

A short (?) introduction to phylogenetic networks

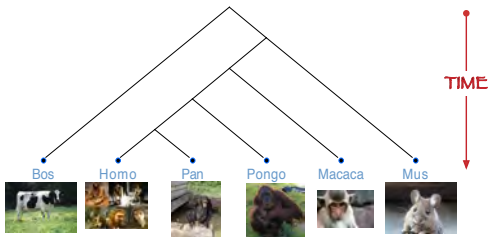
CÉLINE SCORNAVACCA

ISE-M, Equipe Phylogénie & Evolution Moléculaires
Montpellier, France

Rooted species trees ...

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

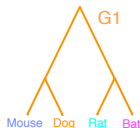
- the leaves or **taxa** represent extant organisms
- internal nodes represent hypothetical ancestors
- each **internal node** represents the lowest common ancestor of all taxa below it (**clade**)
- the only node without ancestor is called **root**



Gene trees

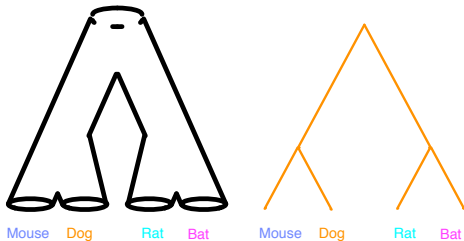
- Gene trees are built by analyzing a **gene family**, i.e., **homologous** molecular sequences appearing in the genome of different organisms.

Mouse	G	G	A	G	C	T	T	G	A	G	C	C	G	G	A	A	T	A	G	T	A	G	G	A	A	C	A	T	C	T	T	A	A	G	A	A	T	T	T	A	A	T	T	C	G	A	G	C	
Dog	G	G	A	A	T	C	T	G	A	A	C	A	G	G	C	T	T	A	G	T	A	G	C	C	A	C	T	A	G	A	A	T	A	A	G	A	C	T	T	T	A	A	T	T	C	G	A	G	C
Bat	G	G	A	A	T	T	T	G	A	A	C	A	G	G	T	T	T	A	G	T	A	G	C	C	A	C	T	A	G	A	A	T	A	A	G	A	C	T	C	T	A	A	T	T	C	G	A	G	C
Rat	G	G	A	A	T	T	T	G	A	A	C	C	G	G	C	T	C	G	T	A	G	C	A	A	C	A	A	G	A	A	T	A	A	G	C	T	T	A	A	T	T	C	G	T	G	C			



Gene trees

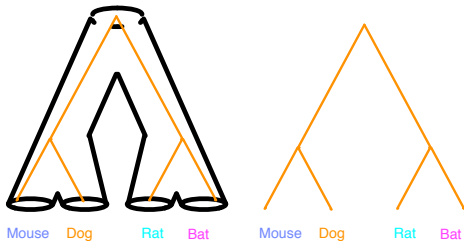
- Gene trees are built by analyzing a gene family.



- Used, among other things, to estimate species trees.

Gene trees

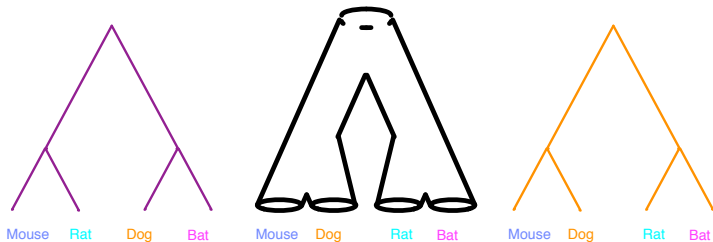
- Gene trees are built by analyzing a gene family.



- Used, among other things, to estimate species trees.

Gene trees

- Gene trees are built by analyzing a gene family.



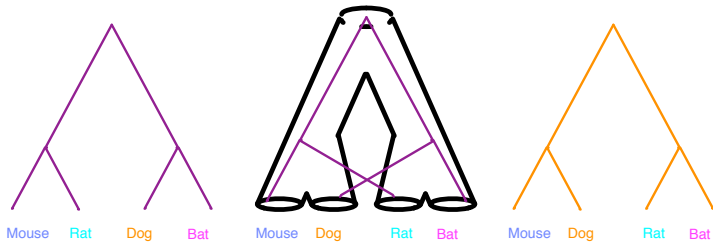
- Used, among other things, to estimate species trees.

Gene trees can significantly differ from the species tree for:

- methodological reasons
- biological reasons

Gene trees

- Gene trees are built by analyzing a gene family.



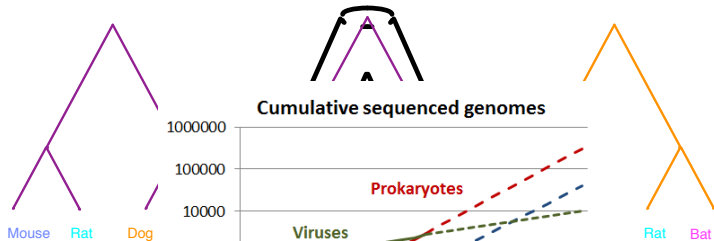
- Used, among other things, to estimate species trees.

Gene trees can significantly differ from the species tree for:

- methodological reasons
- biological reasons
- We usually use several gene families...

Gene trees

- Gene trees are built by analyzing a gene family.



- Used, among o

Gene trees can sig

- methodological
- biological reasons

ee for:

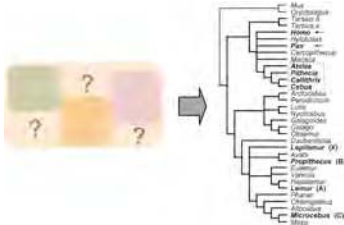
- We usually use several gene families...

<http://sulab.org/2013/06/sequenced-genomes-per-year/>

Reconstruction of phylogenies for multiple datasets

The two main *classic* approaches:

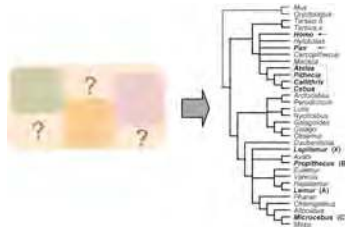
- Supermatrix approach: assembling primary data



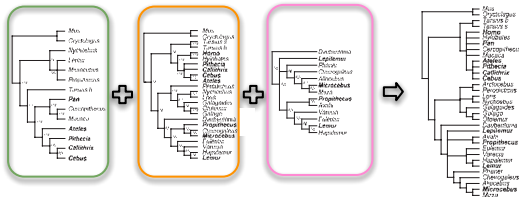
Reconstruction of phylogenies for multiple datasets

The two main *classic* approaches:

- Supermatrix approach: assembling primary data

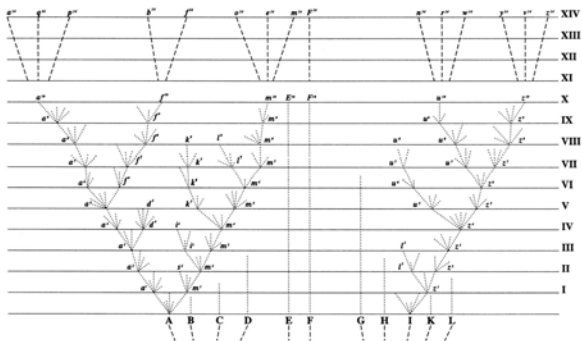


- Supertree approach: assembling trees



An implicit assumption

The implicit assumption of using trees is that, at a macroevolutionary scale, each (current or extinct) species or gene only descends from one ancestor. Darwin described evolution as "descent with modification", a phrase that does not necessarily imply a tree representation...

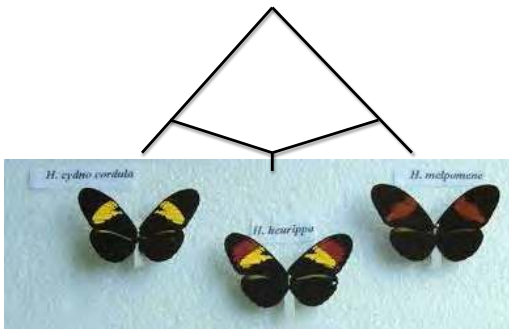


A *new* approach: building phylogenic networks

Why do we need them? Due to reticulate evolutionary phenomena (hybridization, recombination, horizontal gene transfer) the evolution of a set of species sometimes cannot be described using phylogenetic trees.

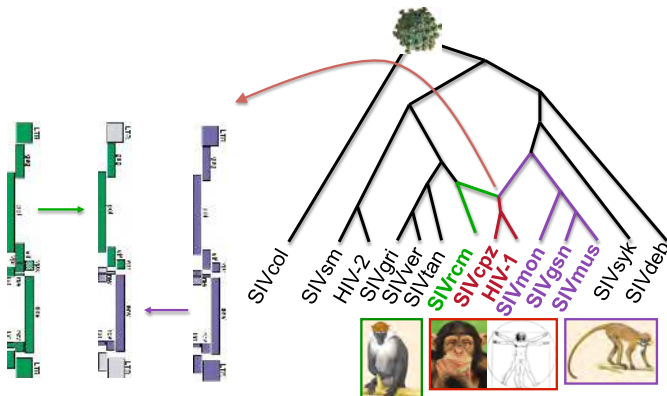
A new approach: building phylogenetic networks

Why do we need them? Due to reticulate evolutionary phenomena (hybridization, recombination, horizontal gene transfer) the evolution of a set of species sometimes cannot be described using phylogenetic trees.



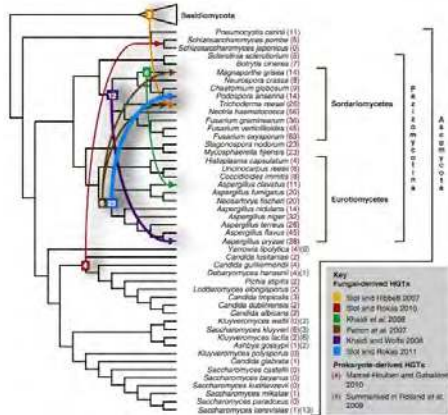
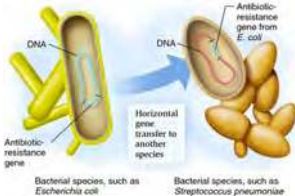
A new approach: building phylogenetic networks

Why do we need them? Due to reticulate evolutionary phenomena (hybridization, recombination, horizontal gene transfer) the evolution of a set of species sometimes cannot be described using phylogenetic trees.

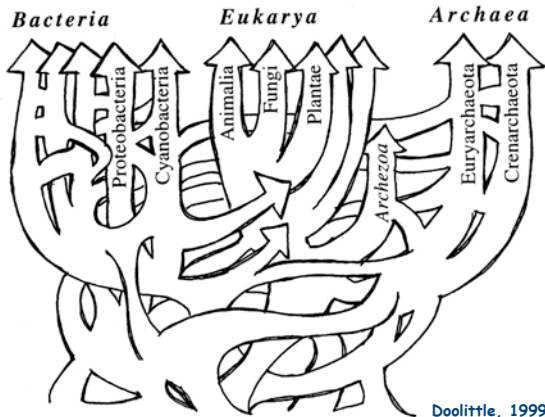
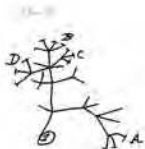


A *new* approach: building phylogenic networks

Why do we need them? Due to reticulate evolutionary phenomena (hybridization, recombination, horizontal gene transfer) the evolution of a set of species sometimes cannot be described using phylogenetic trees.

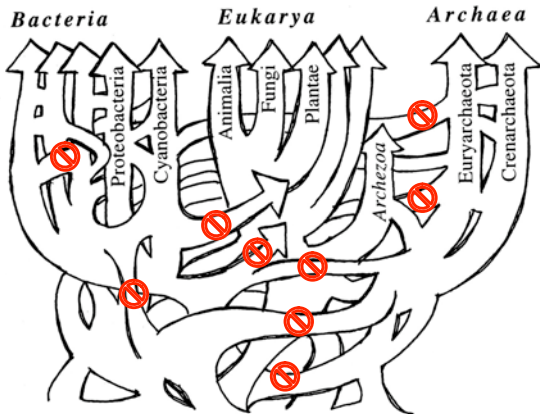
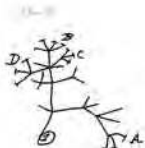


The network of life

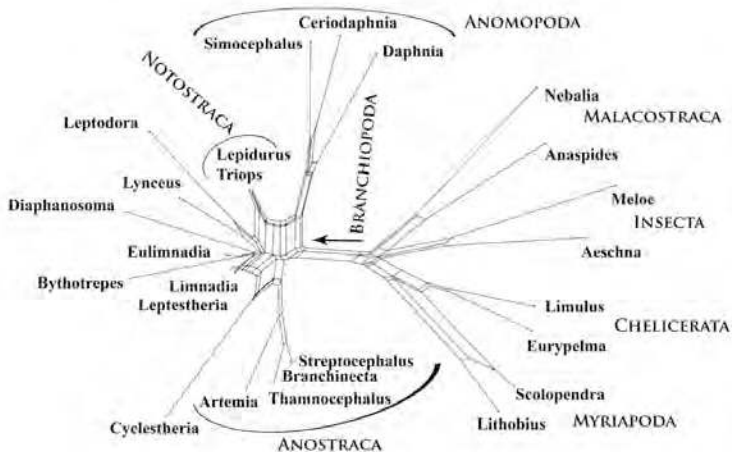


Three different paradigms

We (want to) see only the tree

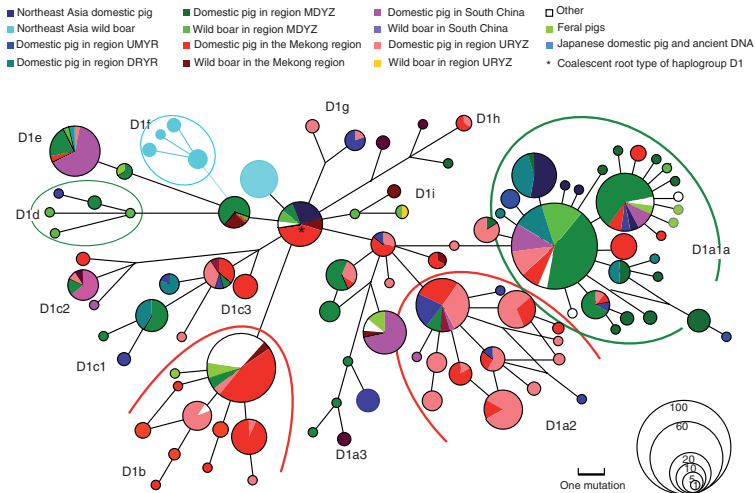


An example - a split network



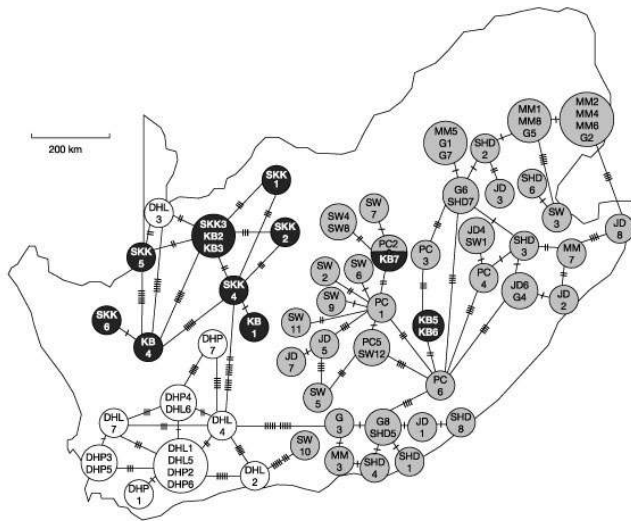
J. Wagele and C. Mayer. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evolutionary Biology, 7(1):147, 2007

An example - a reduced median network



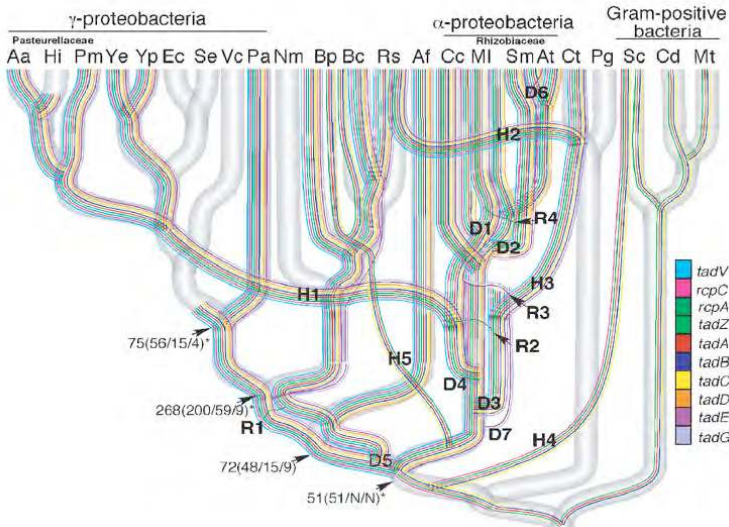
G.-S. Wu, Y.-G. Yao, K.-X. Qu, Z.-L. Ding, H. Li, M. Palanichamy, Z.-Y. Duan, N. Li, Y.-S. Chen, and Y.-P. Zhang. Population phylogenomic analysis of mitochondrial DNA in wild boars and domestic pigs revealed multiple domestication events in East Asia. *Genome Biology*, 8(11):R245, 2007

An example - a minimum spanning network



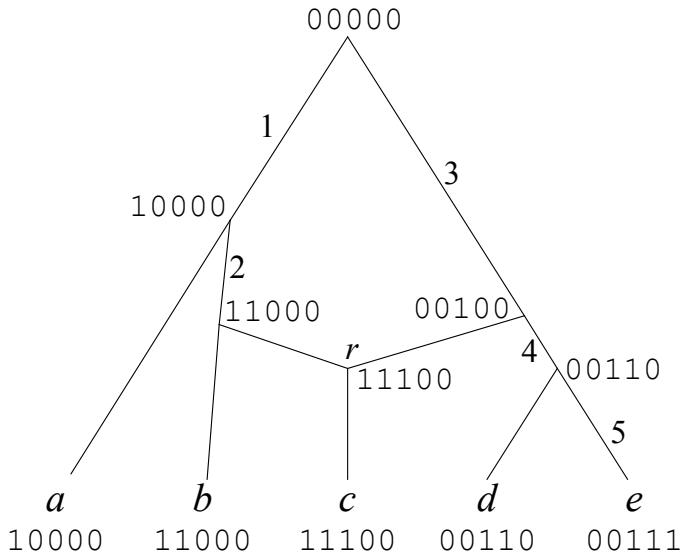
C. M. Miller-Butterworth, D. S. Jacobs, and E. H. Harley. Strong population sub- structure is correlated with morphology and ecology in a migratory bat. *Nature*, 424(6945):187-191, 2003

An example - a DTLR network



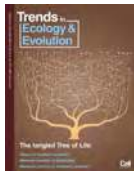
P.J. Planet, S.C. Kachlany, D.H. Fine, R. DeSalle, and D.H. Figurski. The wide spread colonization island of actinobacillus actinomycetemcomitans. Nature Genetics, 34:193–198, 2003.

An example - a recombination network

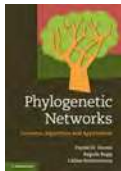


Phylogenetic networks

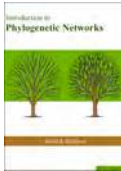
2008



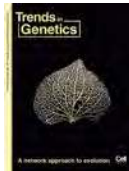
2010



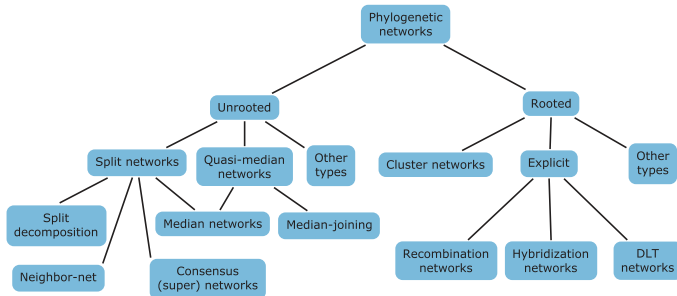
2011



2013

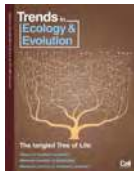


2014

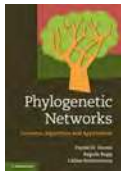


Phylogenetic networks

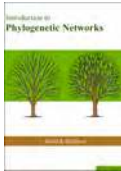
2008



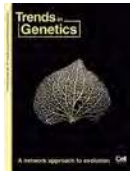
2010



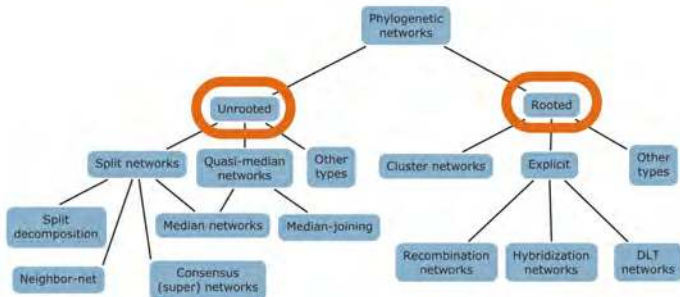
2011



2013

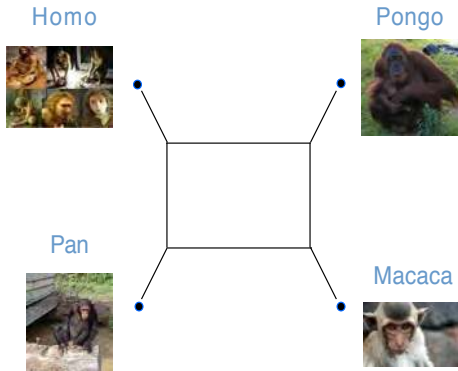


2014



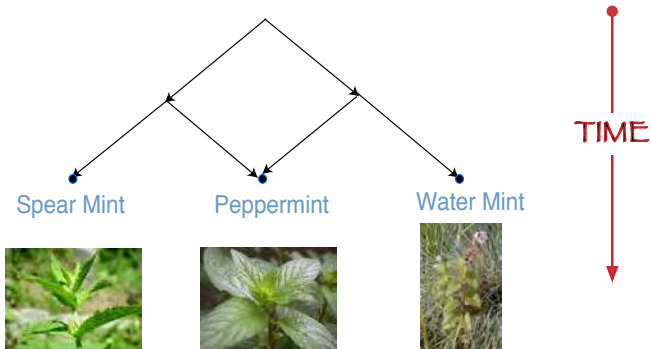
A phylogenetic network ...

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



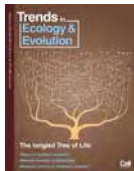
A rooted phylogenetic network ...

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

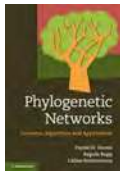


Phylogenetic networks

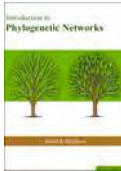
2008



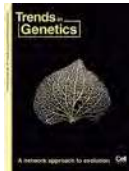
2010



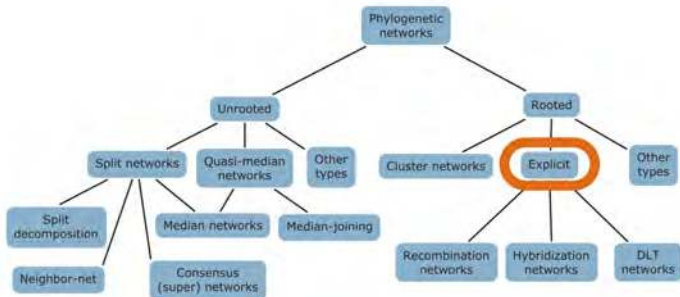
2011



2013

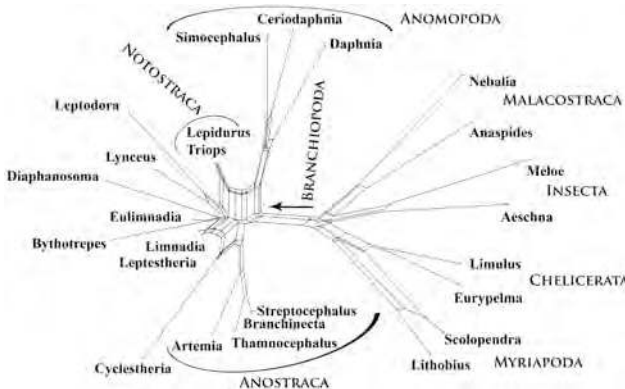


2014



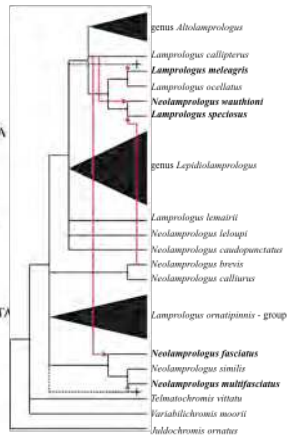
Abstract VS explicit phylogenetic networks

Split network:



Shows conflicting placement of taxa

Hybridization network:



Shows putative hybridization history

The plan of the survey

- ① combinatorial and distance methods not accounting for ILS
 - unrooted networks
 - rooted networks (explicit or not)
- ② methods accounting for ILS (always explicit)

Unrooted phylogenetic networks



SplitsTree4

by Daniel Huson and David Bryant

with contributions from Markus Franz, Miguel Jette',
Tobias Kloepper and Michael Schröder

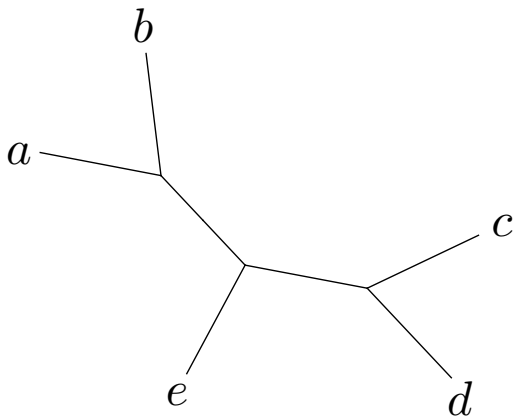
www.splitstree.org

Reconstruction of unrooted phylogenetic networks

- from splits
- from distances (via splits or not)
- from trees (via splits)
- from sequences (via splits or not)

Splits

A *split* $A \mid B$ on $\mathcal{X}$ is a partition of a taxon set $\mathcal{X}$ into two non-empty sets.



Compatible splits

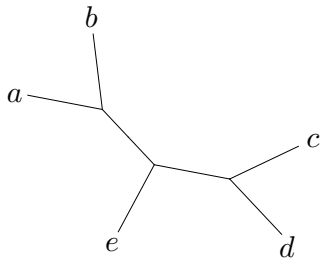
Two splits are $S_1 = A_1|B_1$ and $S_2 = A_2|B_2$ are compatible, if one of the $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$ or $B_1 \cap B_2$ is empty. A set of splits $\mathcal{S}$ is called compatible if all pairs of splits in $\mathcal{S}$ are compatible.

Example

$$\begin{array}{l} \mathcal{S}_1 = \begin{array}{l} \{a\}|\{b, c, d, e\} \\ \{b\}|\{a, c, d, e\} \\ \{c\}|\{a, b, d, e\} \\ \{d\}|\{a, b, c, e\} \\ \{e\}|\{a, b, c, d\} \\ \{a, b\}|\{c, d, e\} \\ \{a, b, e\}|\{c, d\} \end{array} \end{array} \quad \begin{array}{l} \mathcal{S}_2 = \begin{array}{l} \{a, b, d, e, h\} | \{c, f, g\} \\ \{a, c, d, e, g, h\} | \{b, f\} \\ \{a, c, e, g\} | \{b, d, f, h\} \\ \{a, c, g\} | \{b, d, e, f, h\} \\ \{a, c, e, f, g\} | \{b, d, h\} \\ \{a, e, h\} | \{b, c, d, f, g\} \end{array} \end{array}$$

Compatible splits

A set of compatible splits corresponds univocally to a unrooted phylogenetic tree.



(a) Unrooted tree T

$\{a\}|\{b, c, d, e\}$
 $\{b\}|\{a, c, d, e\}$
 $\{c\}|\{a, b, d, e\}$
 $\{d\}|\{a, b, c, e\}$
 $\{e\}|\{a, b, c, d\}$
 $\{a, b\}|\{c, d, e\}$
 $\{a, b, e\}|\{c, d\}$

(b) Split encoding of T

Circular splits

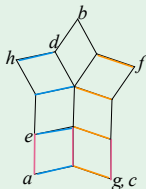
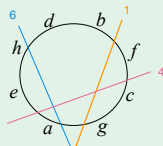
A set of splits $\mathcal{S}$ on $\mathcal{X}$ is called *circular*, if there exists a linear ordering $\pi = (x_1, \dots, x_n)$ of the elements of $\mathcal{X}$ for $\mathcal{S}$ such that each split $S \in \mathcal{S}$ is *interval-realizable*, that is, has the form

$$S = \{x_p, x_{p+1}, \dots, x_q\} \mid (\mathcal{X} \setminus \{x_p, x_{p+1}, \dots, x_q\}),$$

for appropriately chosen $1 < p \leq q \leq n$.

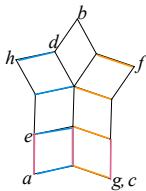
Example

$\{a, b, d, e, h\} \mid \{c, f, g\}$
 $\{a, c, d, e, g, h\} \mid \{b, f\}$
 $\{a, c, e, g\} \mid \{b, d, f, h\}$
 $\{a, c, g\} \mid \{b, d, e, f, h\}$
 $\{a, c, e, f, g\} \mid \{b, d, h\}$
 $\{a, e, h\} \mid \{b, c, d, f, g\}$



Circular splits

A set of circular splits corresponds to a unrooted network that is outer-labeled planar.



(a) Planar network

$$\begin{aligned} &\{a, b, d, e, h\} \mid \{c, f, g\} \\ &\{a, c, d, e, g, h\} \mid \{b, f\} \\ &\{a, c, e, g\} \mid \{b, d, f, h\} \\ &\{a, c, g\} \mid \{b, d, e, f, h\} \\ &\{a, c, e, f, g\} \mid \{b, d, h\} \\ &\{a, e, h\} \mid \{b, c, d, f, g\} \end{aligned}$$

(b) Circular splits

Weakly compatible splits

Three splits $S_1 = \frac{A_1}{B_1}$, $S_2 = \frac{A_2}{B_2}$, and $S_3 = \frac{A_3}{B_3}$ *weakly compatible*, if

- ① at least one of the following four intersections is empty:
 $A_1 \cap A_2 \cap A_3$, $A_1 \cap B_2 \cap B_3$, $B_1 \cap A_2 \cap B_3$ and $B_1 \cap B_2 \cap A_3$,
- ② at least one of the following four intersections is empty:
 $B_1 \cap B_2 \cap B_3$, $B_1 \cap A_2 \cap A_3$, $A_1 \cap B_2 \cap A_3$ and $A_1 \cap A_2 \cap B_3$.

A set of splits $\mathcal{S}$ on $\mathcal{X}$ is called *weakly compatible*, if any *three* distinct splits in $\mathcal{S}$ are weakly compatible.

Example

$$\begin{array}{ll} S_1 = \begin{array}{l} \{a, b, d, e, h\} \mid \{c, f, g\} \\ \{a, c, d, e, g, h\} \mid \{b, f\} \\ \{a, c, e, g\} \mid \{b, d, f, h\} \\ \{a, c, g\} \mid \{b, d, e, f, h\} \\ \{a, c, e, f, g\} \mid \{b, d, h\} \\ \{a, e, h\} \mid \{b, c, d, f, g\} \end{array} & S_2 = \begin{array}{l} \{a, b, d, e, h\} \mid \{c, f, g\} \\ \{a, c, d, e, g, h\} \mid \{b, f\} \\ \{a, c, e, g\} \mid \{b, d, f, h\} \\ \{a, c, d, e\} \mid \{b, f, g\} \\ \{a, b\} \mid \{c, d, e, f, g\} \\ \{a, e, f\} \mid \{b, c, d, g\} \end{array} \end{array}$$

Weakly compatible splits

Phylogenetic networks reconstructed from weakly compatible are easier than the ones reconstructed from generic splits

UPN from splits

or “what to do with the splits?”

PN from splits: the Convex hull algorithm

We start with the start tree and we add a split $S = \frac{A}{B}$ as follows:

- 1 Compute the two convex hulls $H(A)$ and $H(B)$ in N and let M be the graph induced by the nodes in $H(A) \cap H(B)$.
- 2 Create a copy M' of M and denote v' and e' the copies of a node v and an edge e in M .
- 3 Substitute any edge $f = (u, v)$ where u in $H(B) \setminus H(A) \neq \emptyset$ and v in M with edge $f = (u, v')$.
- 4 Connect each pair of nodes v in M and v' in M' by a new edge.

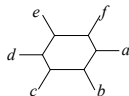
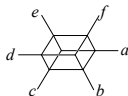
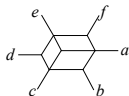
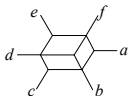
PN from splits: the circular network algorithm

We start with the start tree and we add a split $S = \frac{\{x_p, \dots, x_q\}}{\mathcal{X} \setminus \{x_p, \dots, x_q\}}$ as follows:
(splits have to be considered in a certain order)

- 1 Determine the path $M(x_p, x_q)$ and let $\dot{M}$ denote the path obtained by removing the first and last (leaf) edges from $M(x_p, x_q)$.
- 2 Create a copy $\dot{M}'$ of $\dot{M}$ and denote v' and e' the copies of a node v and an edge e in $\dot{M}$.
- 3 Substitute any edge $f = (u, v)$ where $u = \lambda(x_i)$ and v in $\dot{M}$ with edge $f = (u, v')$, for all $i = p, \dots, q$.
- 4 Connect each pair of nodes v in $\dot{M}$ and v' in $\dot{M}'$ by a new edge.

PN from splits: **attention!!!**

All four different split networks shown below represent the same set of splits.

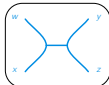


UPN from distances

or “how to get the splits from distances”

PN from distances: the split decomposition

Given a distance matrix D on $\mathcal{X} = \{x_1, \dots, x_n\}$ the split decomposition algorithm [Bandelt and Dress, 1992] starts by computing the isolation index for quartets and splits:



- for any four taxa w, x, y and z with $\{w, x\} \cap \{y, z\} = \emptyset$, :
 $\hat{\alpha}_D(\frac{\{w,x\}}{\{y,z\}}) = \frac{1}{2}(\max\{d(w, x) + d(y, z), d(w, y) + d(x, z), d(w, z) + d(x, y)\} - d(w, x) - d(y, z)).$
- for any (partial) split S : $\alpha_D(S) = \min\{\hat{\alpha}_D(\frac{\{w,x\}}{\{y,z\}}) \mid w, x \in A, y, z \in B\} \geq 0.$



Then, we set $X_0 = \emptyset$ and $\mathcal{S}_0 = \emptyset$. Given the set of splits $\mathcal{S}_i$ on the first i taxa, we obtain $\mathcal{S}_{i+1}$ by, for each split $\frac{A}{B} \in \mathcal{S}_i$ doing:

- 1 Consider $S = \frac{A \cup \{x_{i+1}\}}{B}$. If $\alpha_D(S) > 0$, set $\omega(S) = \alpha_D(S)$ and add S to $\mathcal{S}_{i+1}$.
- 2 Do the same with $S = \frac{A}{B \cup \{x_{i+1}\}}$ and $S = \frac{\mathcal{X}_i}{\{x_{i+1}\}}$

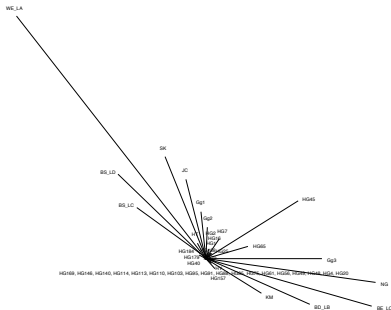
The result is given by $\mathcal{S}_n$.

PN from distances: the split decomposition

- A split S whose isolation index $\alpha_D(S)$ is greater than 0 is called a D -split. D -splits are always weakly compatible.
- It follows from this that the split decomposition always computes a set of weakly compatible splits

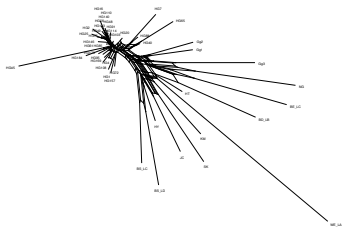
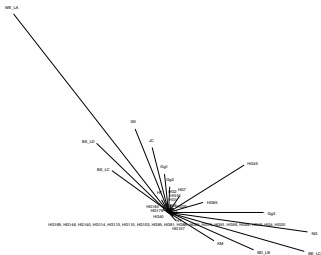
PN from distances: the split decomposition

- A split S whose isolation index $\alpha_D(S)$ is greater than 0 is called a D -split. D -splits are always weakly compatible.
- It follows from this that the split decomposition always computes a set of weakly compatible splits
- The SD is a conservative method
- It can be used for small number of taxa or low divergence



PN from distances: Neighbor-Net

- Given a distance matrix D on $\mathcal{X}$, the Neighbor-Net algorithm [Bryant and Moulton, 2004] computes a circular ordering π of $\mathcal{X}$ from D and then a set of weighted splits $\mathcal{S}$ that are interval-realizable with respect to π :
 - produces circular splits
 - uses together with circular network algorithm to get planar networks
 - can be used for large number of taxa and high divergence



PN from distances

Other algorithms from distances:

- Minimum spanning network
- T-Rex
- ...

A great source of information:

<http://phylnet.univ-mlv.fr/>

Who is Who in Phylogenetic Networks

🏠 Authors Community Keywords Publications **Software** Browse Basket Account Contribute! About Help 🔍

Programs and their Input Data !

How do I interact with the graph ?

Below, you can find all programs present at least **1 time(s)** in [Who is who in phylogenetic networks](#), as well as the links with the data they use as input.

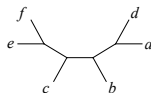
UPN from trees

or “how to get splits from a bunch of trees”

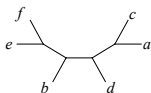
PN from trees: Consensus split networks

Consensus splits [Holland et al, 2004]

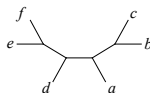
- Input: Trees on identical taxon sets
- Determine splits in more than X% of trees
- For >50%, the result is compatible



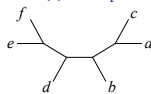
(a) Tree T_1



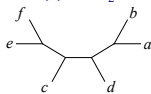
(b) Tree T_2



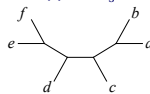
(c) Tree T_3



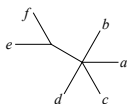
(d) Tree T_4



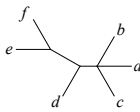
(e) Tree T_5



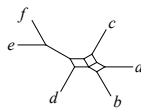
(f) Tree T_6



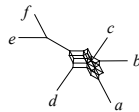
(g) Majority



(h) $d = 2$



(i) $d = 5$



(j) All splits

The Z-closure

- Two partial splits $S_1 = \frac{A_1}{B_1} \in \mathcal{S}$ and $S_2 = \frac{A_2}{B_2} \in \mathcal{S}$ are said to be in *Z-relation* to each other, if exactly one of the four intersections $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$ or $B_1 \cap B_2$ is empty. Then we can create of two new splits (the *Z-operation*)

$$S'_1 = \frac{A_1}{B_1 \cup B_2} \text{ and } S'_2 = \frac{A_1 \cup A_2}{B_2}.$$

- If at least one of the two new splits contains more taxa than its predecessor, the pair of splits is called *productive*.

From a set partial splits $\mathcal{S}$ on $\mathcal{X}$, *Z-closure* method infers a set of *complete splits* on $\mathcal{X}$ as follows: While $\mathcal{S}$ contains a productive pair of splits $\{S_i, S_j\}$, apply the Z-operation to obtain two new splits $\{S'_i, S'_j\}$ and then replace the former pair by the latter pair in $\mathcal{S}$. Finally, add all trivial splits on $\mathcal{X}$.

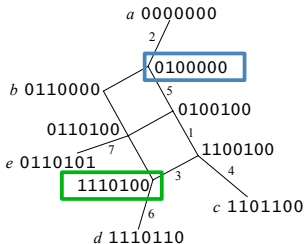
UPN from sequences

Median networks

For a multiple alignment M of binary sequences on $\mathcal{X}$, its **median network** is a phylogenetic network $N = (V, E, \sigma, \lambda)$ whose node set is given by the **median closure** $V = \bar{M}$ and in which any two nodes a and b are connected by an edge e of color $\sigma(e) = i \in E$, if and only if they differ in exactly in their i -th position (as haplotypes). An associated taxon labeling $\lambda : X \rightarrow V$ maps each taxon x onto the node $\lambda(x)$ that represents the corresponding sequence.

a	0000000
b	0110000
c	1101100
d	1110110
e	0110101

(a) Alignment M



(b) Median network N

Quasi median networks

a A A A A A
 b B B A A A
 c A B A B B
 d A A B B C
 e A A C B C

(a) Input M

a 0 0 0 0 0 0 0 0
 b 1 1 0 0 0 0 0 0
 c 0 1 0 0 0 1 1 0
 d 0 0 1 1 0 1 1 0
 e 0 0 1 0 1 1 1 0

(b) Binary expansion M_1

a 0 0 0 0 0 0 0
 b 1 1 0 0 0 0 0
 c 0 1 0 0 0 1 1
 d 0 0 1 1 0 1 0
 e 0 0 1 0 1 1 0

(c) Condensed M_1

0 0 0 0 0 0 0
 1 1 0 0 0 0 0
 0 1 0 0 0 1 1
 0 0 1 1 0 1 0
 0 0 1 0 1 1 0
 0 0 1 0 0 1 0
 0 0 0 0 0 1 0
 0 1 0 0 0 0 0
 0 1 0 0 0 1 0

(d) Median closure M_2

0 0 0 0 0 0 0 0
 1 1 0 0 0 0 0 0
 0 1 0 0 0 1 1 0
 0 0 1 1 0 1 1 0
 0 0 1 0 1 1 1 0
 0 0 1 0 0 1 1 0
 0 0 0 0 0 1 1 0
 0 1 0 0 0 0 0 0
 0 1 0 0 0 1 1 0

(e) Expanded M_2

A A A A A
 B B A A A
 A B A B B
 A A C B C
 A B B B C
 A A * B C
 A A A B *
 A B A A A
 A B A B *

(f) Multi-states M_3

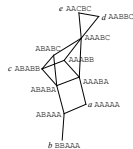
A 0 0 0
 B 1 1 0
 C 1 0 1

$AA*BC = \begin{cases} A A A B C \\ A A B B C \\ A A C B C \end{cases}$
 $AAAB* = \begin{cases} A A A B A \\ A A A B B \\ A A A B C \end{cases}$
 $ABAB* = \begin{cases} A B A B A \\ A B A B B \\ A B A B C \end{cases}$

(g) Expansion of virtual medians

A A A A A
 B B A A A
 A B A B B
 A A C B C
 A A B B C
 A A A B C
 A A A B A
 A A A B B
 A B A A A
 A B A B A
 A B A B C

(h) Final matrix M_4



(i) Quasi-median network N

How to keep the complexity of the network down...

The number of nodes of the quasi-median network can be very large, even for a small number of short sequences. Thus, the quasi-median network is rarely useful in practice. There exist two alternative methods:

- median-joining algorithm, which aims at computing an UPN that is as informative as a quasi-median network, but usually much smaller. The algorithm has a parameter Δ that is used to control how complex the resulting phylogenetic network will be.
- geodesically-pruned quasi-median networks: a method that aims at computing a pruned version of the full quasi-median network by considering only those sequences that lie on a geodesic between two of the original input sequences.

How to keep the complexity of the network down...

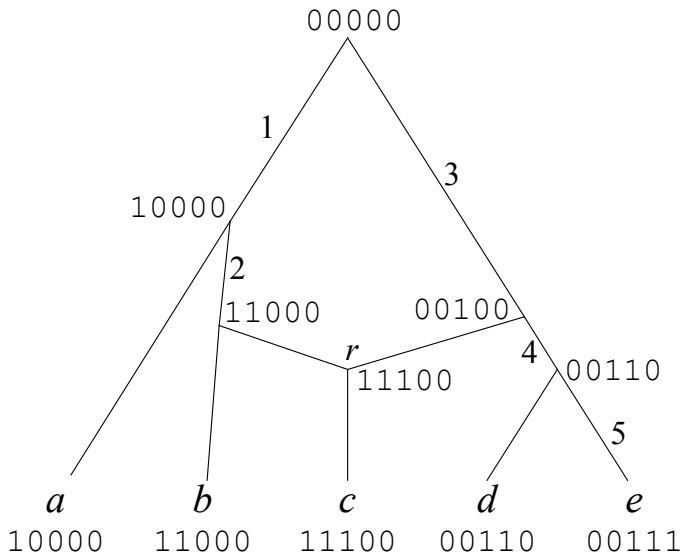
UPN from ...

quartets ... QNet

<http://www2.cmp.uea.ac.uk/~vlm/qnet/>

<http://phylnet.univ-mlv.fr/>

Recombination networks

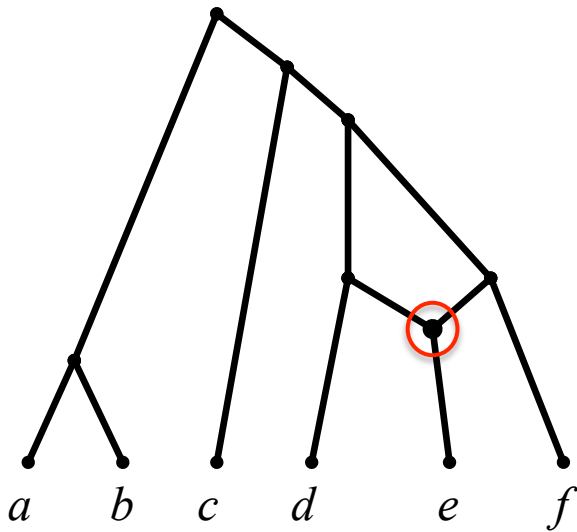


*Methods for reconstructing rooted
phylogenetic networks not
accounting for ILS*

some slides have been kindly provided by Fabio Pardi

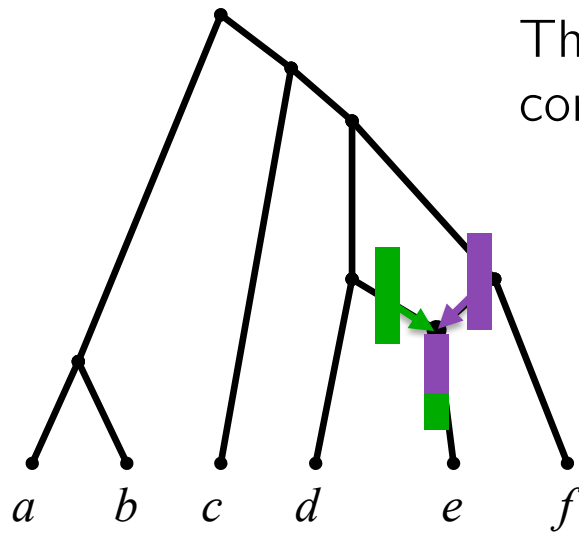
Trees displayed by a network

In a phylogenetic network, a reticulate event is represented as a **reticulation**, where branches converge to give rise to a new lineage:



Trees displayed by a network

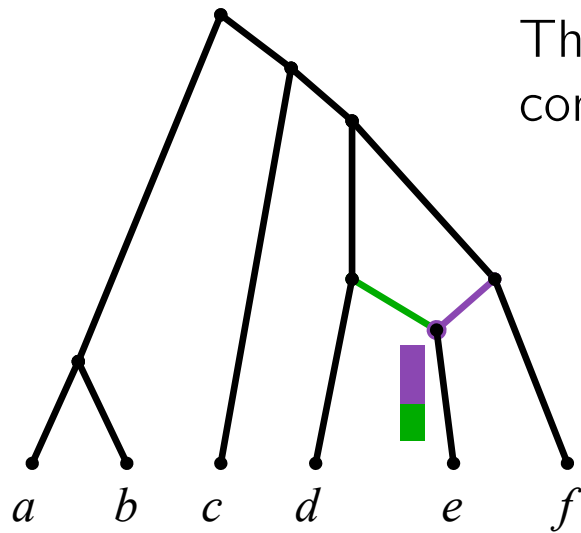
In a phylogenetic network, a reticulate event is represented as a **reticulation**, where branches converge to give rise to a new lineage:



The genome at the start of the new lineage is a composition of those of the parent lineages.

Trees displayed by a network

In a phylogenetic network, a reticulate event is represented as a **reticulation**, where branches converge to give rise to a new lineage:

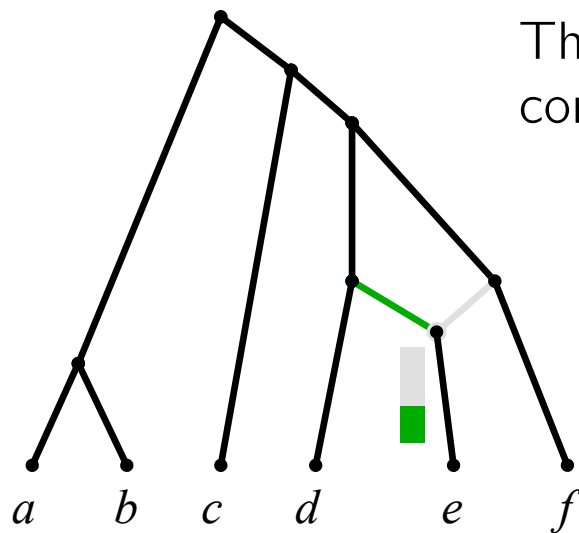


The genome at the start of the new lineage is a composition of those of the parent lineages.

The evolution of each part independently inherited is described by a “*gene*” tree

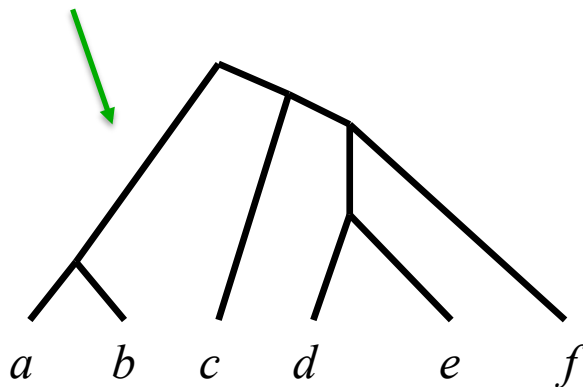
Trees displayed by a network

In a phylogenetic network, a reticulate event is represented as a **reticulation**, where branches converge to give rise to a new lineage:



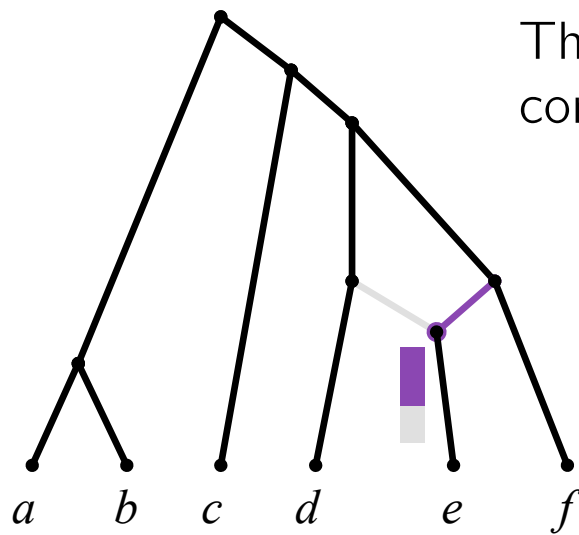
The genome at the start of the new lineage is a composition of those of the parent lineages.

The evolution of each part independently inherited is described by a “*gene*” tree



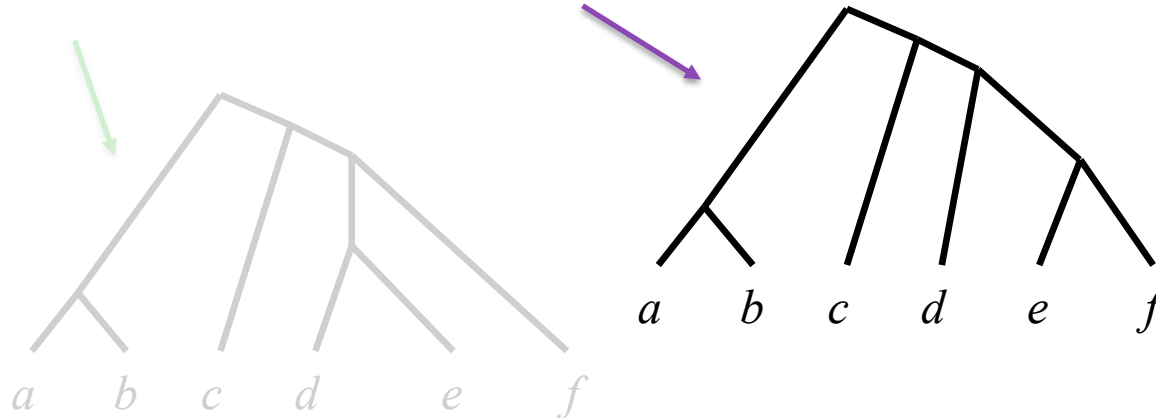
Trees displayed by a network

In a phylogenetic network, a reticulate event is represented as a **reticulation**, where branches converge to give rise to a new lineage:



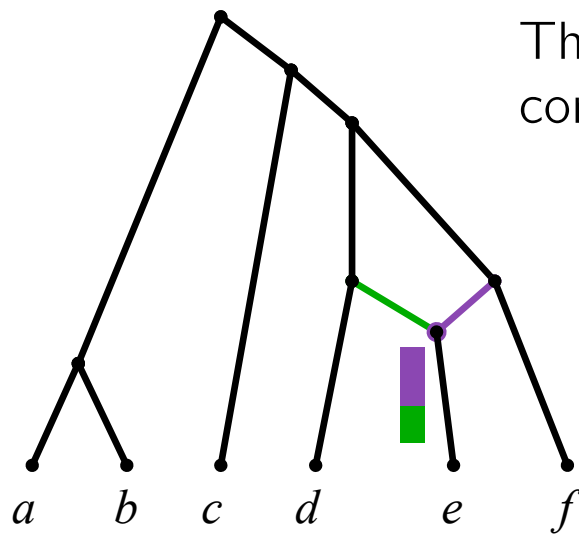
The genome at the start of the new lineage is a composition of those of the parent lineages.

The evolution of each part independently inherited is described by a “*gene*” tree



Trees displayed by a network

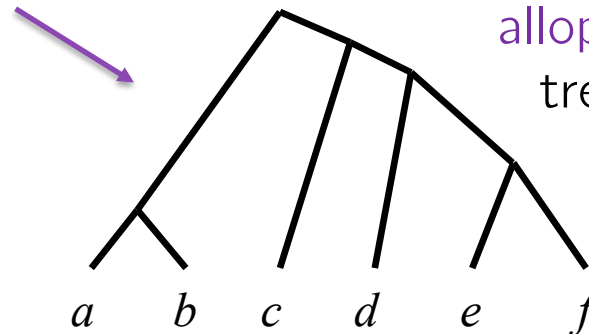
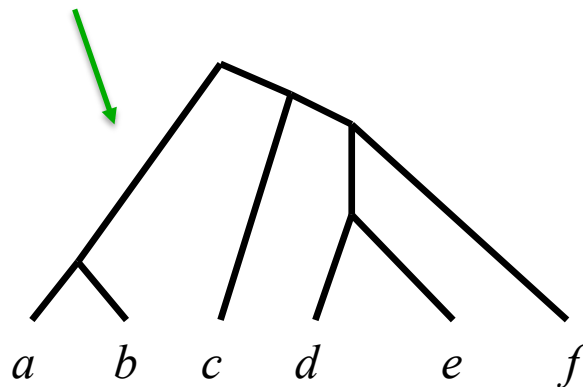
In a phylogenetic network, a reticulate event is represented as a **reticulation**, where branches converge to give rise to a new lineage:



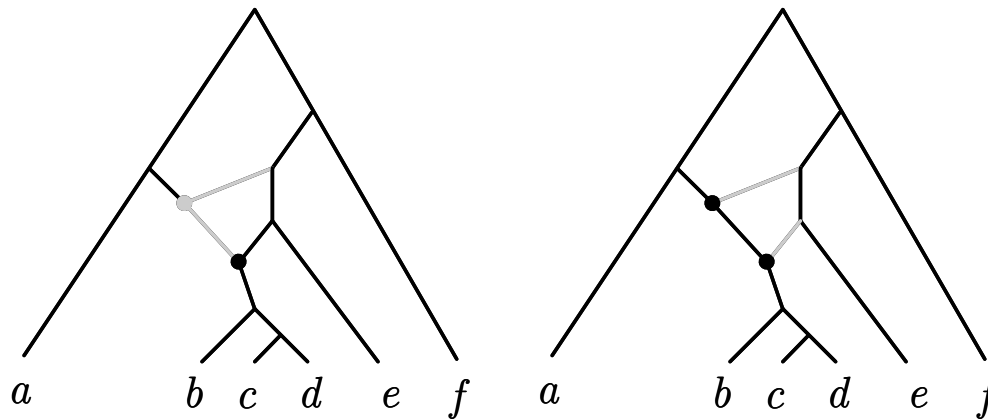
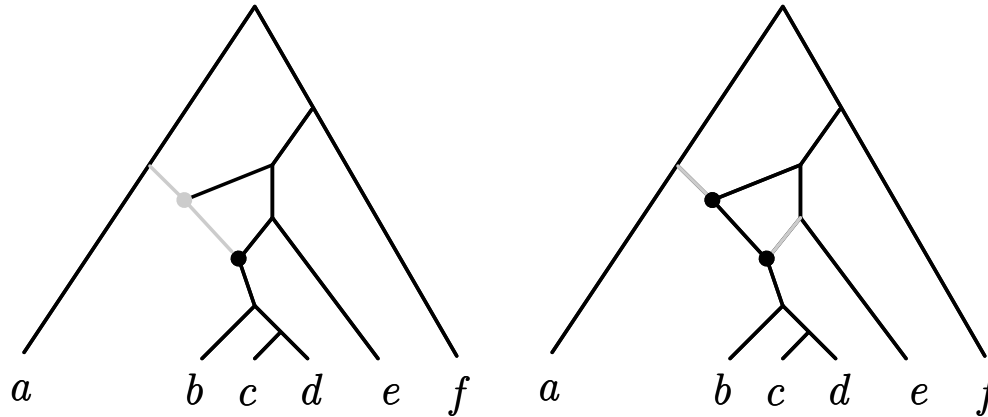
The genome at the start of the new lineage is a composition of those of the parent lineages.

The evolution of each part independently inherited is described by a *“gene” tree*

In the absence of deep coalescence and allopolyploidy, the gene trees are *displayed* by the network

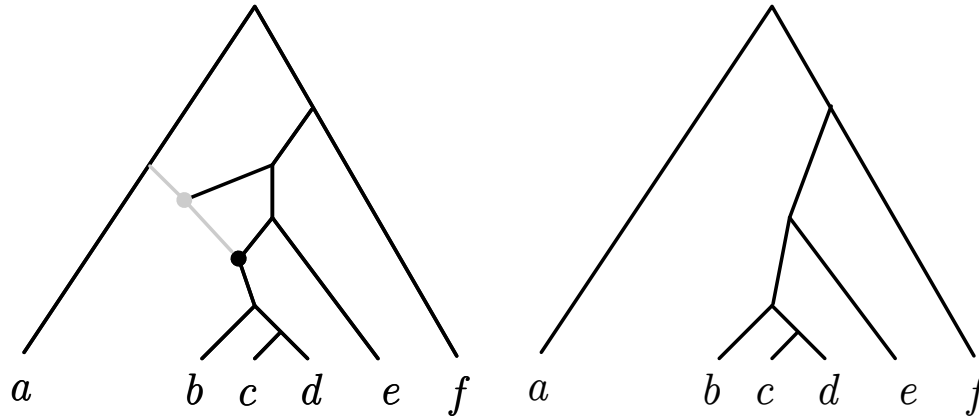


Trees displayed by a network



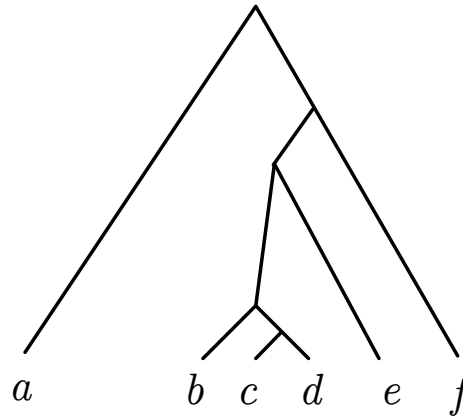
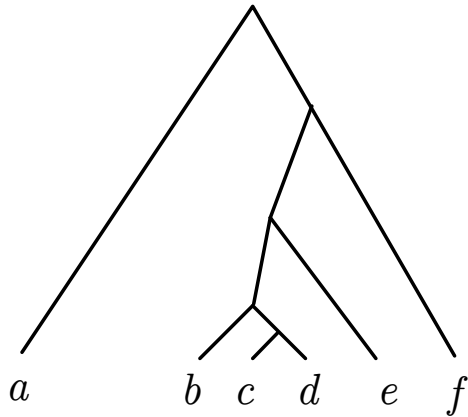
Switch on and
off
reticulated
edged

Trees displayed by a network

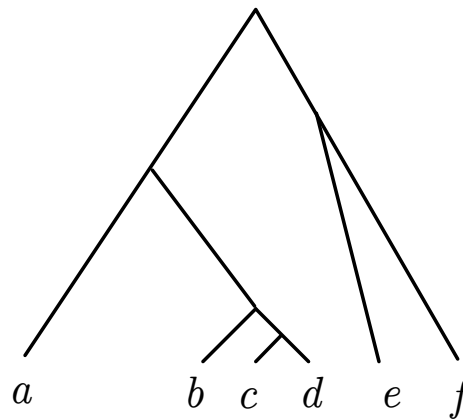
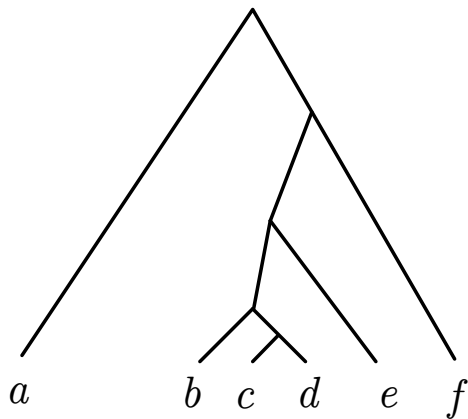


**Delete switched off
edges and unlabelled
leaves and suppress
outdegree-1
indegree-1 nodes**

Trees displayed by a network

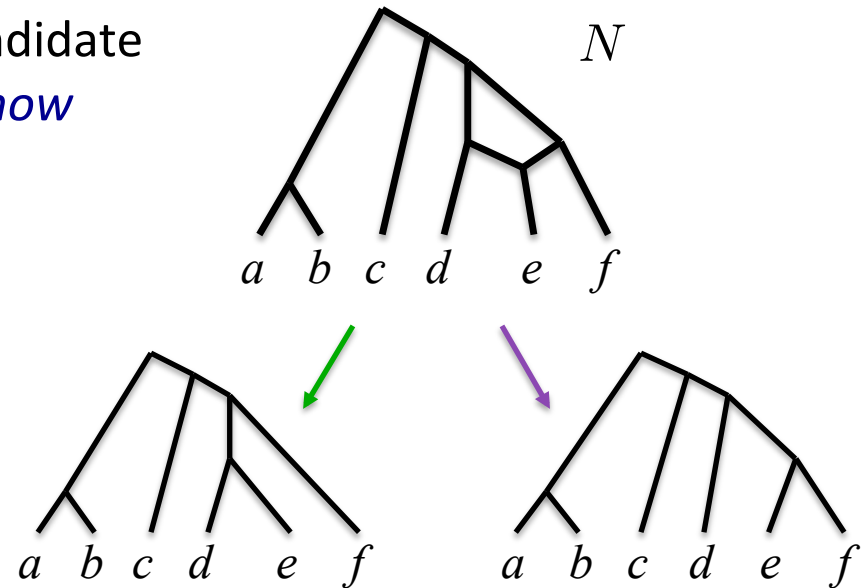


2^r possible trees!!!



Phylogenetic network inference

An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data*:



Many possible formulations:

Data:

Trees with 3 taxa:

(inferred from other data)



Goal:

Find the network N with the lower hybridization number such that the triplets are 'consistent' with one of the trees displayed by N

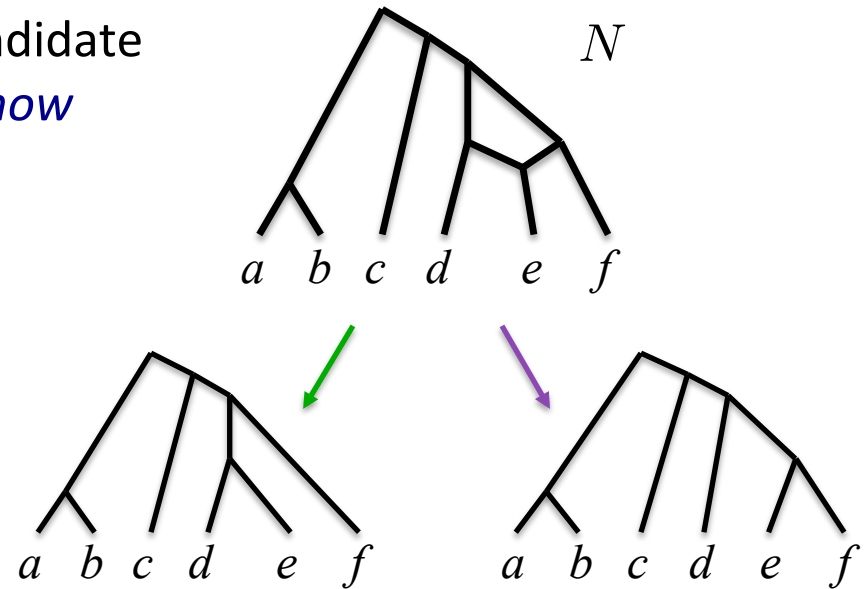
subject to constraints on the complexity of N

Triplets - Software

- **LEV1ATHAN**: A practical algorithm for reconstructing level-1 phylogenetic networks. Combines any set of phylogenetic trees into a level-1 phylogenetic network that is consistent with a large number of the triplets of the input trees.
- **SIMPLISTIC**: Returns a phylogenetic network with minimum level consistent with all input triplets
- **MARLON**: Constructs a level-1 phylogenetic networks with a minimum number of reticulations consistent with a dense set of triplets, if such a network exists
- **LEVEL2**: Constructs a level-2 phylogenetic network consistent with a dense set of triplets, if such a network exists

Phylogenetic network inference

An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data*:



Many possible formulations:

Data:

Clusters of taxa: $\{a, b\}, \{d, e\}, \{d, e, f\}, \{a, b, c, d, e, f\}, \{e, f\}, \{c, d, e, f\}, \dots$

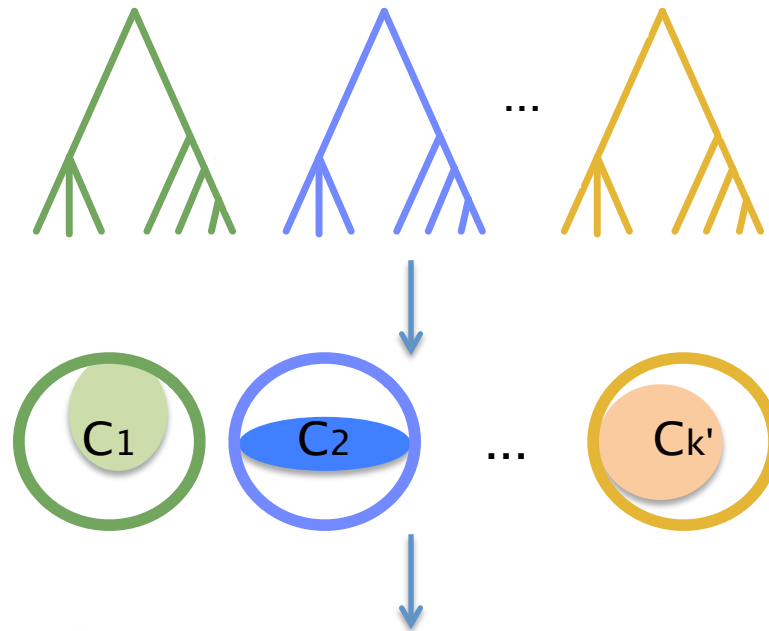
Goal:

Find the network N with the lower hybridization number such that the input clusters are 'explained' by one of the trees displayed by N

subject to constraints on the complexity of N

Clusters

CASS algorithm : search for the level-k network containing a set of clusters

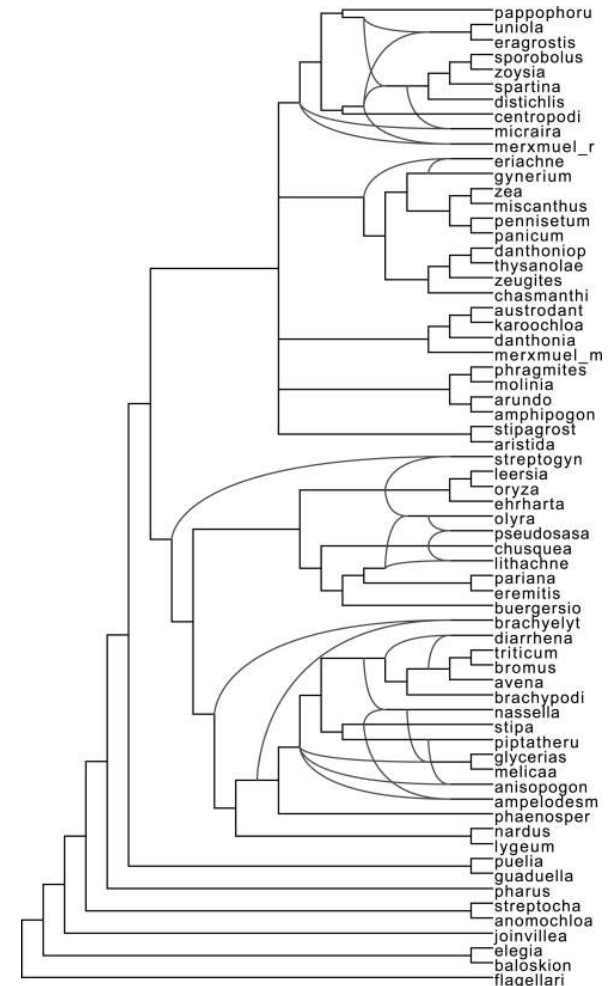


Dendroscope 3

by Daniel H. Huson

with contributions from Benjamin Albrecht,
Philippe Gambette, Leo van Iersel,
Celine Scornavacca and others.

www-ab.informatik.uni-tuebingen.de/software/dendroscope

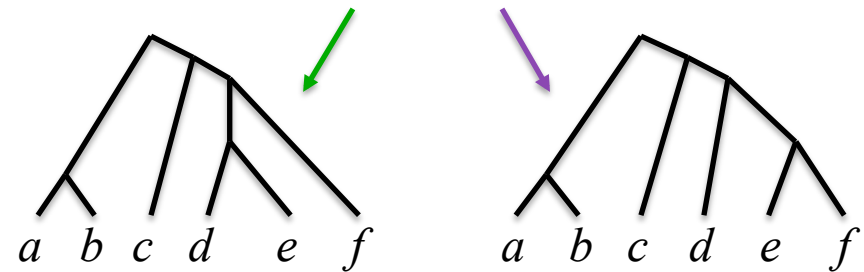


Phylogenetic network inference

An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data*:



Many possible formulations:



Data:

Any trees on the same taxa:
(inferred from other data)



Goal:

Find the network N with the lower hybridization number such that the input trees are 'consistent' with one of the trees displayed by N

subject to constraints on the complexity of N

Software

Hybroscale 1.5



Hybroscale

by
Benjamin Albrecht

www.bio.ifi.lmu.de/software



Dendroscope 3

by Daniel H. Huson

with contributions from Benjamin Albrecht,
Philippe Gambette, Leo van Iersel,
Celine Scornavacca and others.

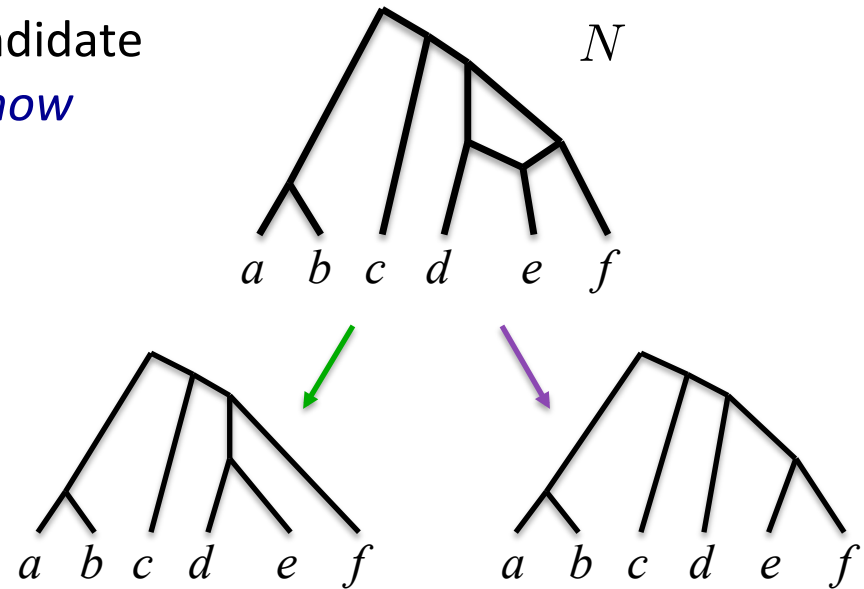
www-ab.informatik.uni-tuebingen.de/software/dendroscope

ultraNet

An UltraFast Tool for Minimum Reticulate Networks

Phylogenetic network inference

An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data*:



Many possible formulations:

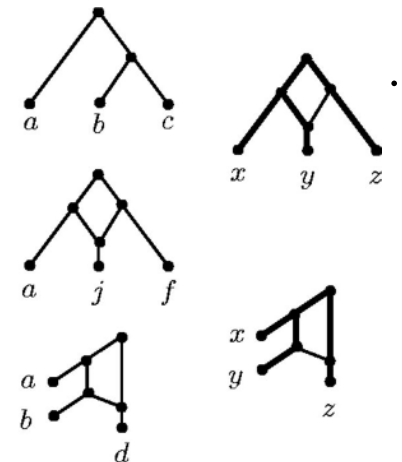
Data:

Any trinetts on the same taxa:
(inferred from other data)

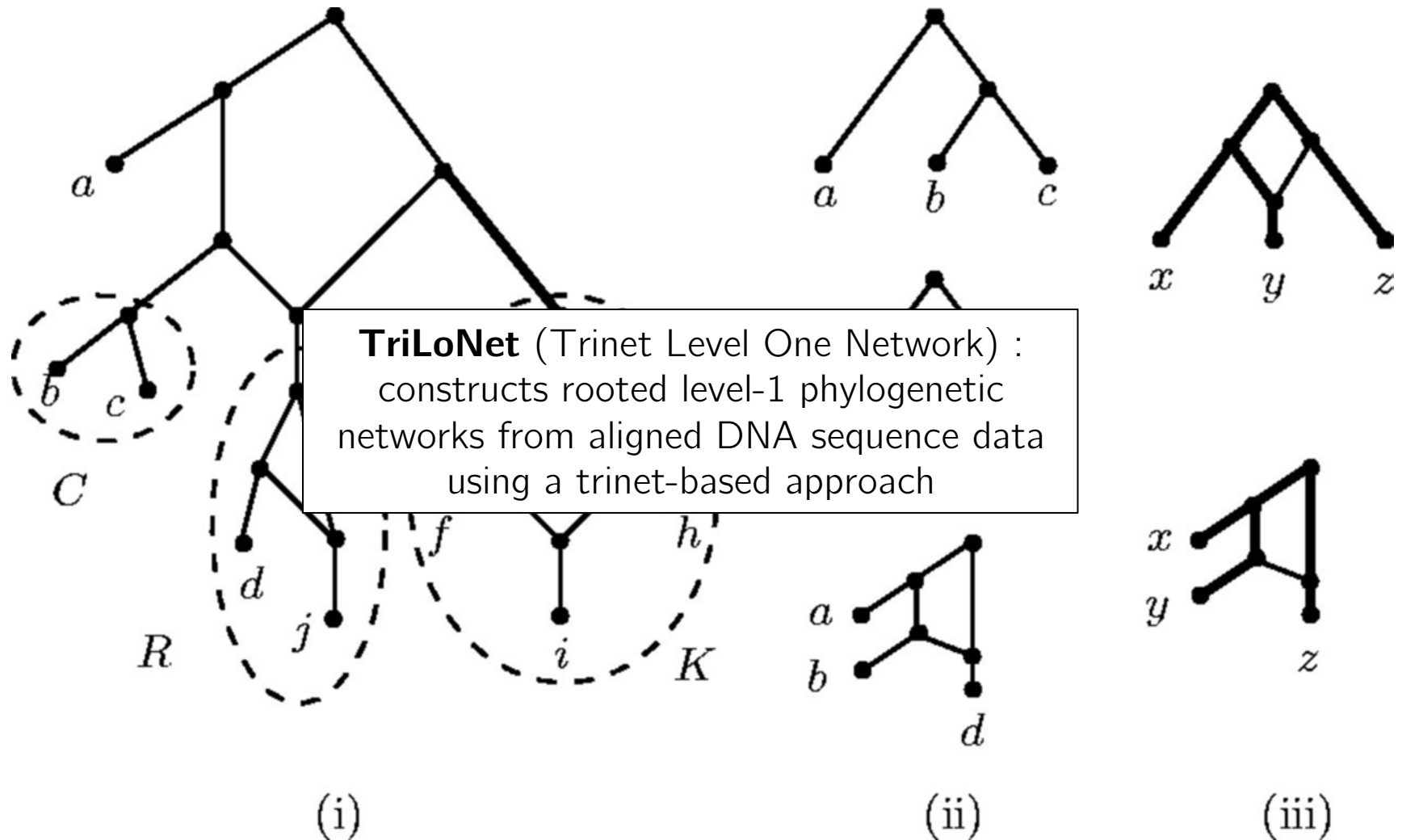
Goal:

Find the network N with the lower hybridization number such that the input trees are 'consistent' with the N

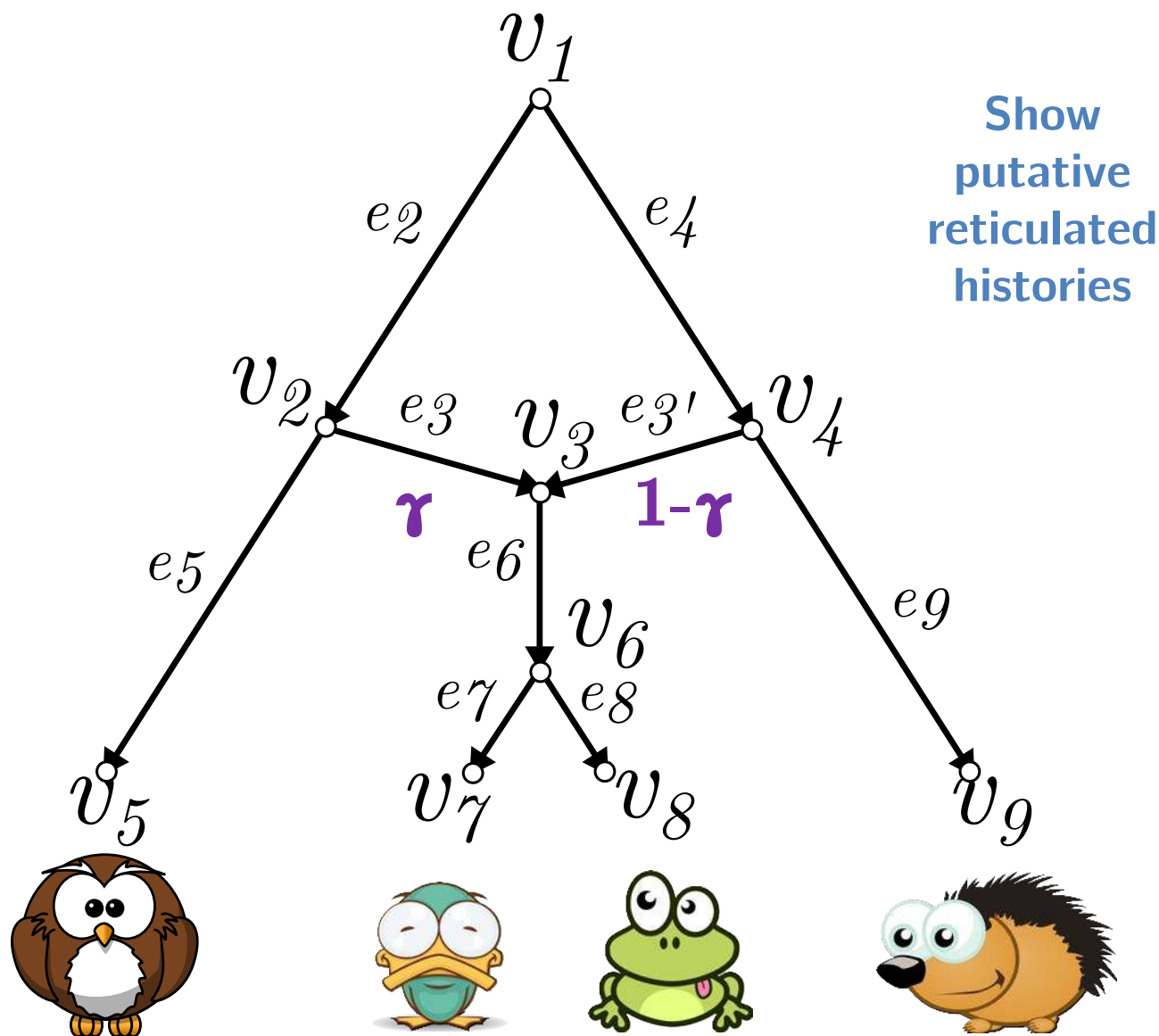
subject to constraints on the complexity of N



Trinets

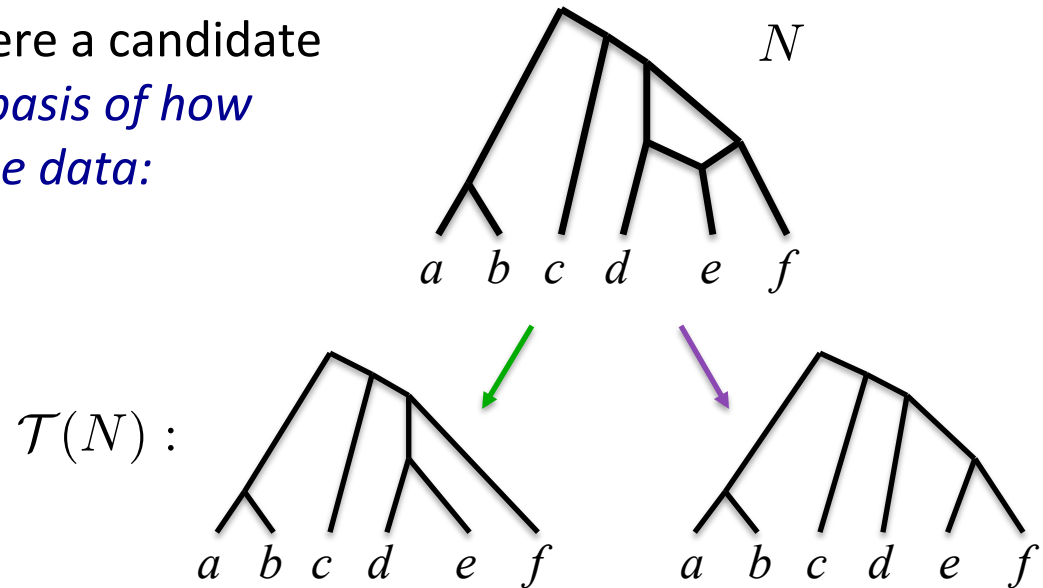


Explicit phylogenetic networks (rDAG)



Phylogenetic network inference

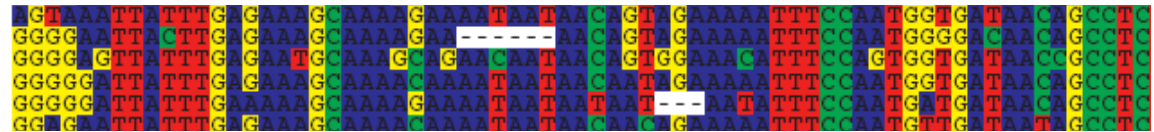
An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data*:



Many possible formulations:

Data:

Sequence alignments:
(typically given in blocks)



Goal:

Find N that minimizes
$$F(N|A_1, A_2, \dots, A_m) = \sum_{i=1}^m \min_{T \in \mathcal{T}(N)} F(T|A_i)$$

subject to constraints on the complexity of N . $F()$ is the parsimony score.

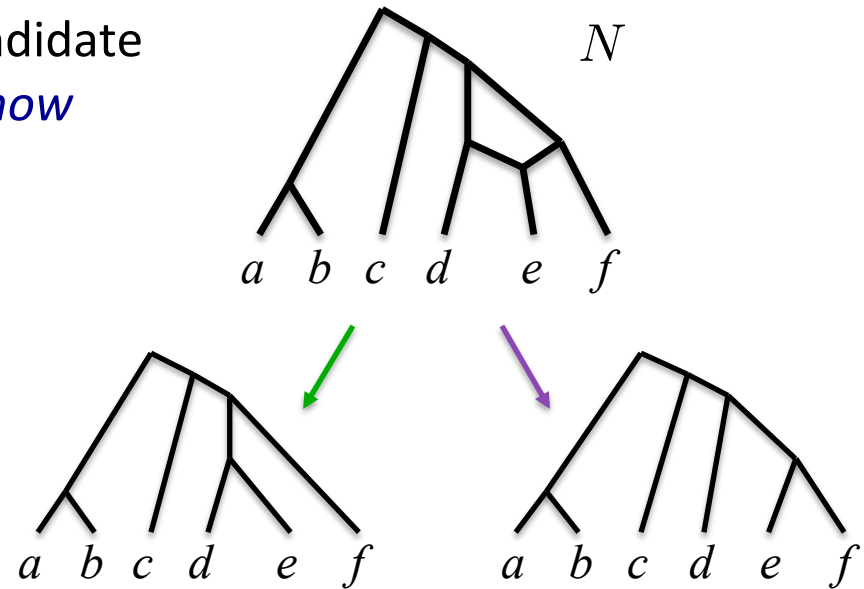
Phylogenetic network inference

An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data*:

NEPAL
Phylogenetic Networks
Parsimony and Likelihood
Toolkit

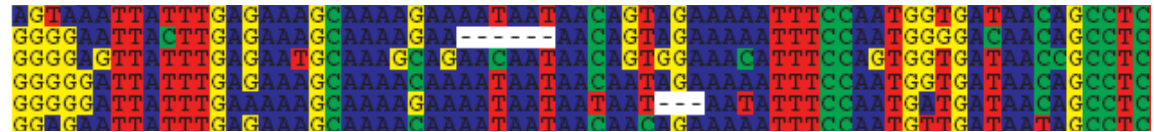
Many possible formulations:

$\mathcal{T}(N)$:



Data:

Sequence alignments:
 (typically given in blocks)



A_1

A_2

$\dots$

A_m

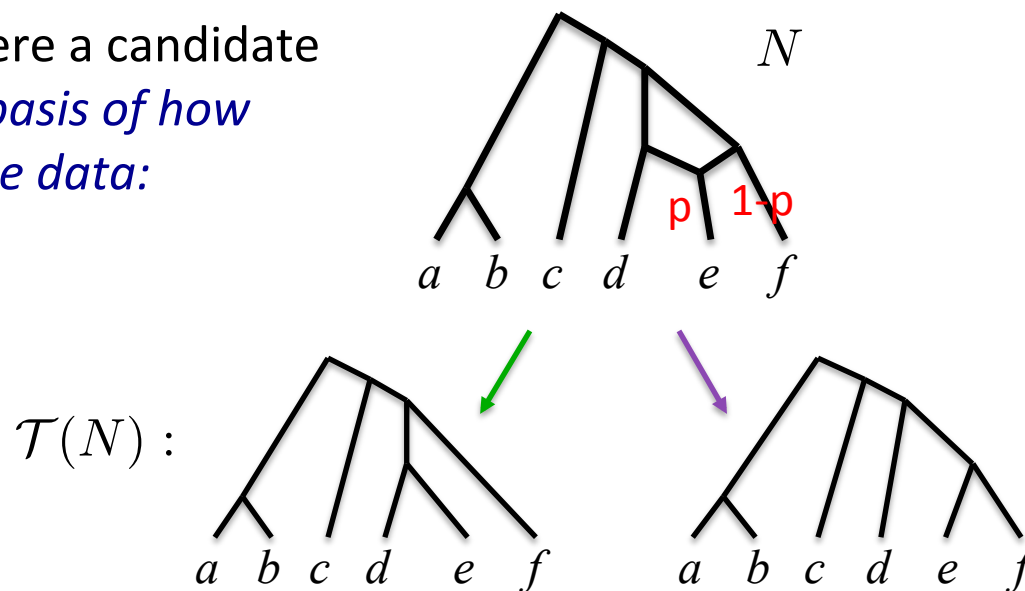
Goal:

Find N that minimizes
$$F(N|A_1, A_2, \dots, A_m) = \sum_{i=1}^m \min_{T \in \mathcal{T}(N)} F(T|A_i)$$

subject to constraints on the complexity of N . $F()$ is the parsimony score.

Phylogenetic network inference

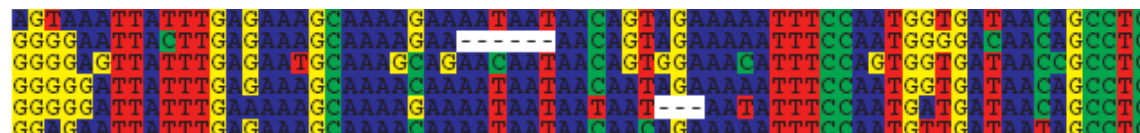
An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data*:



Many possible formulations:

Data:

Sequence alignments:
(typically given in blocks)



Goal:

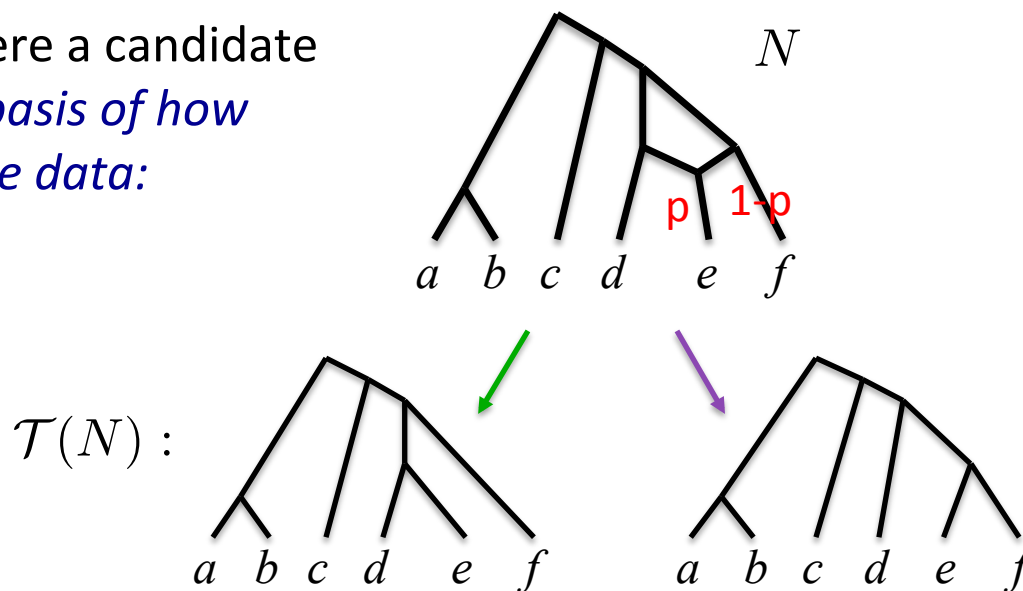
Find N that maximises $\Pr(A_1, A_2, \dots, A_m | N) = \prod_{i=1}^m \Pr(A_i | N) = \prod_{i=1}^m \left(\sum_{T \in \mathcal{T}(N)} \Pr(A_i | T) \Pr(T | N) \right)$

Phylogenetic network inference

An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data*:

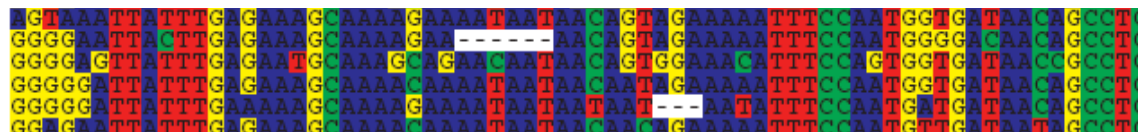
NEPAL
Phylogenetic Networks
Parsimony and Likelihood
Toolkit

Many possible formulations:



Data:

Sequence alignments:
 (typically given in blocks)



A_1

A_2

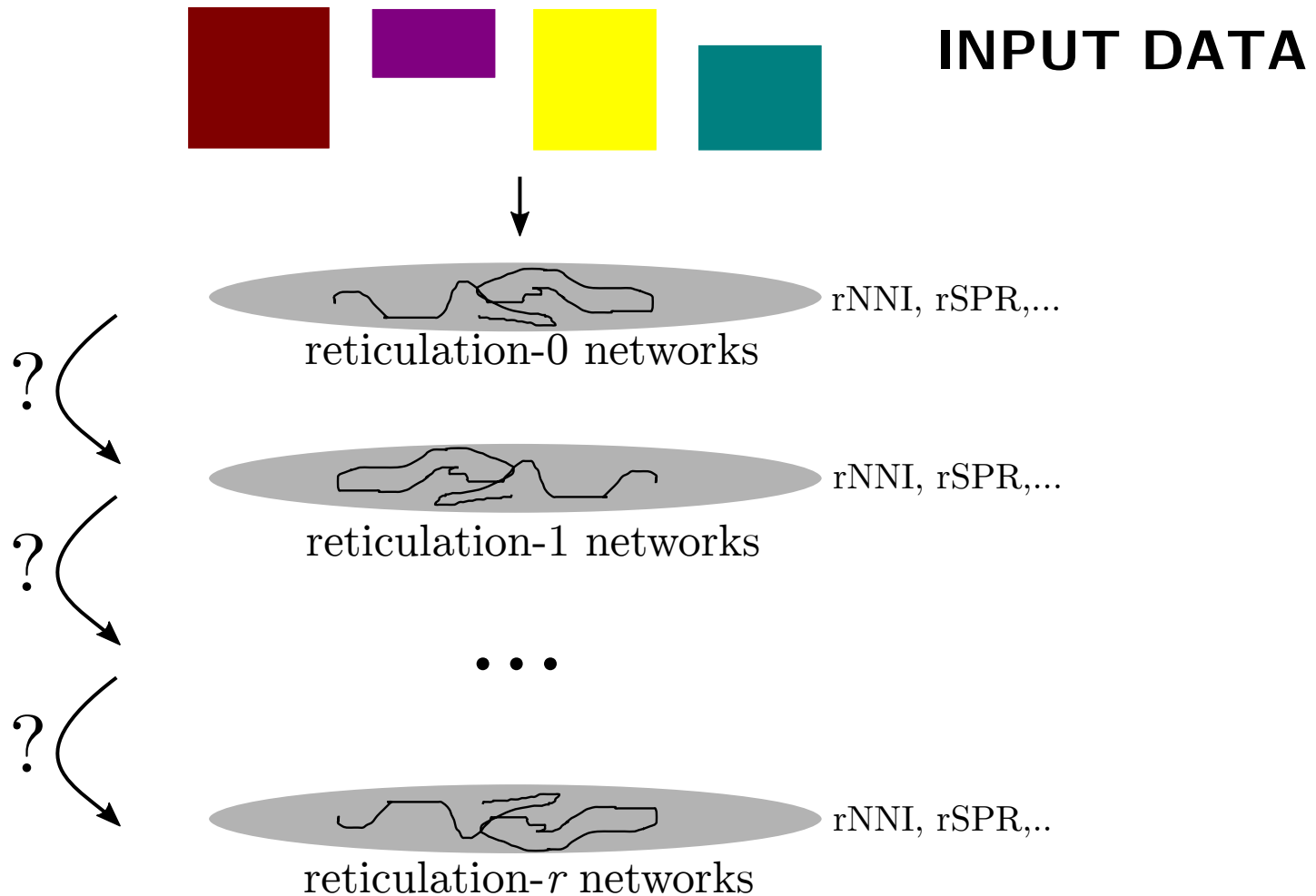
$\dots$

A_m

Goal:

Find N that maximises $\Pr(A_1, A_2, \dots, A_m | N) = \prod_{i=1}^m \Pr(A_i | N) = \prod_{i=1}^m \left(\sum_{T \in \mathcal{T}(N)} \Pr(A_i | T) \Pr(T | N) \right)$

The strategy



Some issues

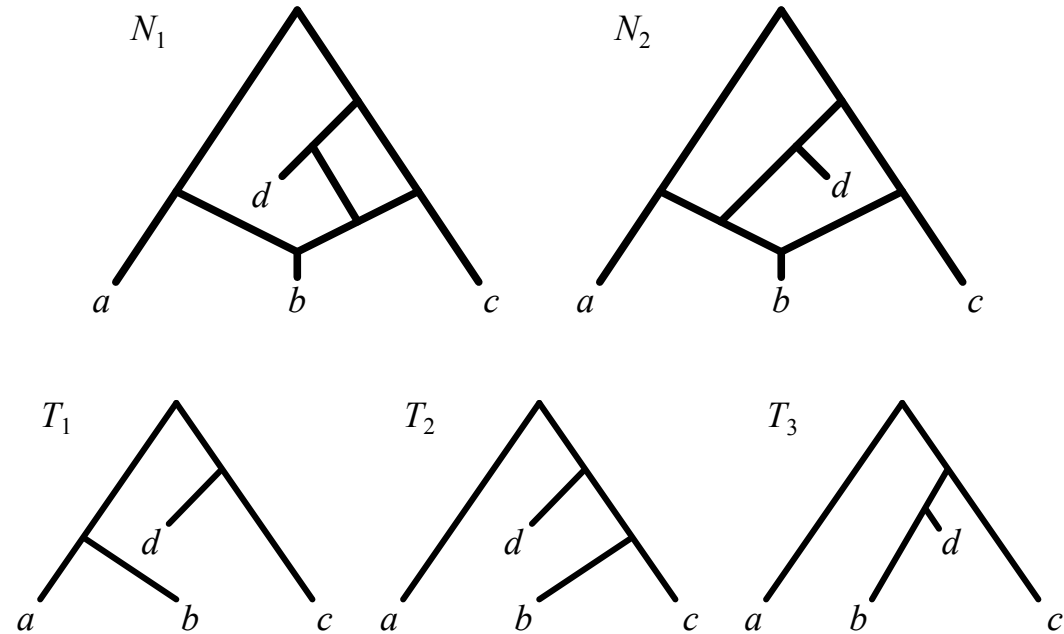
- **Searching the space of phylogenetic networks**
The space of networks with k reticulations is infinite.
- **Controlling for Model Complexity**
Because any network with k reticulations provides a more complex model than any network with $(k-1)$ reticulations, we must handle the model selection problem (AIC, BIC, K-fold cross-validation, ...).
- **Identifiability issues**

$$\Pr(A_1, A_2, \dots, A_m | N) = \prod_{i=1}^m \Pr(A_i | N) = \prod_{i=1}^m \left(\sum_{T \in \mathcal{T}(N)} \Pr(A_i | T) \Pr(T | N) \right)$$

- **Not accounting for ILS and allopolyploidy**

Different networks can display the same trees

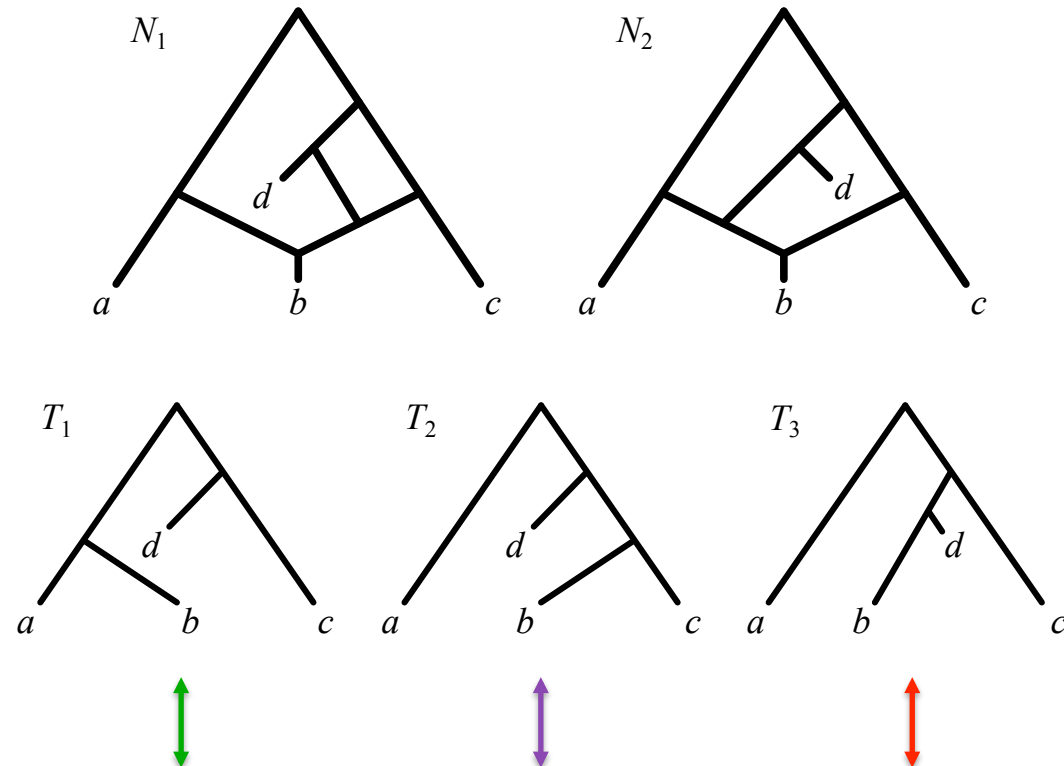
Some networks display exactly the same trees:



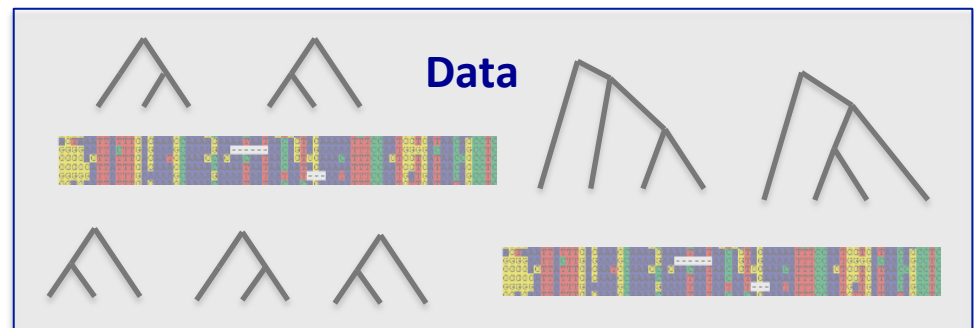
Different networks can display the same trees

Some networks display exactly the same trees:

Because N_1 and N_2 display the same trees, they are equally good to any of the inference methods we saw – *no matter the input data*



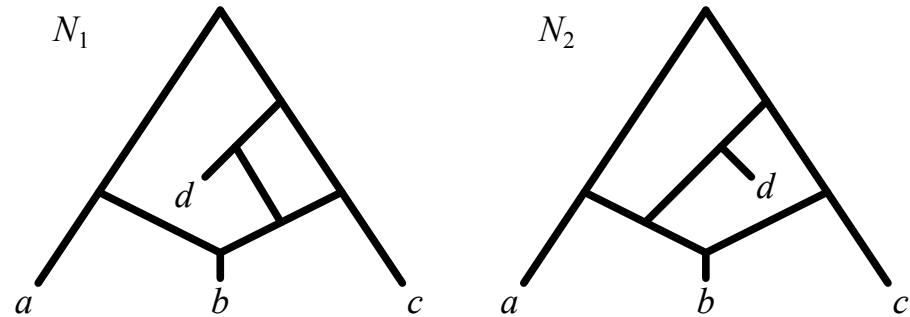
(Recall that a network is evaluated *on the basis of how well the trees it displays fit the data*)



Different networks can display the same trees

Some networks display exactly the same trees:

Because N_1 and N_2 display the same trees, they are equally good to any of the inference methods we saw – *no matter the input data*

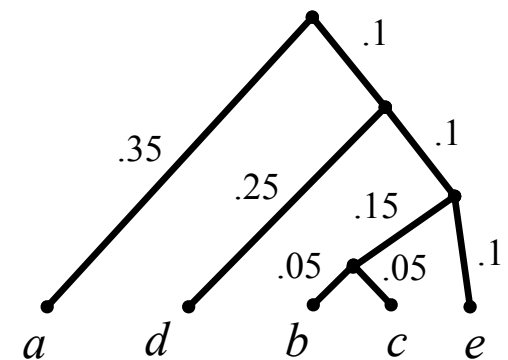
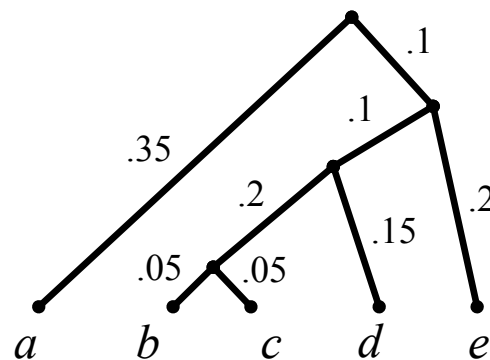
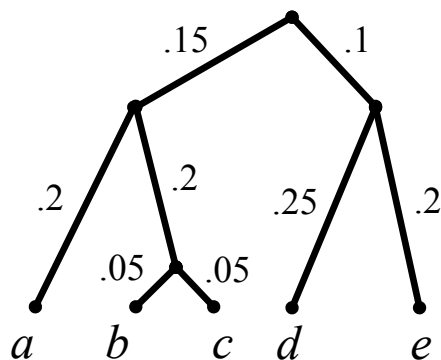
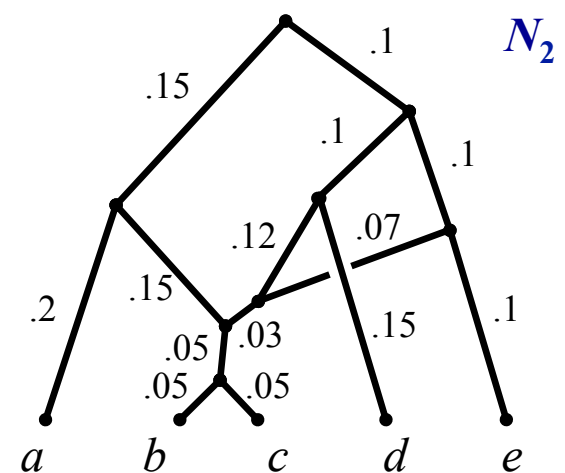
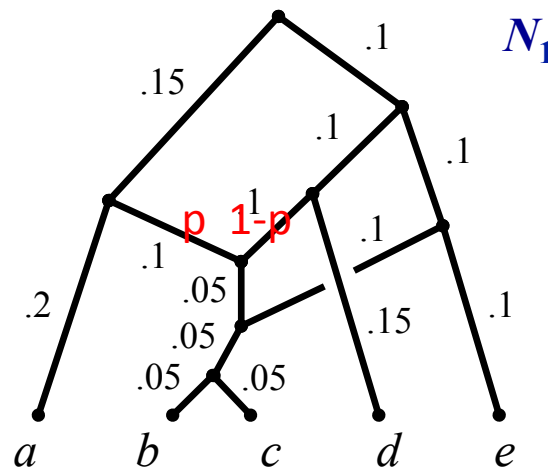


UNIDENTIFIABILITY

Indistinguishable networks

Branch lengths do not eliminate non-identifiability...

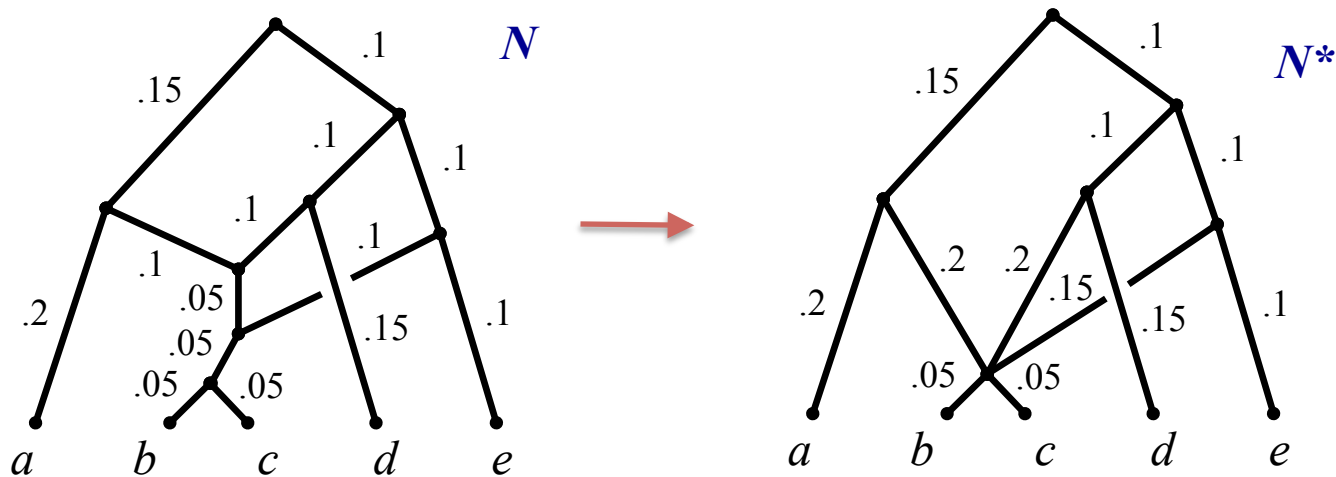
The same hold for inheritance probabilities



N_1 and N_2 display the same trees (i.e. including branch lengths) and are thus *indistinguishable* even to methods accounting for lengths

What it means for the evolutionary biologist

If N is reconstructed by a "classic" inference method, then even assuming perfect and unlimited data, the best you can hope is that the true phylogenetic network is just one of the many that are indistinguishable from N ...



The canonical form of N is a unique representative of the networks indistinguishable from N , that excludes their unrecoverable aspects...

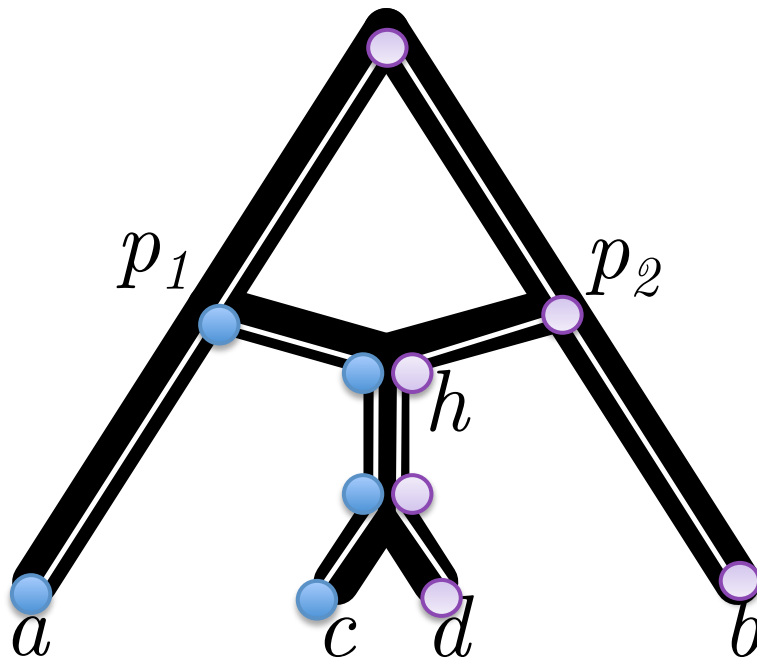
*Methods for reconstructing rooted
phylogenetic networks not
accounting for ILS*

Deep coalescence (ILS)

(a) Population view

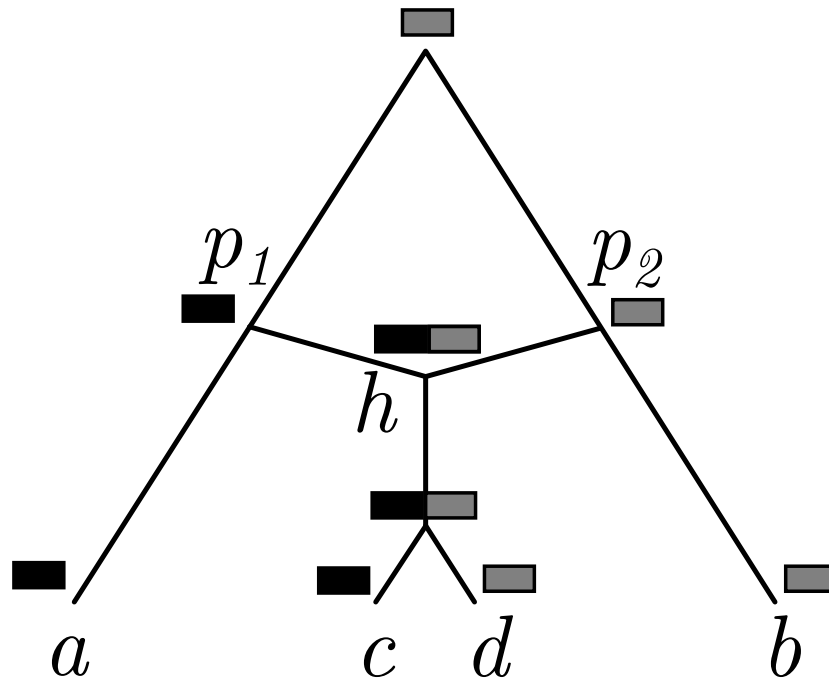
(b) Reconciliation representation

ILS in phylogenetic networks



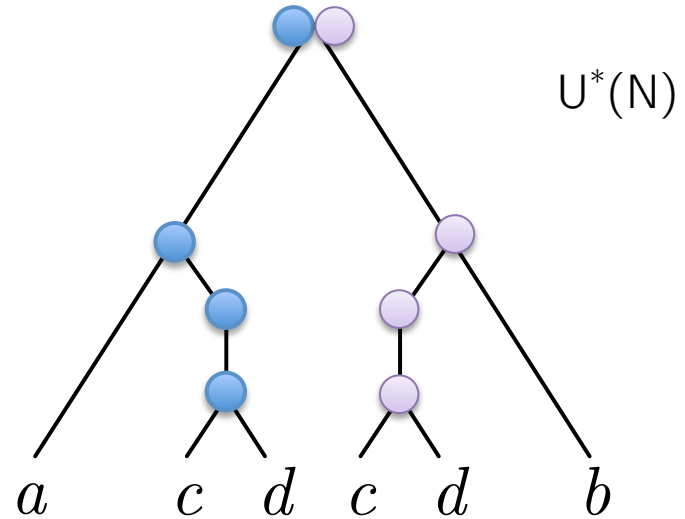
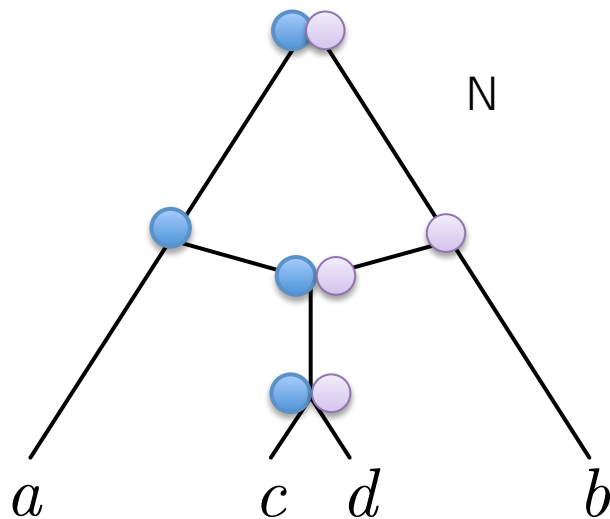
The true gene tree is **not displayed** by the network because it needs to *use* both edges entering the hybrid node

Allopolyploidy



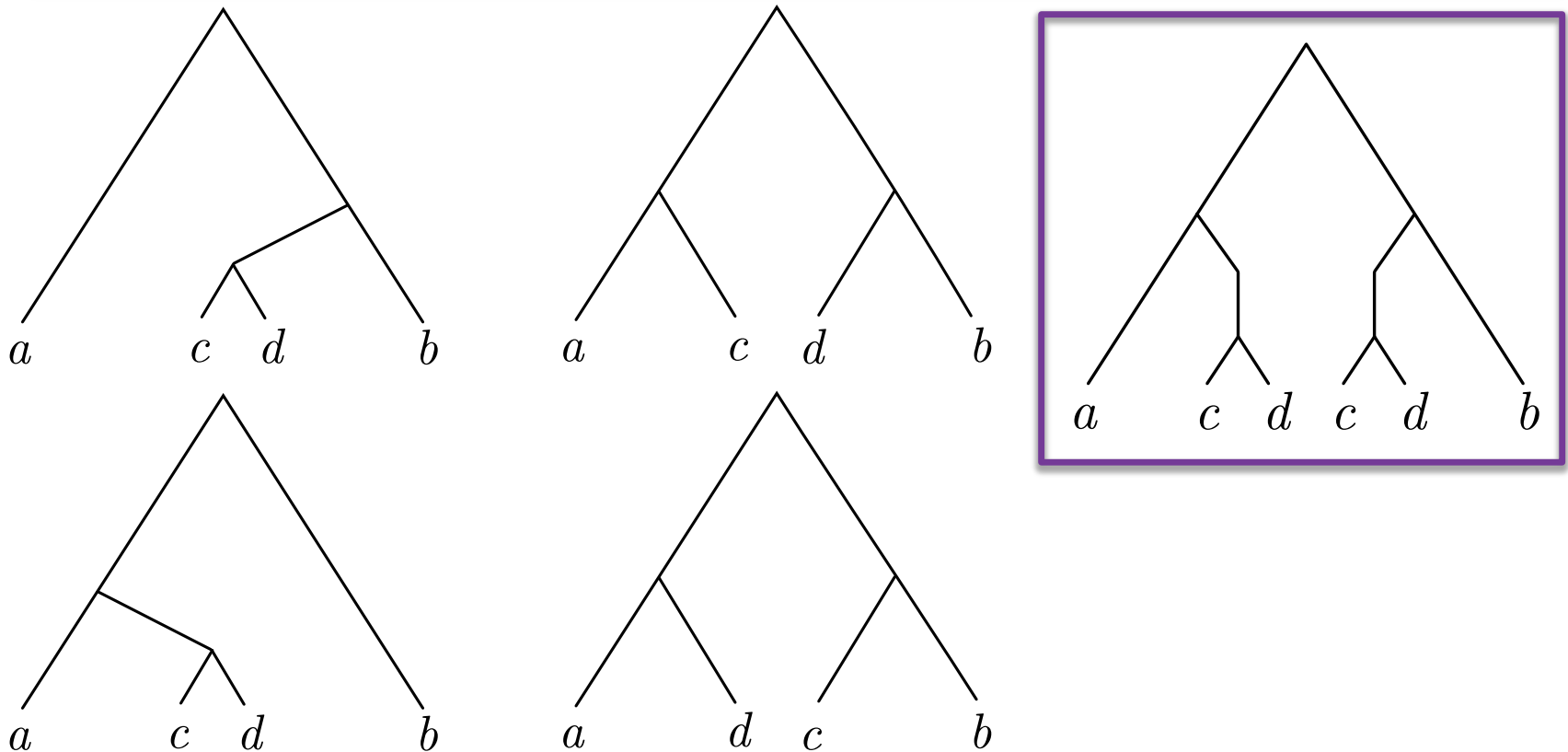
The true gene tree is **not displayed** by the network because it needs to *use* both edges entering the hybrid node

The multi-labelled tree $U^*(N)$



- **nodes** are the directed paths in N starting at $r(N)$
- for each pair of paths p, p' in N , there is an **edge** in $U^*(N)$ from p to p' if and only if $p = p'e$ for some edge e in N
- each node in $U^*(N)$ corresponding to a path in N that starts at $r(N)$ and ends at x in X is **labelled** by x

Parental trees



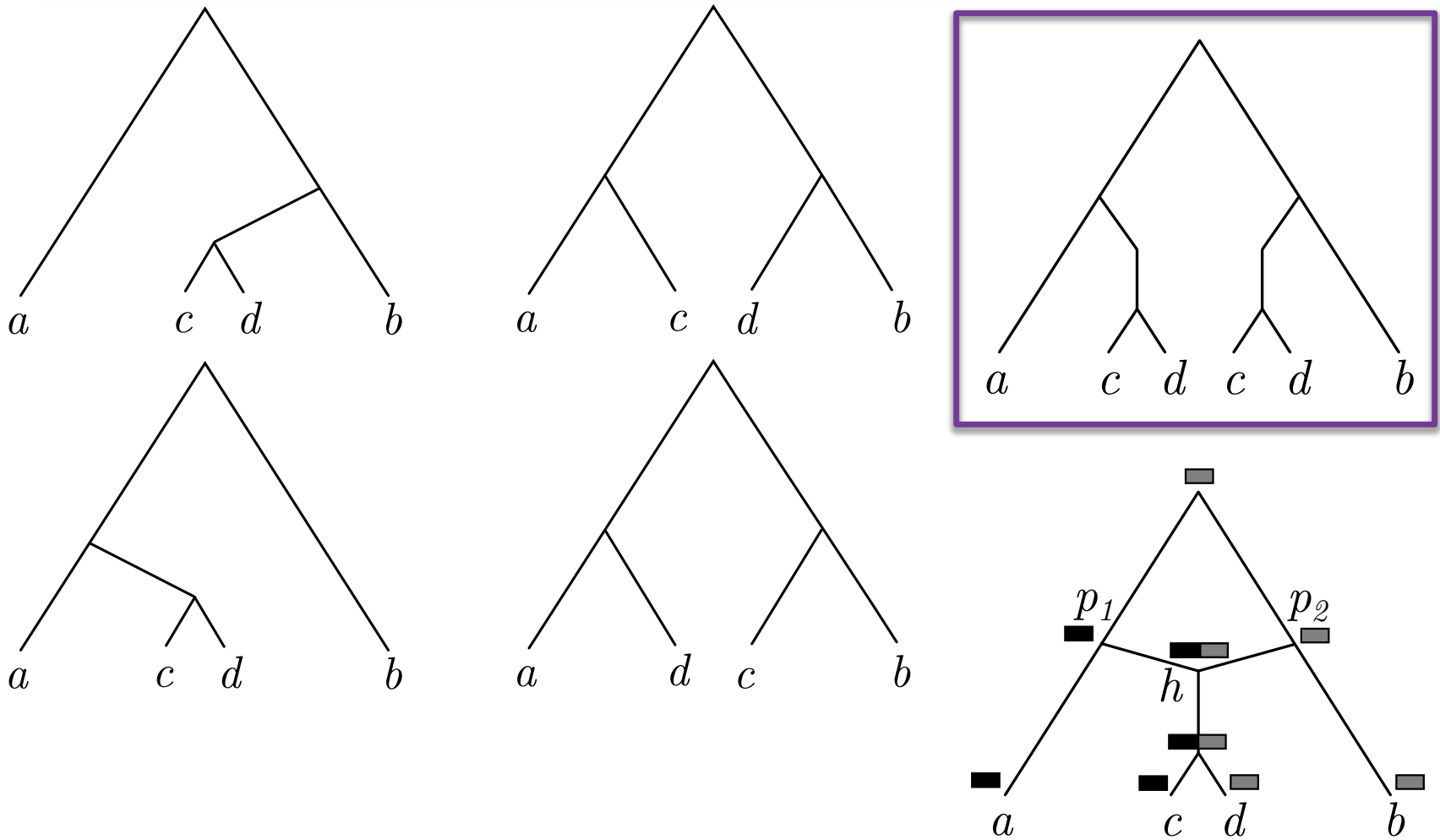
A phylogenetic tree T on X is a parental tree of N if it is displayed by $U^*(N)$

Huber et al. Folding and unfolding phylogenetic trees and networks, 2016 [[weakly displayed](#)]

Zhu al. In the light of deep coalescence: revisiting trees within networks, 2016

Zhu and Degnan. Displayed trees do not determine distinguishability under the network multispecies coalescent, 2016

Parental trees

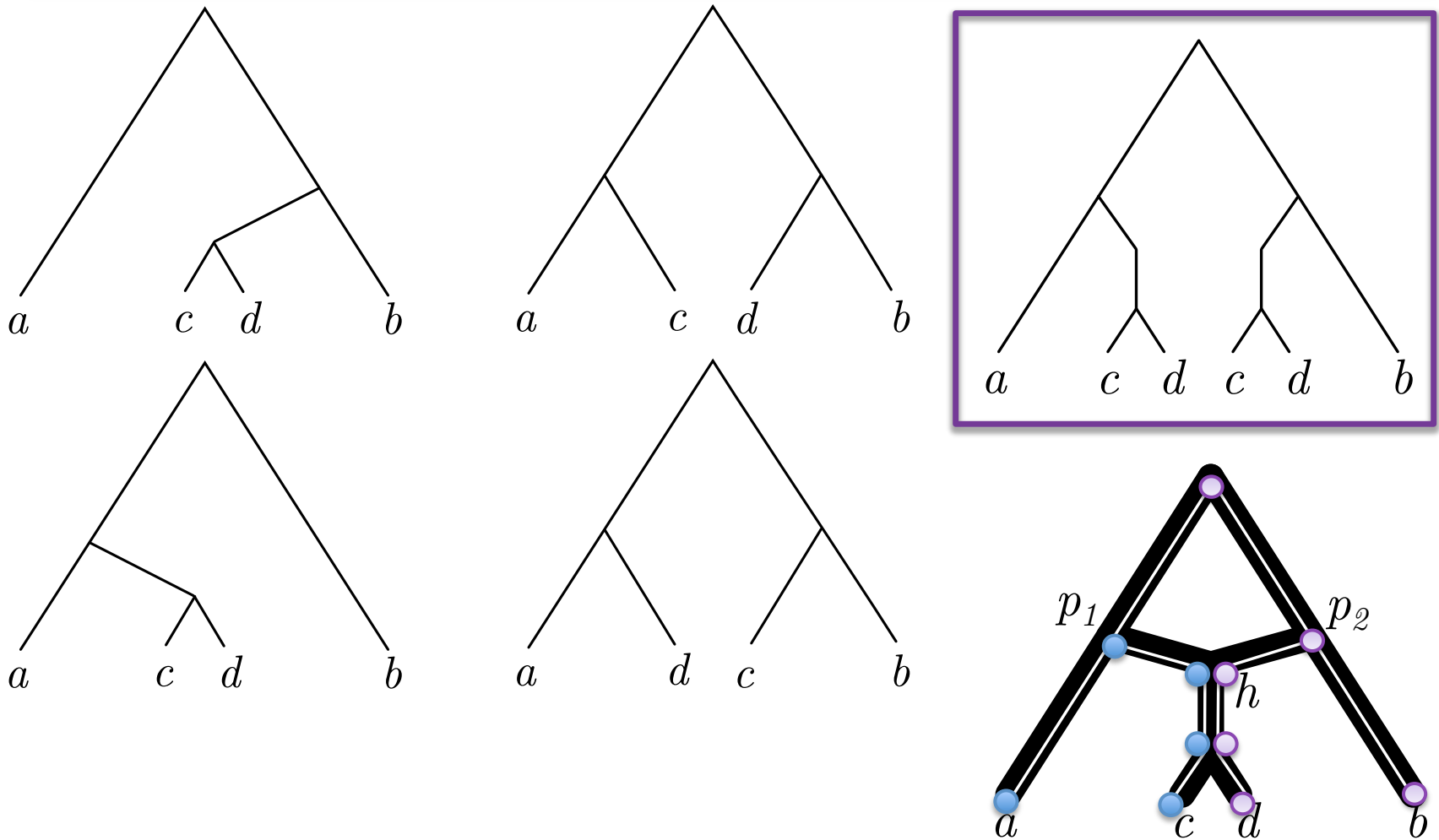


Huber et al. Folding and unfolding phylogenetic trees and networks, 2016 [[weakly displayed](#)]

Zhu al. In the light of deep coalescence: revisiting trees within networks, 2016

Zhu and Degnan. Displayed trees do not determine distinguishability under the network multispecies coalescent, 2016

Parental trees

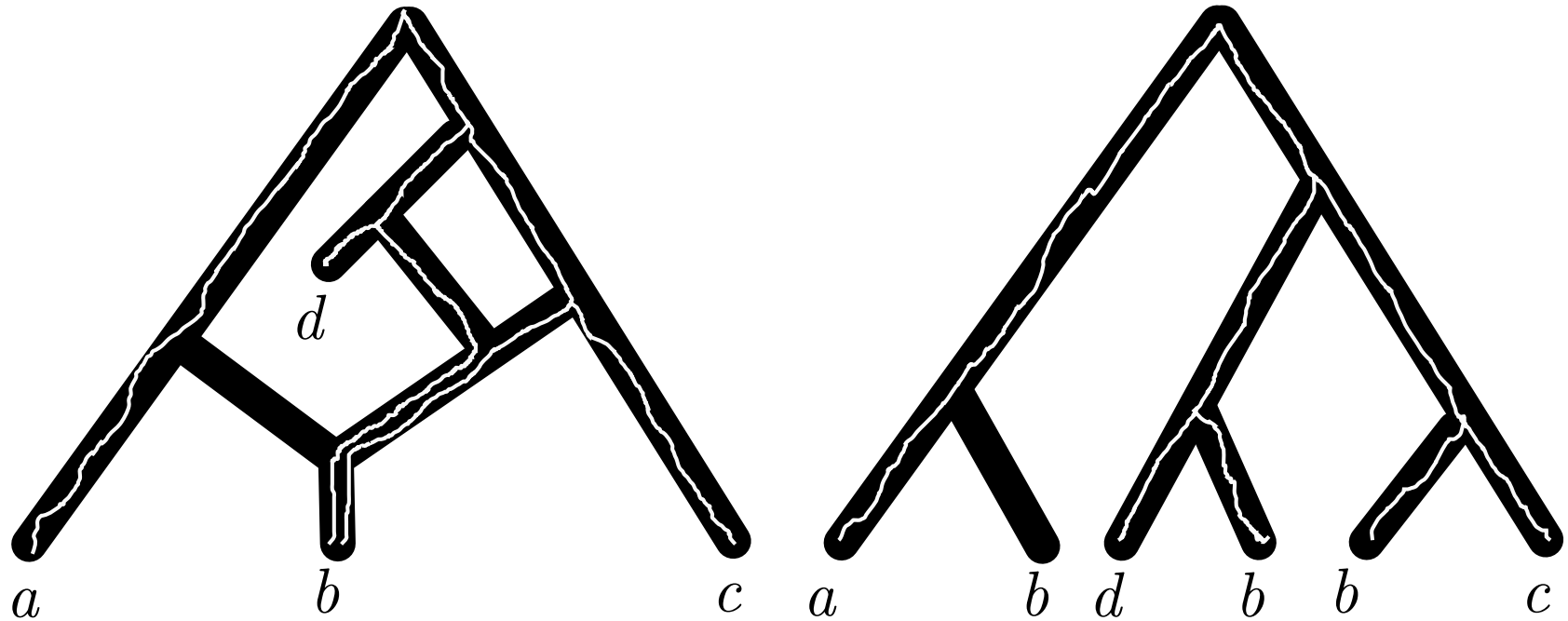


Huber et al. Folding and unfolding phylogenetic trees and networks, 2016 [[weakly displayed](#)]

Zhu al. In the light of deep coalescence: revisiting trees within networks, 2016

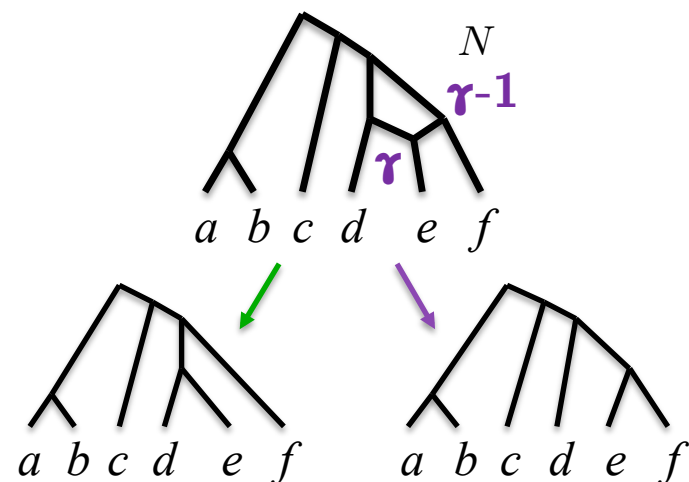
Zhu and Degnan. Displayed trees do not determine distinguishability under the network multispecies coalescent, 2016

Parental trees can be multi-labelled



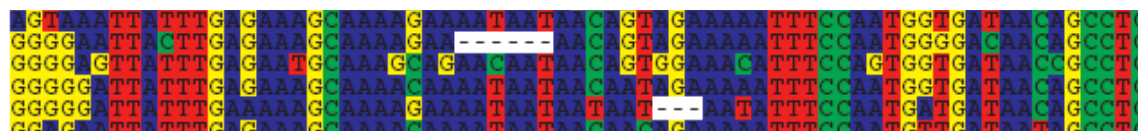
multiple individuals per species are allowed

Scoring schemes based on **parental** trees (NMSC)



Data:

Sequence alignments:
(typically given in blocks)



Goal:

Find N that maximises $\Pr(A_1, A_2, \dots, A_m | N) = \prod_{i=1}^m \Pr(A_i | N) = \prod_{i=1}^m \left(\sum_{T \in \mathcal{T}(N)} \Pr(A_i | T) \Pr(T | N) \right)$

Yu et al. The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection, 2012

Yu et al. Maximum likelihood inference of reticulate evolutionary histories, 2014

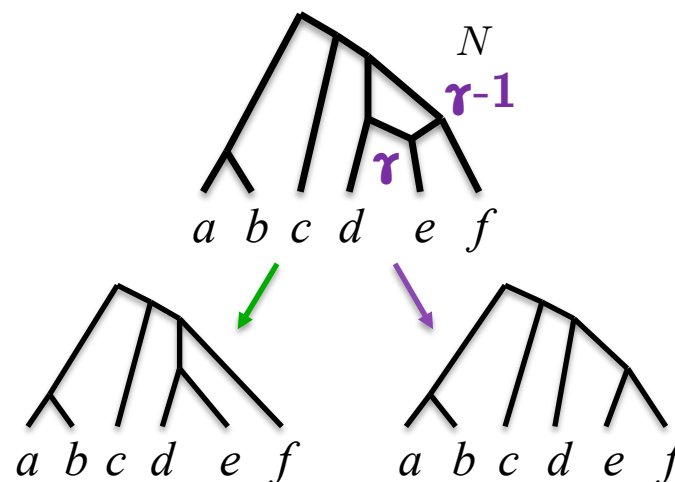
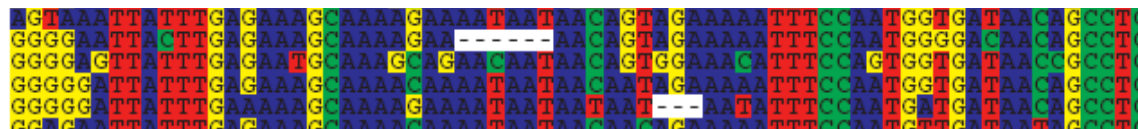
Wen et al. PLOS Genetics 2016 (Bayesian method)

Scoring schemes based on **parental** trees (NMSC)

PhyloNet

Data:

Sequence alignments:
(typically given in blocks)



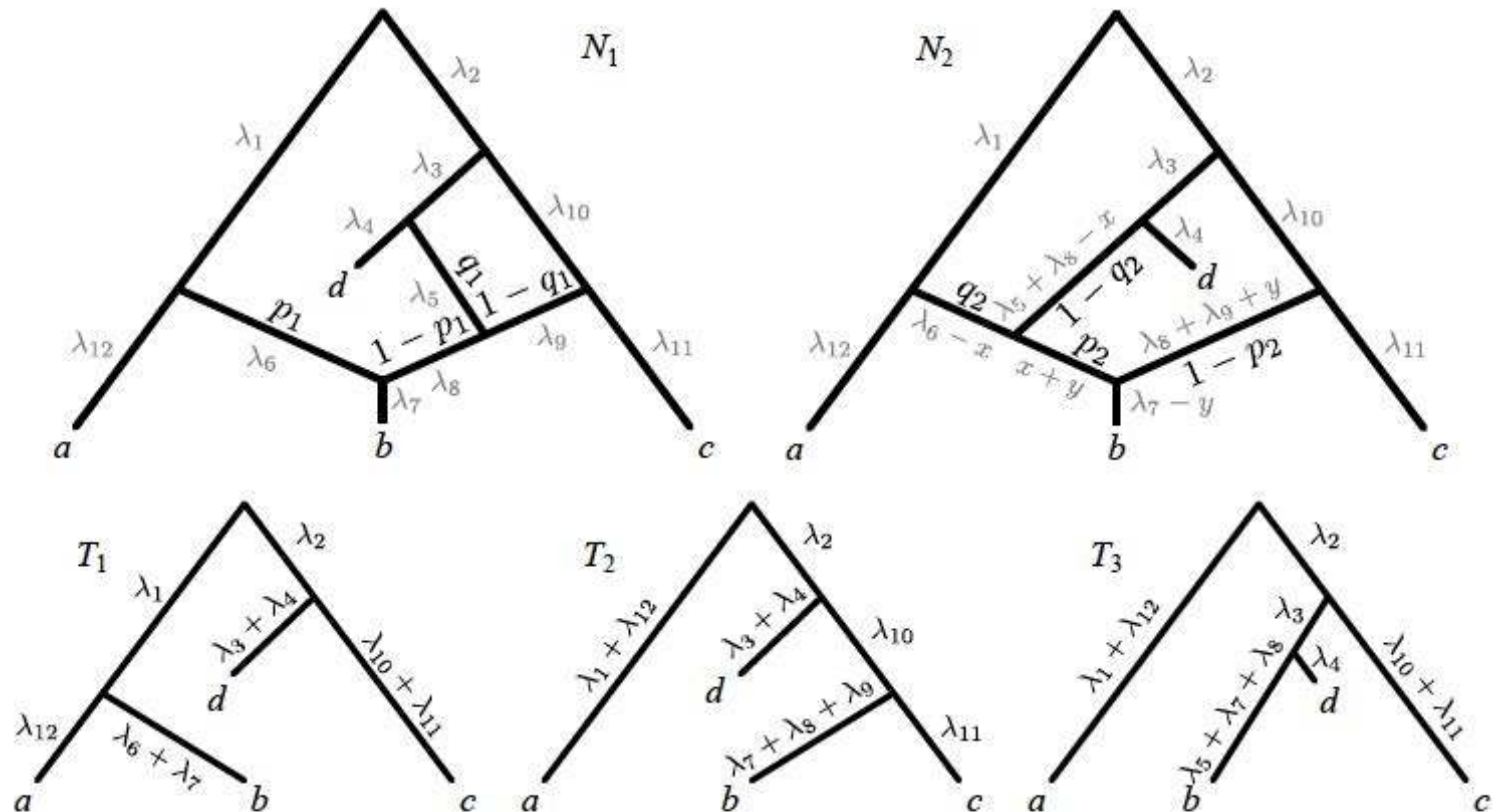
Goal:

Find N that maximises

$$\Pr(A_1, A_2, \dots, A_m | N) = \prod_{i=1}^m p(G_i | N).$$

Zhu and Degnan. Displayed trees do not determine distinguishability under the network multispecies coalescent, 2016

Scoring schemes based on **parental** trees (NMSC)



$$\Pr(T_1|N_1) = p_1$$

$$\Pr(T_2|N_1) = (1 - p_1)(1 - q_1)$$

$$\Pr(T_3|N_1) = (1 - p_1)q_1$$

$$\Pr(T_1|N_2) = p_2q_2$$

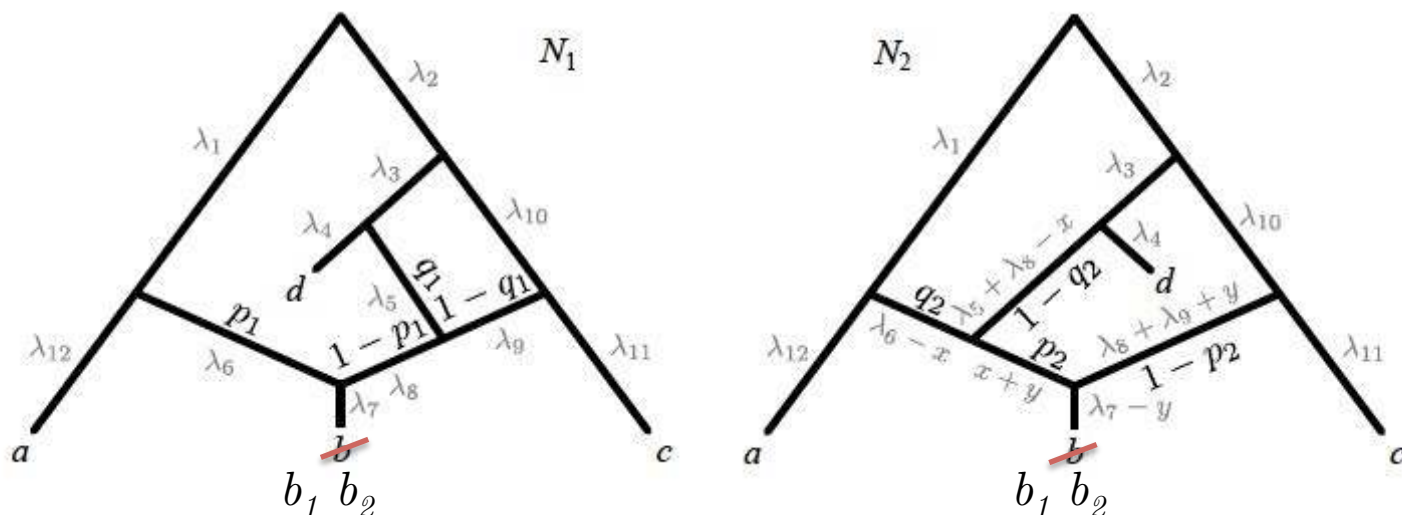
$$\Pr(T_2|N_2) = 1 - p_2$$

$$\Pr(T_3|N_2) = p_2(1 - q_2)$$

$$p_1 = 1/3 \quad p_2 = 2/3 \quad q_1 = 7/9 \quad \text{and} \quad q_2 = 3/7$$

$$x = y = 1/2 \quad \text{and} \quad \lambda_i = 1, \text{ for all } i$$

Scoring schemes based on **parental** trees (NMSC)



$$p_1 = 1/3 \quad p_2 = 2/3 \quad q_1 = 7/9 \quad \text{and} \quad q_2 = 3/7$$

$$x=y = 1/2 \quad \text{and} \quad \lambda_i = 1, \text{ for all } i$$

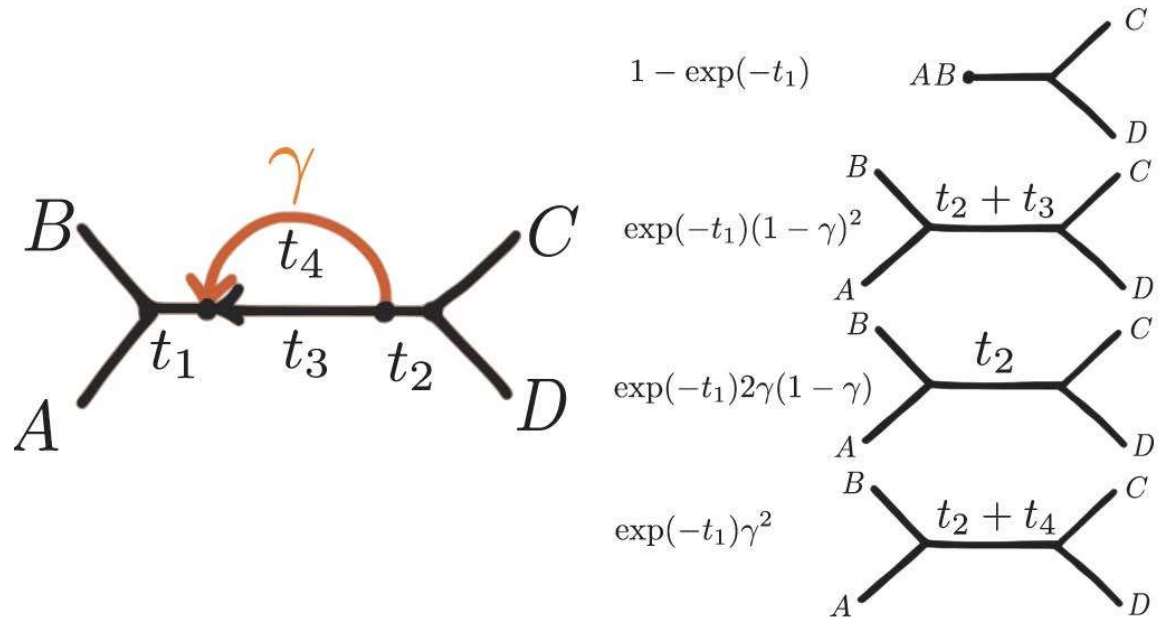
$$g = (((a, d), c), b_1), b_2)$$

$$P(g|N_1) \approx 7.7 \times 10^{-6}, \quad P(g|N_2) \approx 7.6 \times 10^{-6}$$

This may solve the **identifiability issues** for several practical cases but we need more samples per species “**well positioned**” in the phylogeny

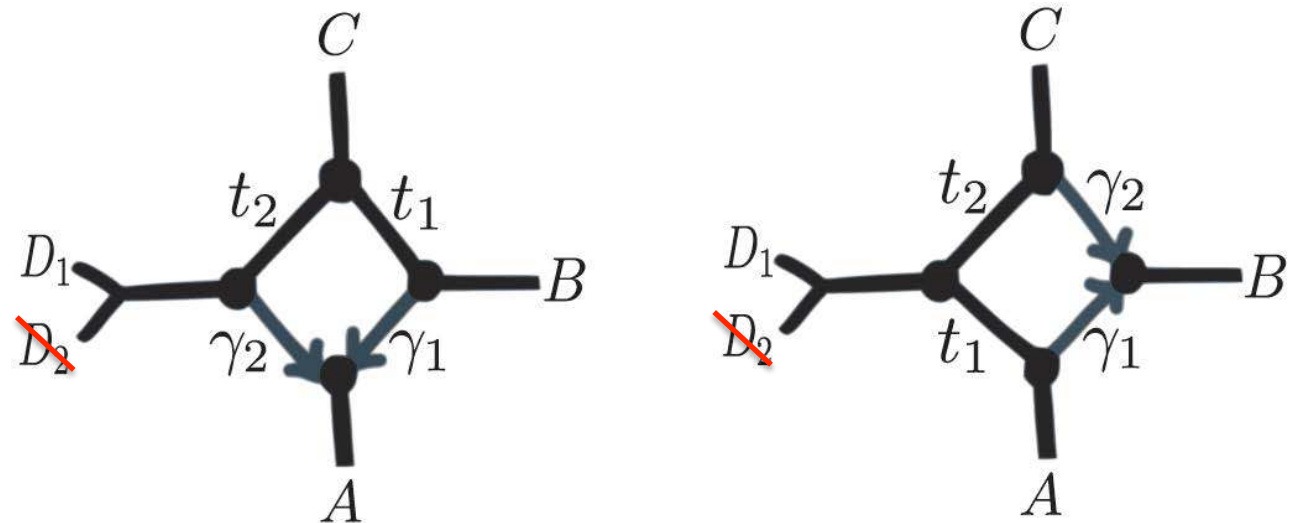
SNaQ (Species Networks applying Quartets) – pseudo-likelihood

Input: quartet CFs
Output: level-1
 semidirected
 networks



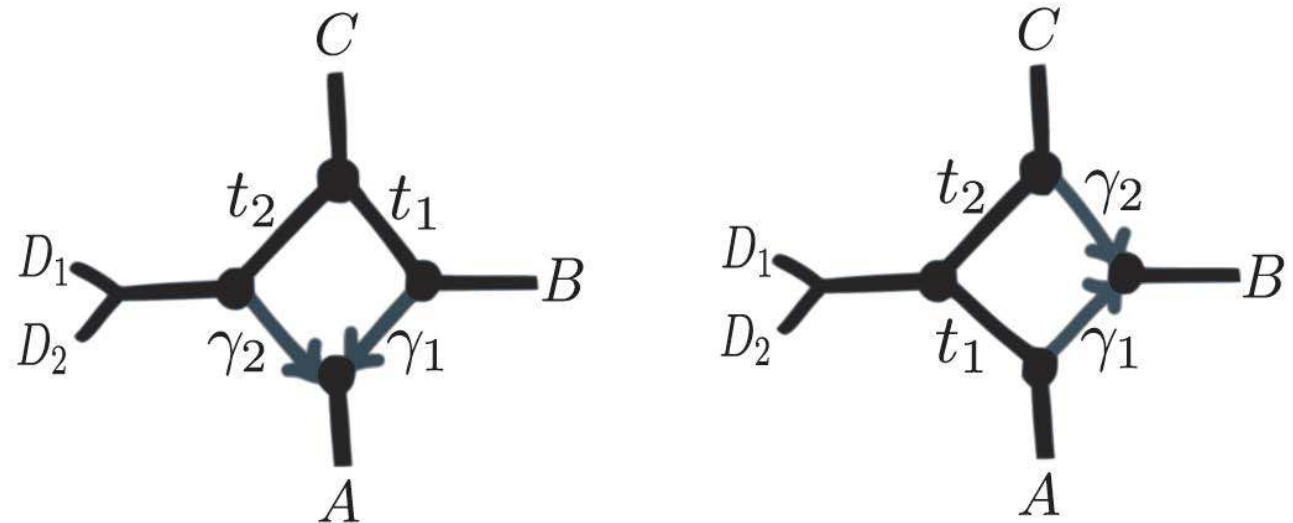
- quartet CFs do not depend on the root placement → semidirected networks
- if $n=4$, $k=2,3$ reticulations cannot be detected because equivalent to a tree

SNaQ (Species Networks applying Quartets) – an example of how to cope with indistinguishability



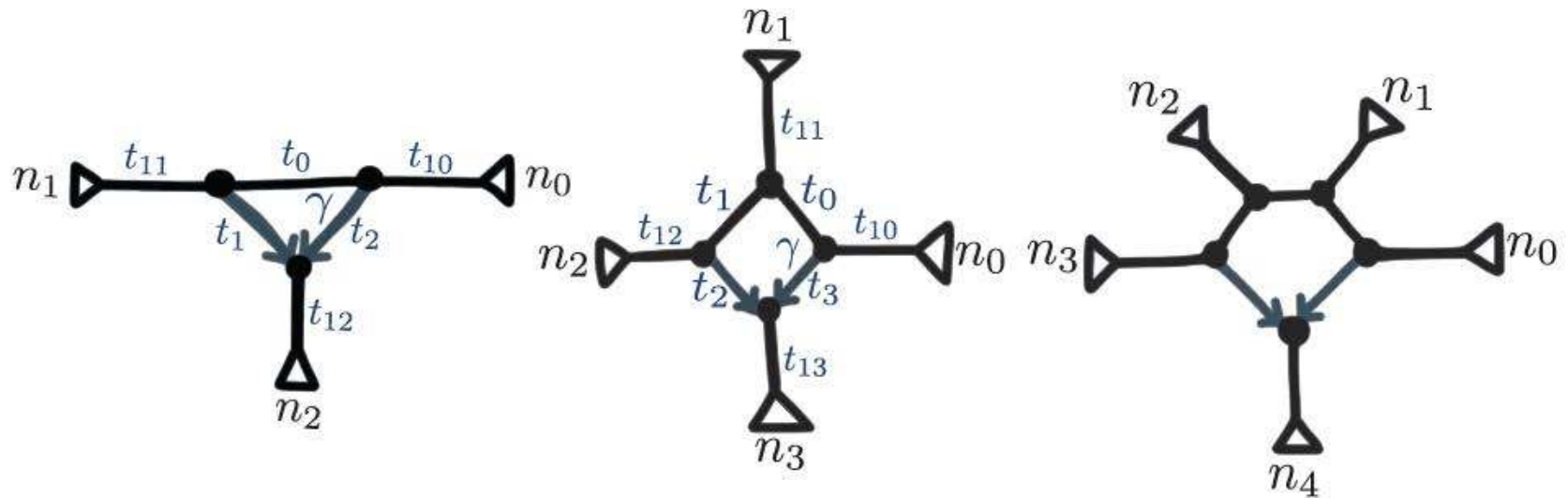
- quartet CFs do not depend on the root placement → semidirected networks
- if $n=4$, $k=2,3$ reticulations cannot be detected because equivalent to a tree
- if $n=4$, $k=4$, reticulations can be detected but not the “placement”

SNaQ (Species Networks applying Quartets) – an example of how to cope with indistinguishability



- quartet CFs do not depend on the root placement $\rightarrow$ semidirected networks
- if $n=4$, $k=2,3$ reticulations cannot be detected because equivalent to a tree
- if $n=4$, $k=4$, reticulations can be detected but not the “placement”
- for $n \geq 4$, $k=2$ reticulations are not detectable, $k=3$ sometimes and $k=4$ yes in general if $n \geq 5$, along with the placement

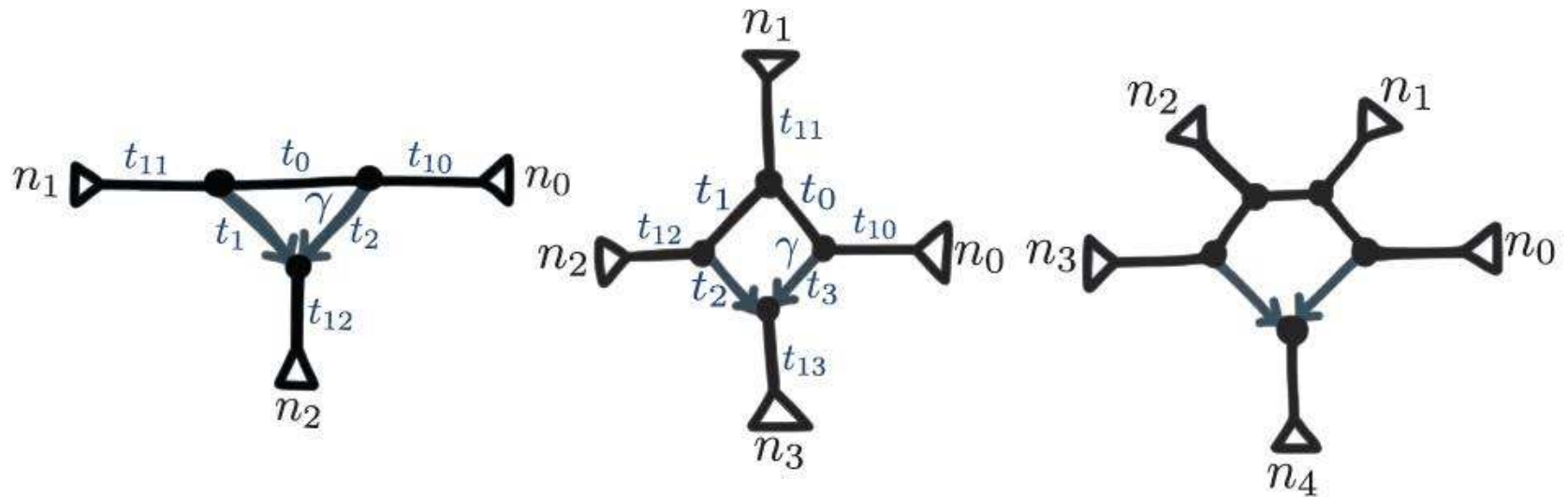
SNaQ (Species Networks applying Quartets) – an example of how to cope with indistinguishability



With only 4 taxa, there are more parameters than equations (3 quartet CFs), so focus on the case $n \geq 5$.

- If $k=3$, parameters are identifiable if $n_1, n_2, n_3 \geq 2$, and setting $t_{12} = 0$.
- If $k = 4$, parameters are identifiable if either $n_0 \geq 2$ (or n_2 , symmetrically), or if both n_1 and $n_3 \geq 2$. Parameters are not all identifiable in the remaining 2 cases (bad diamonds I & II)
- If $k=5$, all the parameters are identifiable.

SNaQ (Species Networks applying Quartets) – an example of how to cope with indistinguishability

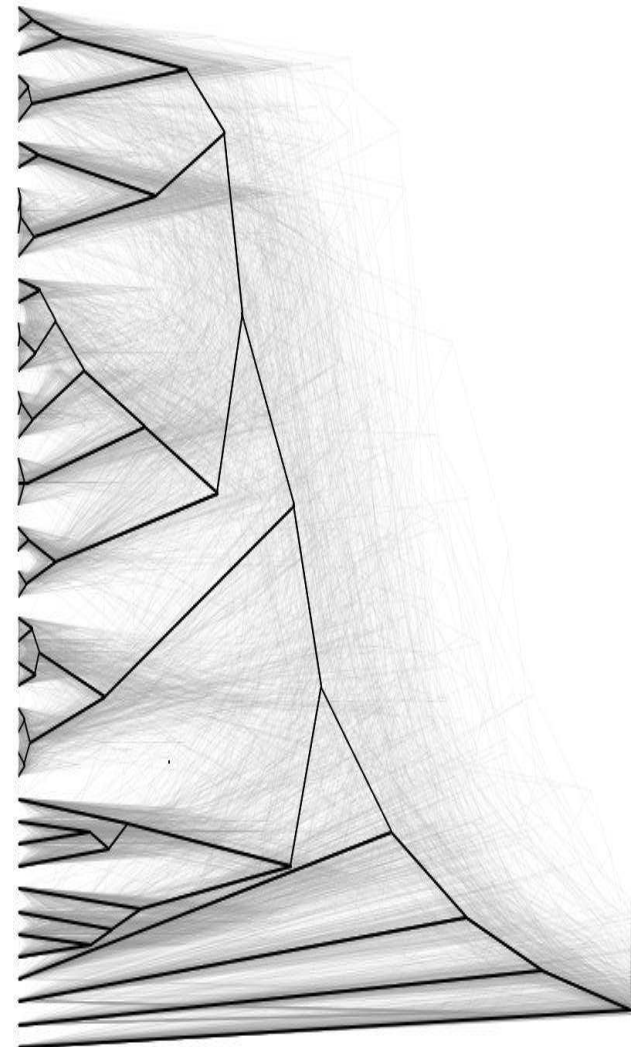
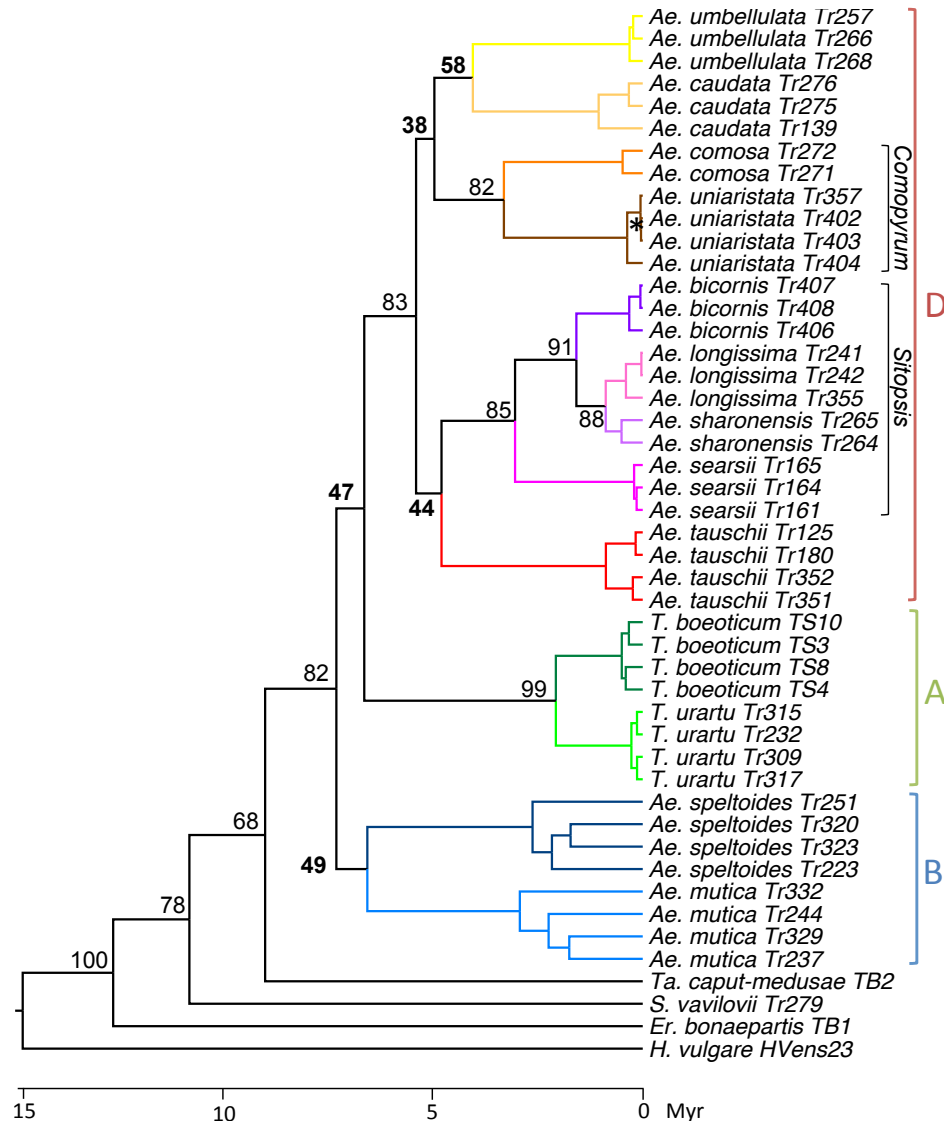


They search only in the space of identifiable networks:

- $k = 2$ not considered
- $k = 3$, only $n_1, n_2, n_3 \geq 2$, and setting $t_{12} = 0$
- For bad diamonds I, they reparametrized the 3 nonidentifiable values (γ, t_1, t_0) into 2 identifiable ones $\gamma(1-e^{-t_0})$ and $(1-\gamma)(1-e^{-t_1})$. For bad diamonds II, they set $t_{13} = 0$ and kept the other 5 parameters $(\gamma, t_0, t_1, t_2, t_3)$.

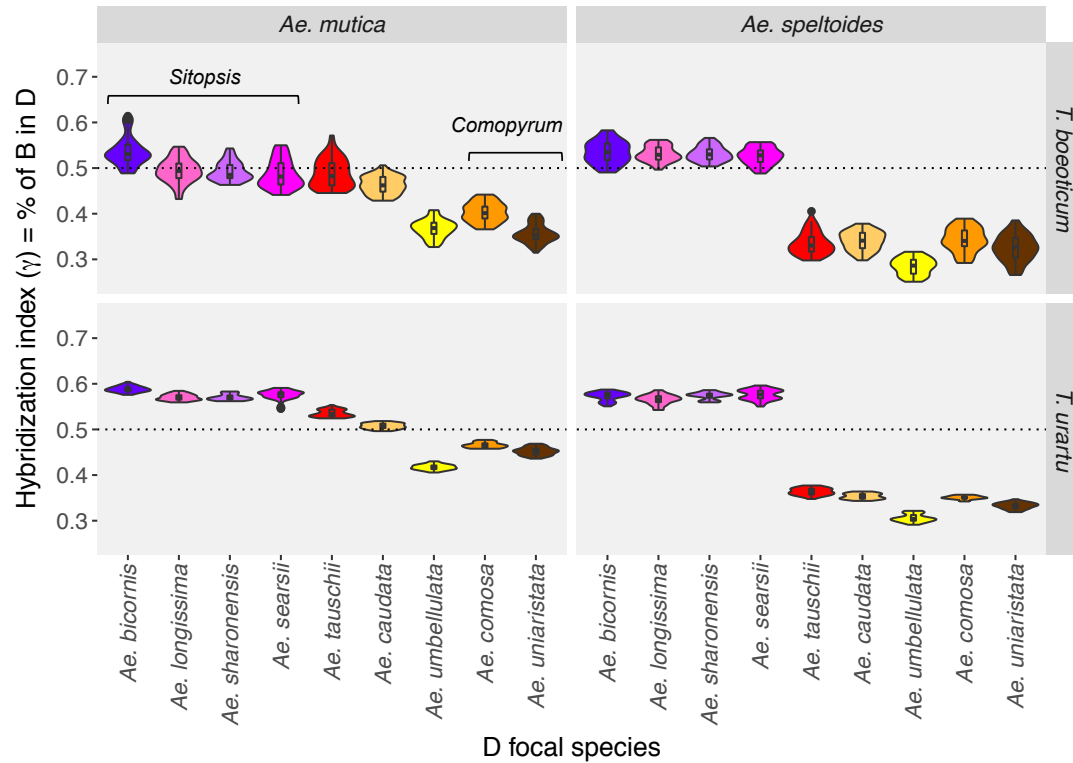
Thank you for your attention

Another (home made) approach

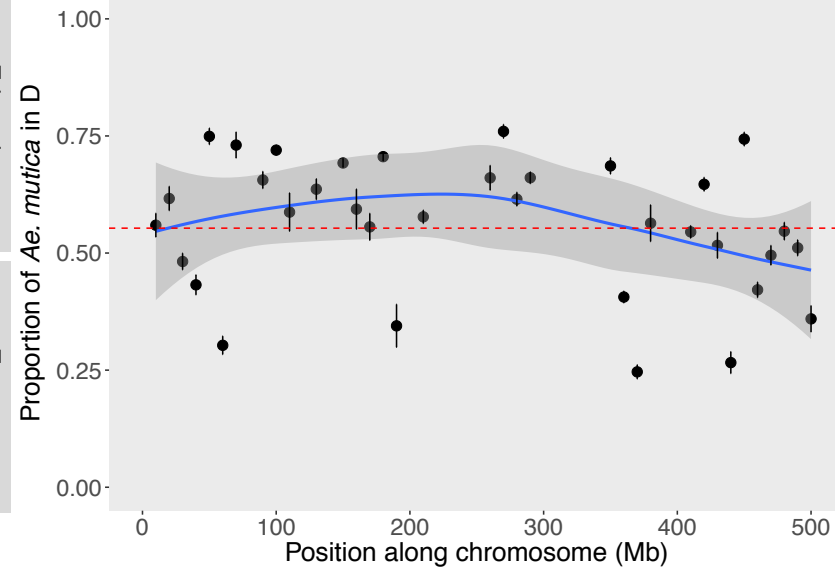


Another (home made) approach

D species as hybrids between A and B lineages

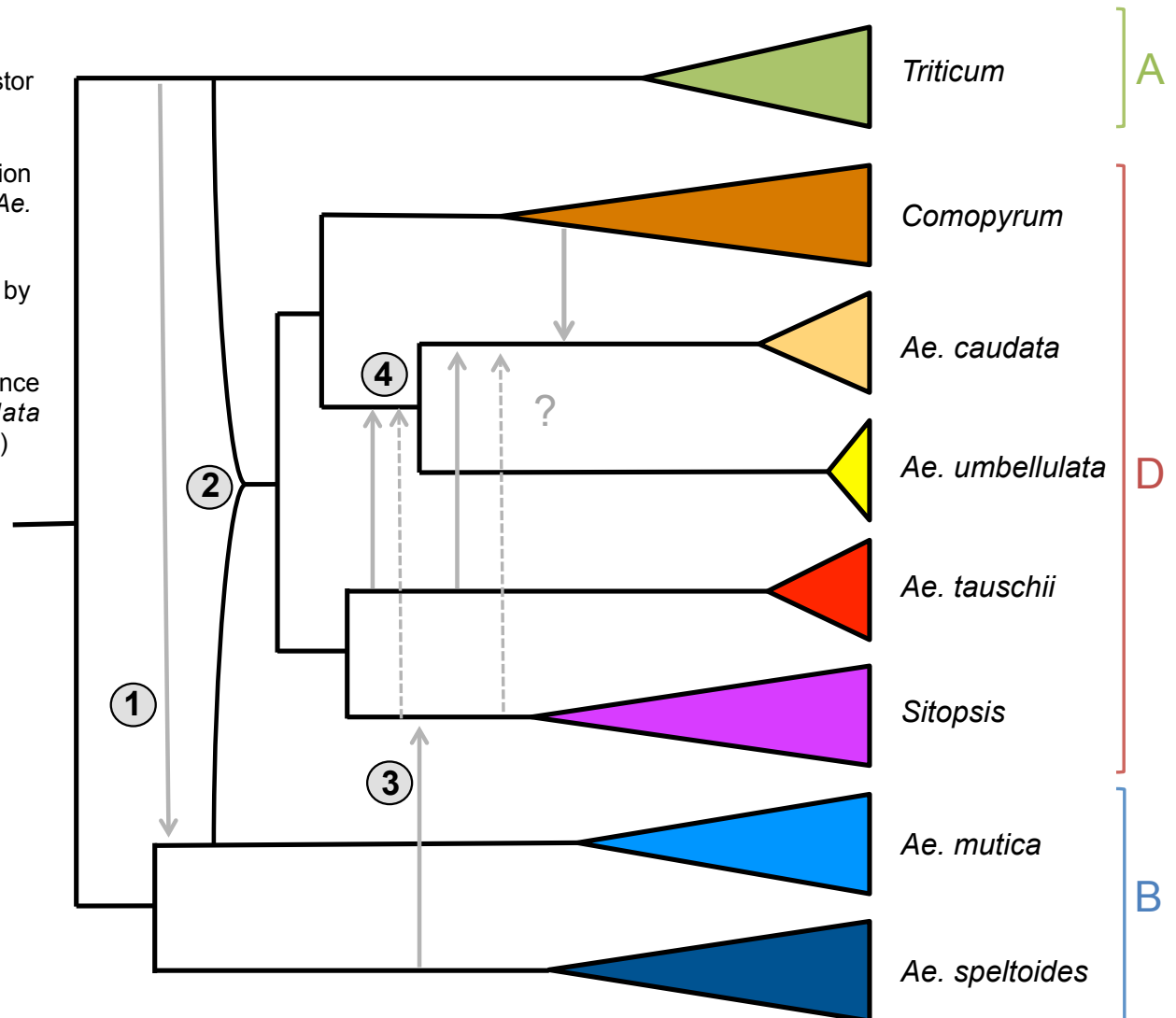


Hybridization index (γ) along chromosome 3



Another (home made) approach

- ① Introgression from the *Triticum* ancestor into the *Ae. mutica* ancestor
- ② Origin of the D clade by hybridization between the A clade ancestor and the *Ae. mutica* ancestor
- ③ Introgression of the *Sitopsis* ancestor by the *Ae. speltoides* ancestor
- ④ Complex gene flows during the divergence of *Ae. caudata* and *Ae. umbellulata* probably involving three events (or more)

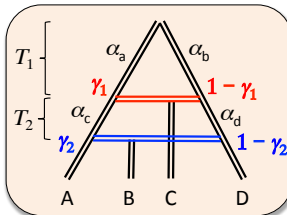


Another (home made) approach

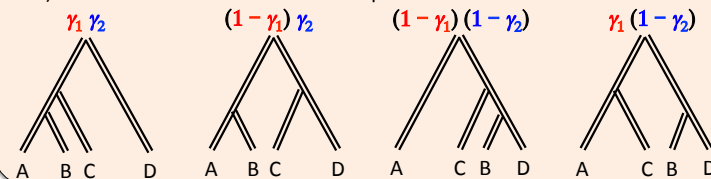
Data

Out	ATTGCTACGTCAT
A	A...
B	G...A...
C	.C.....
D	.C...G.....

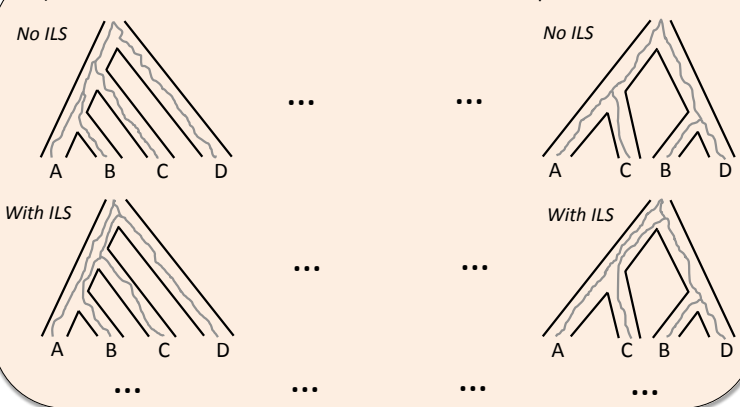
Scenario



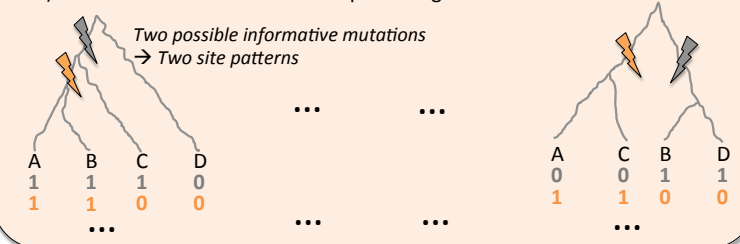
1) Probabilities of the network's components



2) Probabilities of coalescent trees embedded into components



3) Probabilities of informative site patterns given a coalescent tree



Site patterns

Out	A	B	C	D
0	0	0	1	1
0	0	1	0	1
...	...	...	...	...
0	1	1	0	0

Counts

v_1
 v_2
...
 v_{10}

Expectations

p_1
 p_2
...
 p_{10}

Likelihood of the scenario

→ Choice of the best scenario