Evènements d'évolution réticulée: quand les arbres deviennent des réseaux

Céline Scornavacca

ZBIT- Centre pour la Bioinformatique de Tübingen
Université de Tübingen
http://www-ab.informatik.uni-tuebingen.de/people/scornavacca

Séminaire équipe SaAB

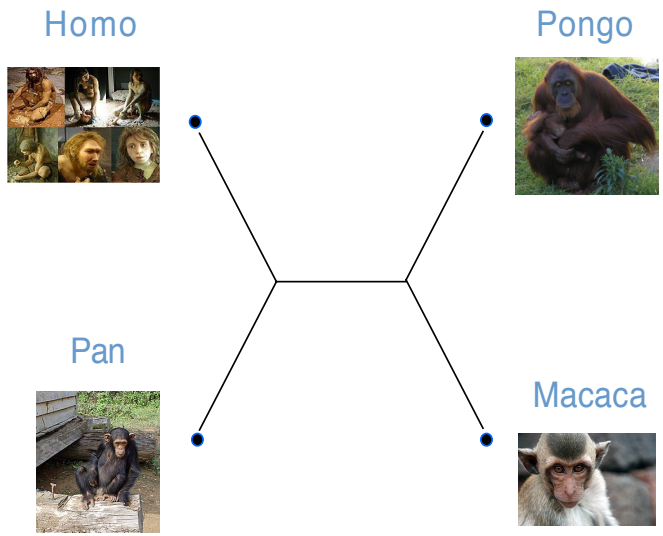
31 mai 2011

Outline :

- Definition of a phylogenetic tree
- Definition of a phylogenetic network
- Overview of types of phylogenetic networks
- Unrooted phylogenetic networks
- Rooted phylogenetic networks

Phylogenetic trees

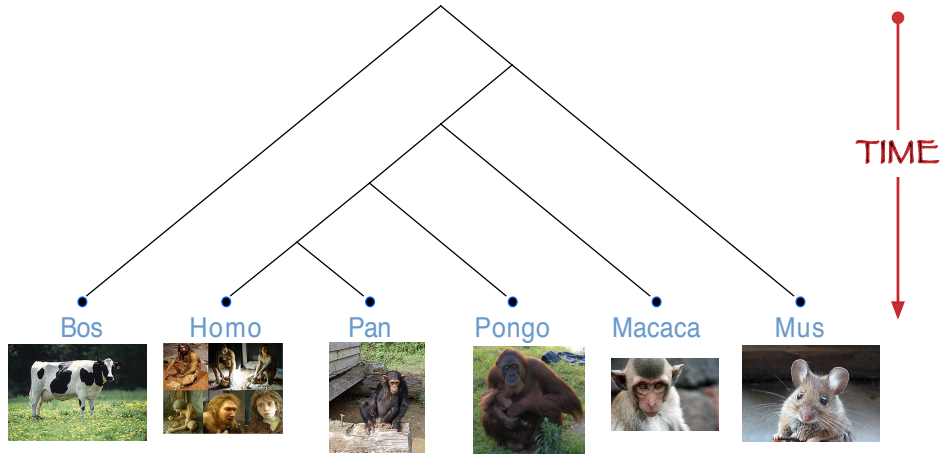
connected and acyclic graphs, where terminal nodes are associated to a set of species.



Rooted phylogenetic trees

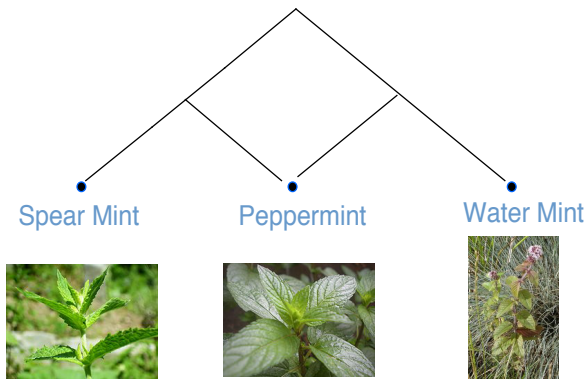
oriented, connected and acyclic graphs, where terminal nodes are associated to a set of species.

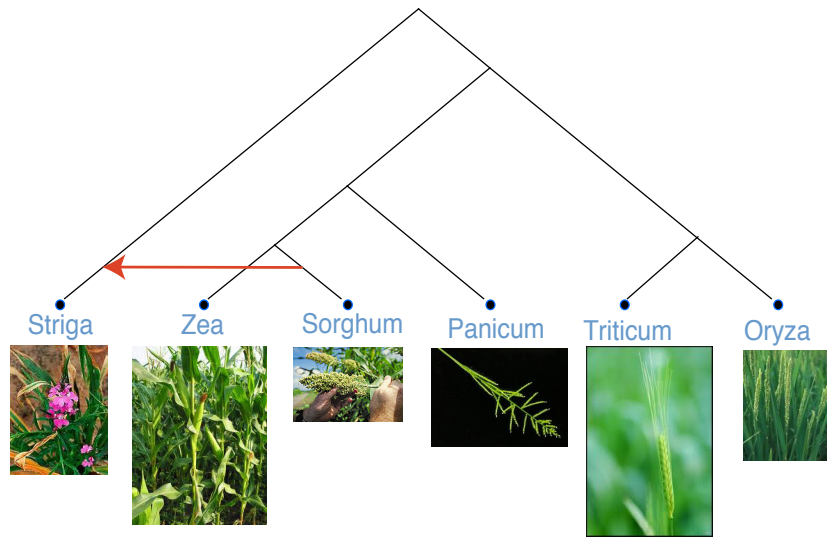
- the leaves or taxa represent extant organisms
- internal nodes represent hypothetical ancestors
- the only node without ancestor is called root
- each internal node represents the lowest common ancestor of all taxa below it (cluster)



But...

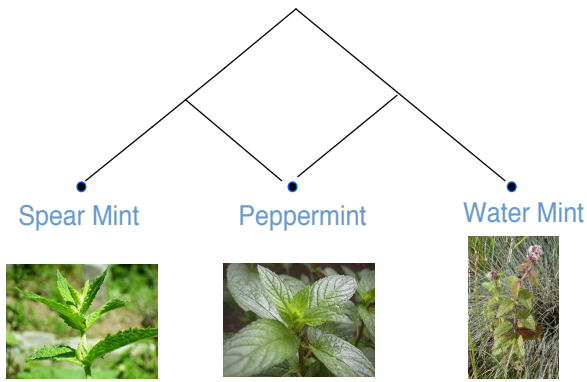
due to reticulate evolutionary phenomena (hybridization, recombination, horizontal gene transfer) the evolution of a set of species sometimes cannot be described using phylogenetic trees.

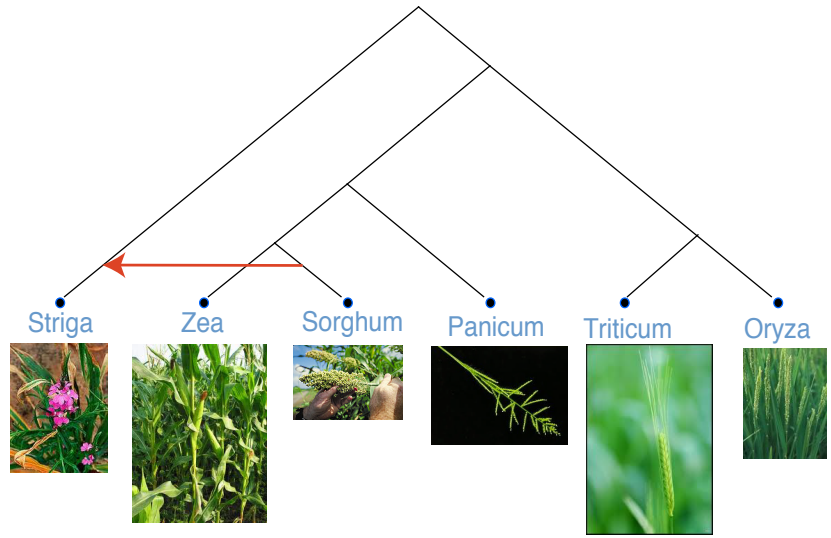




But...

due to reticulate evolutionary phenomena (hybridization, recombination, horizontal gene transfer) the evolution of a set of species sometimes cannot be described using phylogenetic trees.



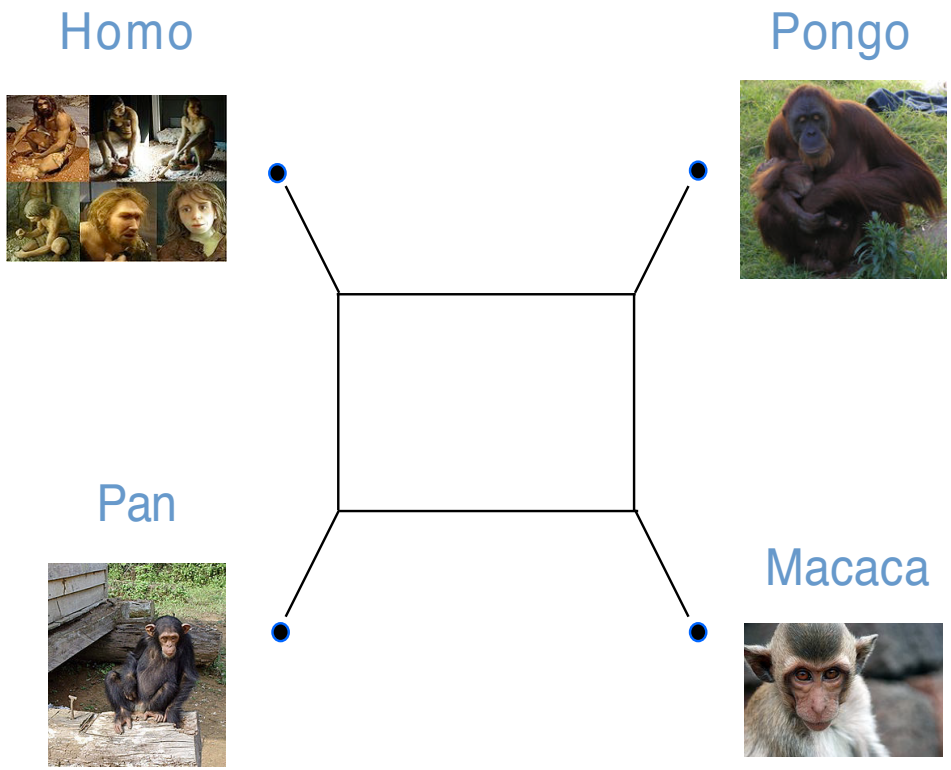


in these cases we use ...

phylogenic networks

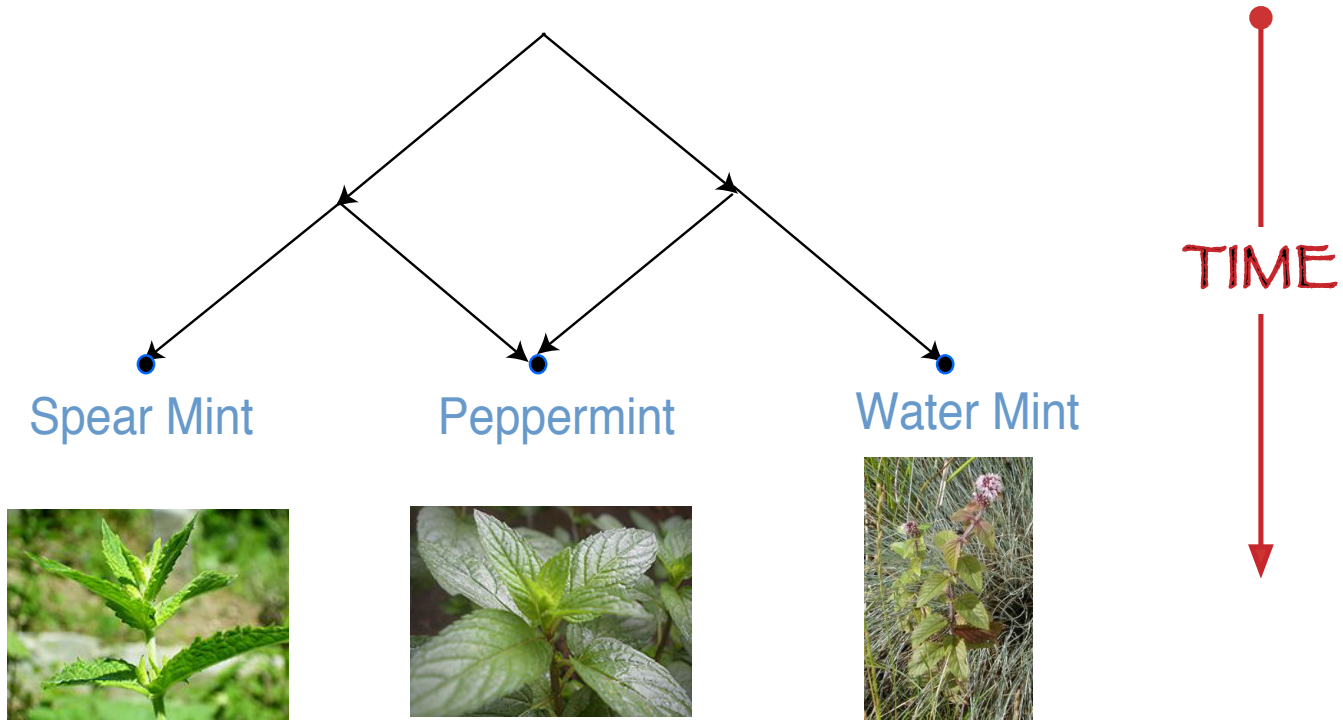
Phylogenetic networks

any connected graph, where terminal nodes are associated to a set of species.



Rooted phylogenetic networks

any rooted directed acyclic graph, where terminal nodes are associated to a set of species.



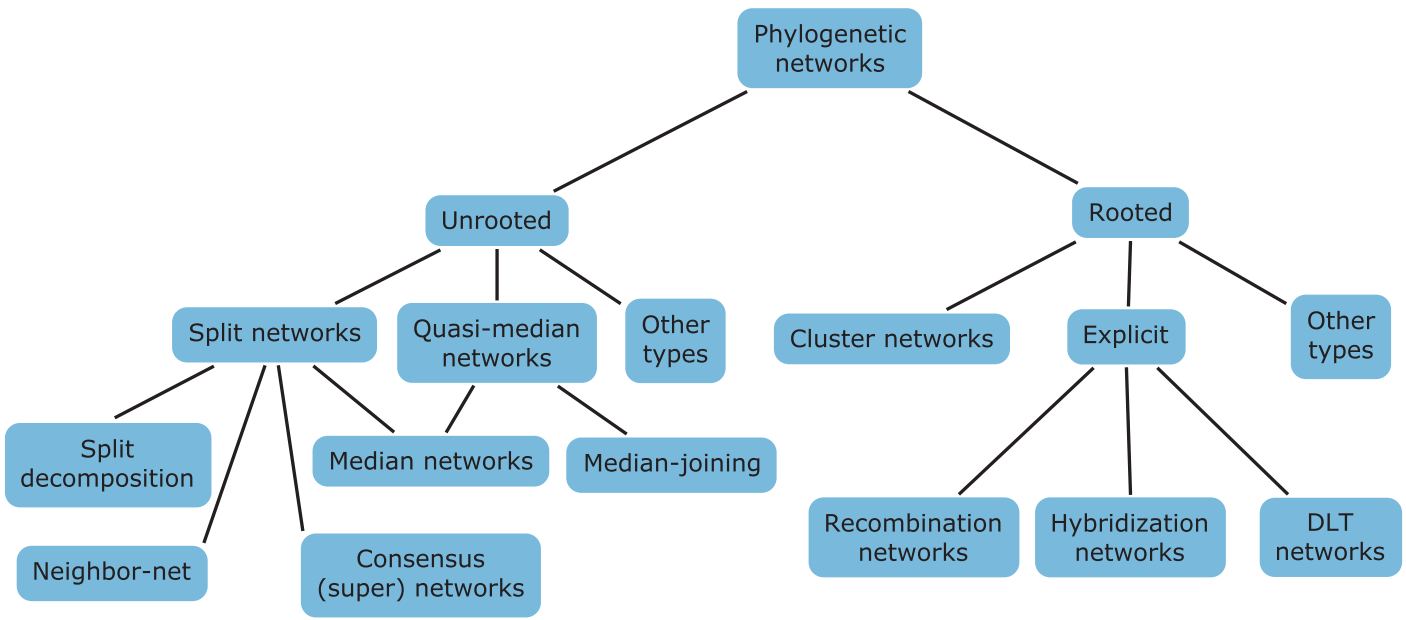
Phylogenetic networks

Abstract networks :

Visualize conflicting signals (also called data-display networks)

Explicit networks :

Show evolutionary scenario involving reticulate events (also called evolutionary networks)



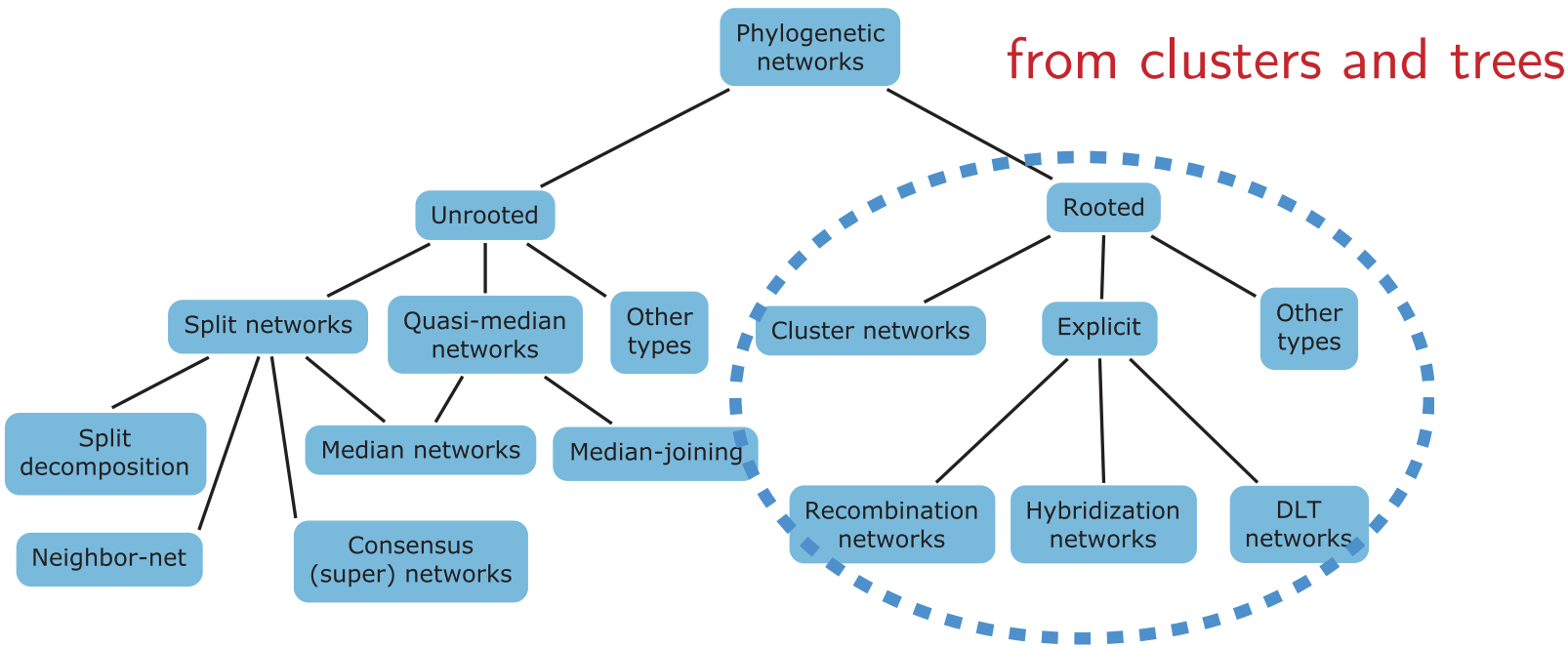
Phylogenetic networks

Abstract networks :

Visualize conflicting signals (also called data-display networks)

Explicit networks :

Show evolutionary scenario involving reticulate events (also called evolutionary networks)

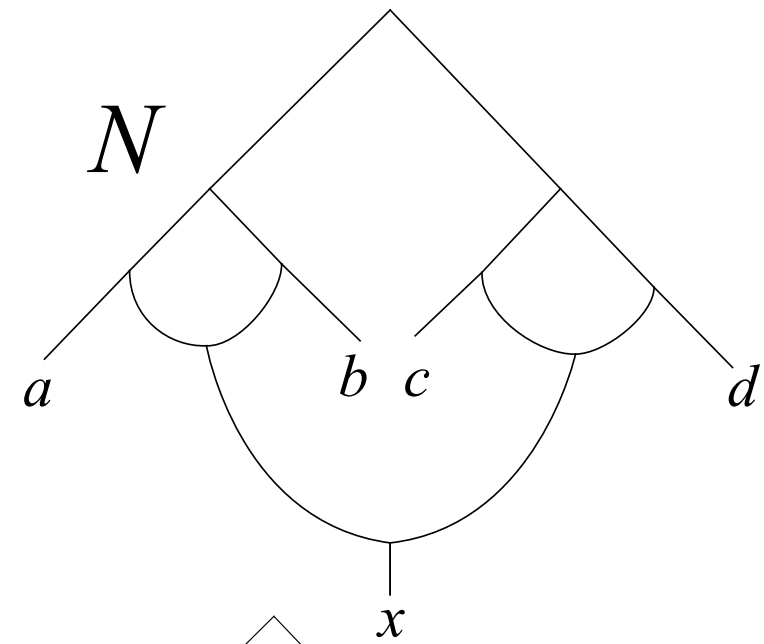


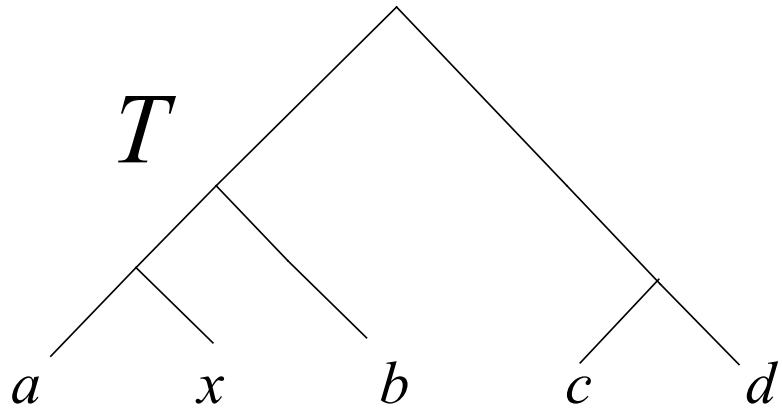
When a phyl. network N represents a tree T ?

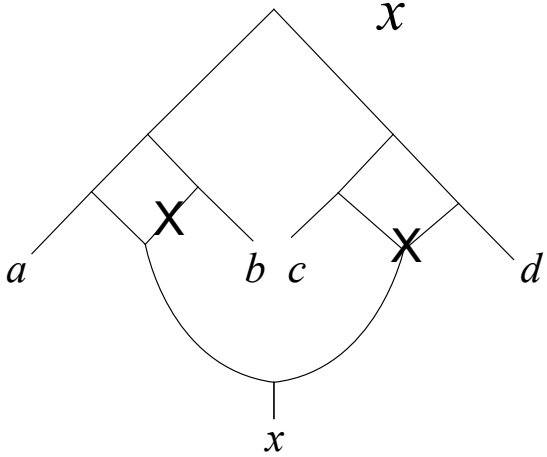
if T can be obtained from N by performing a series of node deletions, edge deletions and node suppressions

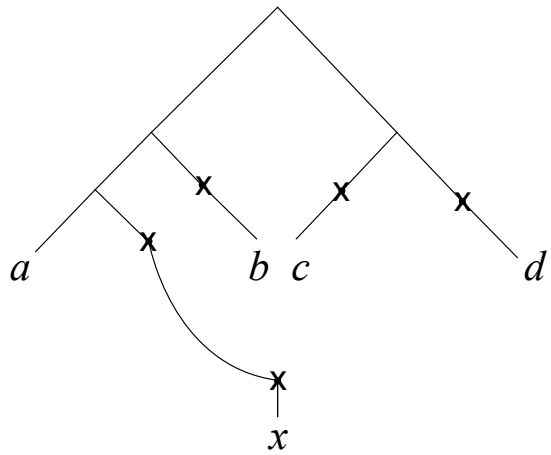
When a phyl. network N represents a tree T ?

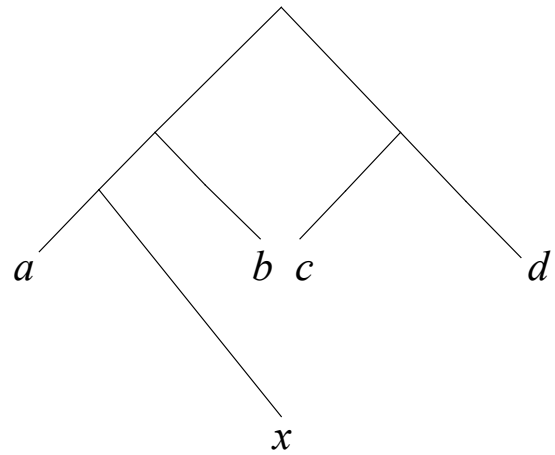
if T can be obtained from N by performing a series of node deletions, edge deletions and node suppressions





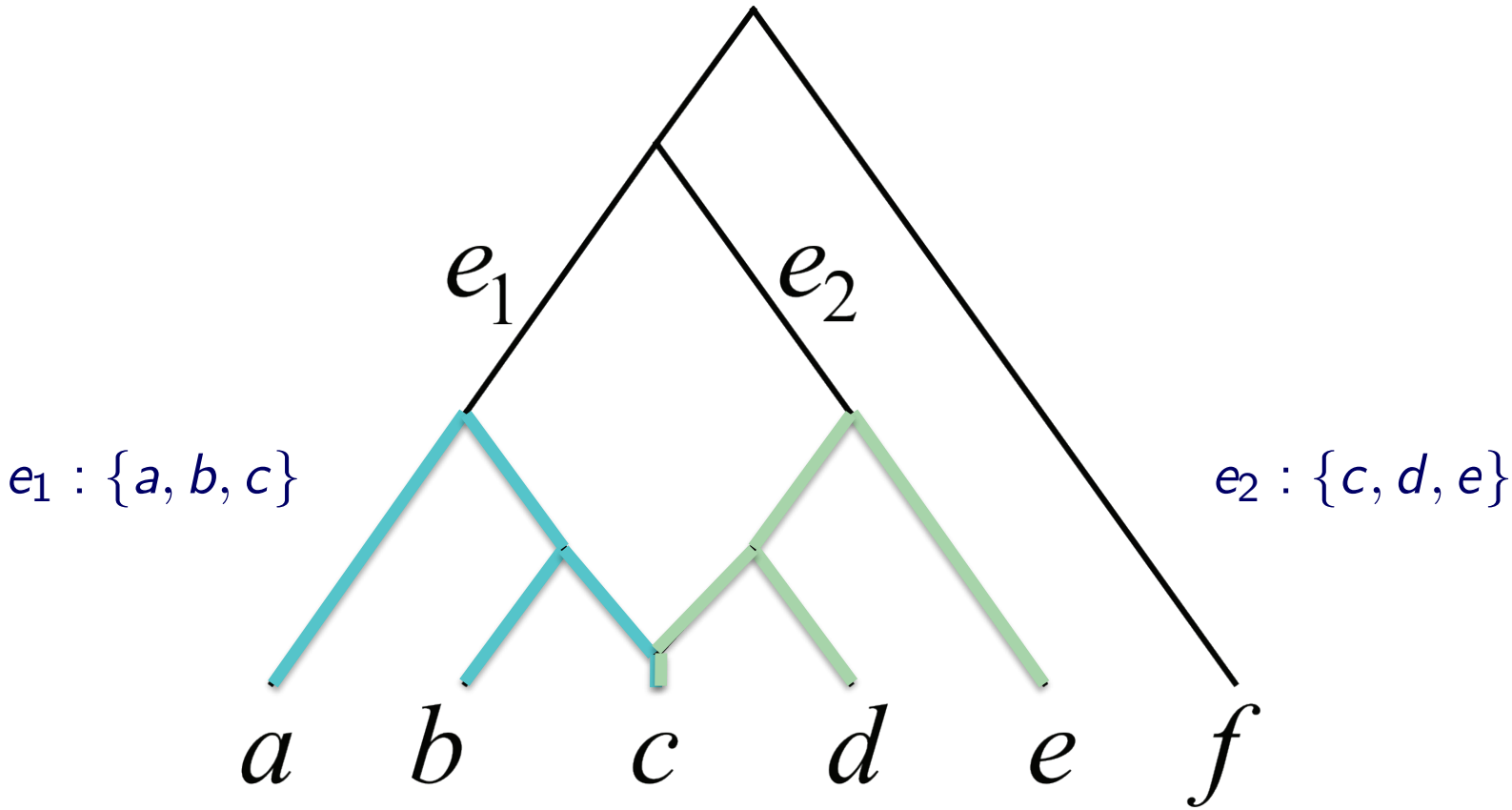






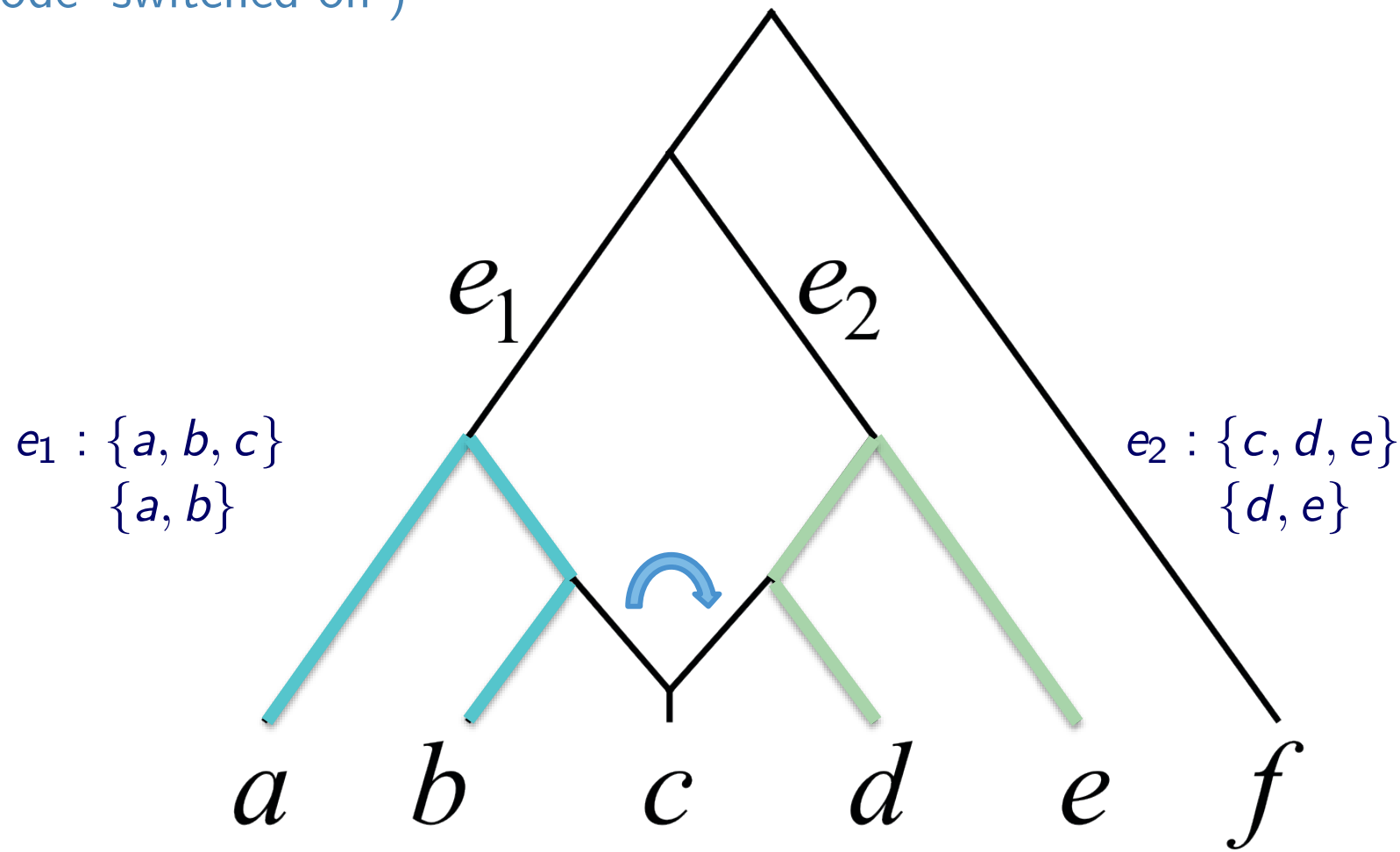
When a phyl. network N represents a cluster C ?

HARDWIRED SENSE : if there exists a tree edge of N such that the set of all taxa below the edge equals C



When a phyl. network N represents a cluster C ?

SOFTWIRED SENSE : if there exists a tree edge of N such that the set of all taxa below the edge equals C (with one edge per reticulation node "switched on")

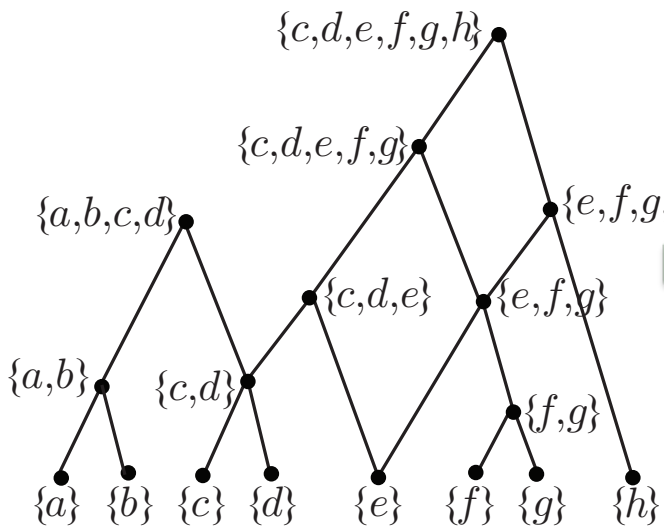


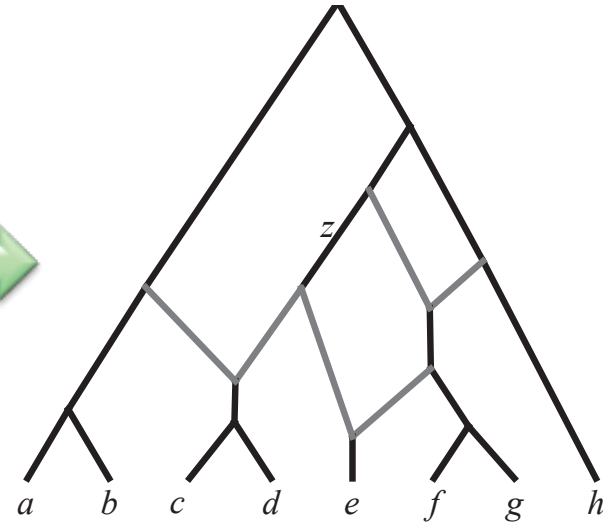
Networks form clusters

Constructing minimal hardwired networks

cluster popping algorithm

{c,d,e,f,g,h}
{c,d,e,f,g}
{e,f,g,h}
{c,d,e}, {e,f,g}
{a,b,}, {c,d}, {f,g}
{a}, {b}, ..., {h}





Clusters **Hasse diagram** **Network**

number of edges required ...

quadratic in $|C|$

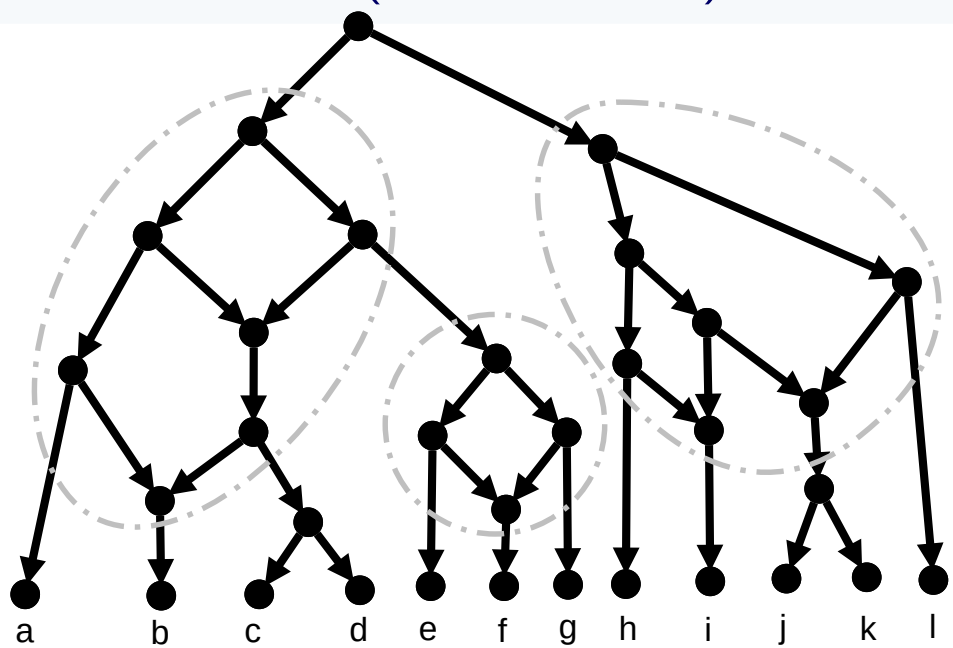
Constructing minimal softwired networks

- cluster containment : NP-hard
- minimization : NP-hard, APX-hard

A possible solution ... topological constraints :

- galled trees
- galled networks
- level-k networks : if the maximum reticulation number among the biconnected components of N is k (still NP-hard)

DECOMPOSABLE !



Constructing minimal softwired networks

- cluster containment : NP-hard
- minimization : NP-hard, APX-hard

A possible solution ... topological constraints :

- galled trees
- galled networks
- level-k networks : if the maximum reticulation number among the biconnected components of N is k (still NP-hard)

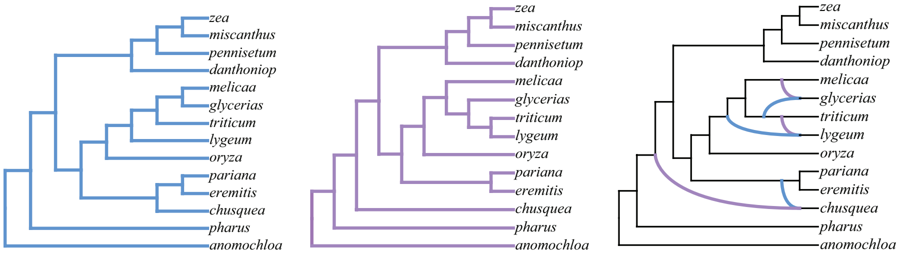
Breaking news :

- minimizing the level is FPT in k
- the CASS algorithm (van Iersel et al, 2010) is not always optimal

Networks form trees

Reconstructing hybridization networks (explicit)

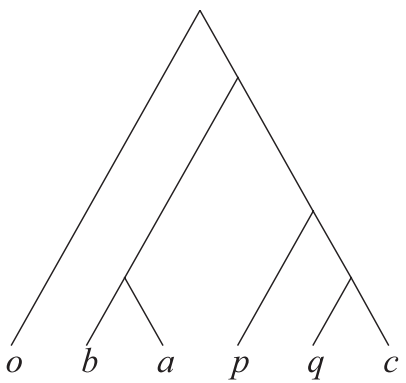
Goal : Find a phylogenetic network that displays a set of tree $\mathcal{T}$ with minimum number of reticulations (called hybrid number of $\mathcal{T}$).



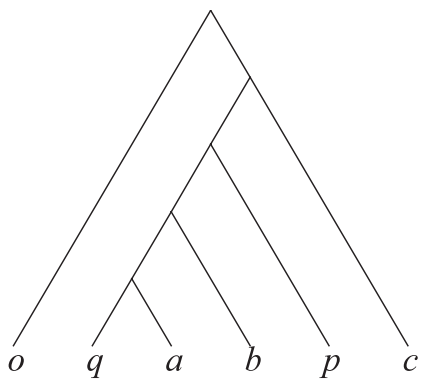
Reconstructing hybridization networks (explicit)

Goal : Find a phylogenetic network that displays a set of tree $\mathcal{T}$ with minimum number of reticulations (called hybrid number of $\mathcal{T}$).

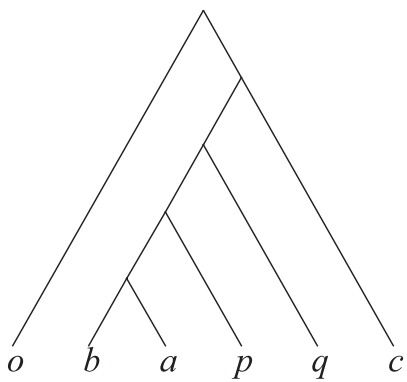
- minimization : NP-hard, FPT (via reductions)
- a networks displaying $\mathcal{C}(\mathcal{T})$ does not in general display $\mathcal{T}$



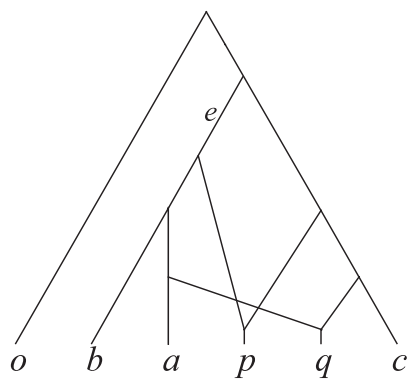
(a) Tree T_1



(b) Tree T_2



(c) Tree T_3



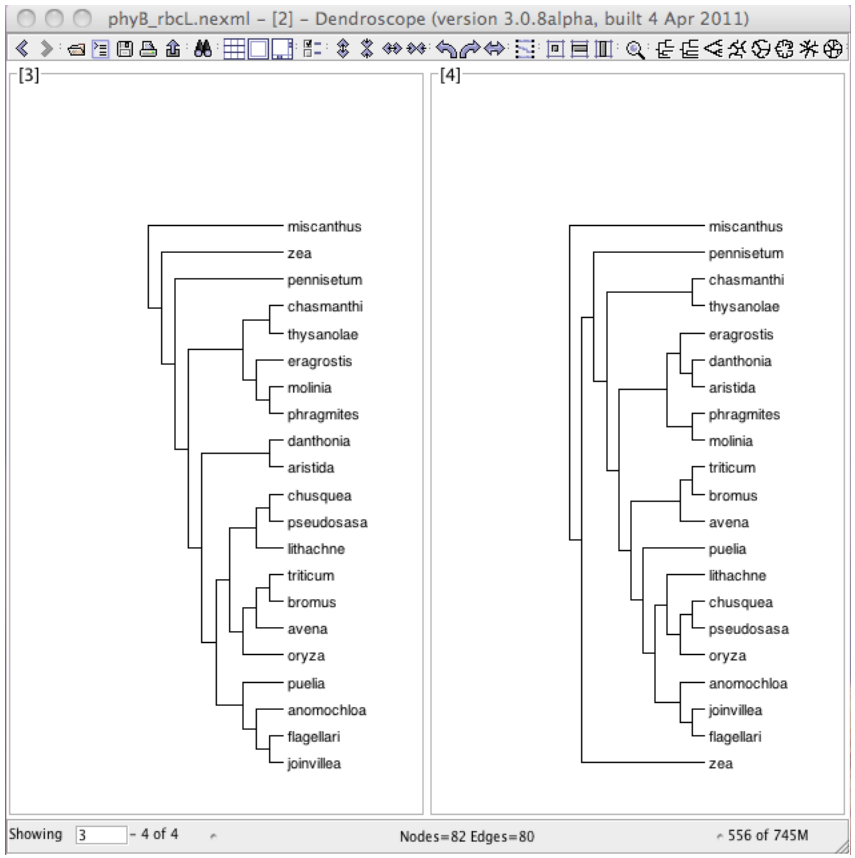
(d) Network N

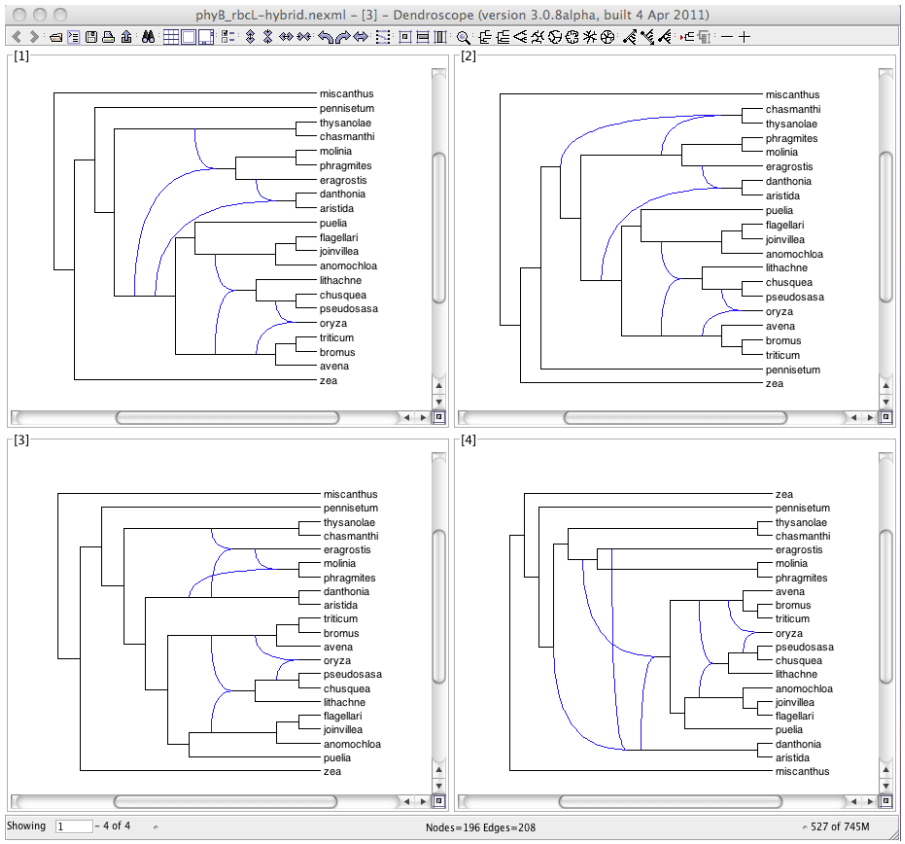
Reconstructing hybridization networks (explicit)

Goal : Find a phylogenetic network that displays a set of tree $\mathcal{T}$ with minimum number of reticulations (called hybrid number of $\mathcal{T}$).

- minimization : NP-hard, FPT (via reductions)
- a networks displaying $\mathcal{C}(\mathcal{T})$ does not in general display $\mathcal{T}$
- minimum number of reticulations required for representing $\mathcal{C}(\mathcal{T}) \leq$ hybrid number of $\mathcal{T}$
- these numbers are equal for 2 trees

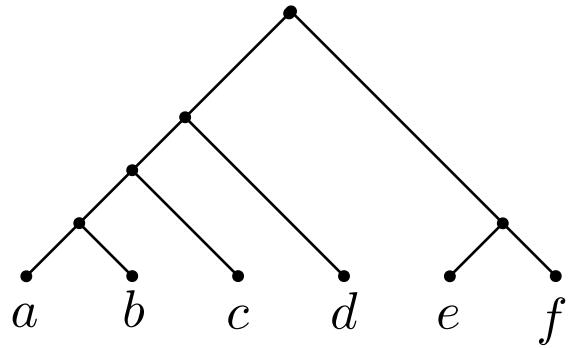
Reconstructing all hybridization networks for two binary trees

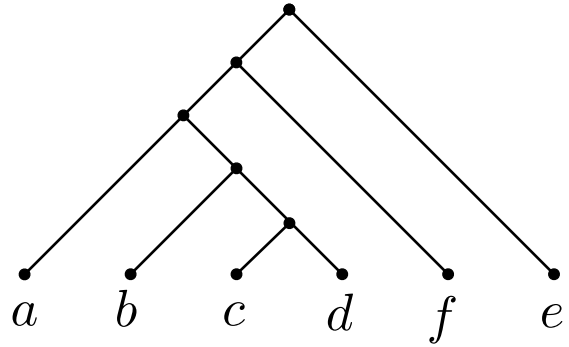




Agreement Forests

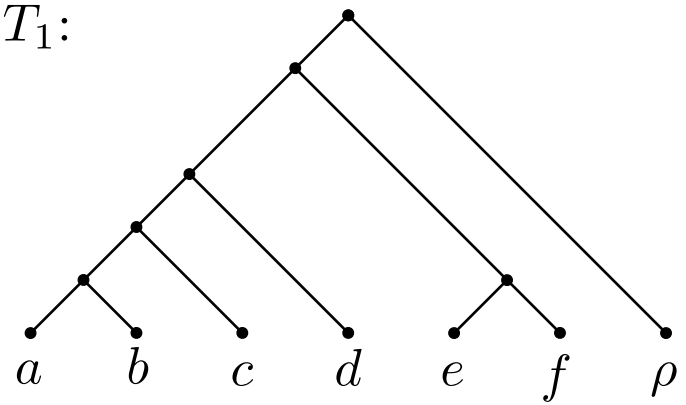
Given two trees...



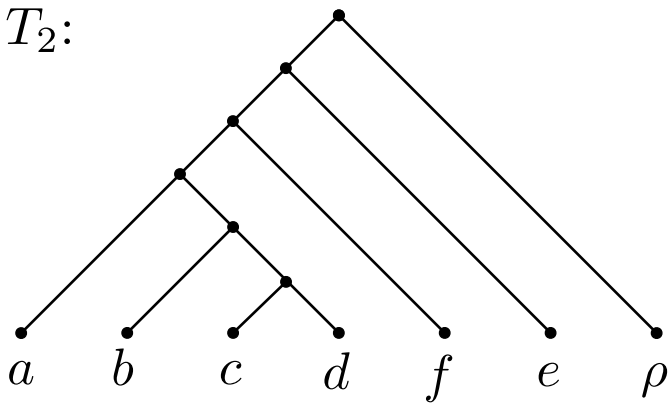


... in a first step an outgroup ρ is attached to the root nodes.

T_1:



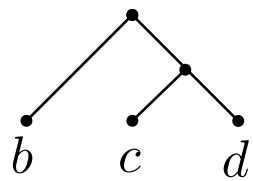
T_2:



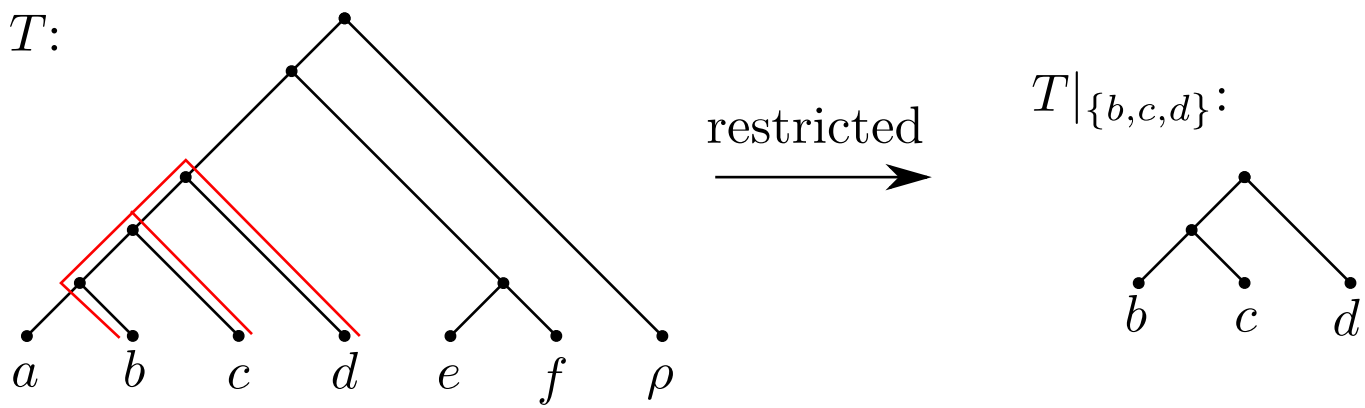
Agreement Forests

An agreement forest for two rooted bifurcating phylogenetic trees T_1 and T_2 on $\mathcal{X} \cup \rho$ is a set of components $\mathcal{F} = \{F_\rho, F_1, \ldots, F_n\}$ on $\mathcal{X} \cup \rho$ such that...

1. each component F_i is a restricted subtree of T_1 and T_2



... is **not** a restricted subtree of the tree T.

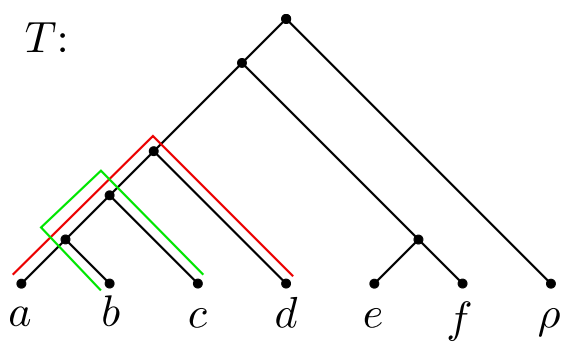


Agreement Forests

An agreement forest for two rooted bifurcating phylogenetic trees T_1 and T_2 on $\mathcal{X} \cup \rho$ is a set of components $\mathcal{F} = \{F_\rho, F_1, \ldots, F_n\}$ on $\mathcal{X} \cup \rho$ such that...

1. each component F_i is a restricted subtree of T_1 and T_2
2. the trees in $\{T_1(\mathcal{X}_i | i = \rho, 1, \ldots, n)\}$ and $\{T_2(\mathcal{X}_i | i = \rho, 1, \ldots, n)\}$ are node disjoint subtrees of T_1 and T_2, respectively

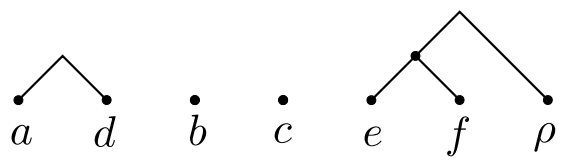
... are **not** node disjoint subtrees in the tree T.

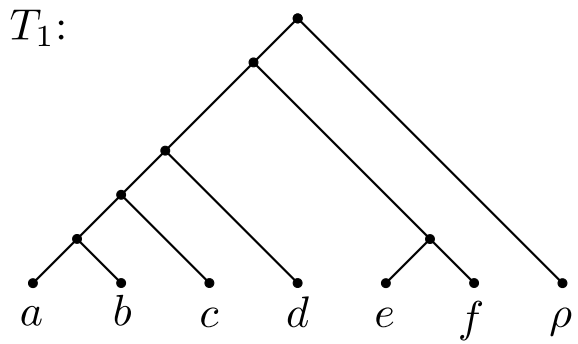


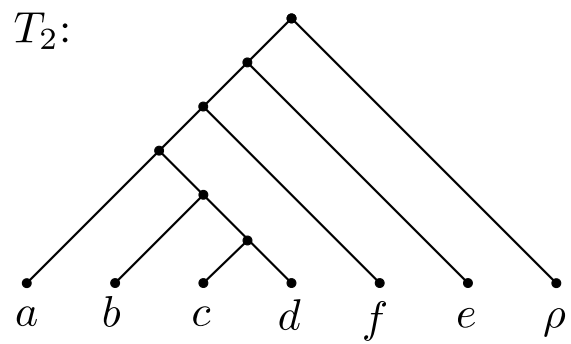
Agreement Forests

An agreement forest for two rooted bifurcating phylogenetic trees T_1 and T_2 on $\mathcal{X} \cup \rho$ is a set of components $\mathcal{F} = \{F_\rho, F_1, \ldots, F_n\}$ on $\mathcal{X} \cup \rho$ such that...

1. each component F_i is a restricted subtree of T_1 and T_2
2. the trees in $\{T_1(\mathcal{X}_i | i = \rho, 1, \ldots, n)\}$ and $\{T_2(\mathcal{X}_i | i = \rho, 1, \ldots, n)\}$ are node disjoint subtrees of T_1 and T_2, respectively
3. the taxon ρ is contained in F_ρ







Agreement Forests

An agreement forest for two rooted bifurcating phylogenetic trees T_1 and T_2 on $\mathcal{X} \cup \rho$ is a set of components $\mathcal{F} = \{F_\rho, F_1, \ldots, F_n\}$ on $\mathcal{X} \cup \rho$ such that...

1. each component F_i is a restricted subtree of T_1 and T_2
2. the trees in $\{T_1(\mathcal{X}_i | i = \rho, 1, \ldots, n)\}$ and $\{T_2(\mathcal{X}_i | i = \rho, 1, \ldots, n)\}$ are node disjoint subtrees of T_1 and T_2, respectively
3. the taxon ρ is contained in F_ρ

MAF

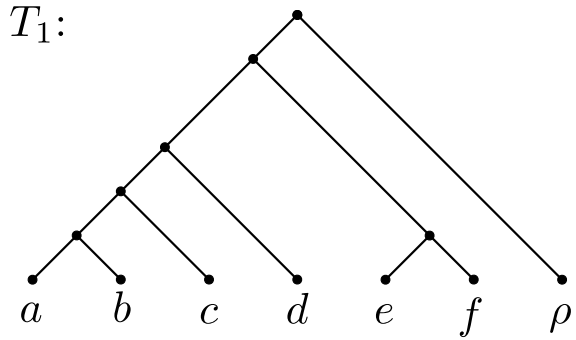
A **maximal** agreement forest, denoted by *MAF*, is any agreement forest $\mathcal{F}(T_1, T_2)$ of **minimal** size. Moreover, we have that

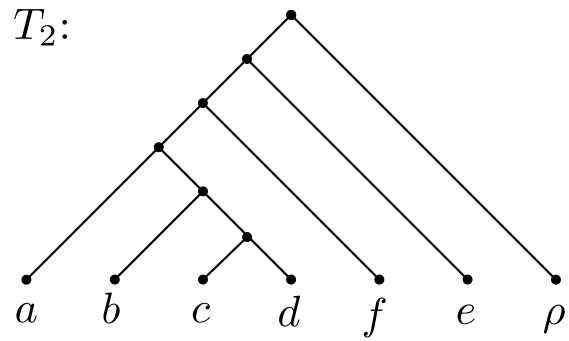
$$d_{\text{rSPR}}(T_1, T_2) = |\mathcal{F}(T_1, T_2)| - 1$$

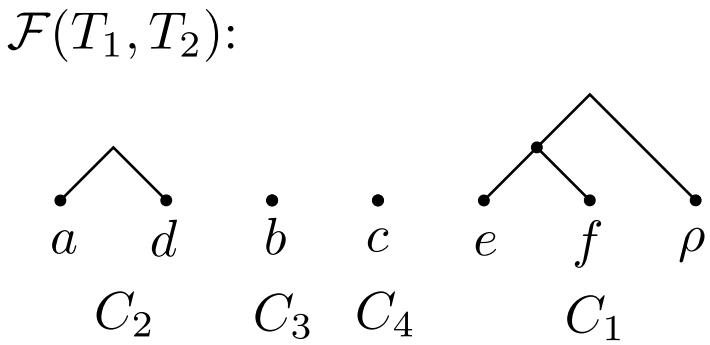
Usually, the number of MAFs of two trees is greater than one.

Acyclic Agreement Forest

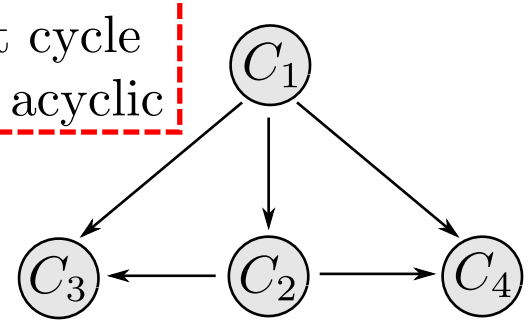
An agreement forest $\mathcal{F}(T_1, T_2)$ for T_1 and T_2 is called **acyclic**, if the components can be numbered such that, if the root of one component F is an ancestor of the root of some other component F', then the number assigned to F is **lower** than the number assigned to F' for all pairs of components F and F' in $\mathcal{F}(T_1, T_2)$.







No direct cycle
⇒ AF is acyclic



Acyclic Agreement Forest

An agreement forest $\mathcal{F}(T_1, T_2)$ for T_1 and T_2 is called **acyclic**, if the components can be numbered such that, if the root of one component F is an ancestor of the root of some other component F', then the number assigned to F is **lower** than the number assigned to F' for all pairs of components F and F' in $\mathcal{F}(T_1, T_2)$.

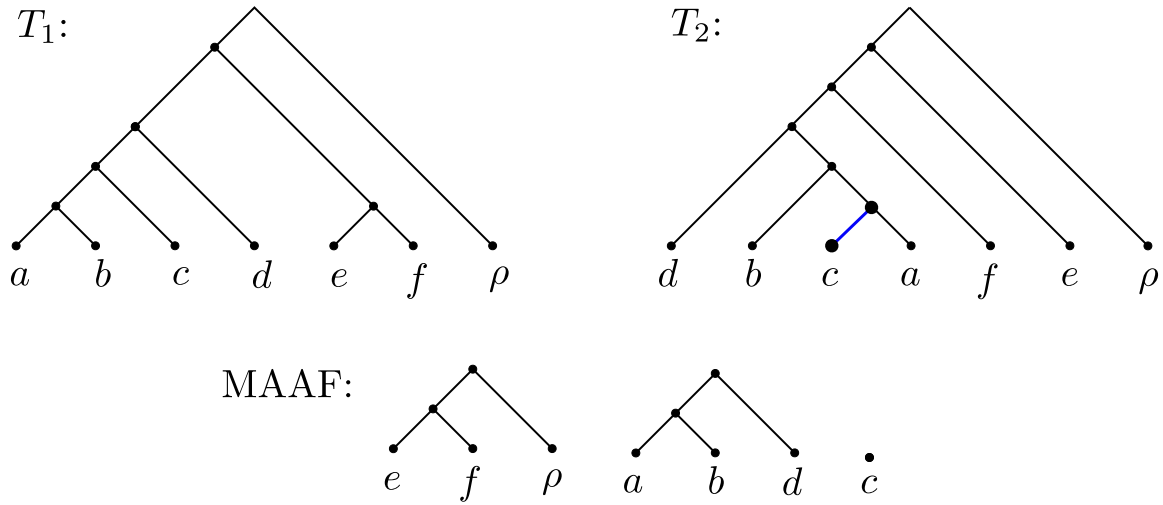
MAF

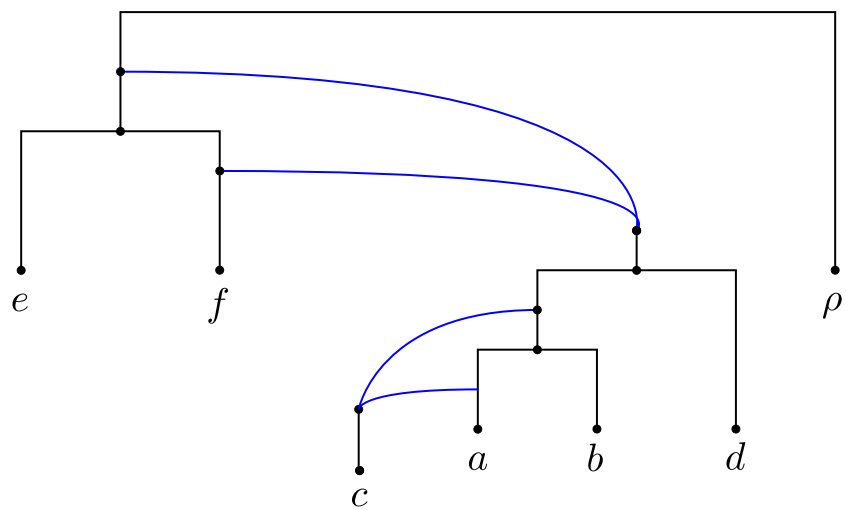
A **maximal** acyclic agreement forest, denoted by *MAAF*, is any acyclic agreement forest $\mathcal{F}(T_1, T_2)$ of **minimal** size. Moreover, we have that

$$h(T_1, T_2) = |\mathcal{F}(T_1, T_2)| - 1$$

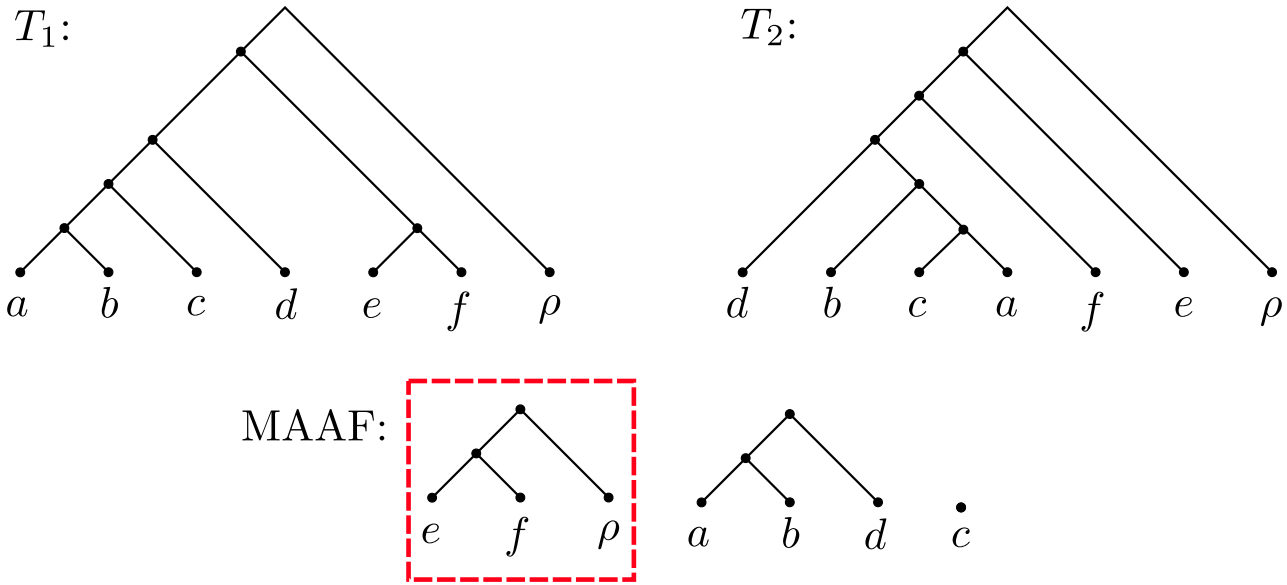
Usually, the number of MAAFs of two trees is greater than one.

MAAFs in hybridization networks

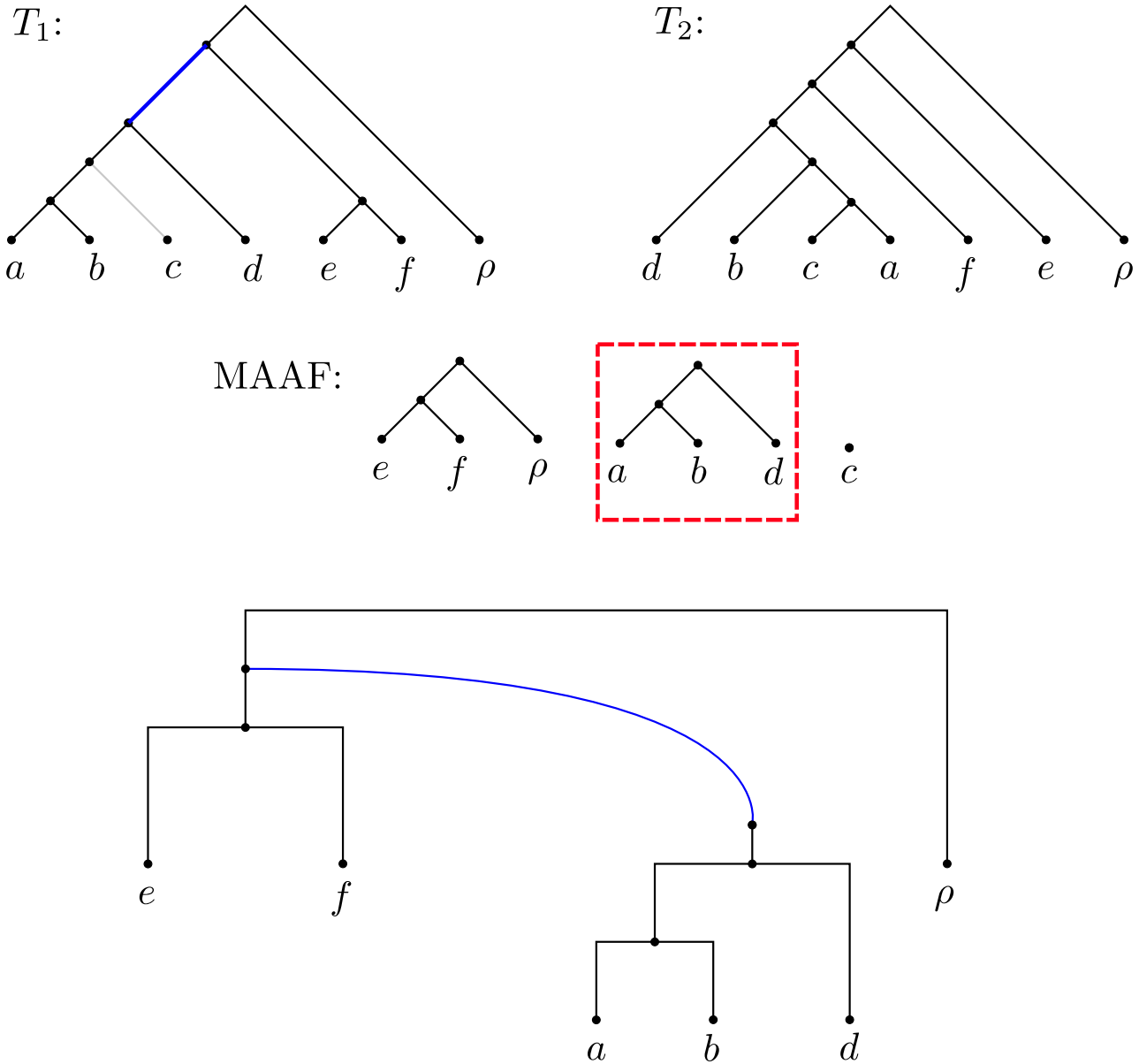




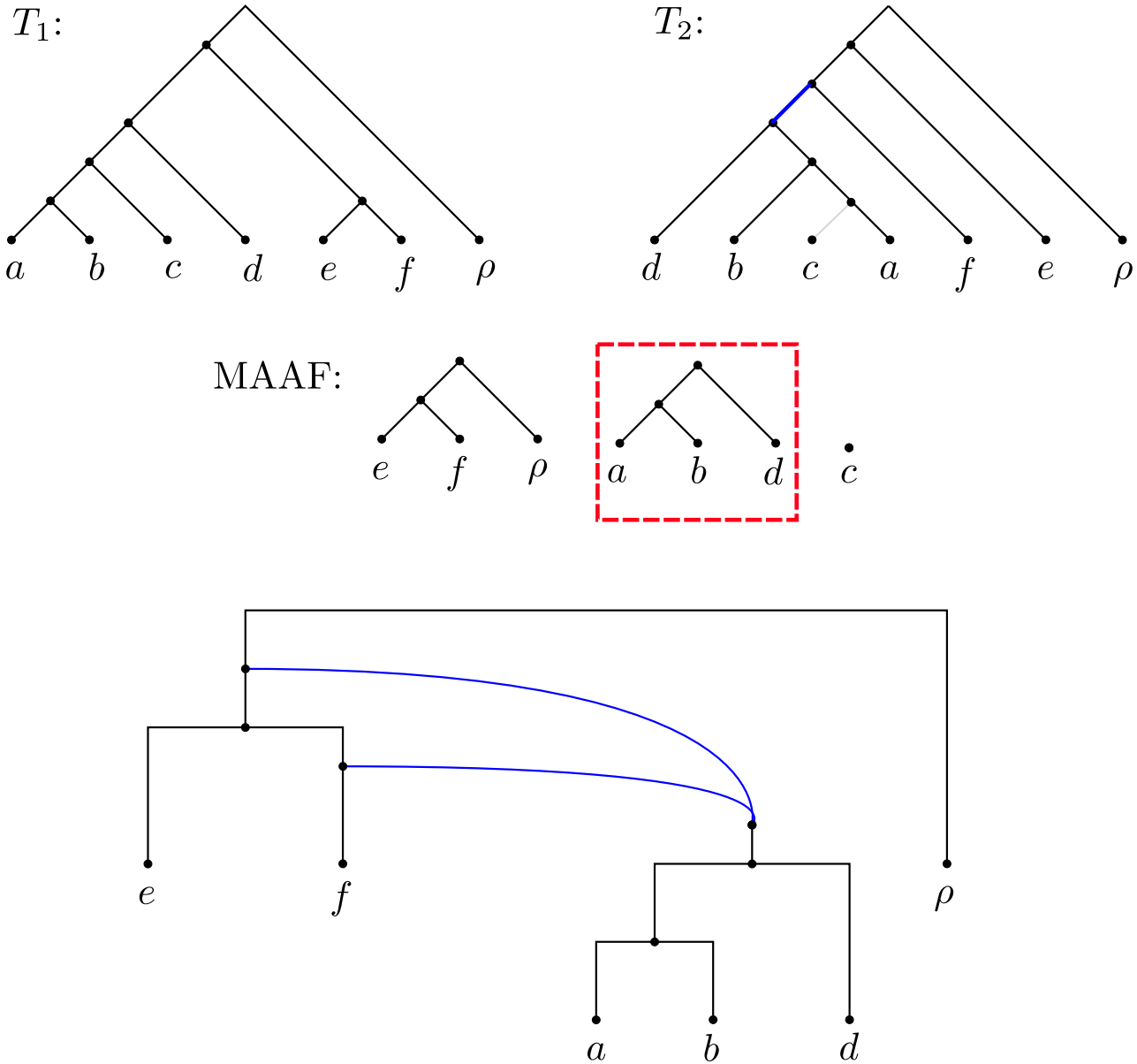
MAAFs in hybridization networks



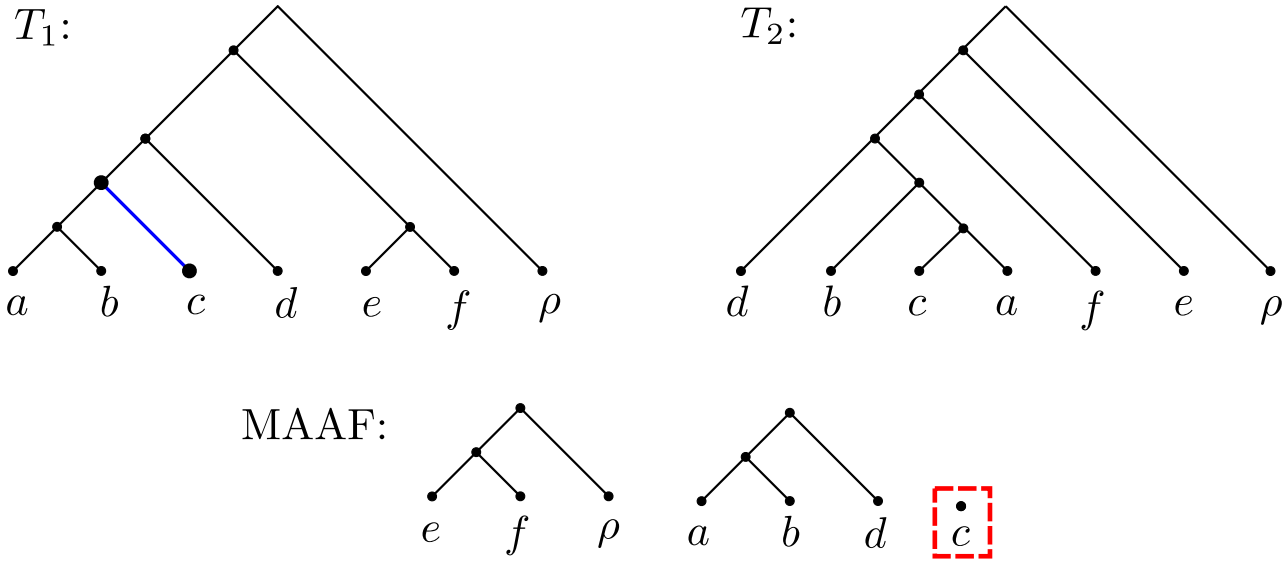
MAAFs in hybridization networks

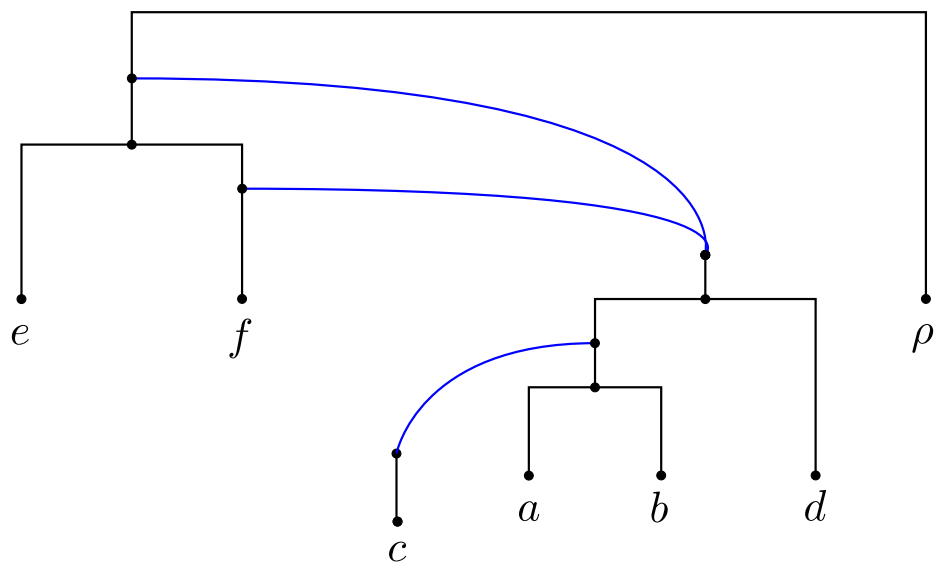


MAAFs in hybridization networks



MAAFs in hybridization networks





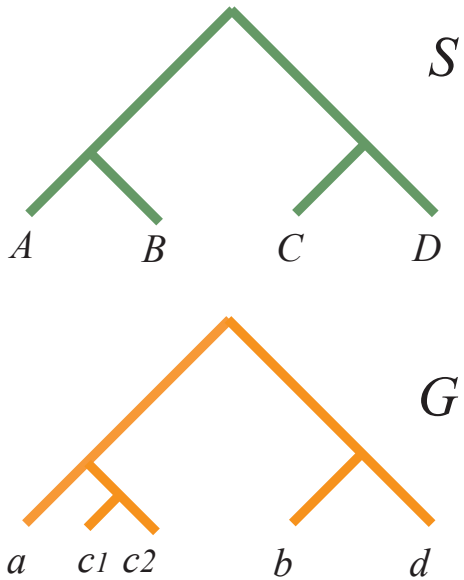
Our approach to construct all hybridization networks

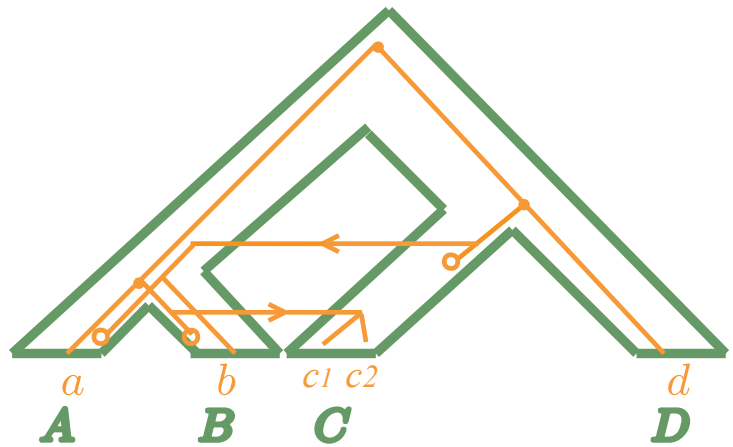
- construct all MAAFs
  - naif approach $O(n^k nlog(n))$: not suitable for huge input trees !)
  - our approach $O(3^k nlog(n))$

Our approach to construct all hybridization networks

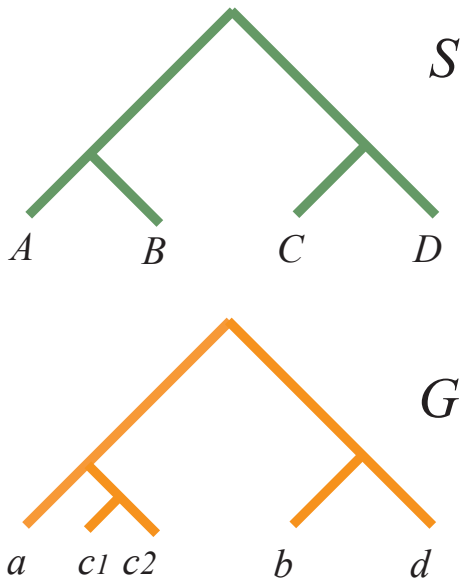
- construct all MAAFs
  - naif approach $O(n^k nlog(n))$: not suitable for huge input trees !)
  - our approach $O(3^k nlog(n))$
- reconstruct a (?) phylogenetic network form each MAAF
  - work in progress (motivated by Baroni et al., 2005) ...

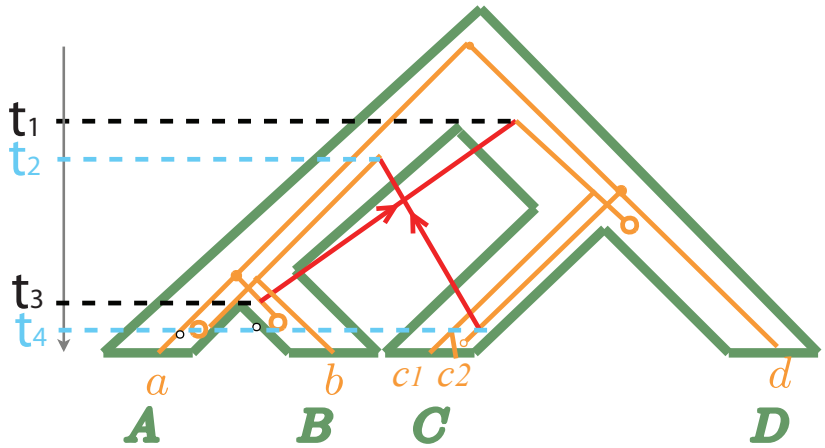
Reconciliation of (binary) phylogenetic trees





Reconciliation of (binary) phylogenetic trees

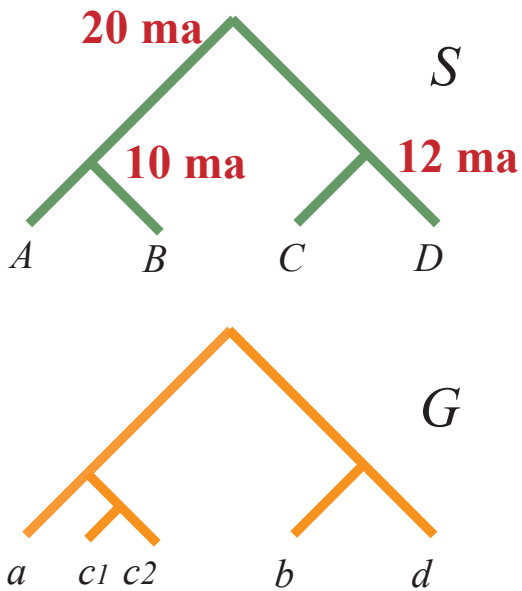


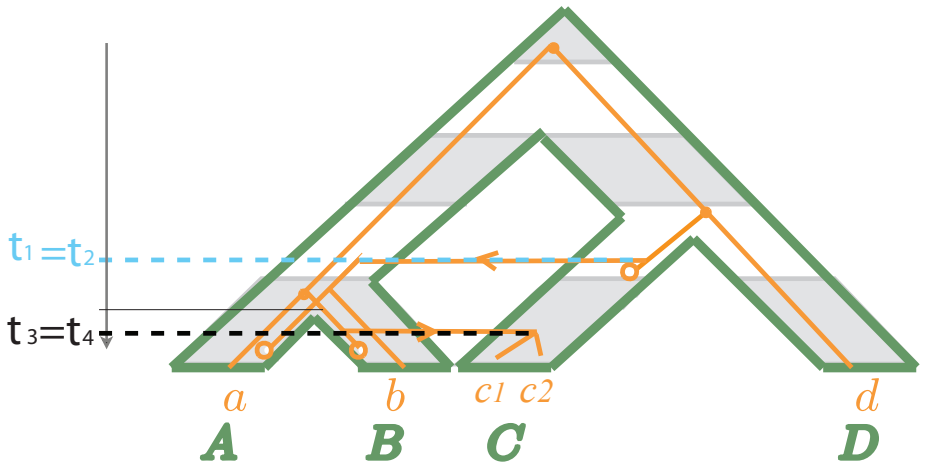


Problem

When taking into account HGTs, the problem is NP-hard.

Reconciliation of (binary) phylogenetic trees

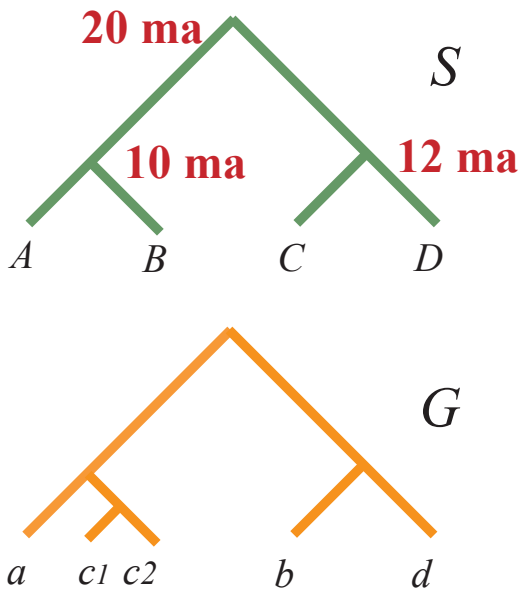


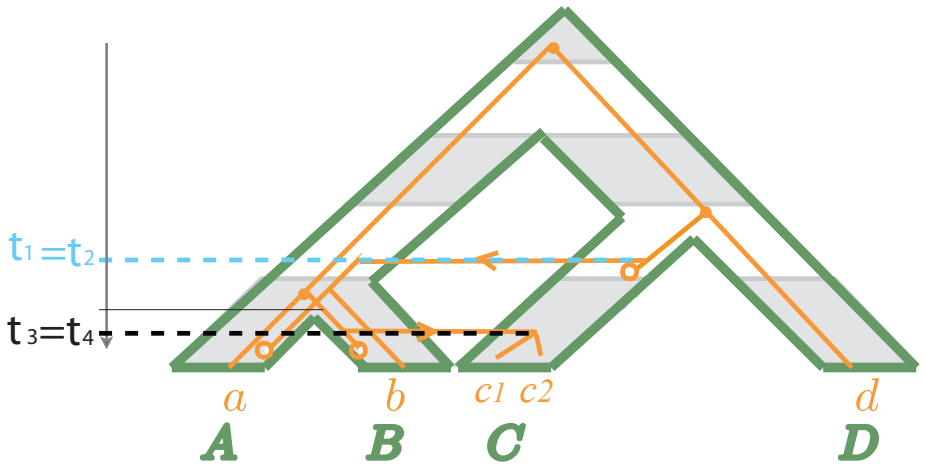


One possible solution

Dated species tree : polynomial and still realistic restriction of the NP-hard problem

Reconciliation of (binary) phylogenetic trees





Contribution

- formal definition of the underlying biological problem
- combinatorial modelling of the problem
- improvement of the complexity (by dynamic programming in $O(|S^2| \cdot |G|)$ instead of $O(|S|^4 \cdot |G|^4)$ or $O(|S|^8 \cdot |G|)$

- Doyon JP, **Scornavacca C**, Szöllősi G.J., Ranwez V et Berry V. LNCS, Springer-Verlag, 2010.
- 1 publication en préparation
  - Logiciels : MPR

Publications and international collaborations :

1. Huson, Rupp et **Scornavacca**. *Phylogenetic Networks*. Cambridge University Press 2011.
2. Escobar, **Scornavacca**, Cenci, Guilhaumon, Santoni, Douzery, Ranwez, Glémin et David. 2011. *Combining supermatrix and supertree in Triticeea*. MInor revisions : BMC Evolutionary Biology.
3. **Scornavacca** et Huson. *Tanglegrams for rooted phylogenetic trees and networks*. Accepted to ISMB 2011.
4. **Scornavacca** et Huson. A survey of combinatorial methods for phylogenetic networks. Genome Biology and Evolution 2011.
5. **Scornavacca**, Berry et Ranwez. *Building species trees from larger parts of phylogenomic databases*. Information and Computation 2011.
6. Jean-Philippe Doyon, **Celine Scornavacca**, Gergely J. Szöllősi , Vincent Ranwez et Vincent Berry. *An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers.* Eighth Annual RECOMB Satellite Workshop on Comparative Genomics 2010 LNCS,.
7. Jean-Philippe Doyon, **Celine Scornavacca**, Gergely J. Szöllősi , Vincent Ranwez et Vincent Berry. *Un algorithme de parcimonie efficace pour la réconciliation d'arbres de gènes/espèces avec pertes, duplications et transferts*. JOBIM 2010.
8. **Scornavacca**, Berry, Douzery et Ranwez. *PhySIC_IST : cleaning source trees to infer more informative supertrees*. BMC Bioinformatics 2009.
9. **Celine Scornavacca**, Vincent Berry et Vincent Ranwez. *From Gene Trees to Species Trees through a Supertree Approach*. Third International Conference on Language and Automata Theory and Applications 2009 LNCS.
10. Braga, Sagot, **Scornavacca** et Tannier. *Exploring the solution space of sorting by reversals and an application in evolution*. IEEE/ACM TCBB 2008.
11. Ranwez, Berry, Criscuolo, Fabre, Guillemot, **Scornavacca** et Douzery. *PhySIC : a Veto Supertree Method with Desirable Properties*. Syst. Biol. 2007.
12. Marília D. V. Braga, Marie-France Sagot, **Celine Scornavacca** et Eric Tannier. *The solution space of sorting by reversals*. Third International Symposium on Bioinformatics Research and Applications 2007 LNCS.