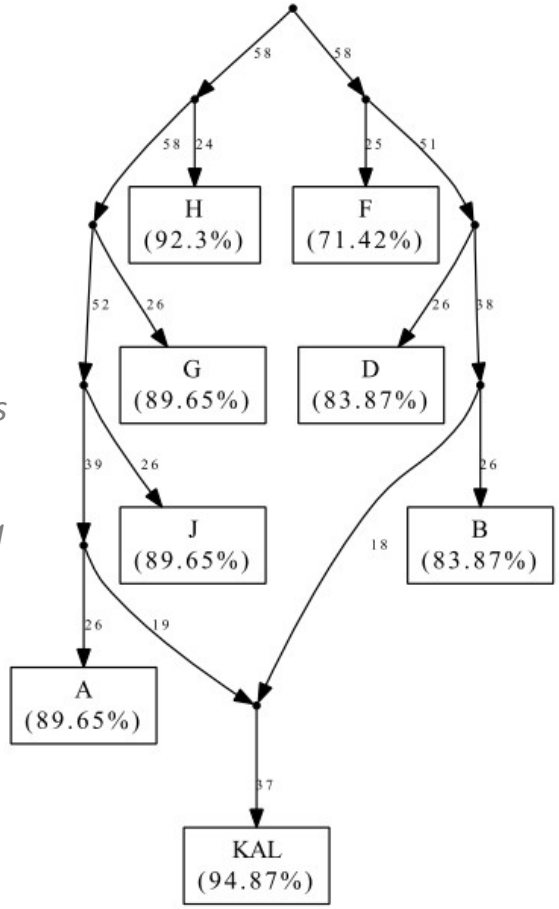
N'
level-1
network

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

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

Program Lev1athan

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



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

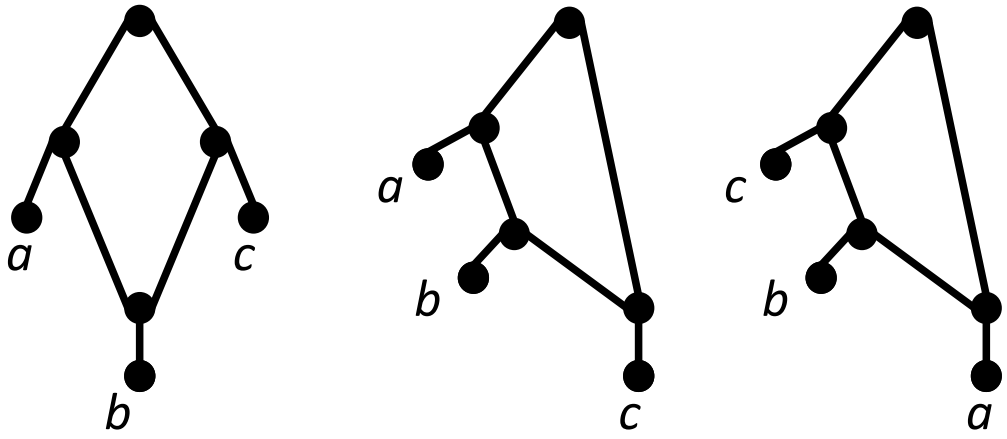
Outline

- 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)

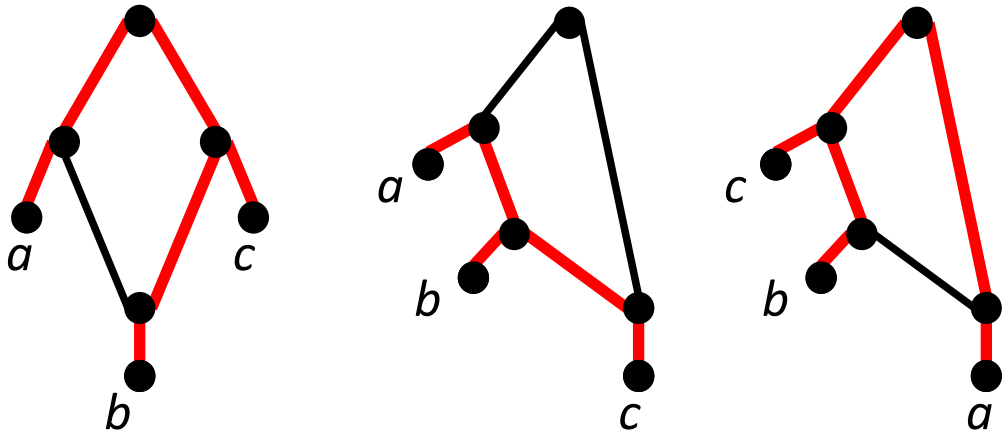
Gambette & Huber, 2011

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.

a|bc



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

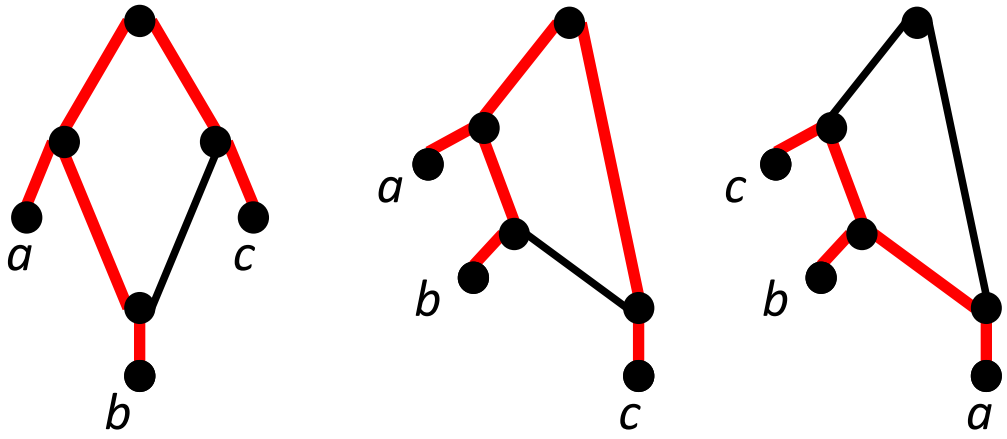
Gambette & Huber, 2011

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.

c|ab



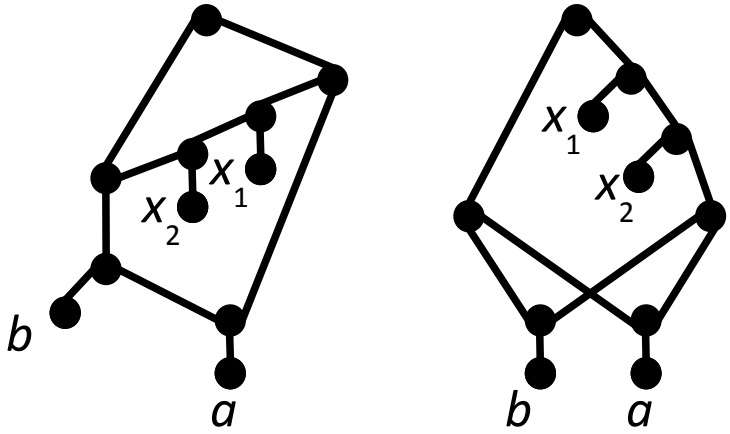
Characterization for level-1 networks :
The only ambiguous cases have such blobs (< 5 vertices)

Gambette & Huber, 2011

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.



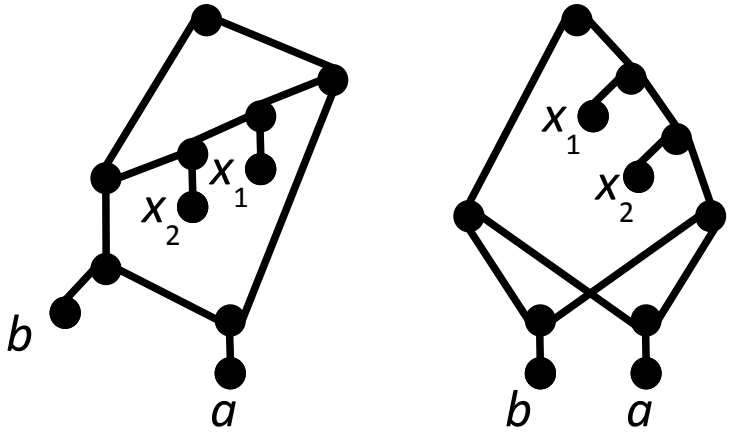
2 level-2 networks with exactly the same triplet set

Gambette & Huber, 2011

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!

Gambette & Huber, 2011

Practical use

existing methods
still work to do!

Conditions for use	Available data	Possible processings
rooted trees	unrooted trees	*Rooting with a reference species tree or topological constraints*
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) *Data editing : solution containing most of the input data*
Complete data (density for triplet sets, complete clusters)	partial data, deleted genes	*Selection of a large number of trees with a large number of common species* *Selection of the maximal number of taxa with triplet density* NP-complete problems

Outline

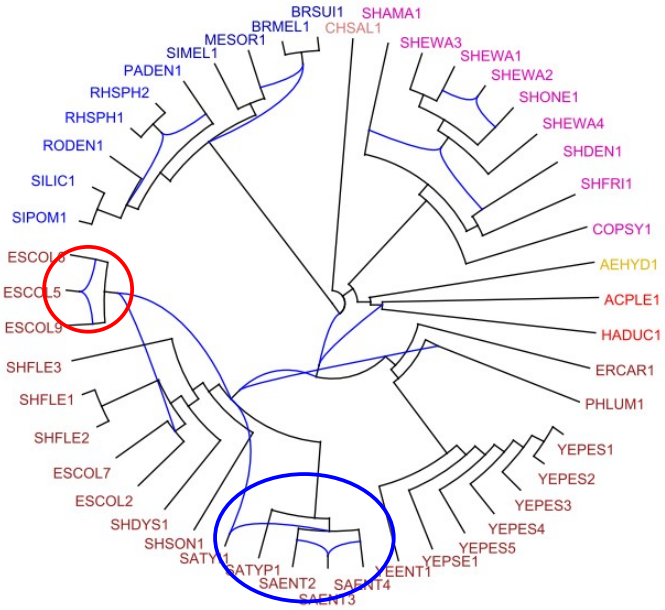
- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- Combinatorial reconstruction methods
- Practical use
- Illustrations
- Perspectives

Illustrations

16 trees on 47 taxa from the HOGENOM database (proteobacteria)

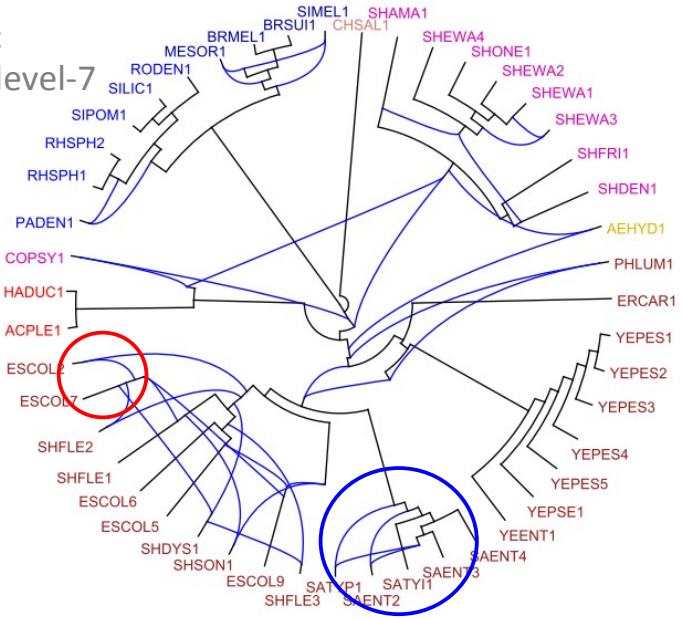
24 Enterobacteriales
2 Pasteurellales
1 Aeromonadales
9 Alteromonadales
1 Oceanospirillales
6 Rhodobacterales
4 Rhizobiales

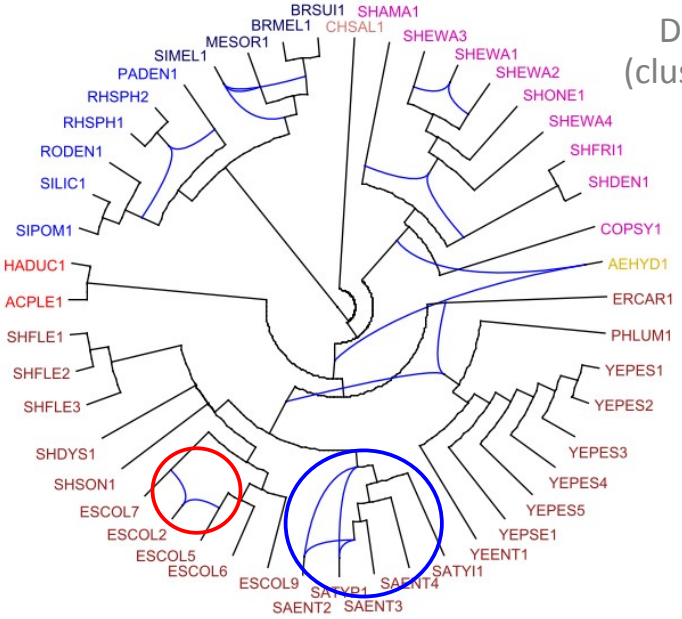
Networks containing triplets, softwired clusters, present in at least 20% of the trees



Lev1athan (heuristic triplets, level-1) 24 sec.

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





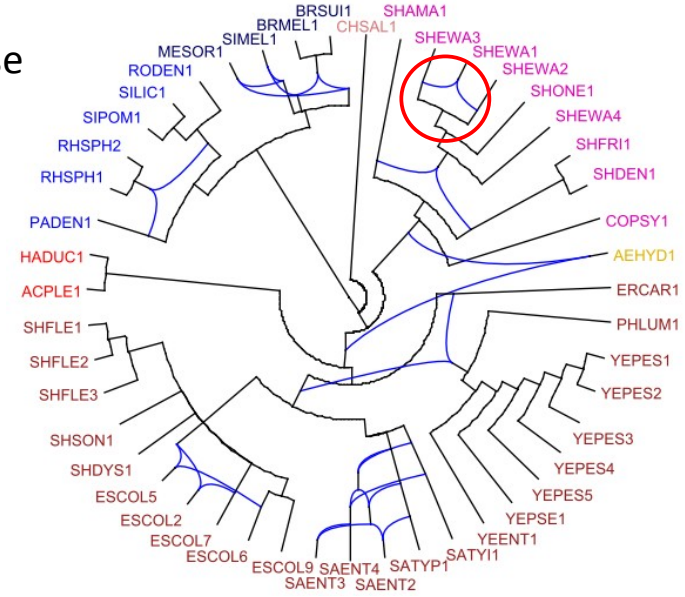
Dendroscope (clusters, galled network) <1 sec.

Illustrations

16 trees on 47 taxa from the HOGENOM database (proteobacteria)

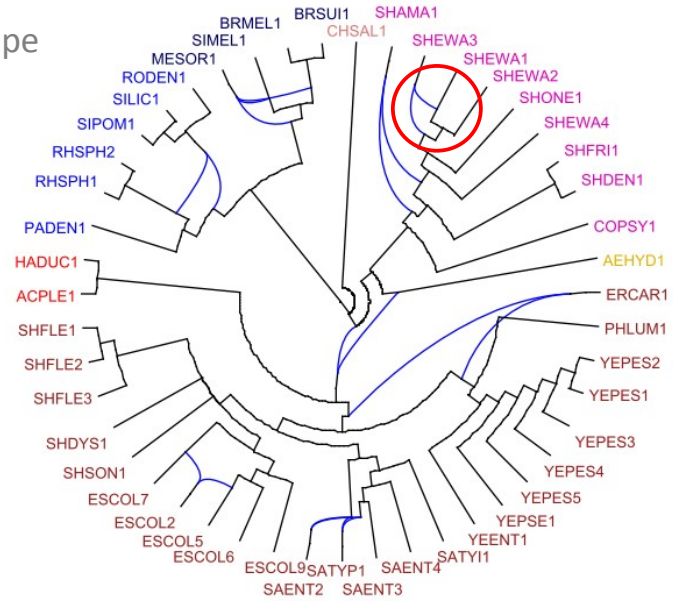
24 Enterobacteriales
2 Pasteurellales
1 Aeromonadales
9 Alteromonadales
1 Oceanospirillales
6 Rhodobacterales
4 Rhizobiales

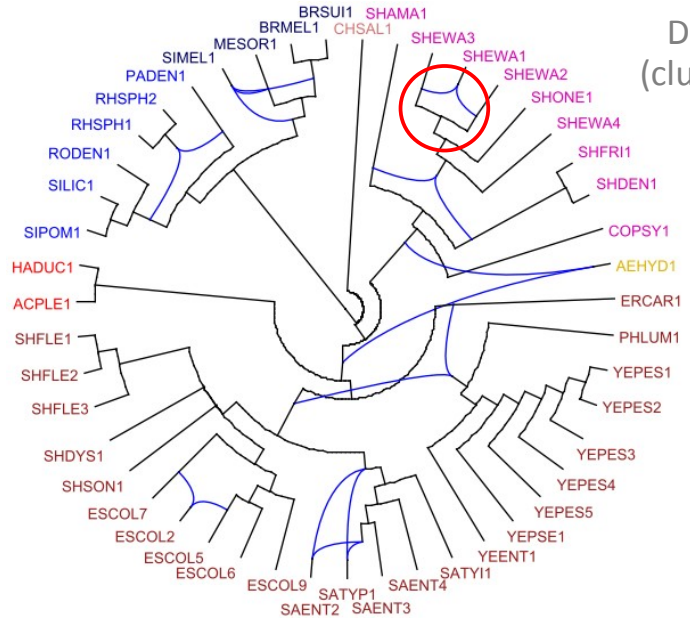
Networks containing triplets, softwired clusters, present in at least 20% of the trees



Dendroscope (clusters, cluster network, level 1) <1 sec.

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



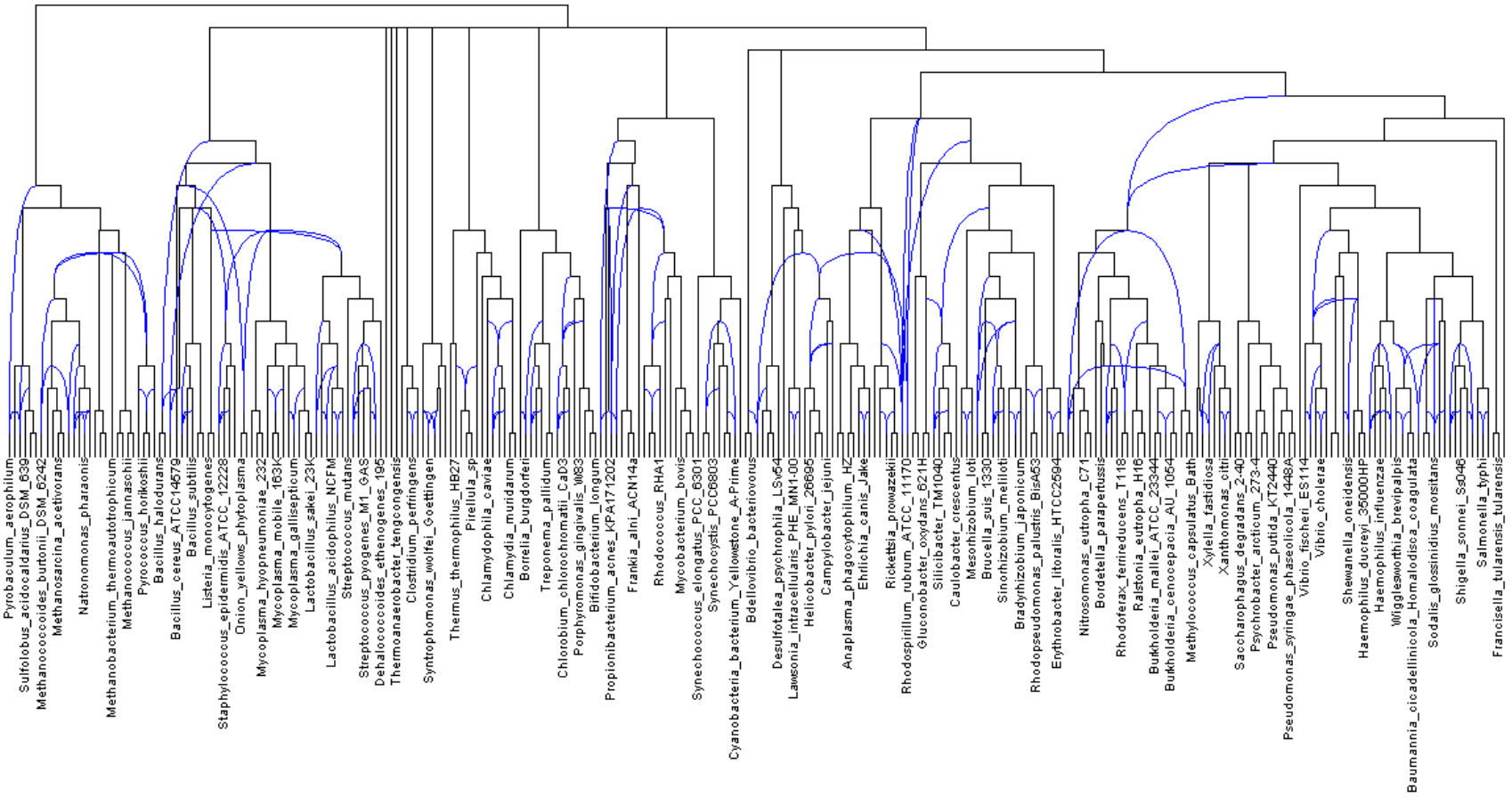


Dendroscope (clusters,galled network) <1 sec.

Illustrations

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

Auch, Steigele, Huson & Henz, 2009

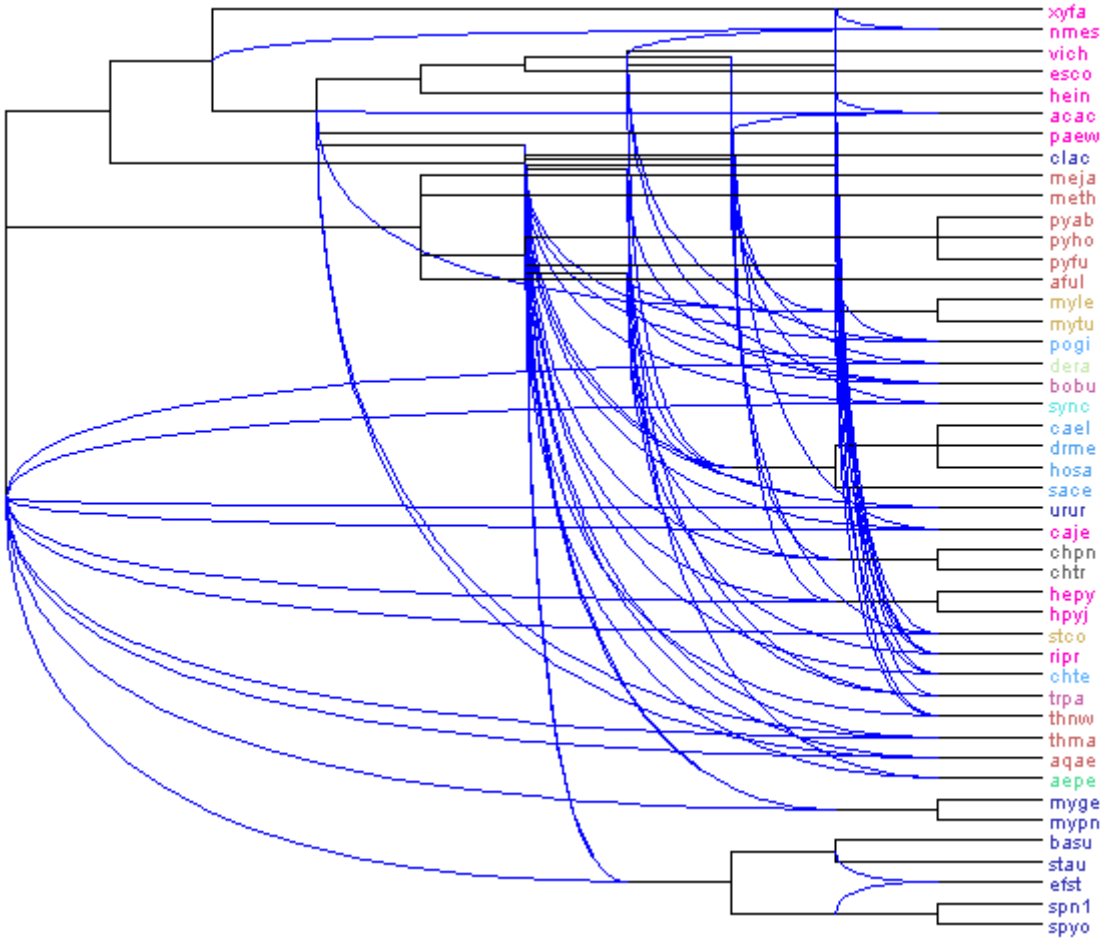


Dendroscope
(clusters, galled network)
2 sec.

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.

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

Illustrations

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

Dendroscope (level-3 network) <1 sec.

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

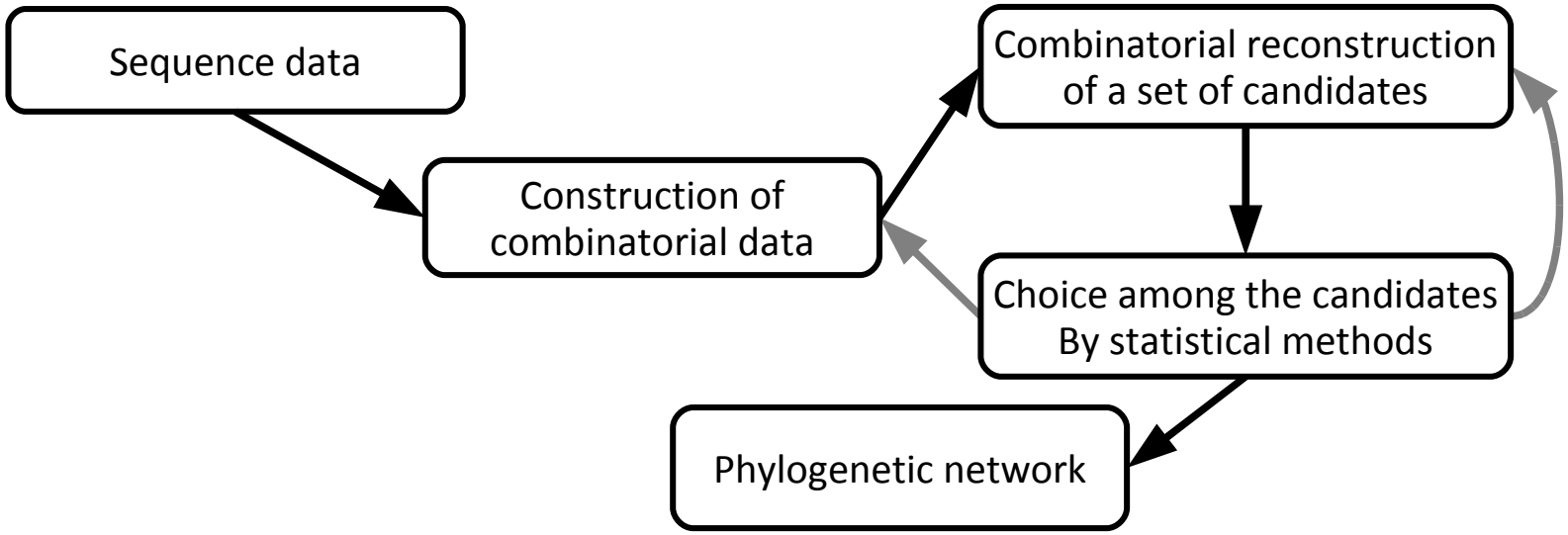
Outline

- 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

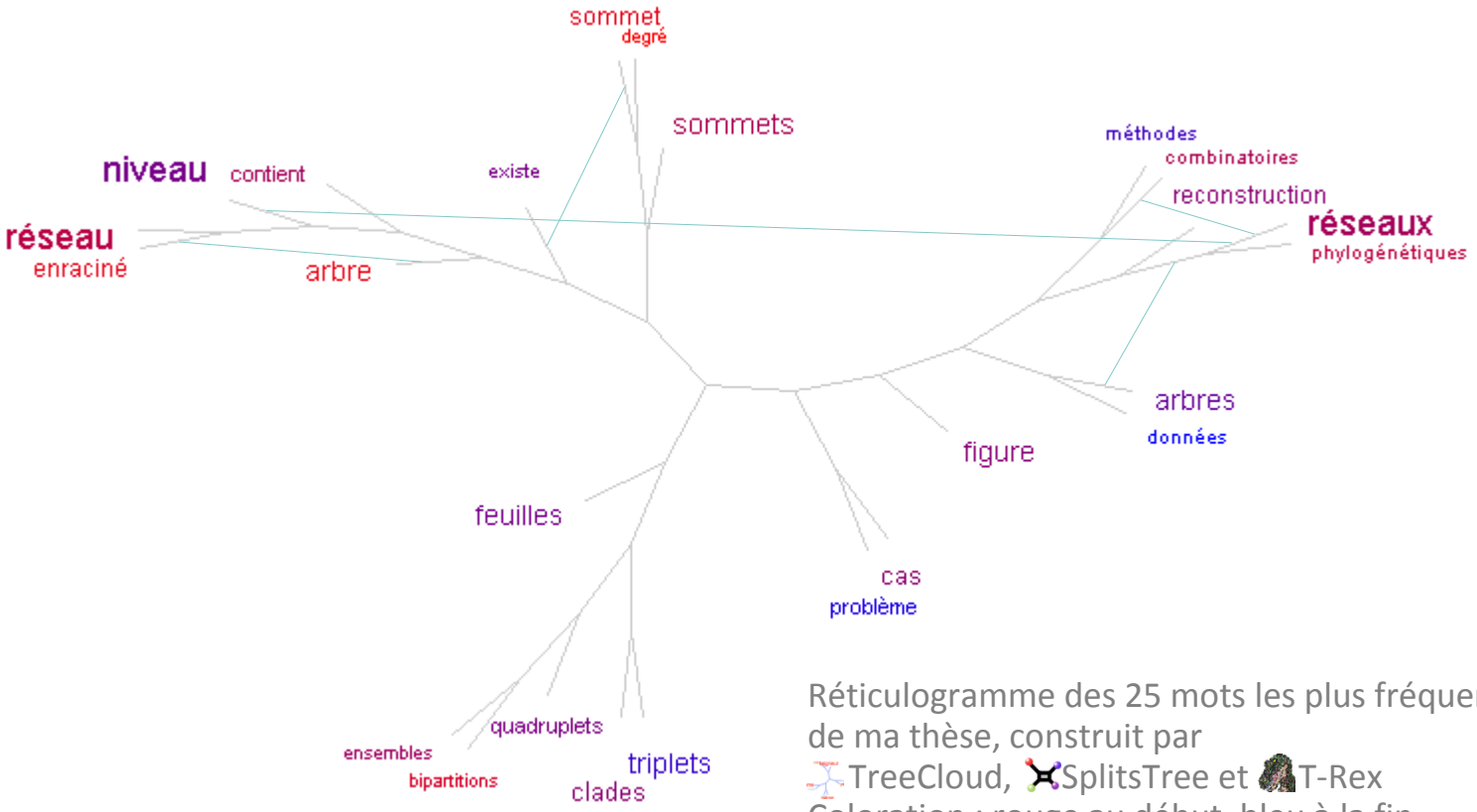
Bioinformatics :
- Function of transferred genes (“transfer highways”)
- Integration of combinatorial methods in a statistical framework



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

http://www.treecloud.org