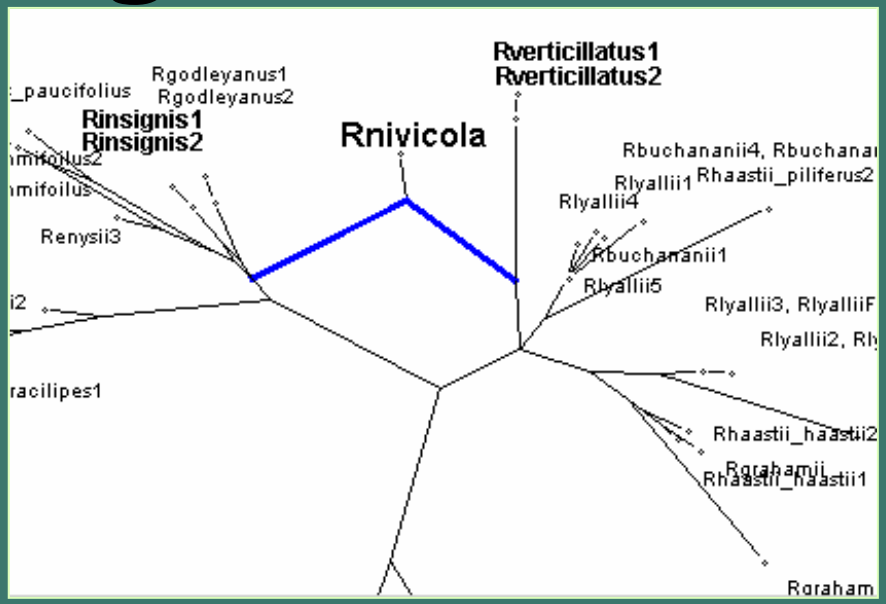
Splits and Phylogenetic Networks



Daniel H. Huson

Paris, June 21, 2005

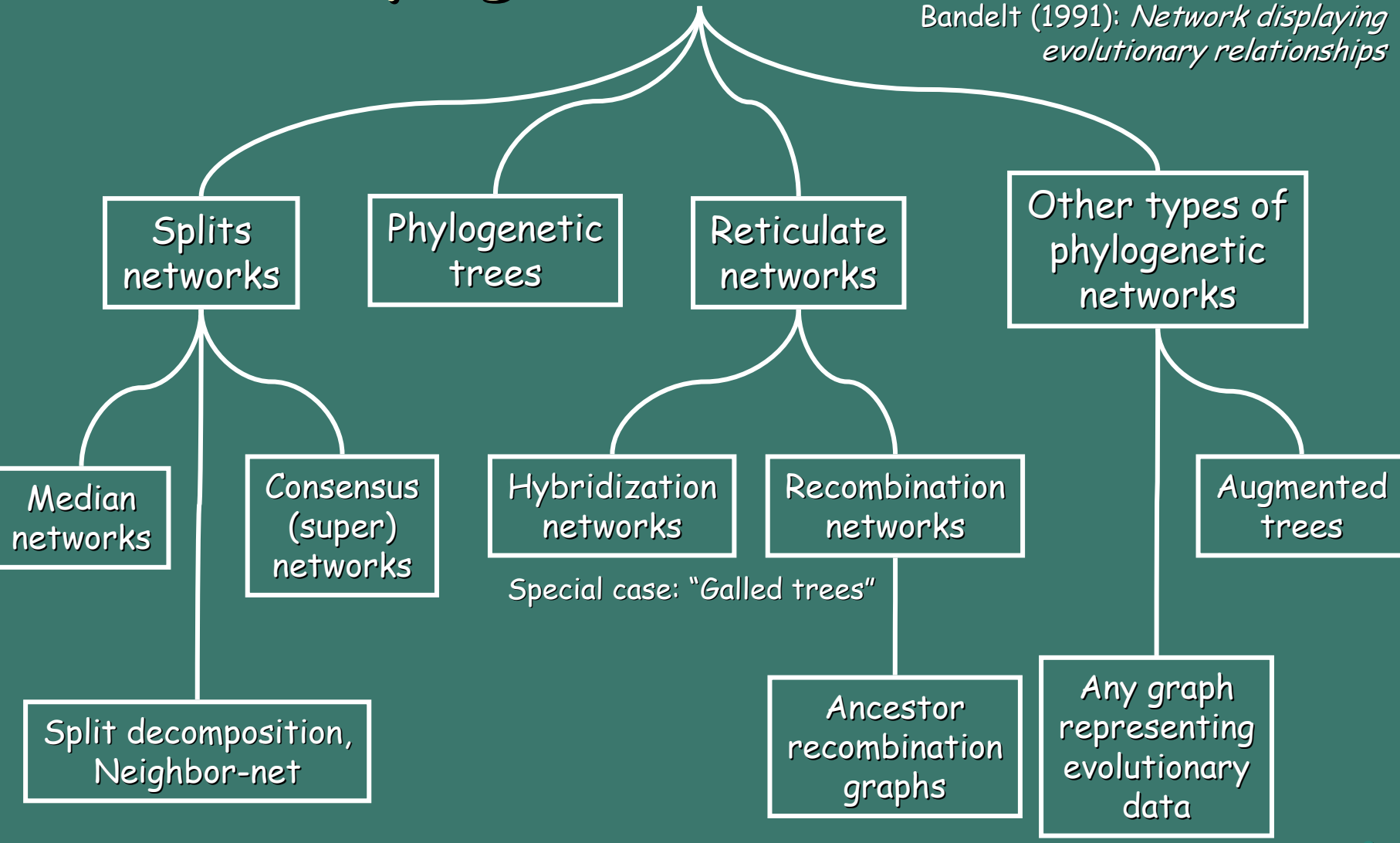
Eberhard Karls
Universität
Tübingen

Contents

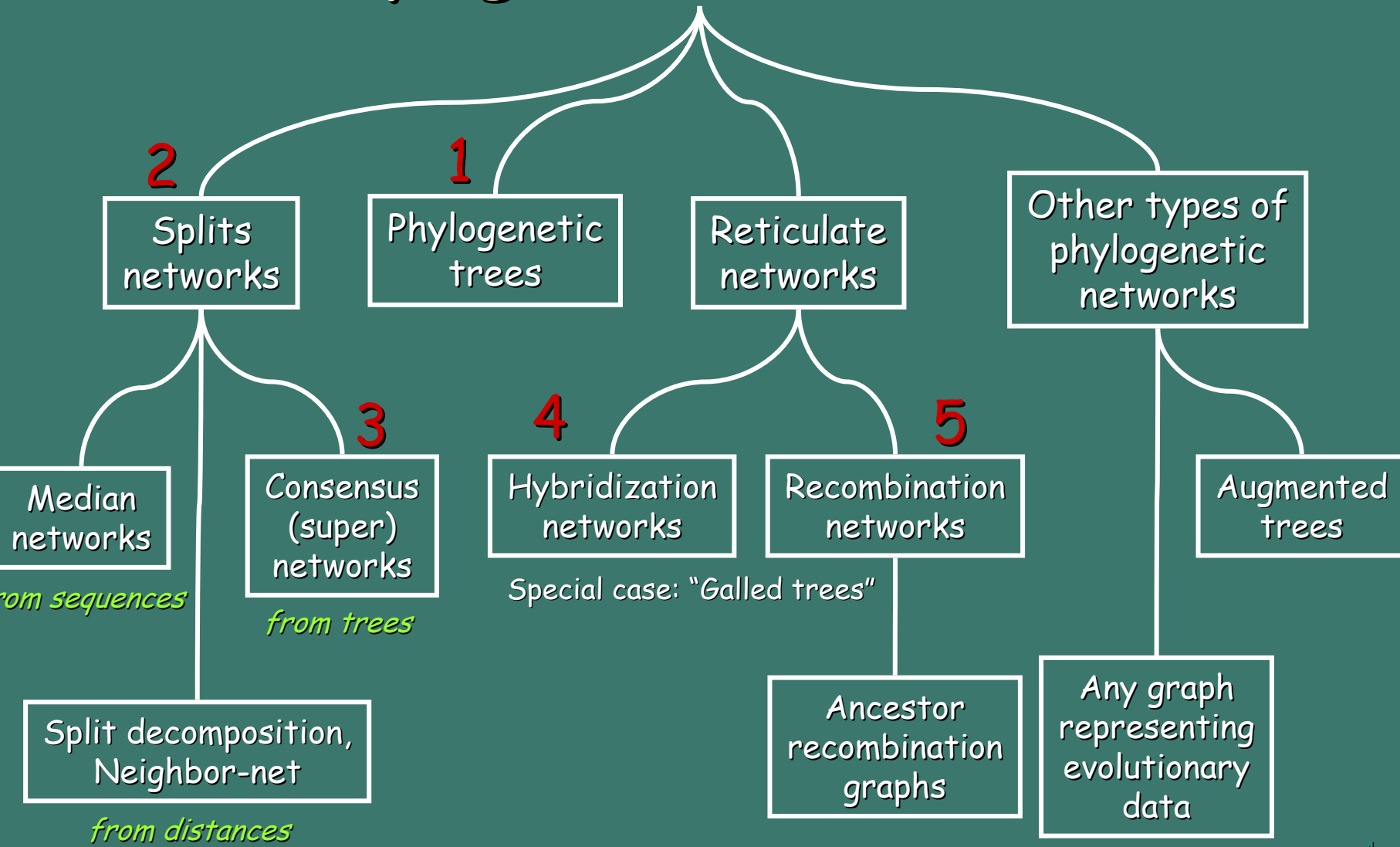
1. Phylogenetic trees
2. Splits networks
3. Consensus networks
4. Hybridization and reticulate networks
5. Recombination networks

Phylogenetic Networks

Bandelt (1991): *Network displaying evolutionary relationships*

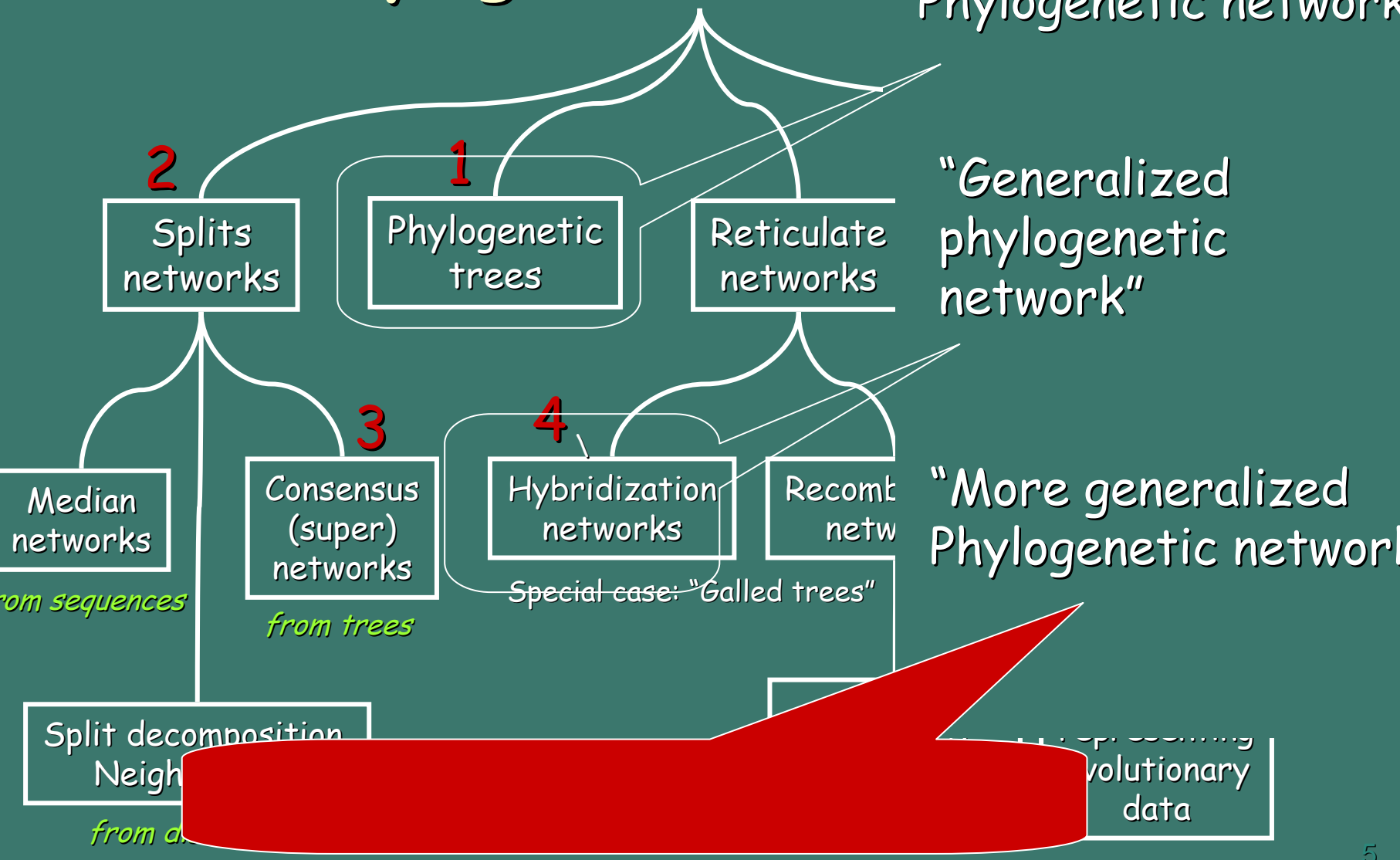


Phylogenetic Networks



Phylogenetic Ne

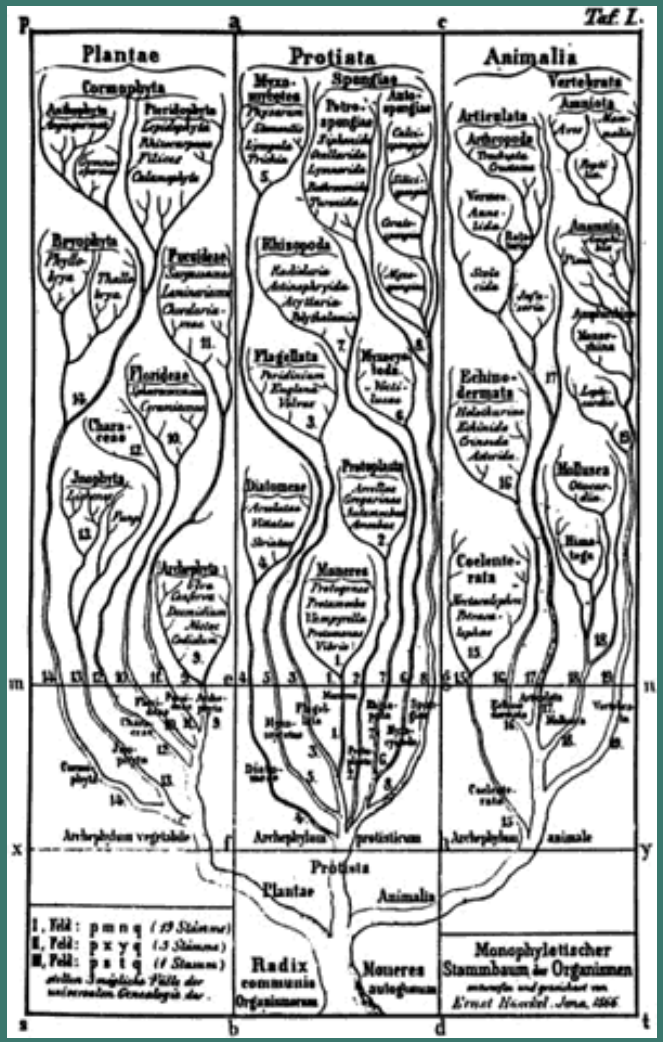
Dan Gusfield:
"Phylogenetic network



Part I

1. Phylogenetic trees
2. Splits networks
3. Consensus networks
4. Hybridization and reticulate networks
5. Recombination networks

Phylogenetic Trees

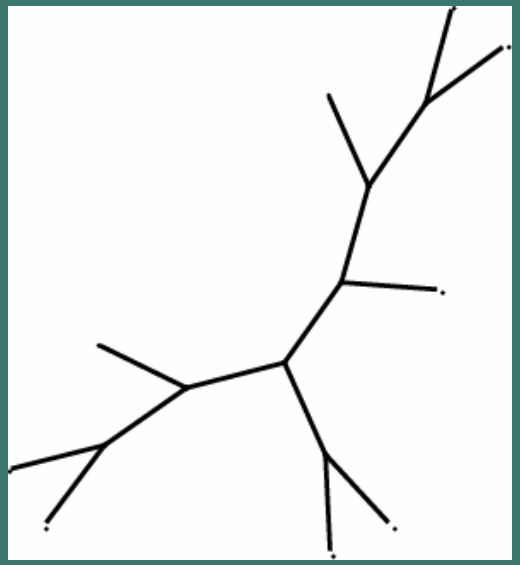


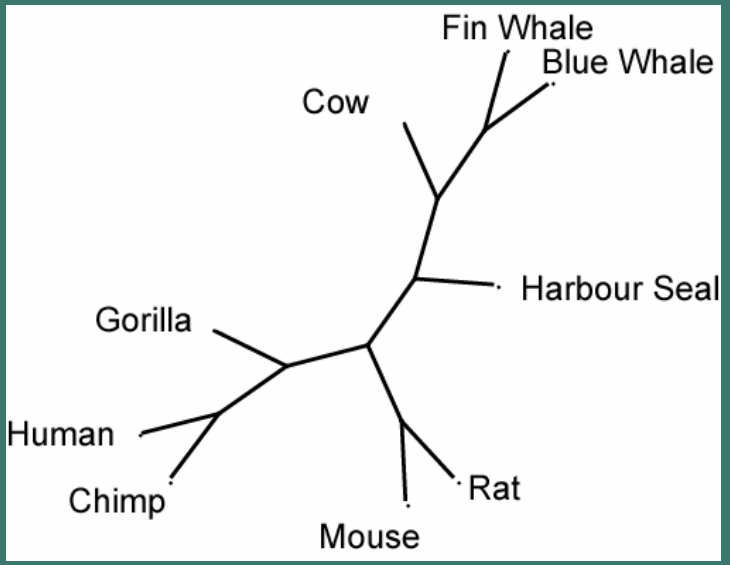
Ernst Haeckel,
Tree of Life
1866

Phylogenetic Trees

- Let $X = \{x_1,\ldots,x_n\}$ denote a set of taxa.
- *A phylogenetic tree T* (or *X-tree*) is given by labeling the leaves of a tree by the set X:

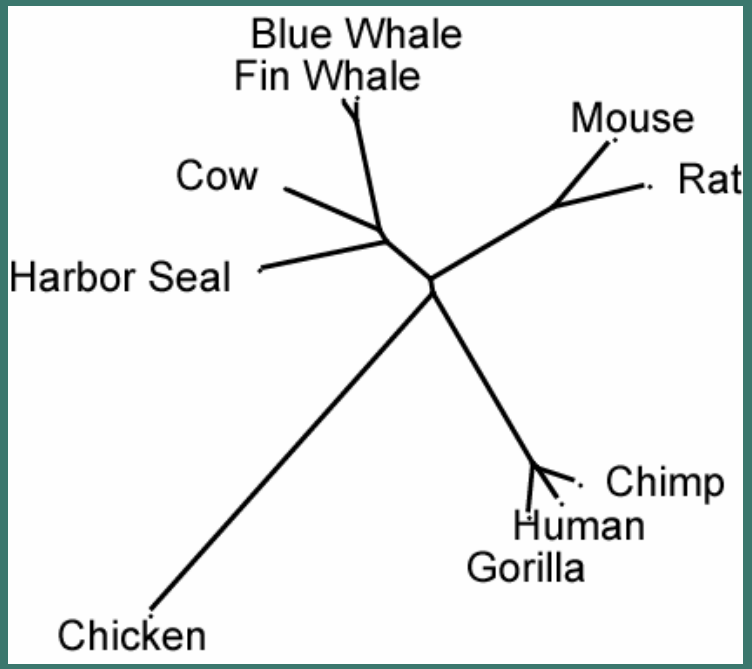
Cow
Fin Whale
Blue Whale
Habor Seal
Rat
Mouse
Chimp
Human
Gorilla



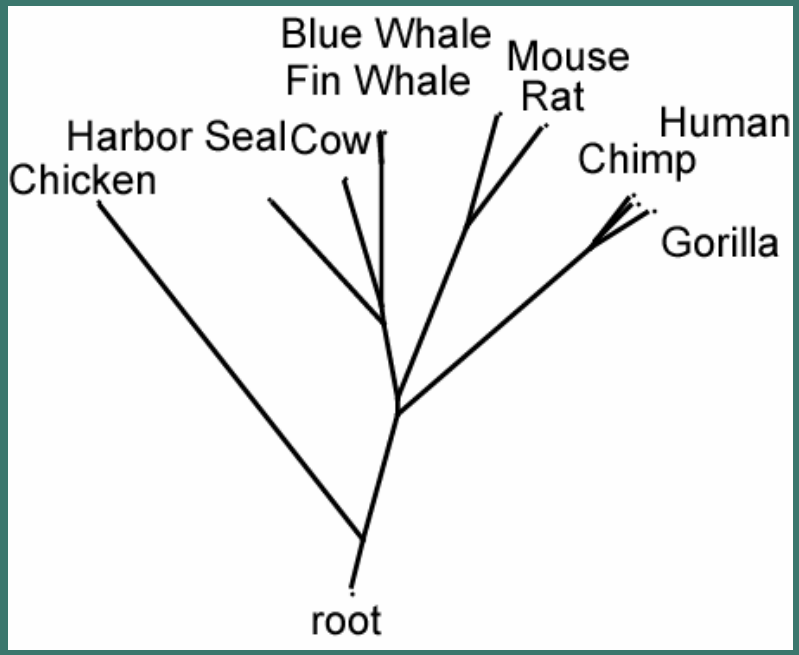


Taxa + tree ⇒ phylogenetic tree

Unrooted vs Rooted Trees



Unrooted tree

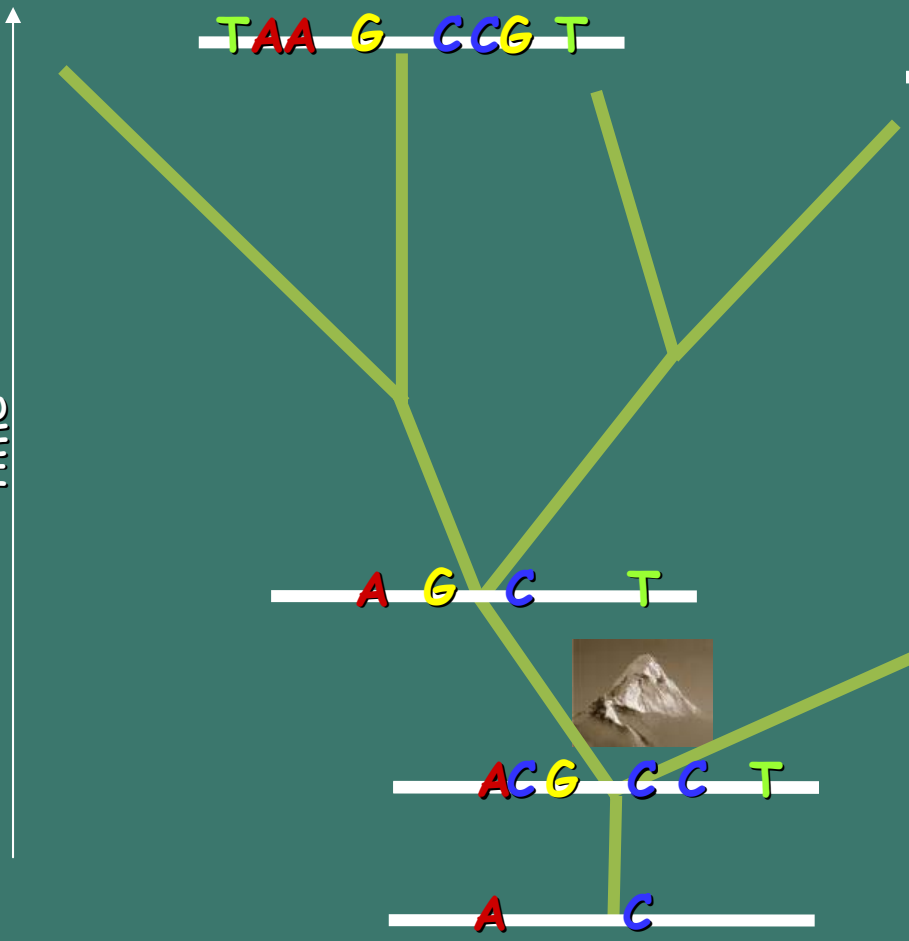


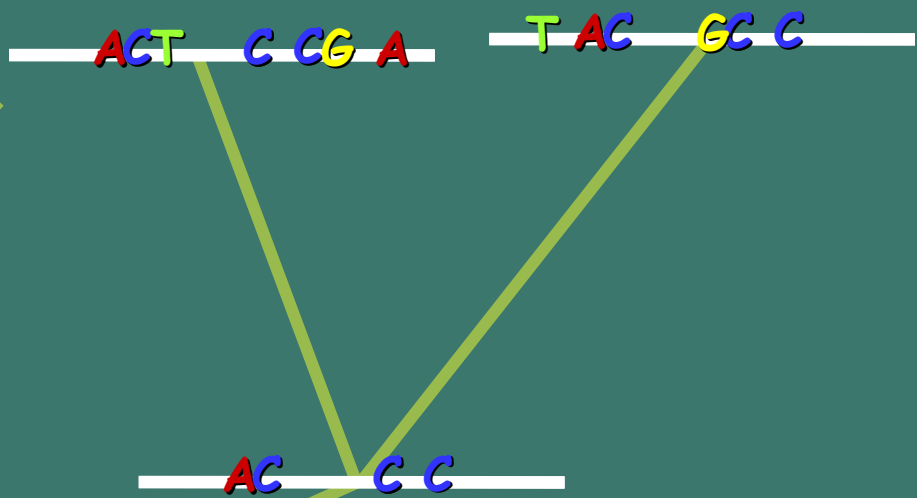
Rooted tree, rooted using Chicken as *outgroup*

mathematically and algorithmically easier to deal with

biologically relevant, defines clades of related taxa

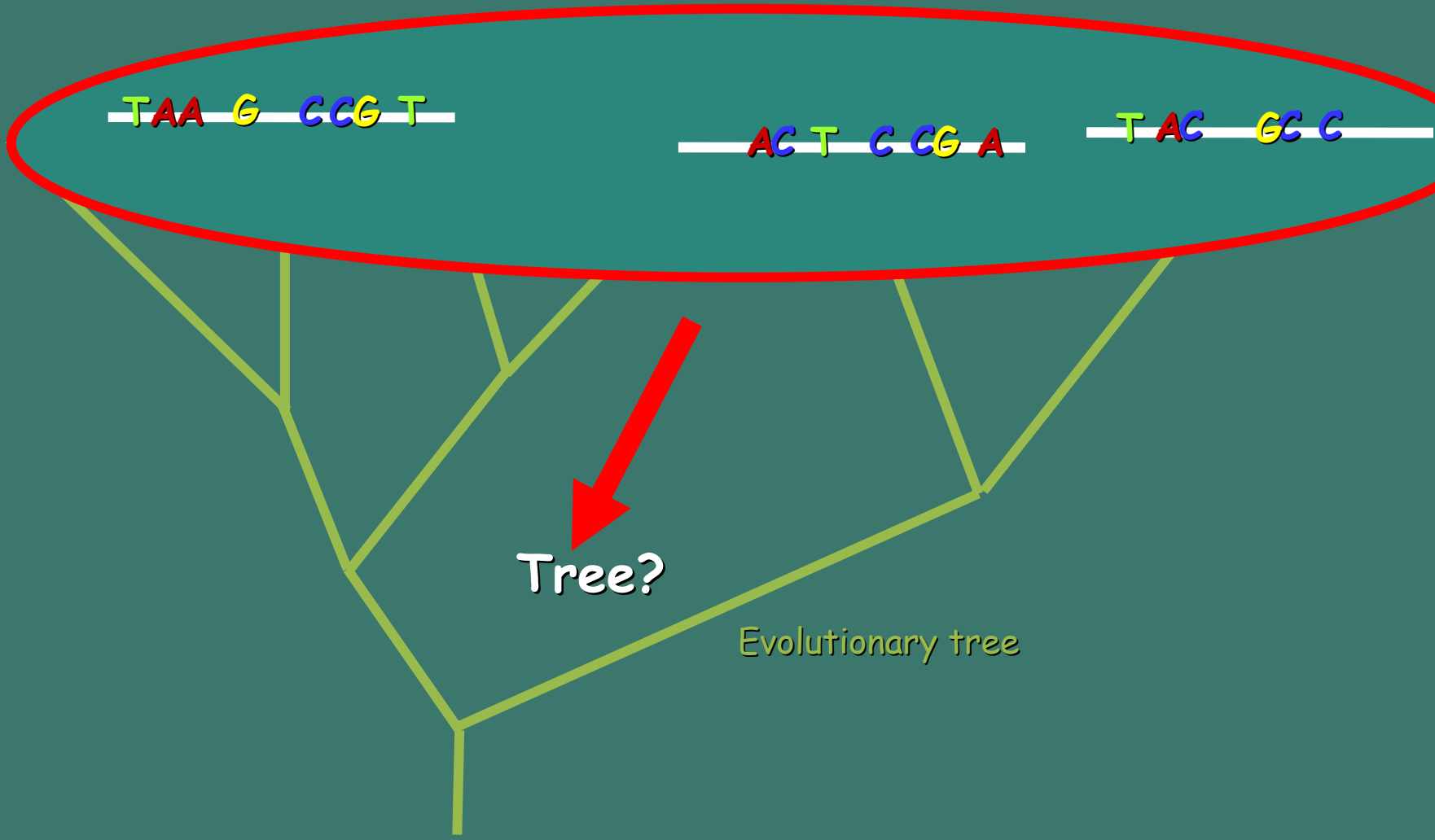
A Simple Model of Evolution



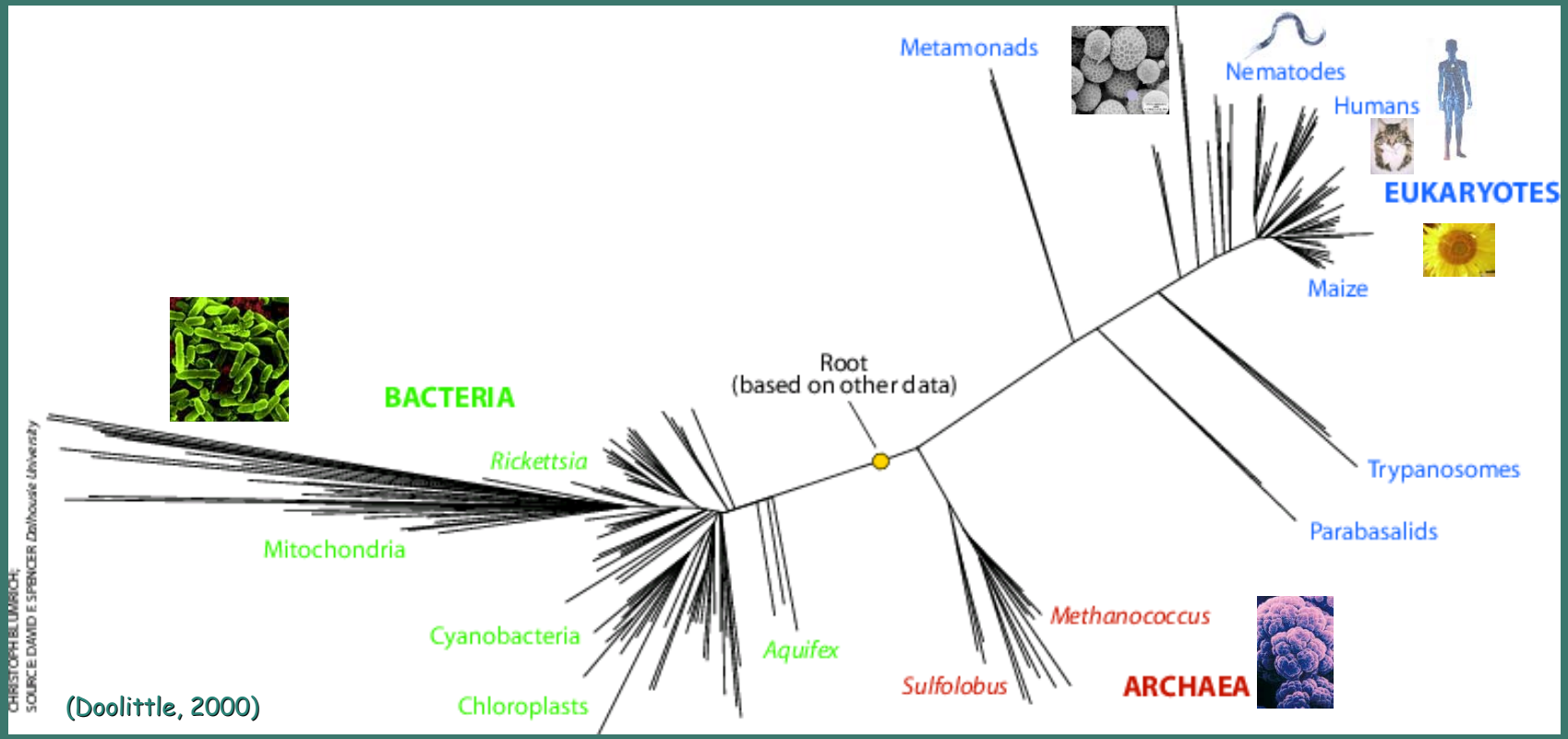


- Mutations along branches
- Speciation events at nodes

Tree Reconstruction Problem



Tree of Life Based on 16S rRNA



(Doolittle, 2000)

Part II

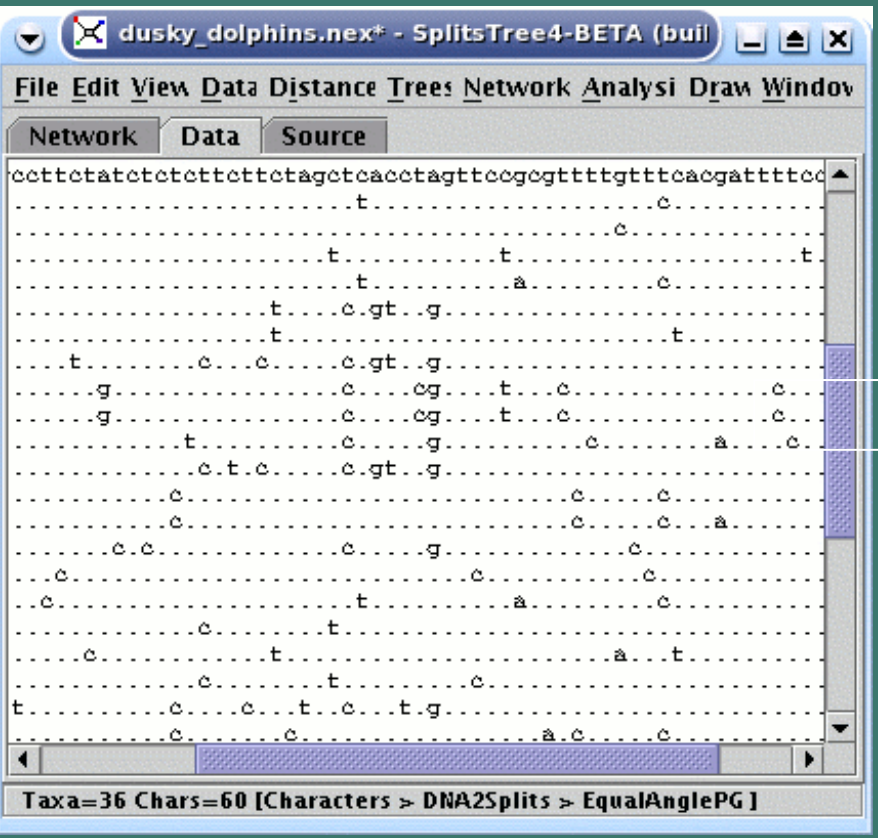
1. Phylogenetic trees
2. Splits networks
3. Consensus networks
4. Hybridization and reticulate networks
5. Recombination networks

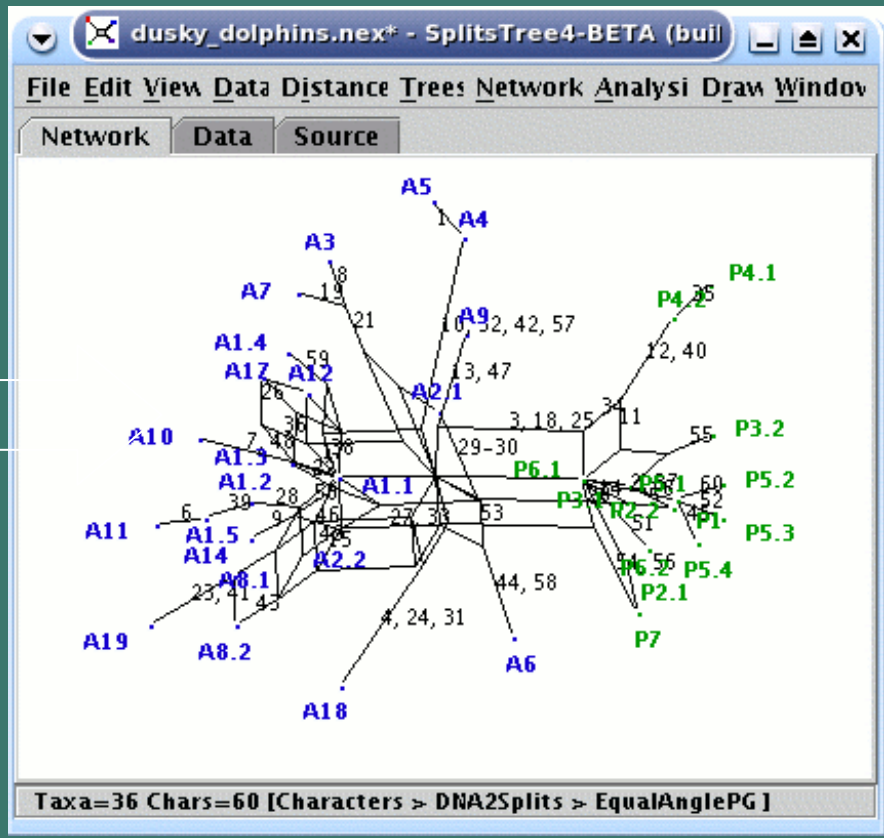
Splits Networks

Represents incompatible signals in data, from:

- Sequences, e.g.:
  - Median network (Bandelt *et al* 1994)
  - Spectral analysis (Hendy and Penny 1993)
- Distances, e.g.:
  - E.g. Split decomposition (Bandelt and Dress 1992)
  - Neighbor-Net (Bryant and Moulton 2002)
- Trees, e.g.:
  - Consensus network (Holland and Moulton 2003)
  - Super network (H., Dezulian, Kloepper and Steel 2004)
  - Bootstrap network (H., implemented in SplitsTree4)

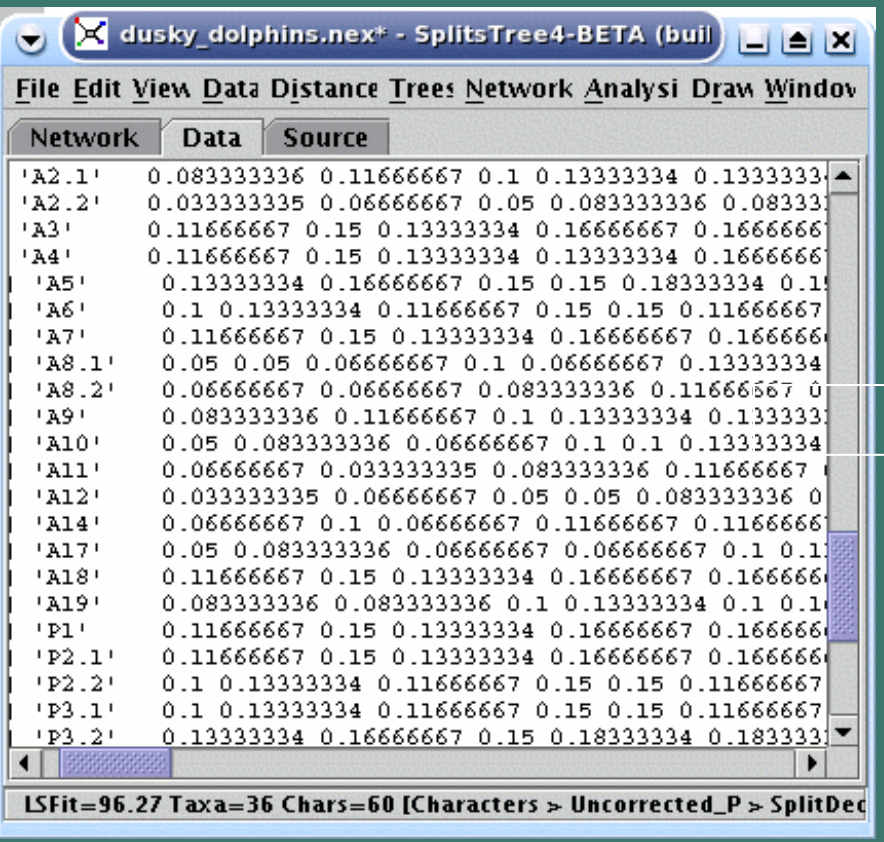
Sequences to Splits Network

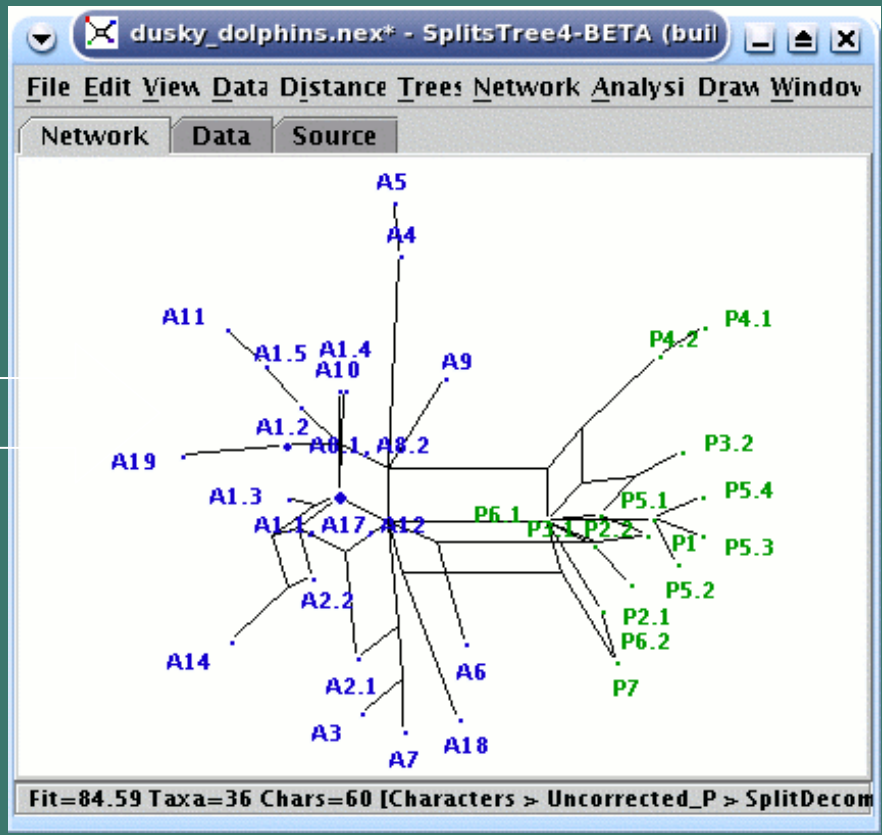




If characters have only 2 states and not too conflicting: interpret columns as splits and draw full splits network

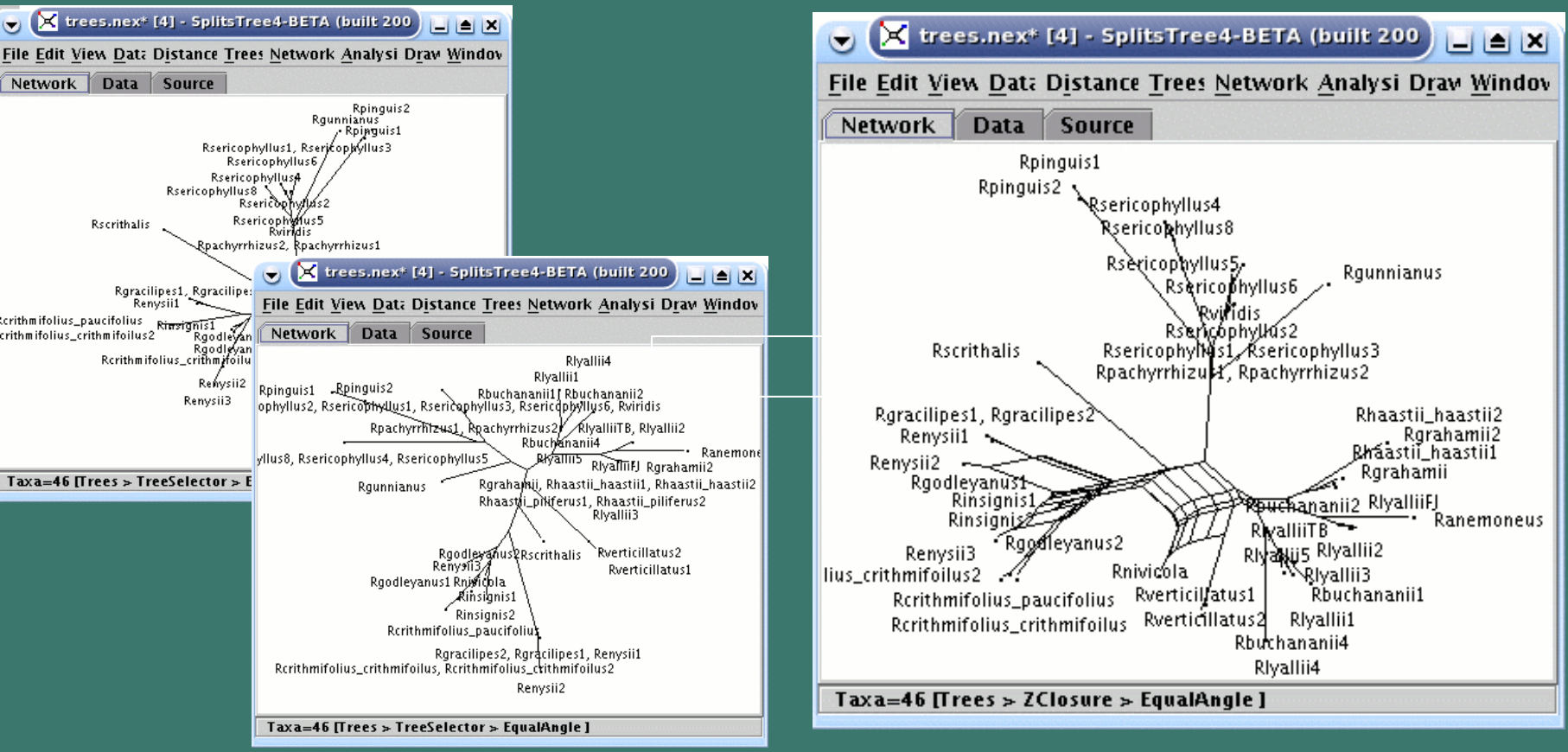
Distances to Splits Network





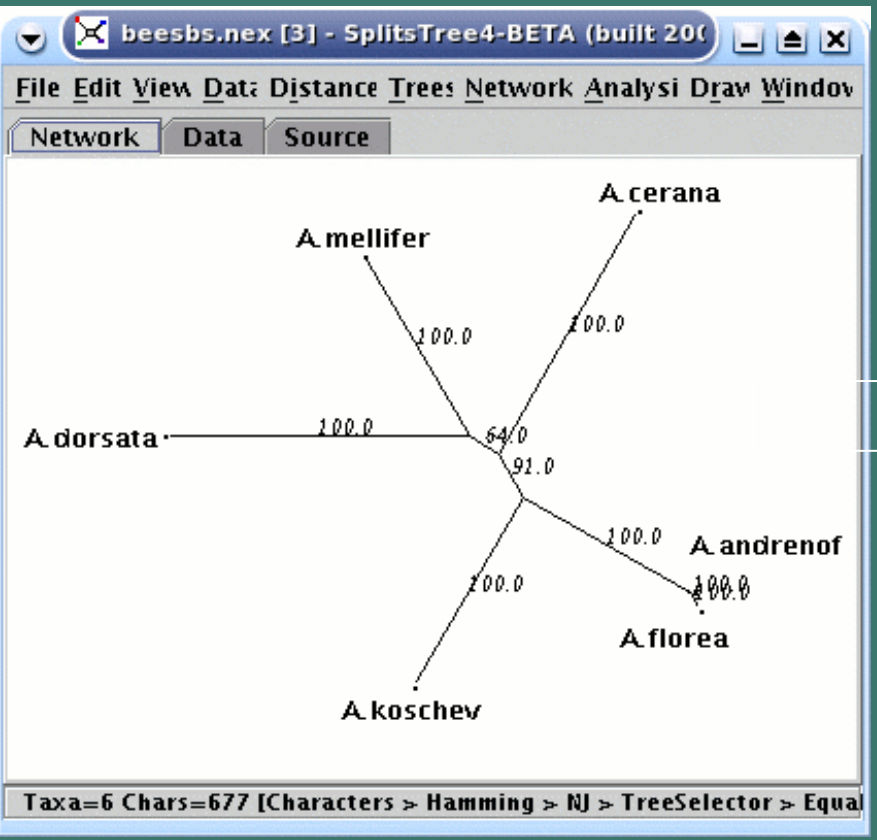
Split decomposition or Neighbor-Net produces network from distances

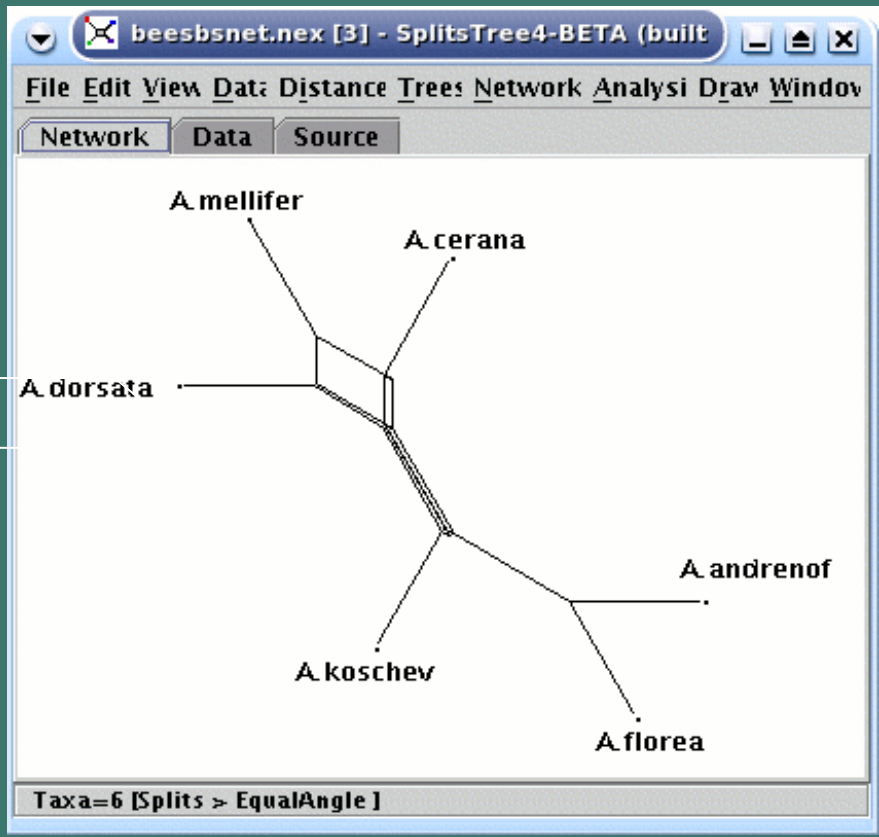
Trees to Splits Network



A collection of trees can be represented by a consensus network or super network

Bootstrap Network

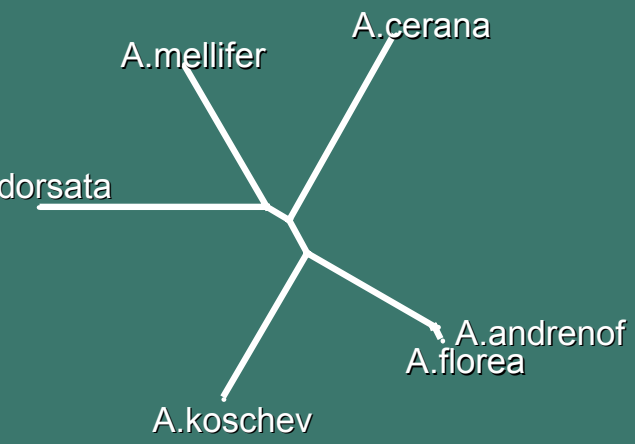




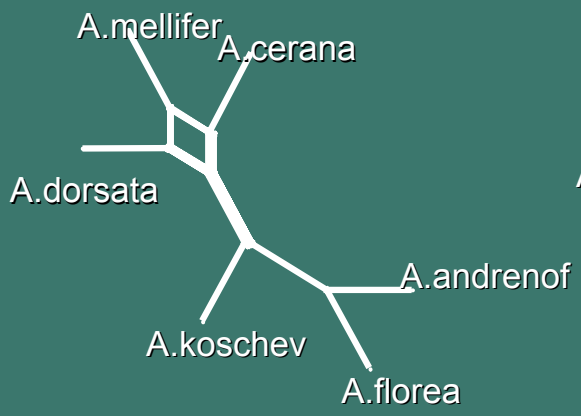
Draw all splits that have positive bootstrap score

Split Decomposition & Bootstrap Network

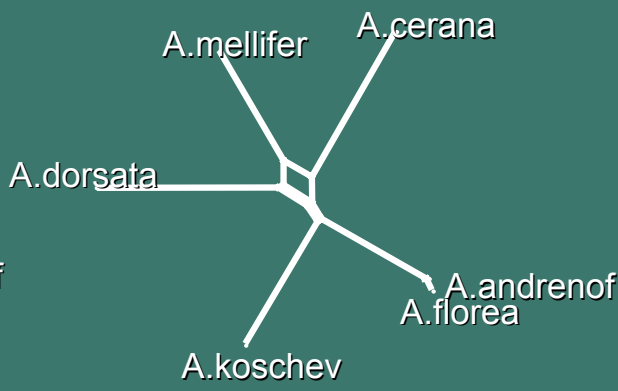
- Compare the result of Split Decomposition with an NJ tree and bootstrap network:



Bio-NJ tree



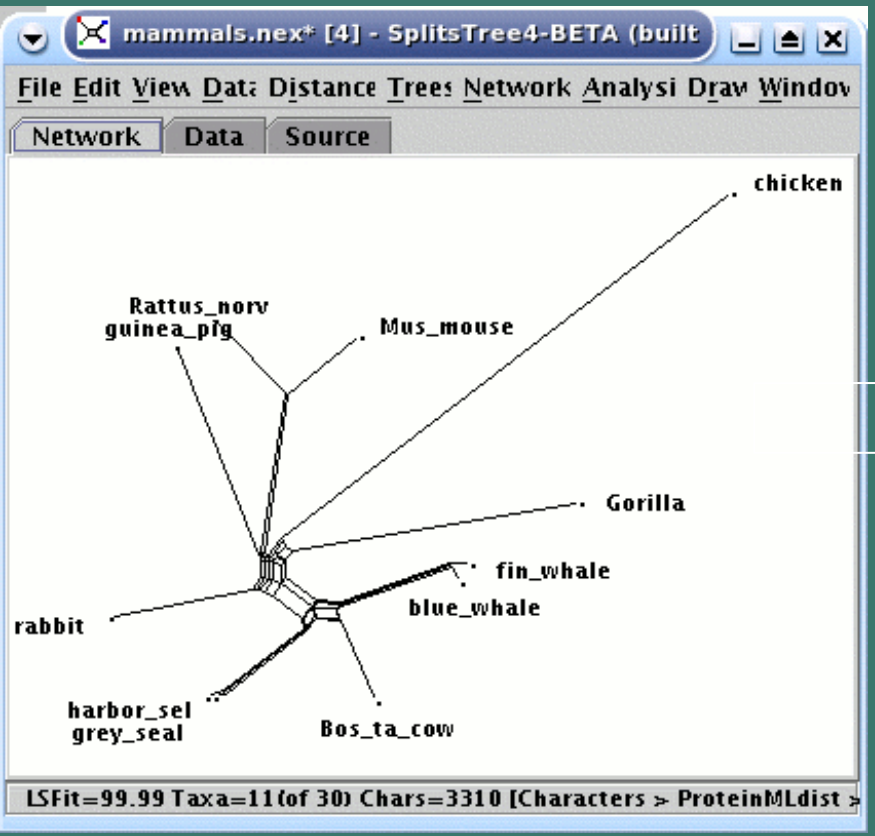
Bootstrap network

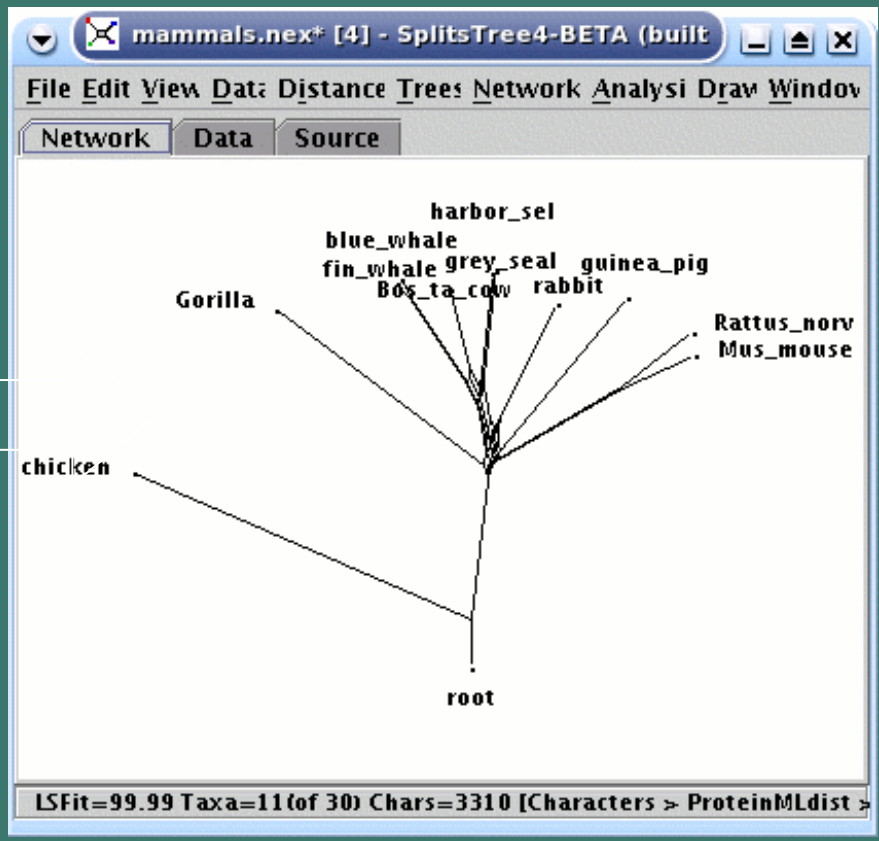


Splits network obtained via the Split Decomposition

Bill Martin: Splits networks show which signals tree reconstruction methods are fighting over…

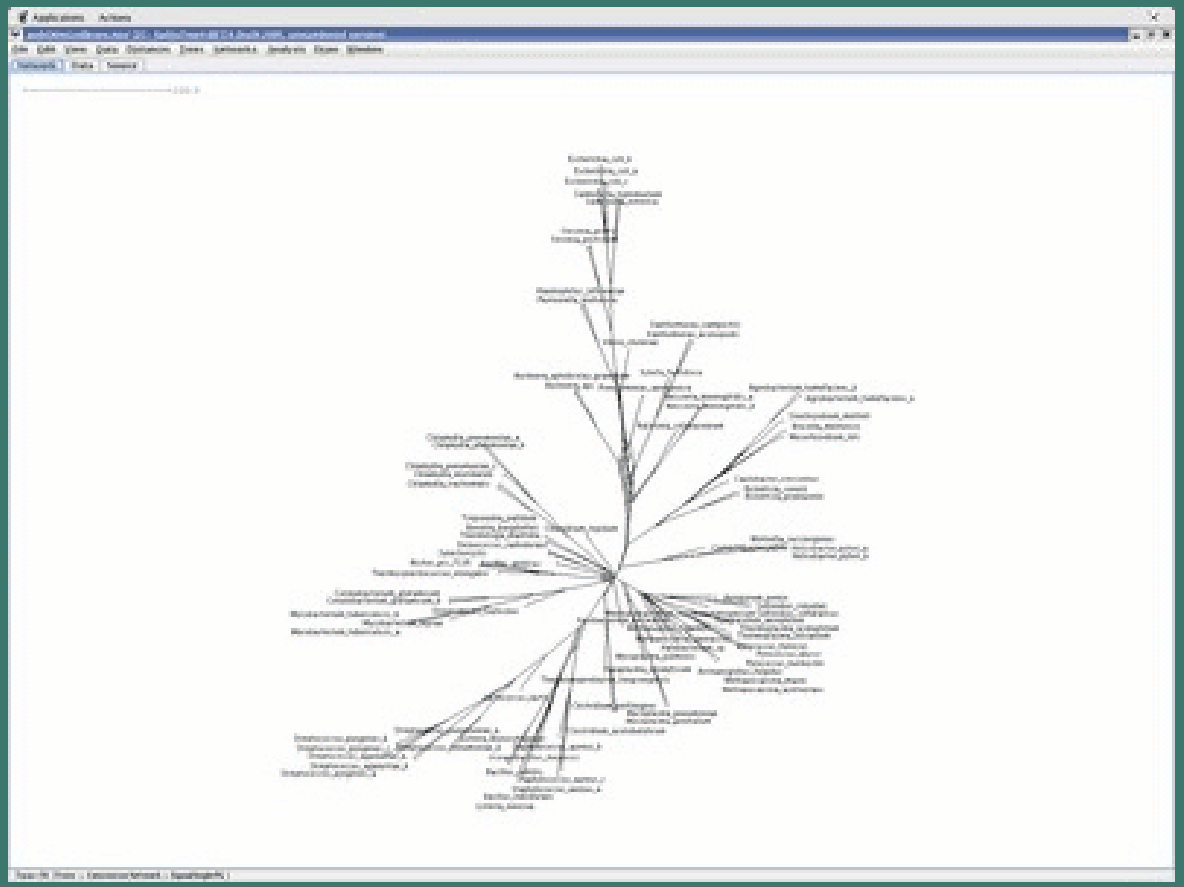
Rooted Splits Networks





Splits network can be "rooted" e.g. using an outgroup
(Gambette and H, manuscript)

Better layout of splits networks



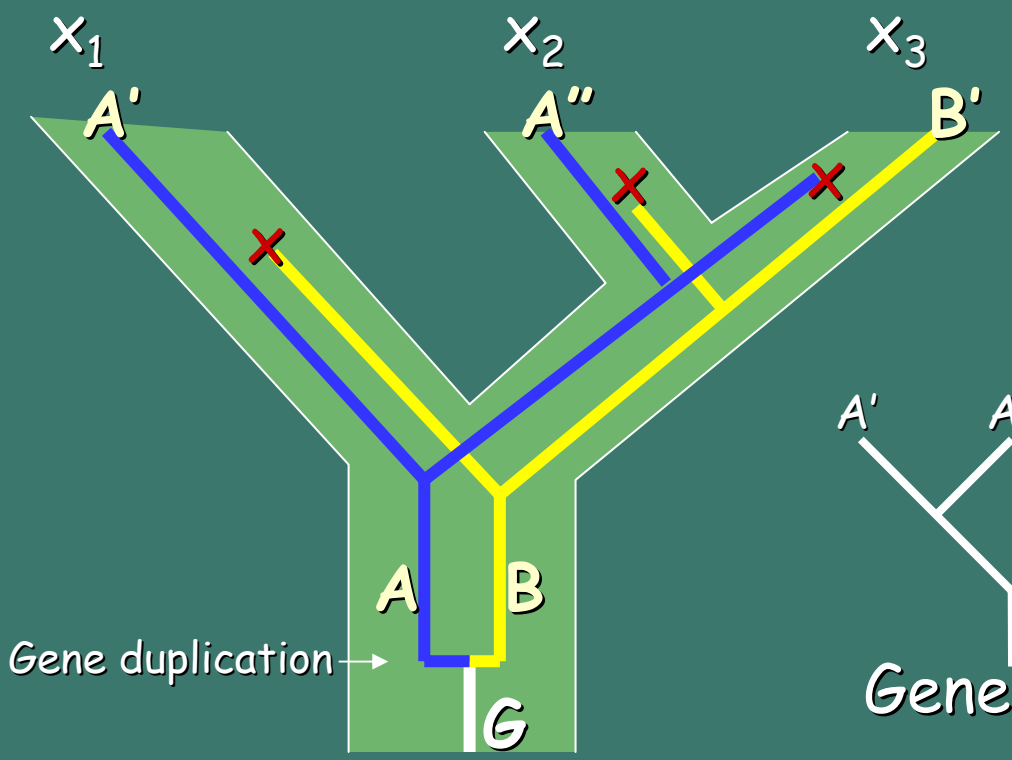
Philippe Gambette, currently visiting Tuebingen from ENS Cachan

Part III

1. Phylogenetic trees
2. Splits networks
3. Consensus networks
4. Hybridization and reticulate networks
5. Recombination networks

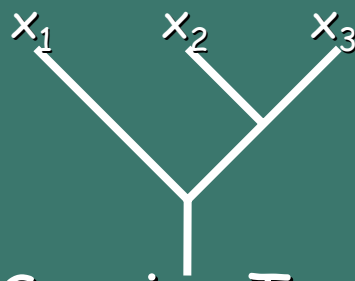
Gene Trees Can Differ

Also allow gene duplication and loss:



Gene Tree

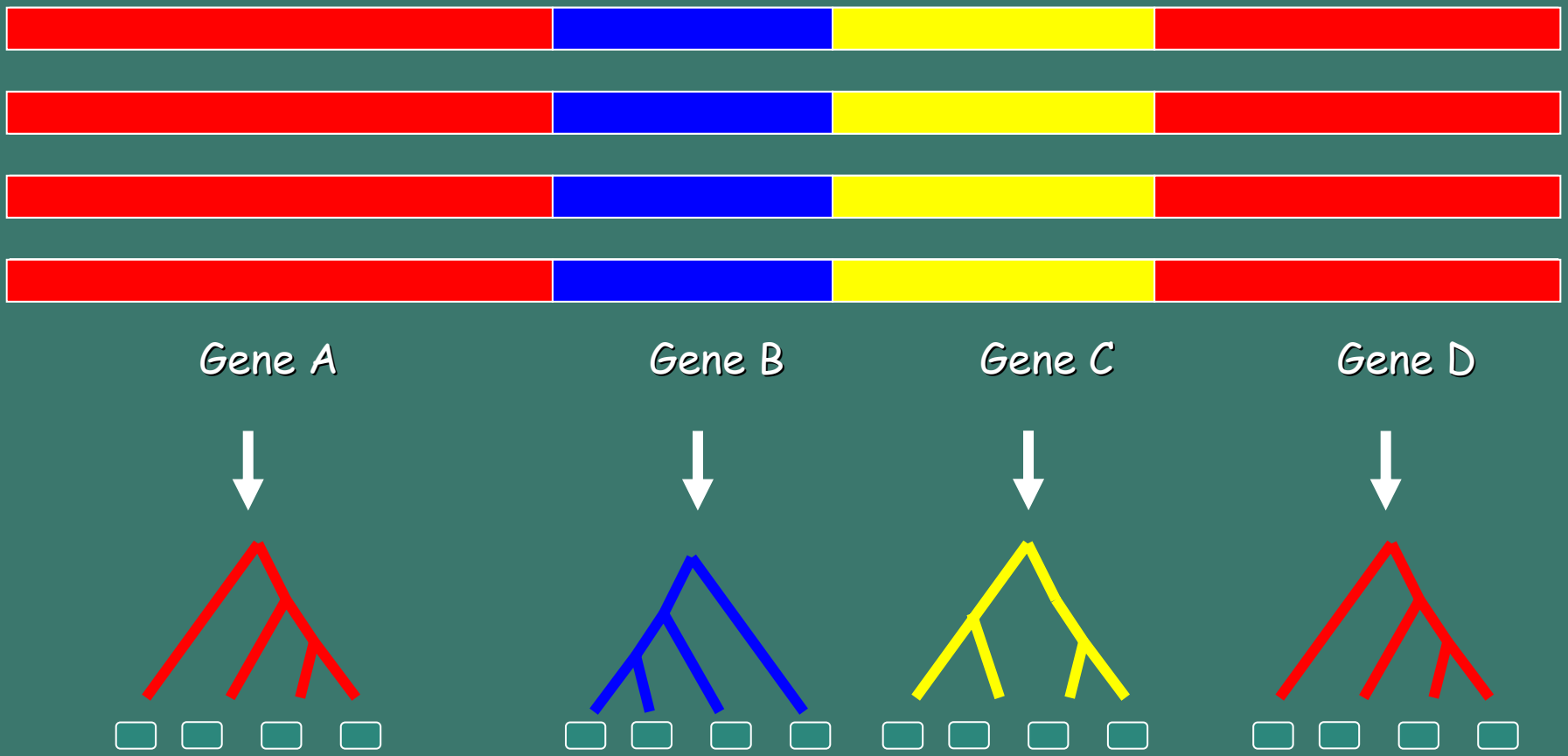
≠



Species Tre

Gene Trees vs Species Trees

Differing gene trees give rise to "mosaic sequences"

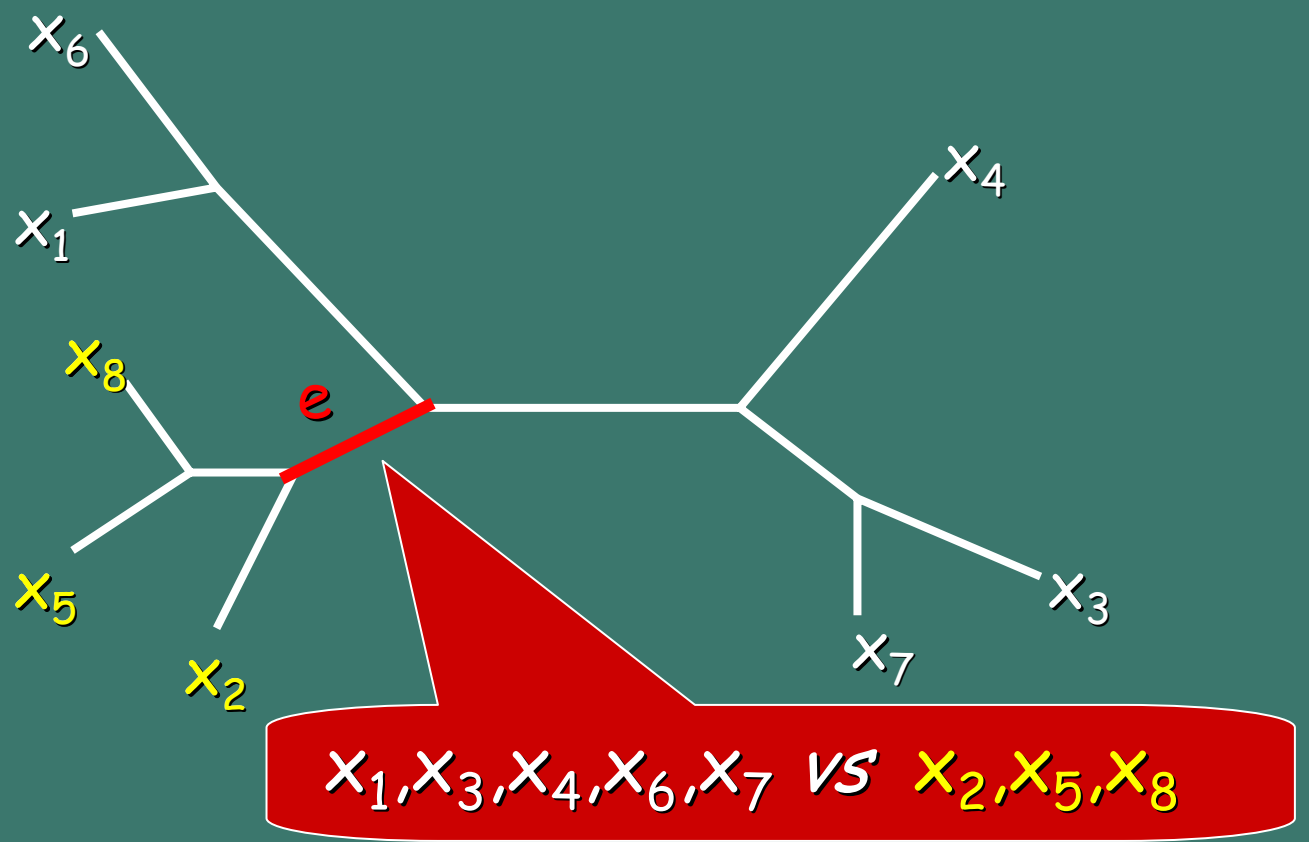


Consensus of Different Gene Trees

- For a given set of species, different genes lead to different trees
- How to form a consensus of the trees?
  - Consensus trees
  - Consensus networks
  - Consensus super networks

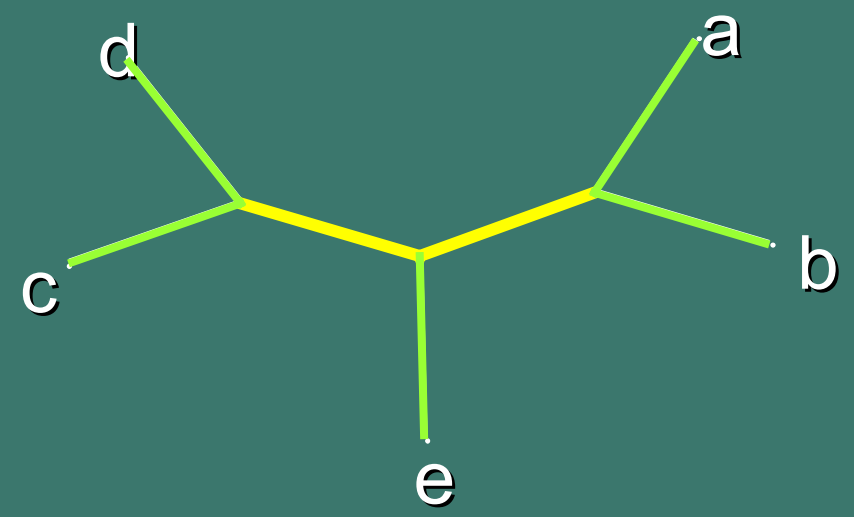
The Splits of a Tree

- Every edge of a tree defines a split of the taxon set X:



The Split Encoding of a Tree

Tree T:



Split encoding $\Sigma(T)$:

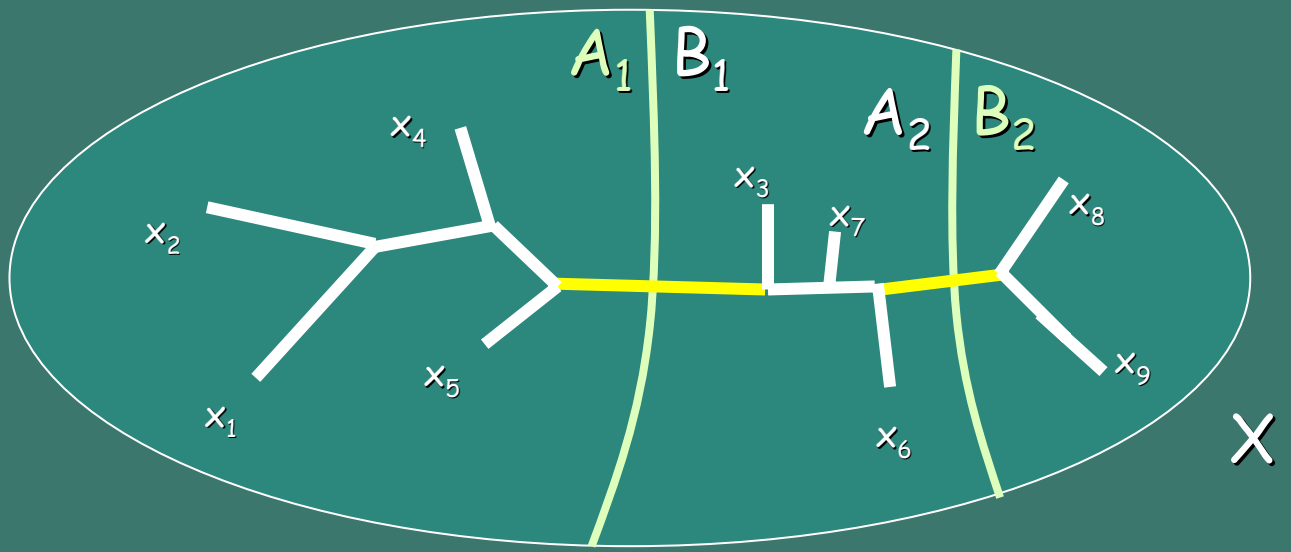
5 trivial splits: $\frac{\{a\}}{\{b,c,d,e\}}, \frac{\{b\}}{\{a,c,d,e\}}, \frac{\{c\}}{\{a,b,d,e\}}, \frac{\{d\}}{\{a,b,c,e\}}$ and $\frac{\{e\}}{\{a,b,c,d\}}$

2 non-trivial splits: $\frac{\{a,b,e\}}{\{c,d\}}$ and $\frac{\{a,b\}}{\{c,d,e\}}$.

Compatibility

- Two splits $A_1|B_1$ and $A_2|B_2$ of X are *compatible*, if $\emptyset \in \{A_1 \cap A_2, A_1 \cap B_2, B_1 \cap A_2, B_1 \cap A_2\}$

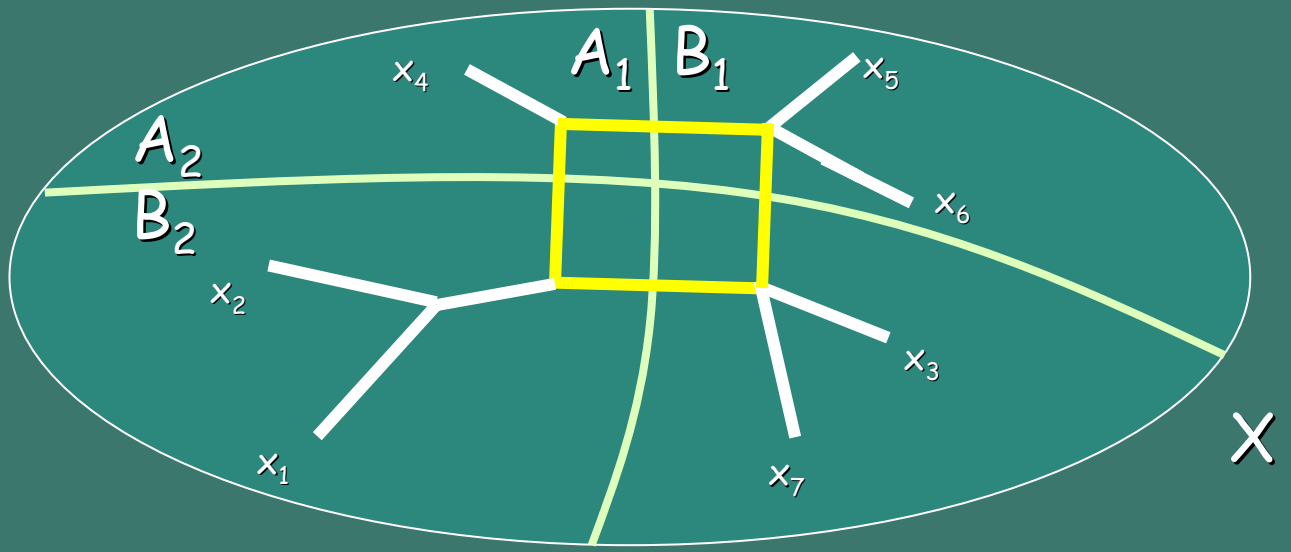
Two compatible splits:



Compatibility

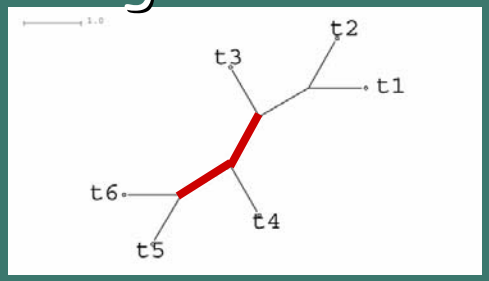
- Two splits $A_1|B_1$ and $A_2|B_2$ of X are *compatible*, if $\emptyset \in \{A_1 \cap A_2, A_1 \cap B_2, B_1 \cap A_2, B_1 \cap A_2\}$

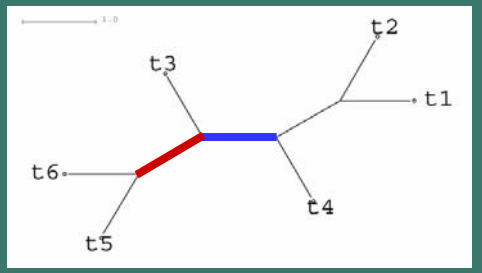
Two **incompatible** splits:

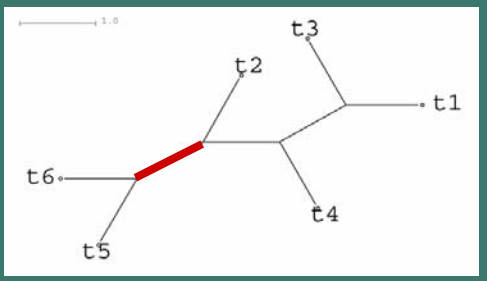


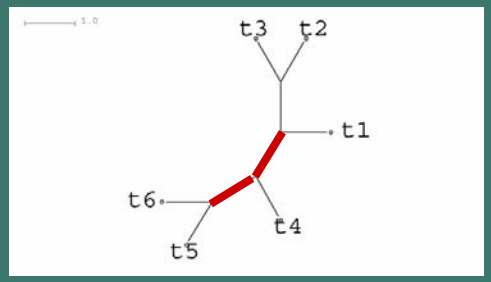
Consensus of Trees

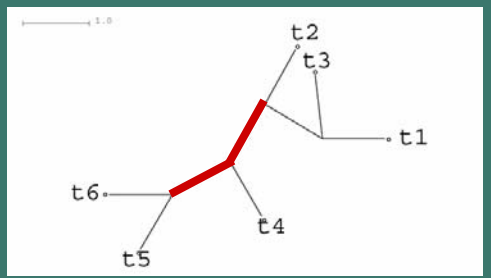
Six gene trees:

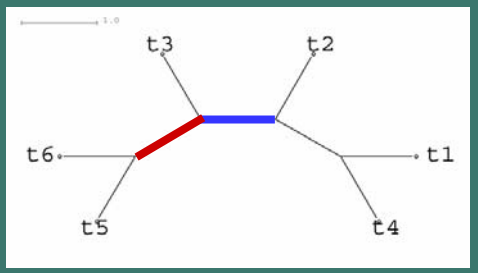




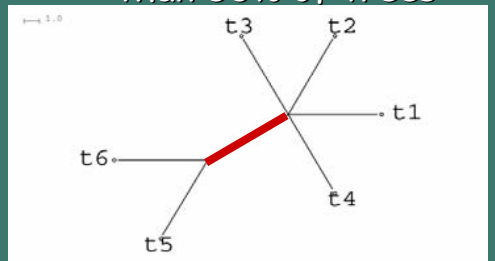




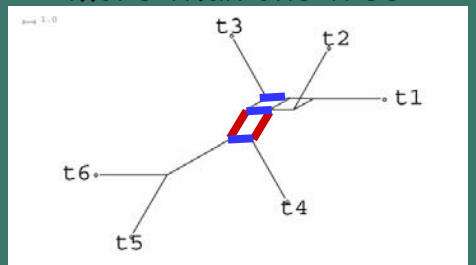




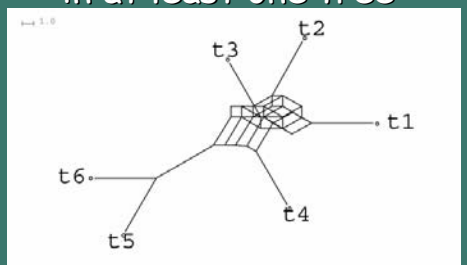
Σ(1/2): majority consensus: splits contained in more than 50% of trees



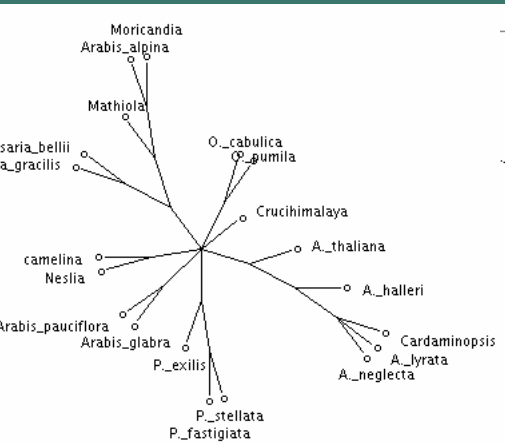
Σ(1/6): splits contained in more than one tree

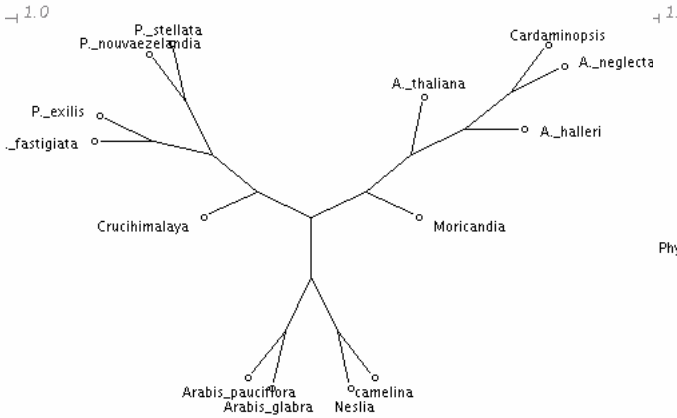


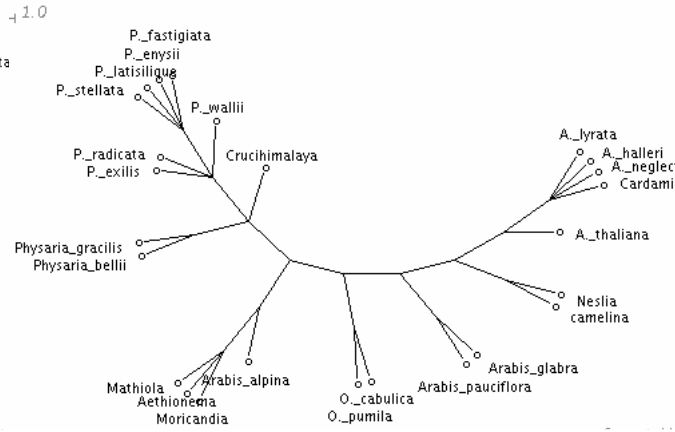
Σ(0): splits contained in at least one tree

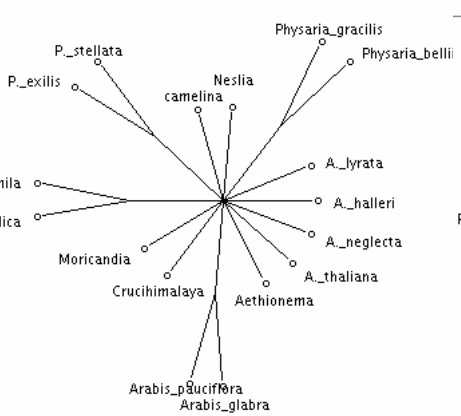


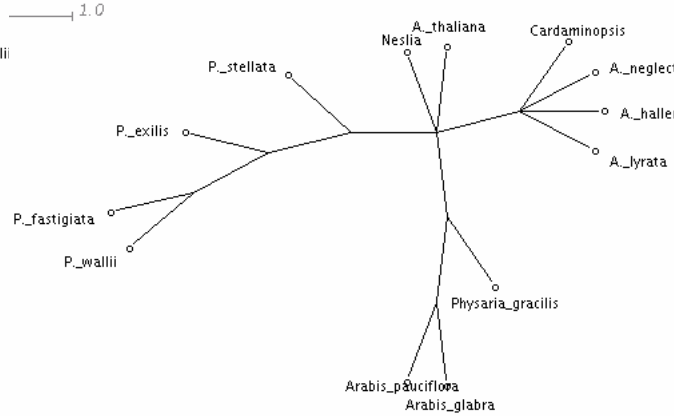
Example of A Super Network (Plants)











artial trees for five plant genes

Super network

Z-Closure Method

- Idea: Extend partial splits.
- Z-rule:

$$\frac{A_1 \cap A_2}{B_1 \quad B_2} \rightarrow \frac{A_1}{B_1 \cup B_2}, \frac{A_1 \cup A_2}{B_2}$$

- Repeatedly apply to completion.
- Return all full splits.

[Huson, Dezulian, Kloepper and Steel, 2004]

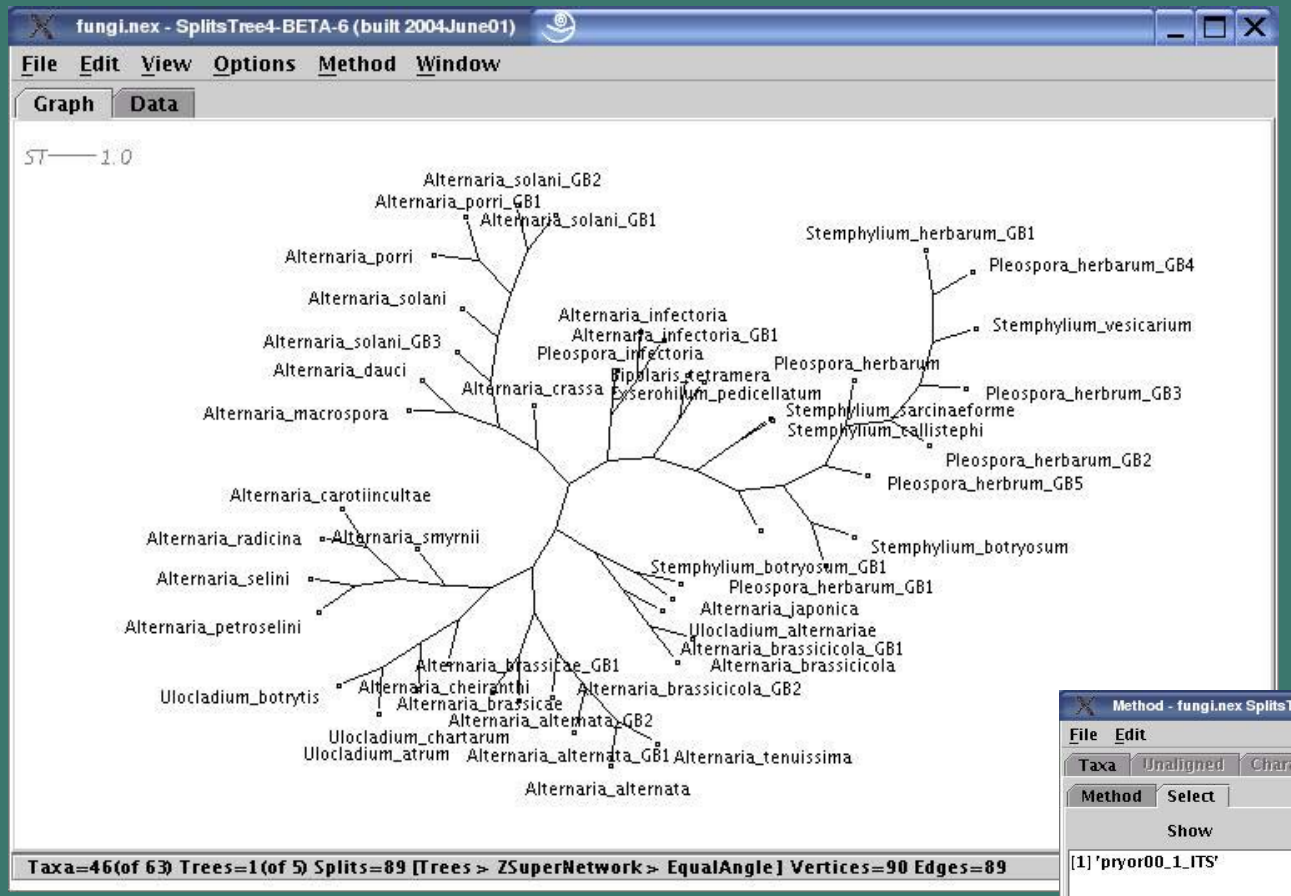
Example

- Five fungal trees from [Pryor, 2000] and [Pryor, 2003]
- Trees:
  - ITS (two trees)
  - SSU (two trees)
  - Gpd (one tree)
- Numbers of taxa differ: "partial trees"

Example

4beta25: User manual now available from www.splitstree.org!

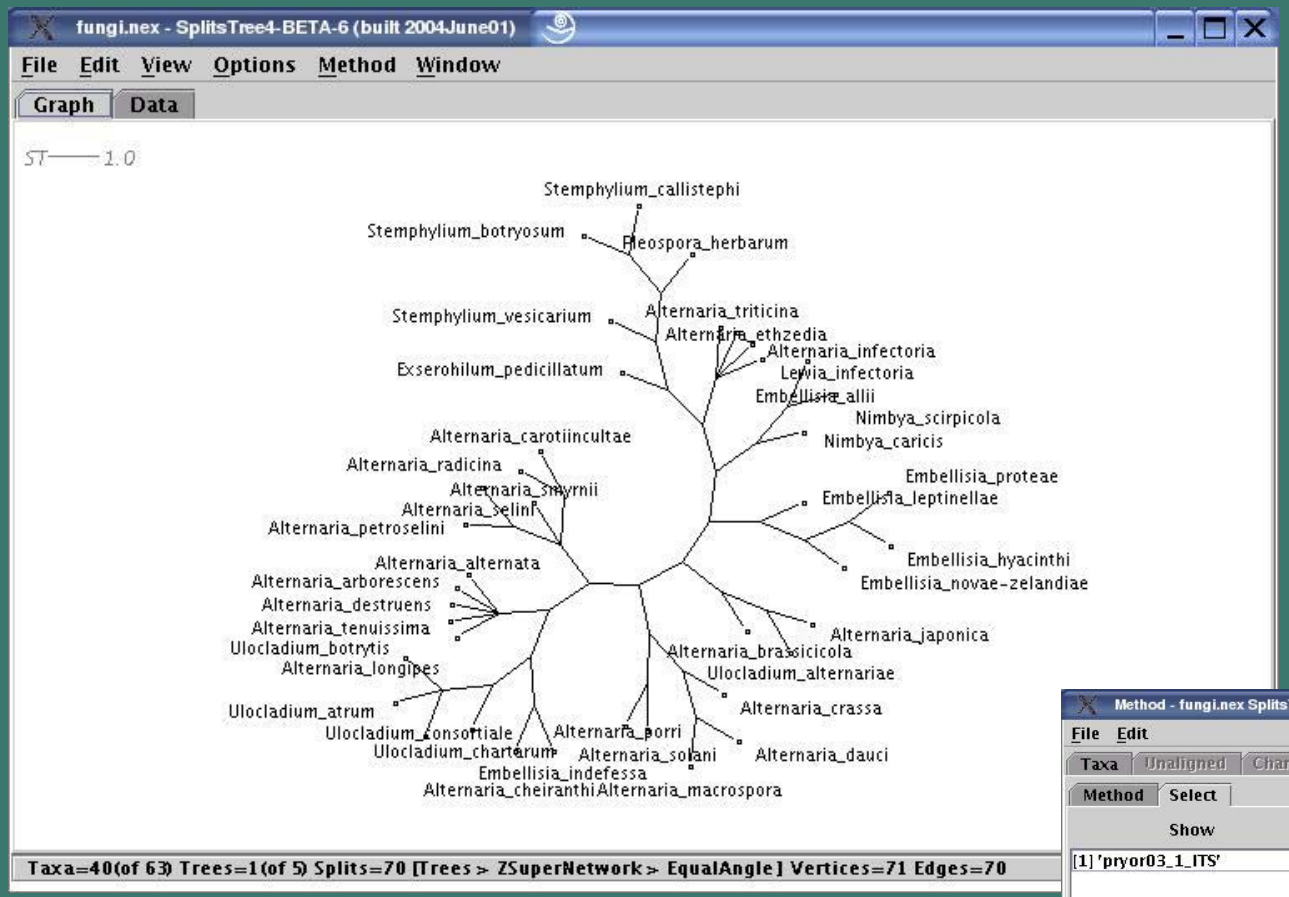
Individual Gene Trees



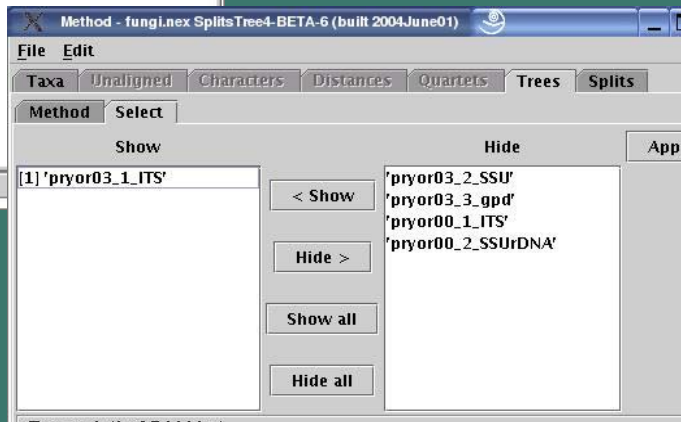
ITS00

46 taxa

Individual Gene Trees

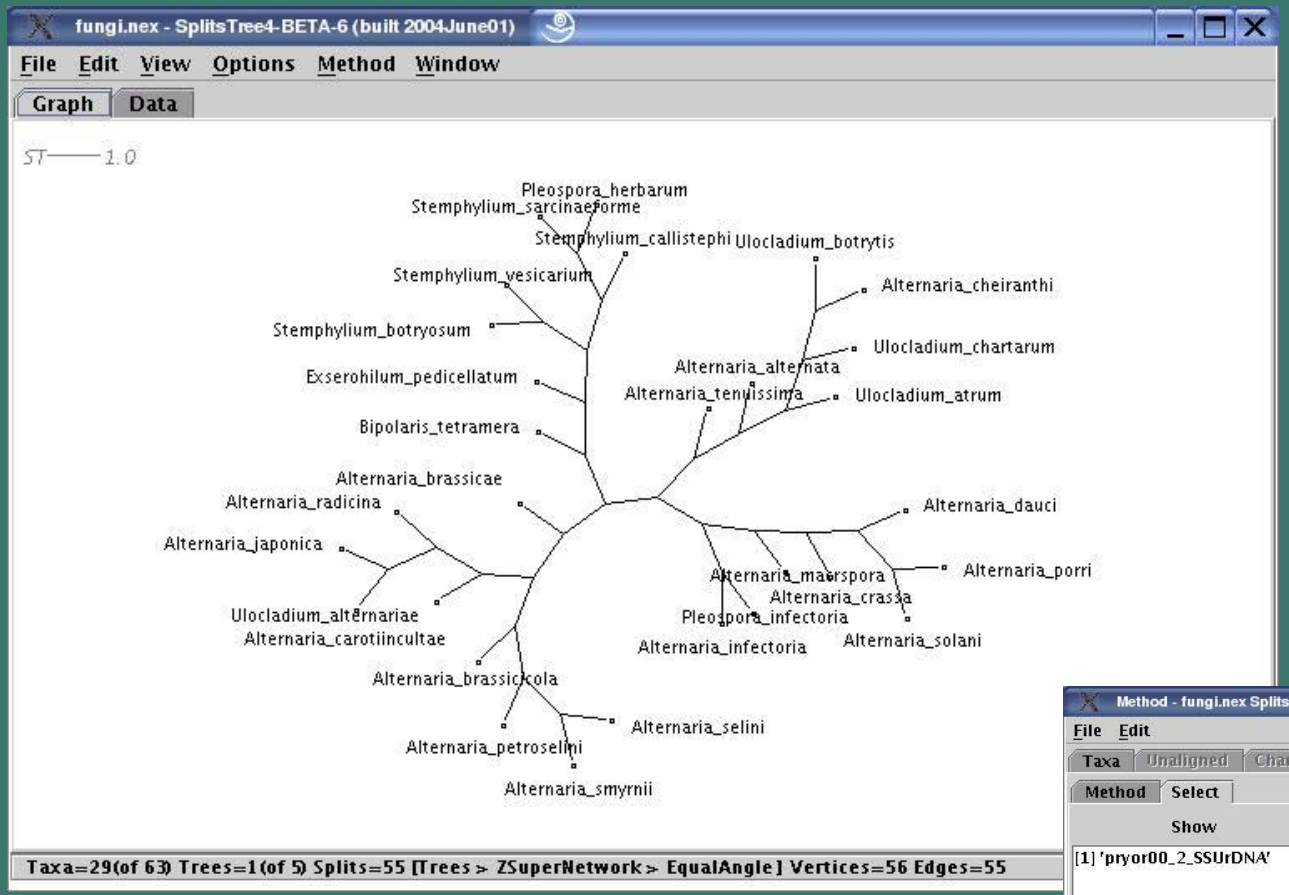


ITS03



40 taxa

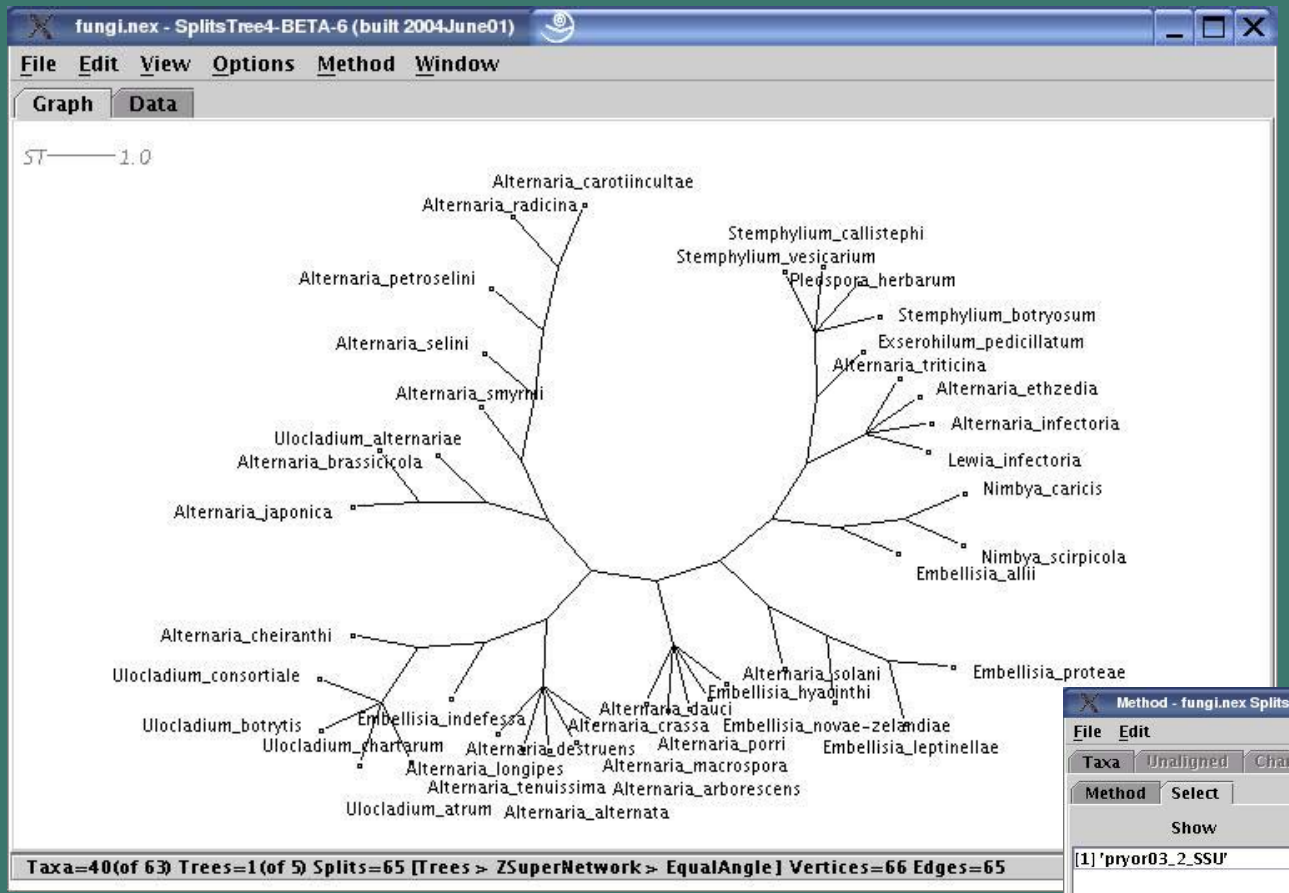
Individual Gene Trees



SSU00

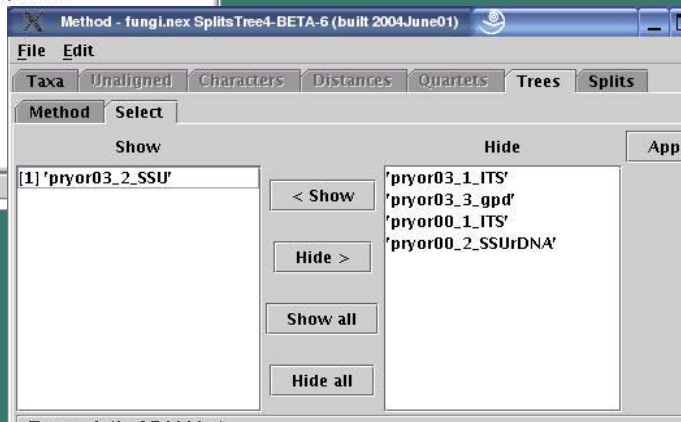
29 taxa

Individual Gene Trees

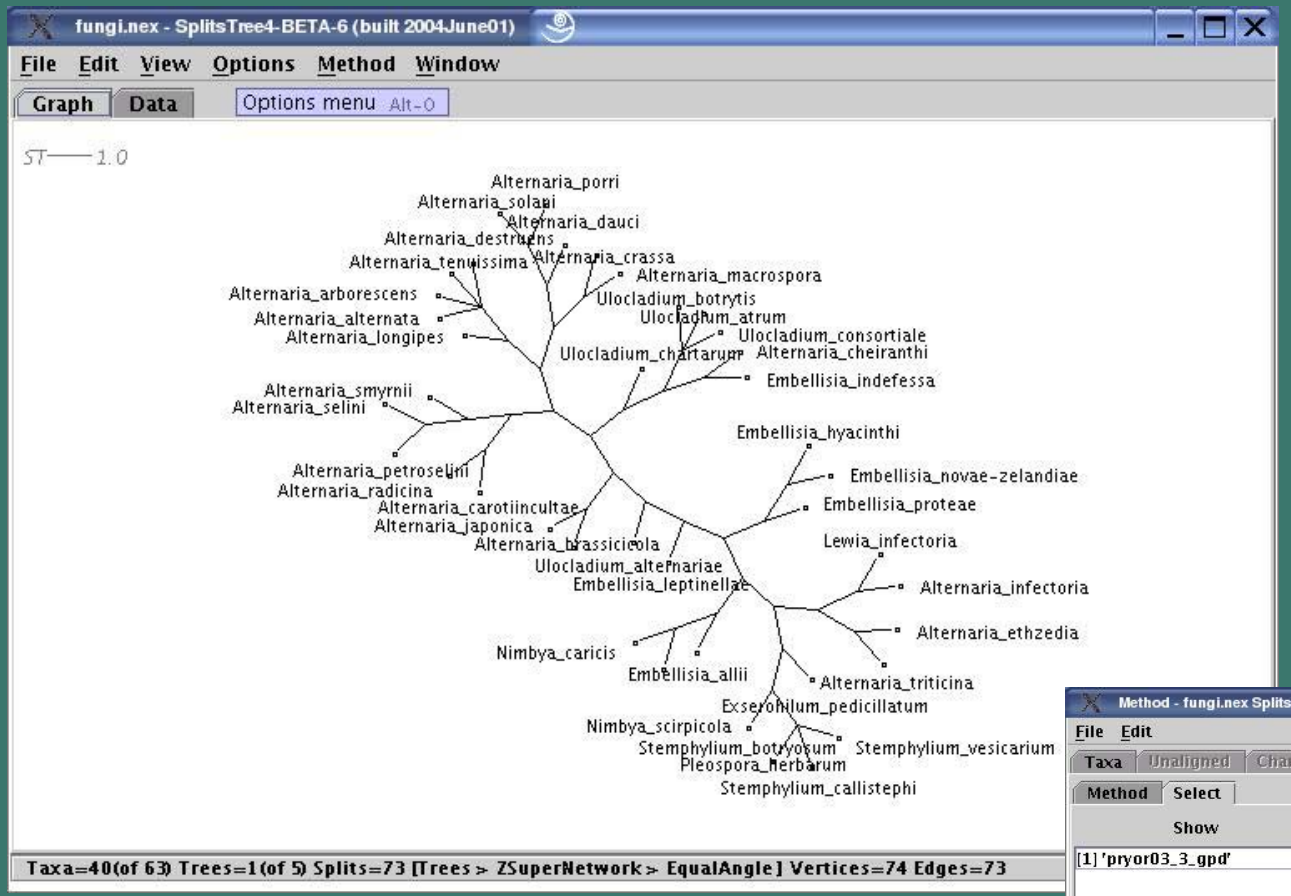


SSU03

40 taxa



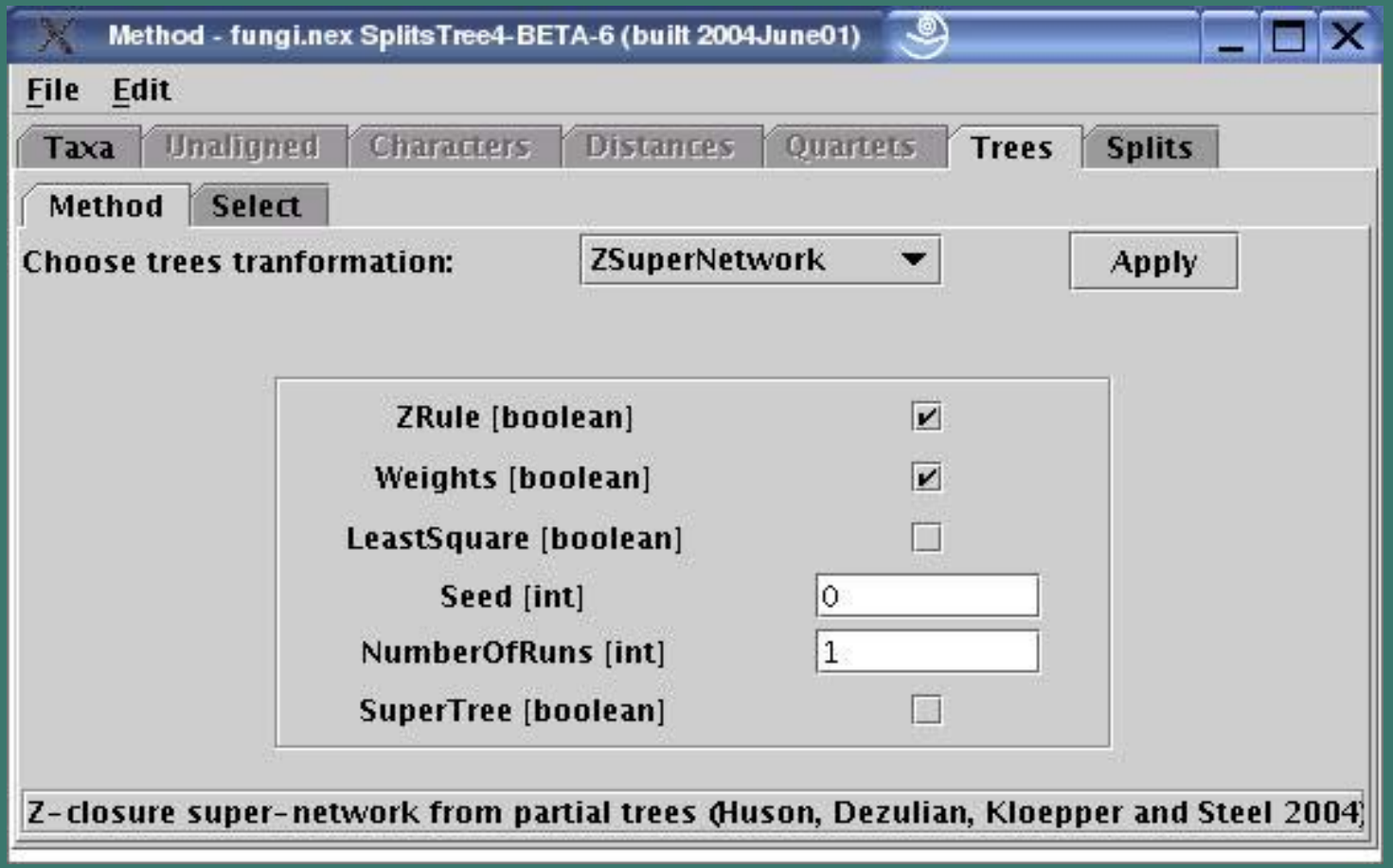
Individual Gene Trees



Gpd03

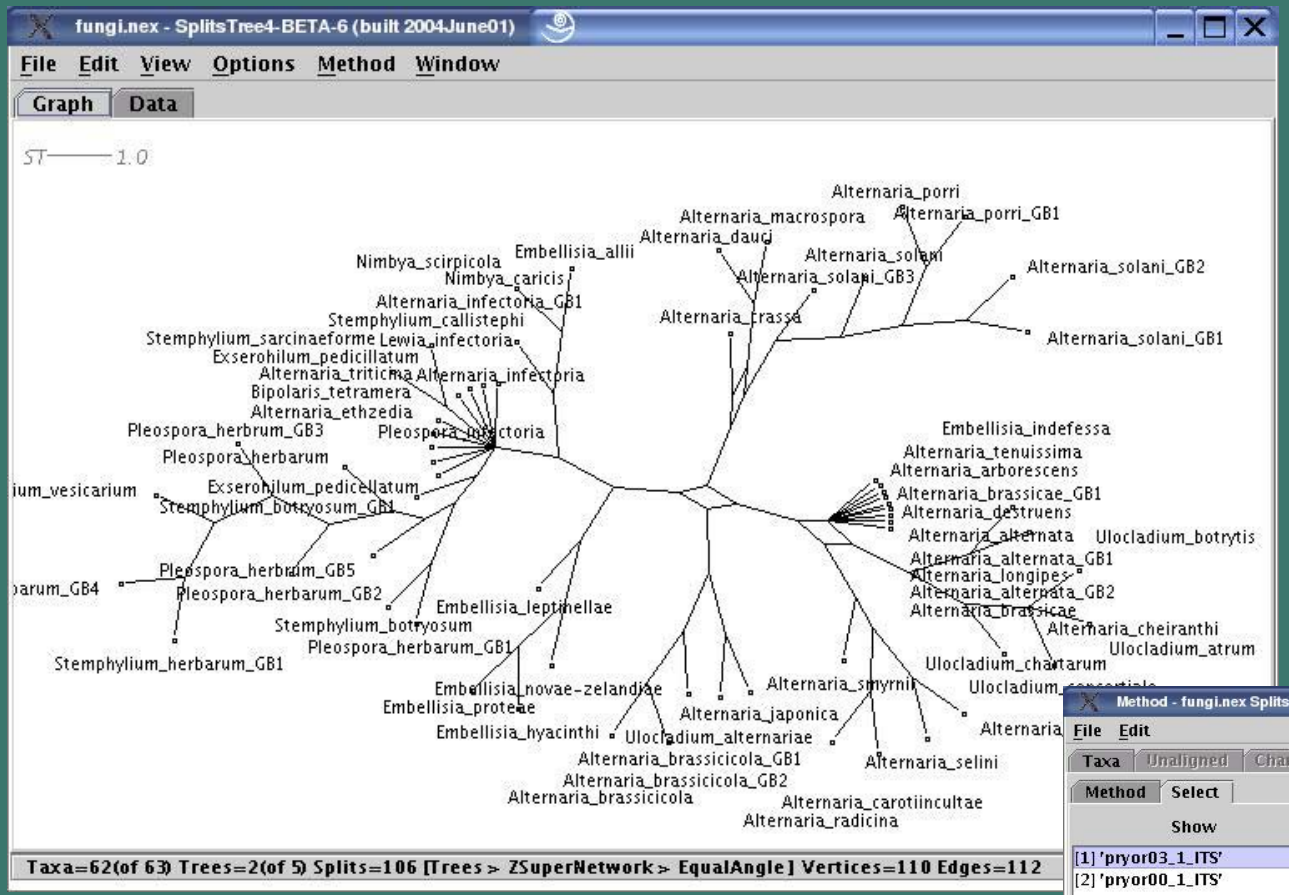
40 taxa

Gene Trees as Super Network

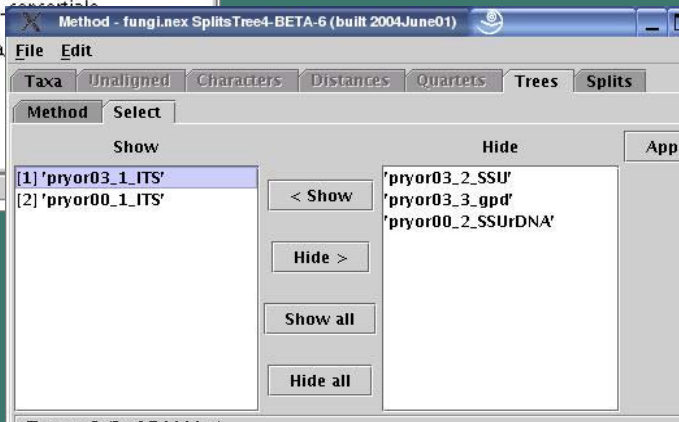


Z-closure: a fast super-network method

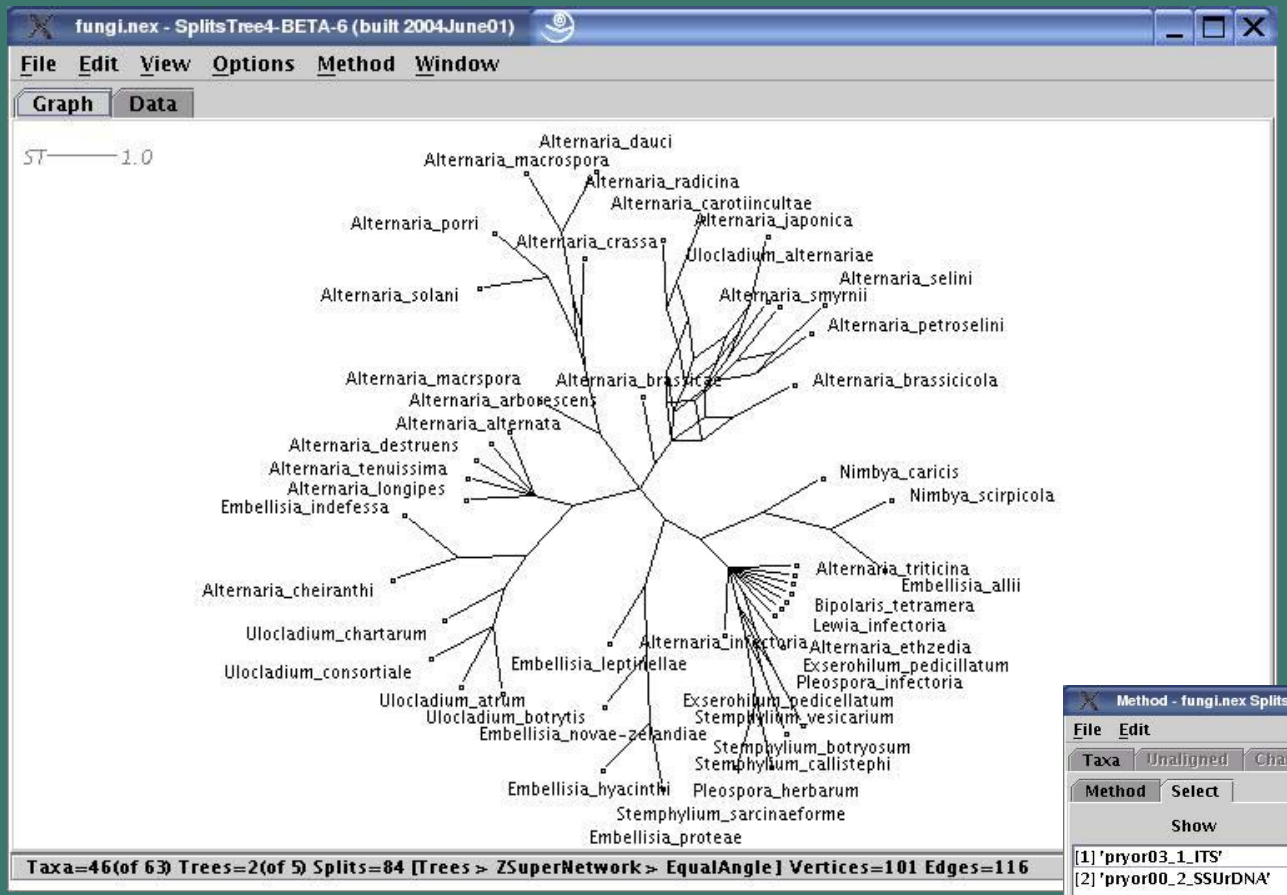
Gene Trees as Super Network



ITS00+
ITS03

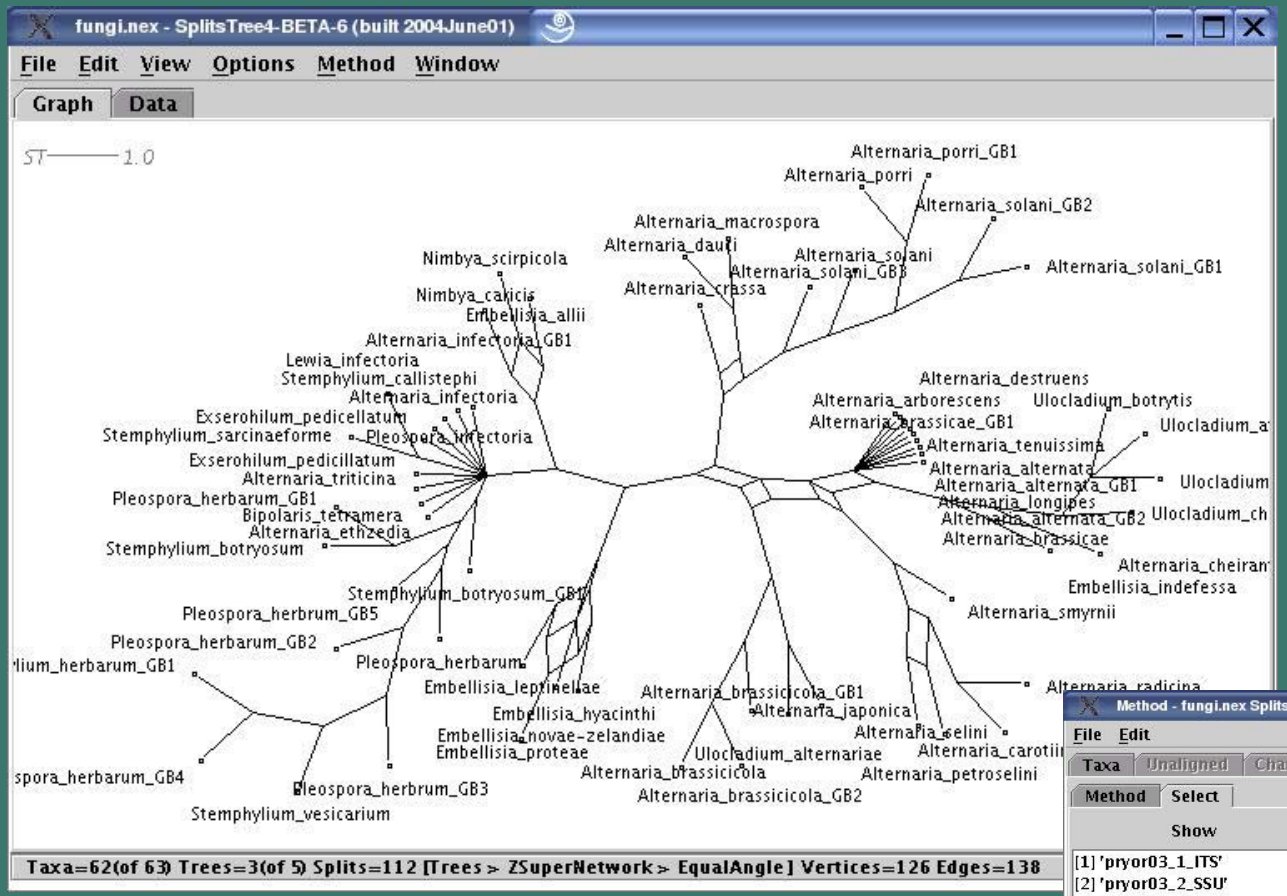


Gene Trees as Super Network



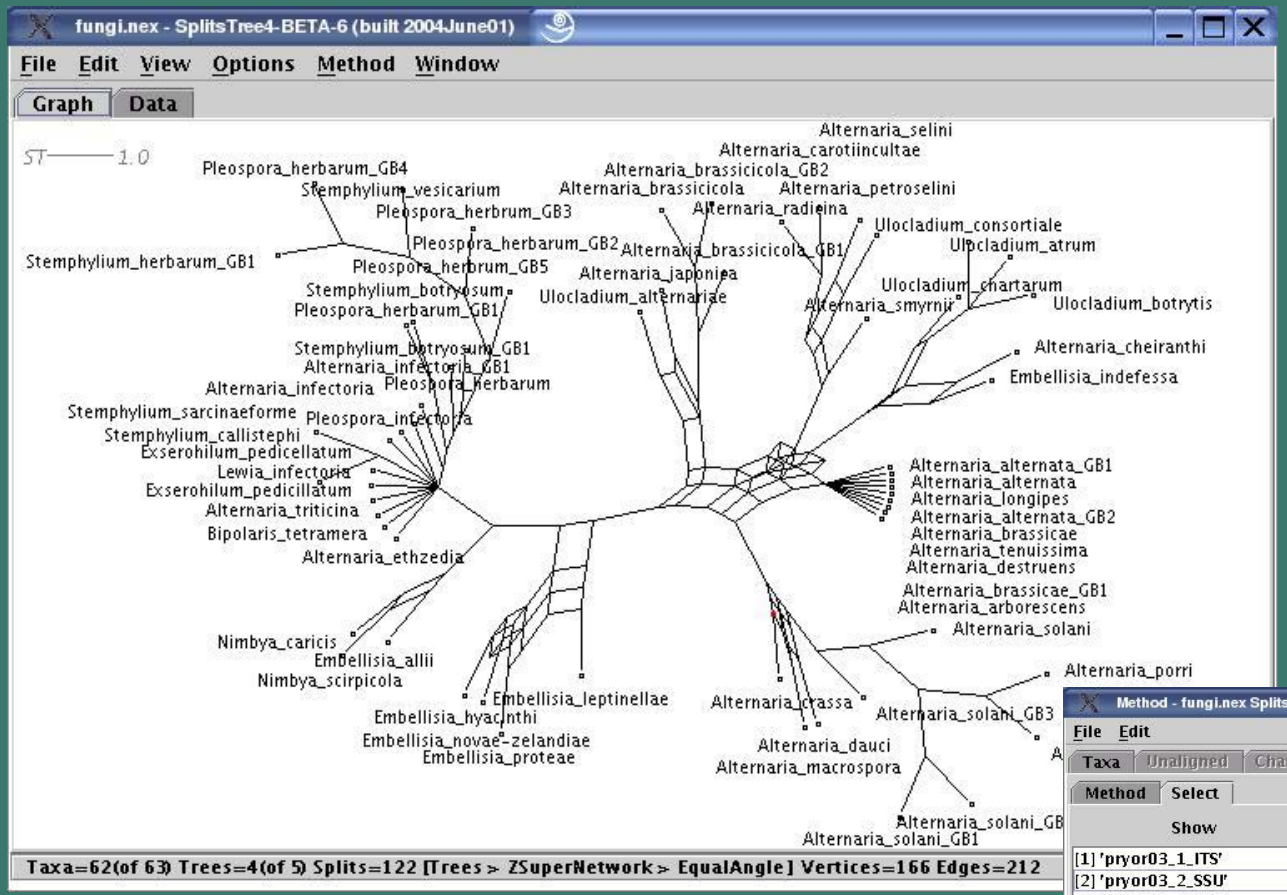
ITS03+
SSU00

Gene Trees as Super Network



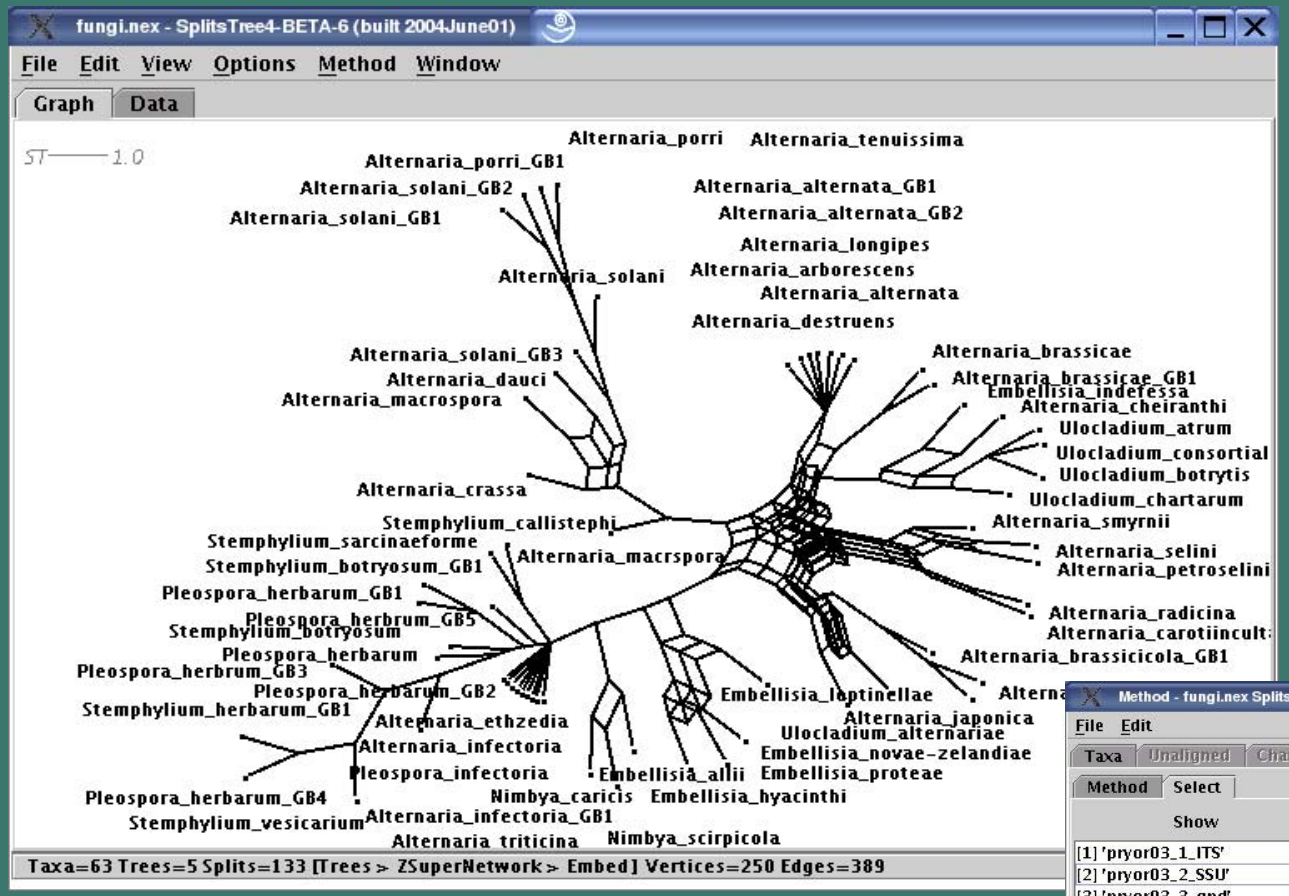
ITS00+
ITS00+
SSU03

Gene Trees as Super Network



ITS00+
ITS03+
SSU03+
Gpd03

Gene Trees as Super Network



ITS00+
ITS03+
SSU00+
SSU03+
Gpd03

Part IV

1. Phylogenetic trees
2. Splits networks
3. Consensus networks
4. Hybridization and reticulate networks
5. Recombination networks

Hybridization

- Occurs when two organisms from different species interbreed and combine their genomes

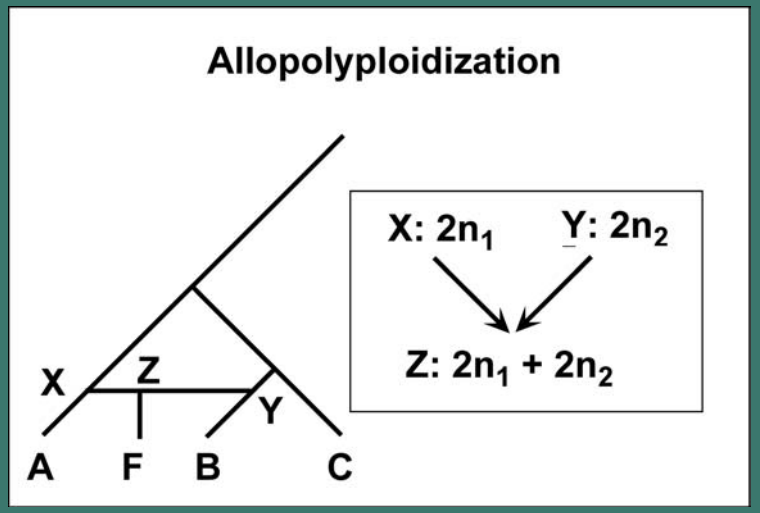
Water hemp

Hybrid

Pigs weed

Speciation by Hybridization 1

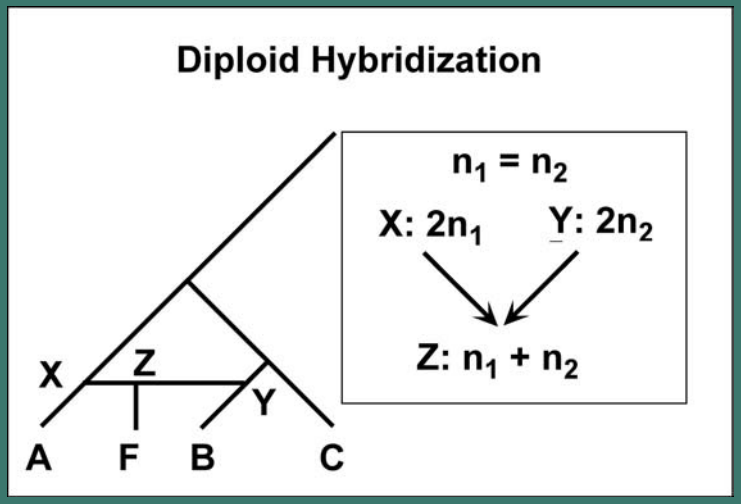
- In allopolyploidization, two different lineages produce a new species that has the complete nuclear genomes of both parental species:



Linder *et al.* 2004

Speciation by Hybridization 2

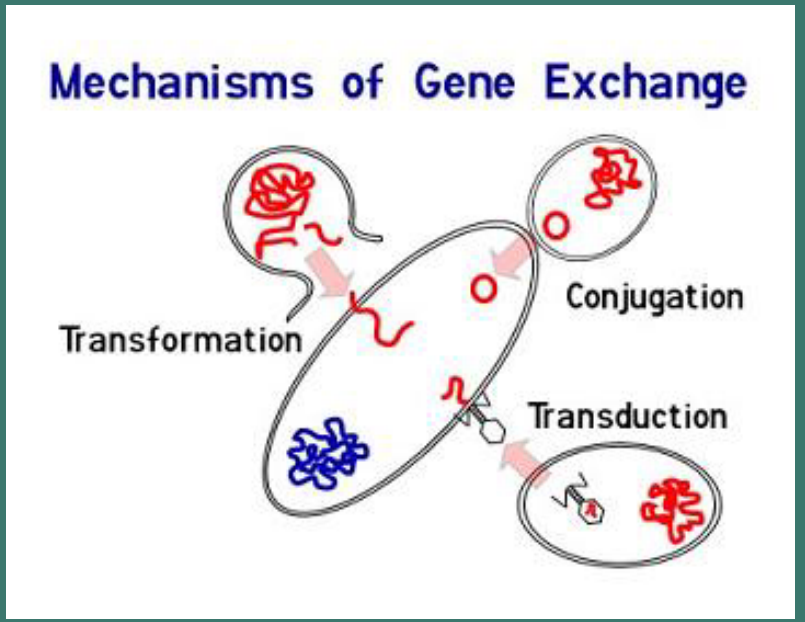
- In diploid (or homoploid) hybrid speciation, each of the parents produces normal gametes (haploid) to produce a normal diploid hybrid:



Linder *et al.* 2004

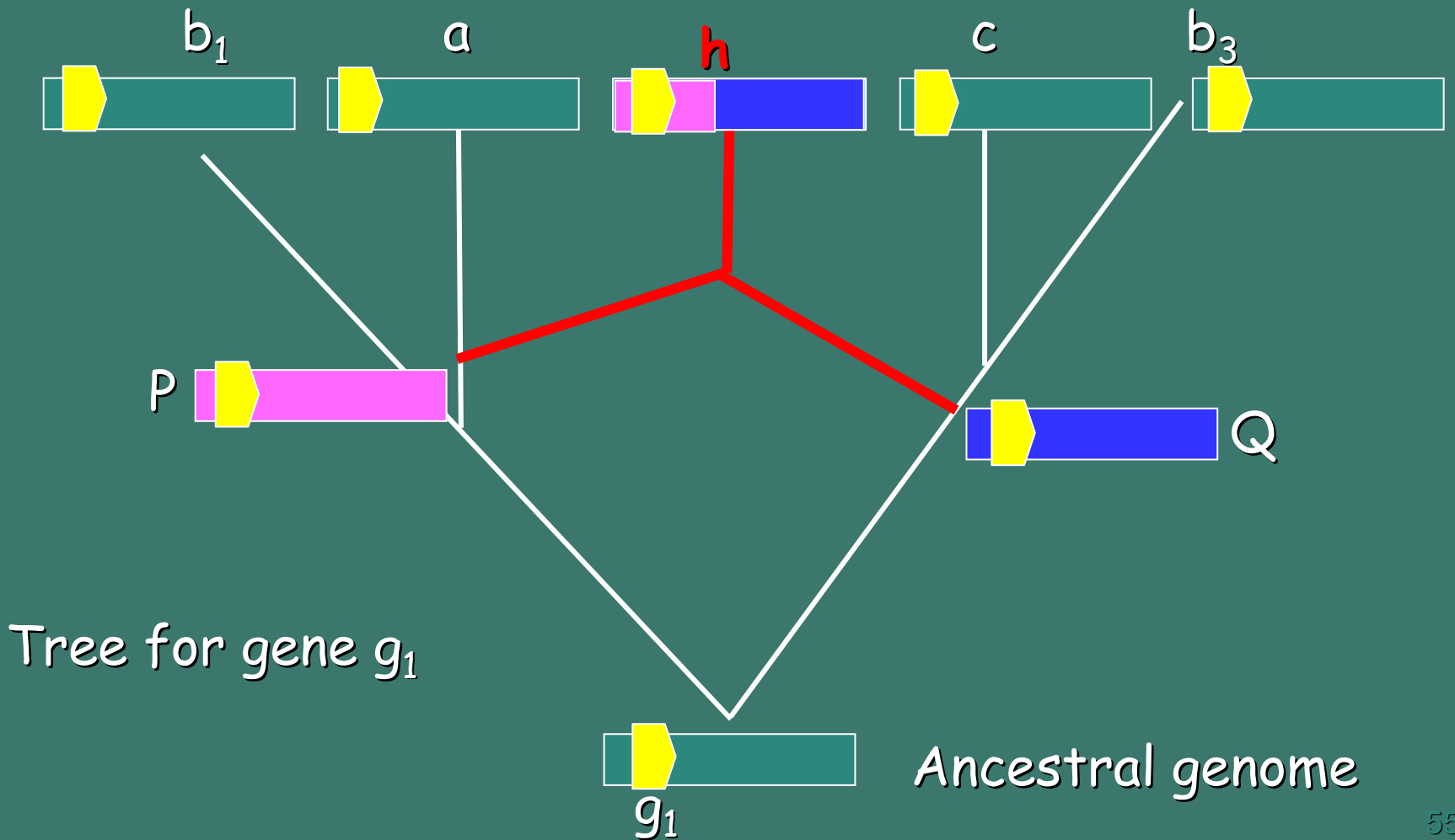
Horizontal Gene Transfer

- There are a number of known mechanisms by which bacteria can exchange genes
  - Transformation
  - Conjugation
  - transduction

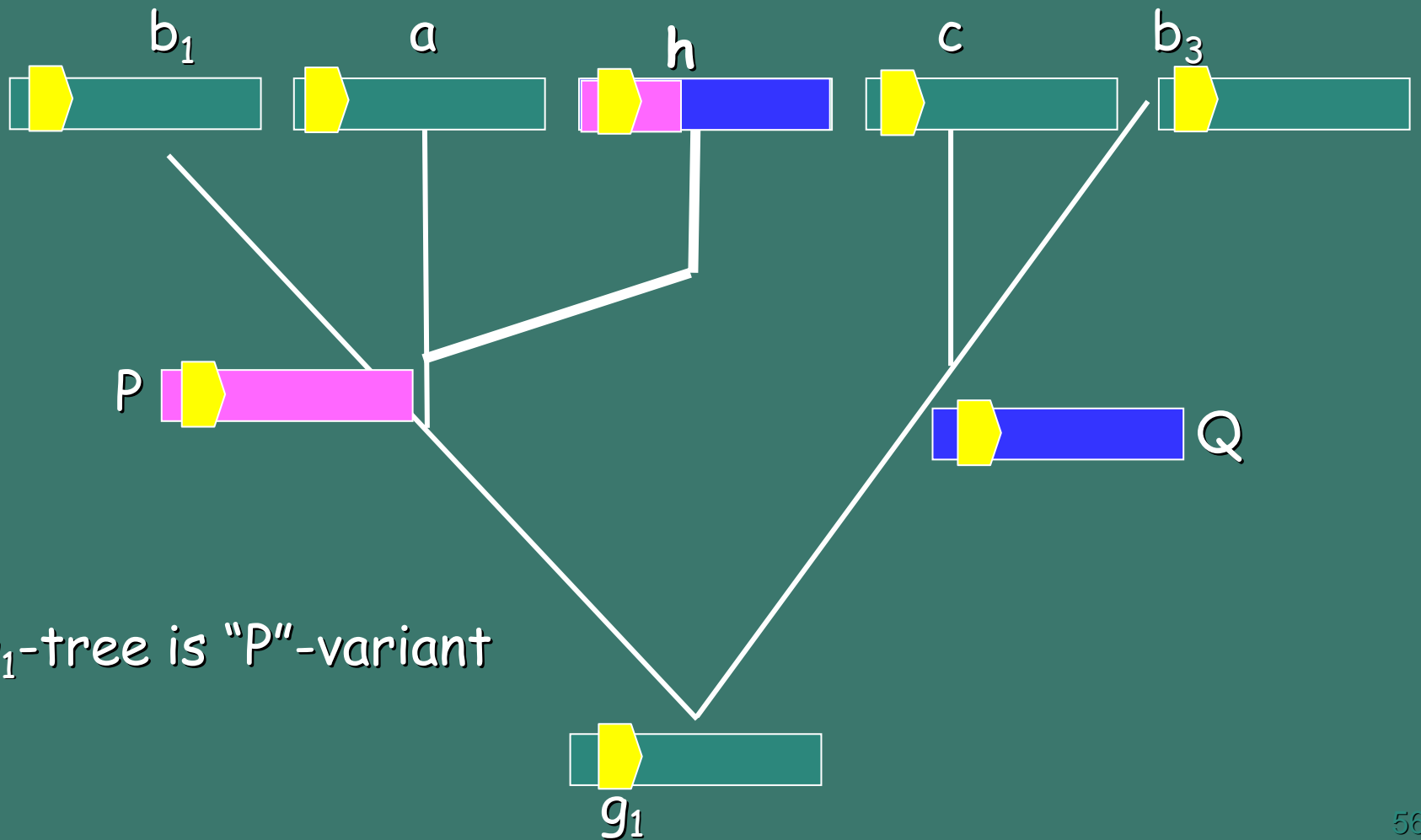


http://www.pitt.edu/~heh1/research.html

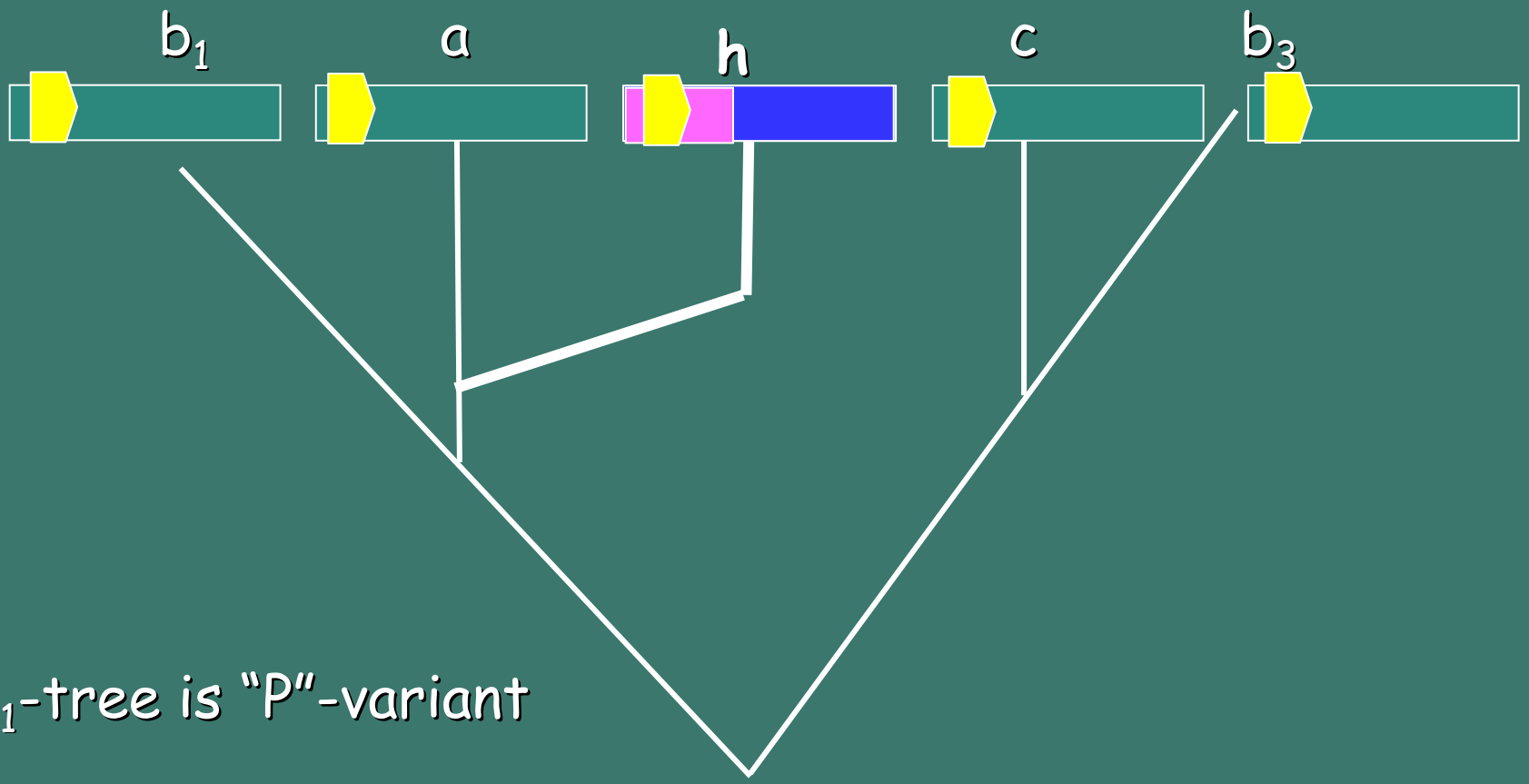
A Simple Model of Reticulate Evolution



A Simple Model of Reticulate Evolution

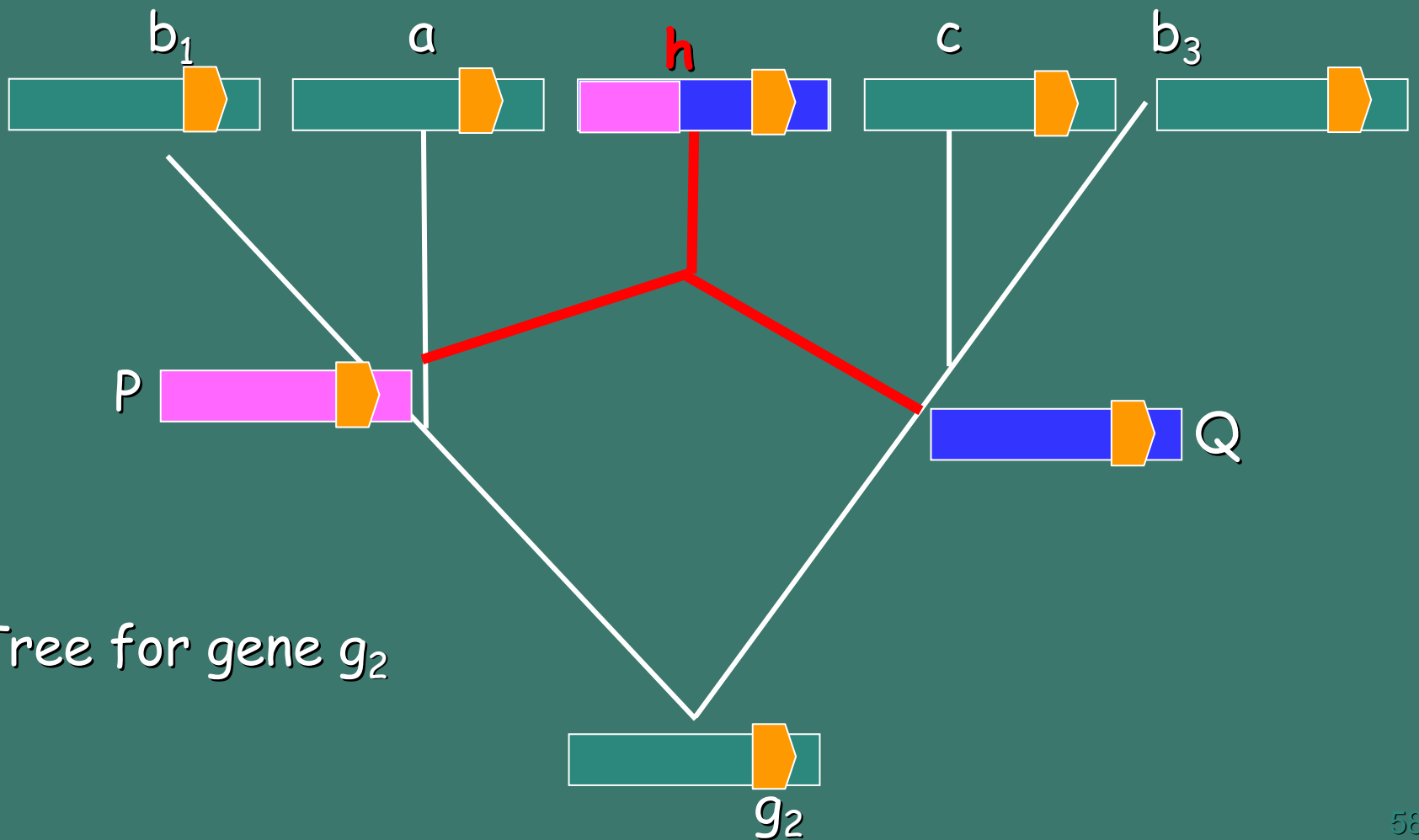


A Simple Model of Reticulate Evolution



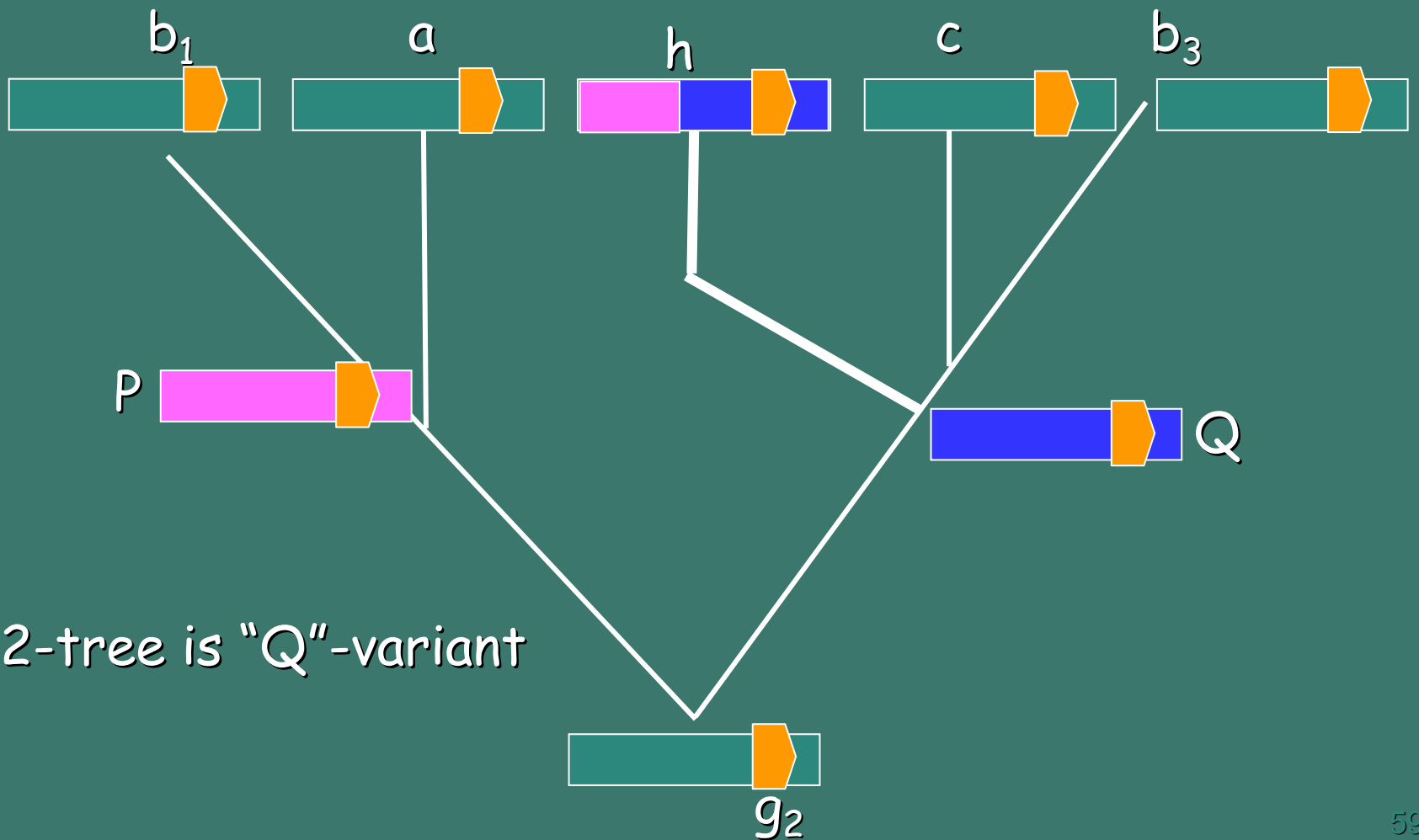
g_1-tree is "P"-variant

A Simple Model of Reticulate Evolution



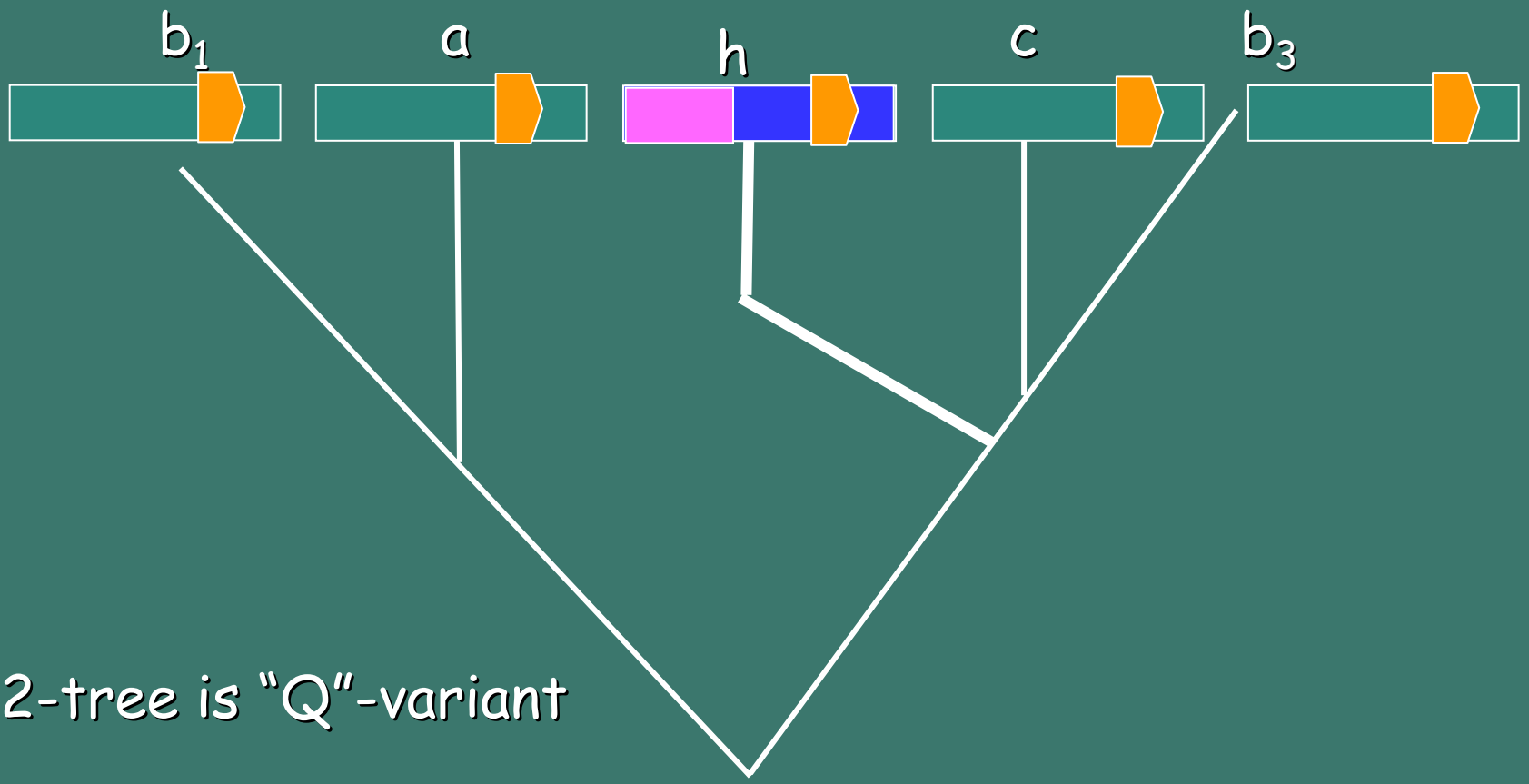
Tree for gene g_2

A Simple Model of Reticulate Evolution



g2-tree is "Q"-variant

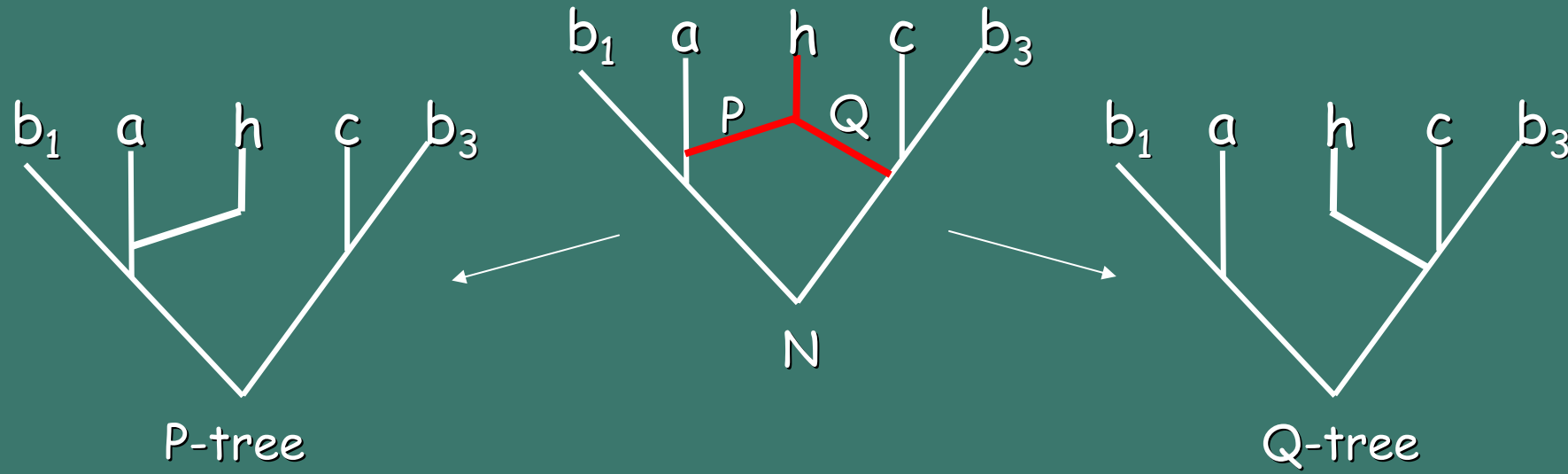
A Simple Model of Reticulate Evolution



g2-tree is "Q"-variant

Reticulate Networks and Trees

- The evolutionary history associated with any given gene is a tree
- A network N with k reticulations gives rise to 2^k different gene trees

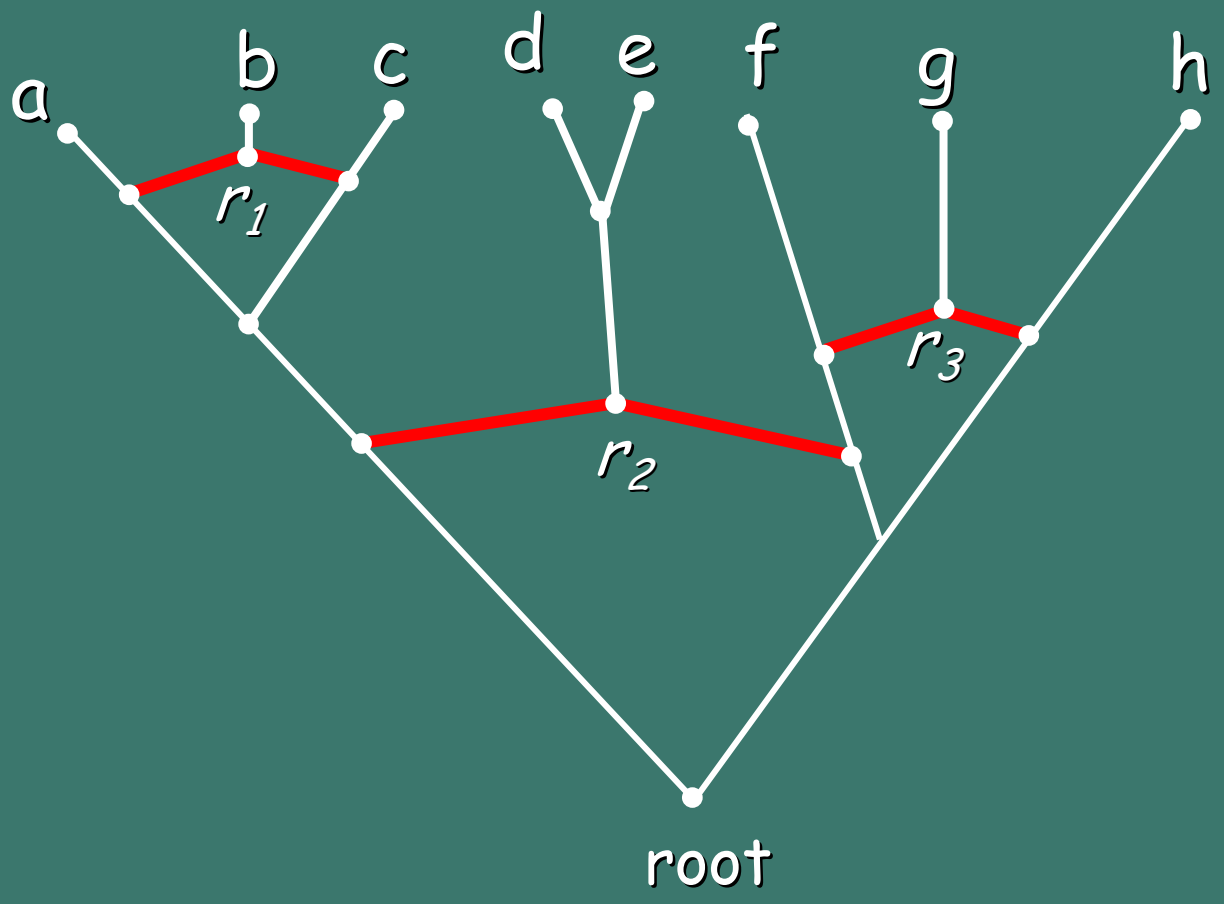


Rooted Reticulate Network

Definition Let X be a set of taxa. A rooted *reticulate network* N on X is a connected, directed acyclic graph with:

- precisely one node of indegree 0, the *root*,
- all other nodes are *tree nodes* of indegree 1, or *reticulation nodes* of indegree 2,
- every edge is a *tree edge* joining two tree nodes, or a *reticulation edge* from a tree node to a reticulation node, and
- the set of leaves consists of tree nodes and is labeled by X.

Rooted Reticulate Network



Most Parsimonious Network Problem:

- Given a set of trees $\mathcal{T}$, determine a reticulate network N such that

  $$\mathcal{T} \subseteq T(N)$$

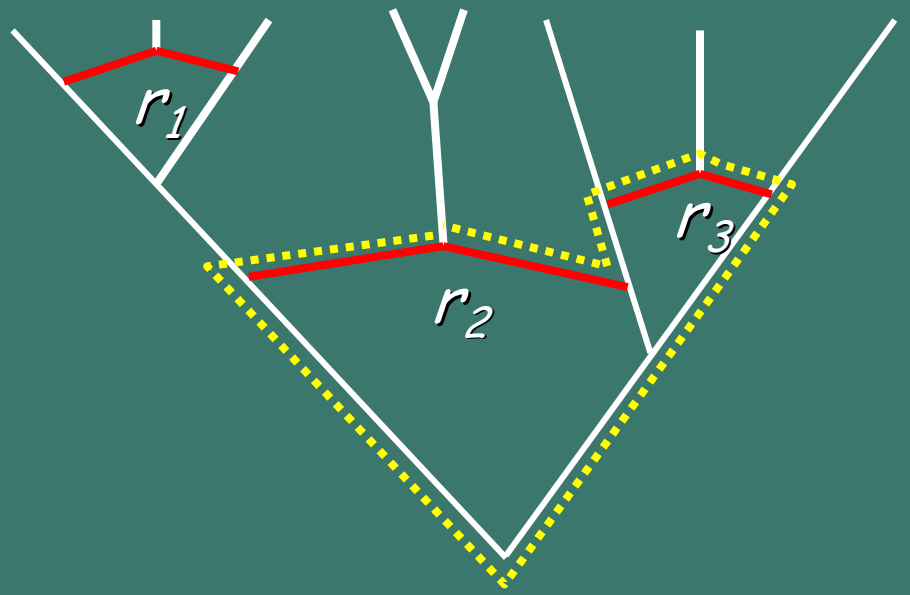
  and N contains a minimum number of reticulation nodes.

- In fully generality, this is known to be a computationally hard problem

  [Wang *et al* 2001, Bordewich and Semple 2004].

Independent Reticulations

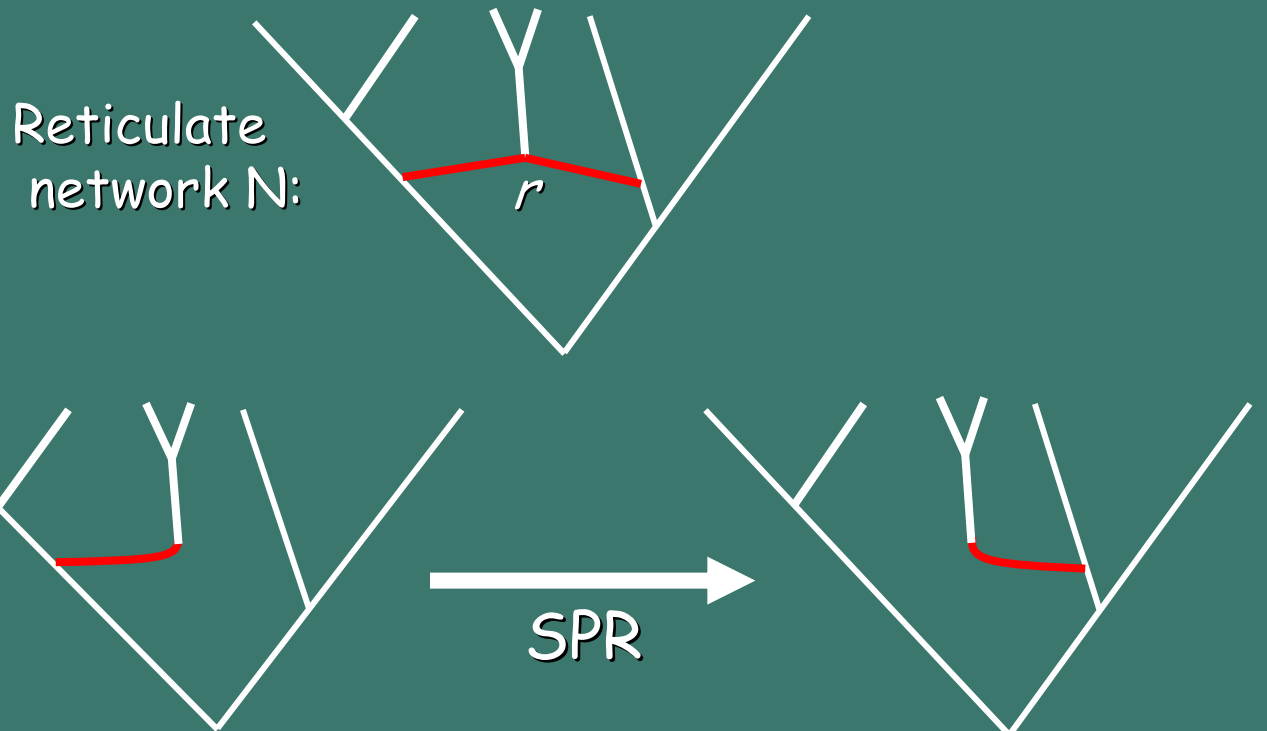
- Reticulation nodes $r_i, r_j \in N$ are *independent*, if they are not contained in a common cycle:



- Independent reticulations also called *galls* and a network only containing galls is also called a *galled tree* [Gusfield *et al.* 2003]

SPR's and Reticulations

Observation [Maddison 1997]: If N contains only one reticulation r, then it corresponds to a "sub-tree prune and regraft" operation:



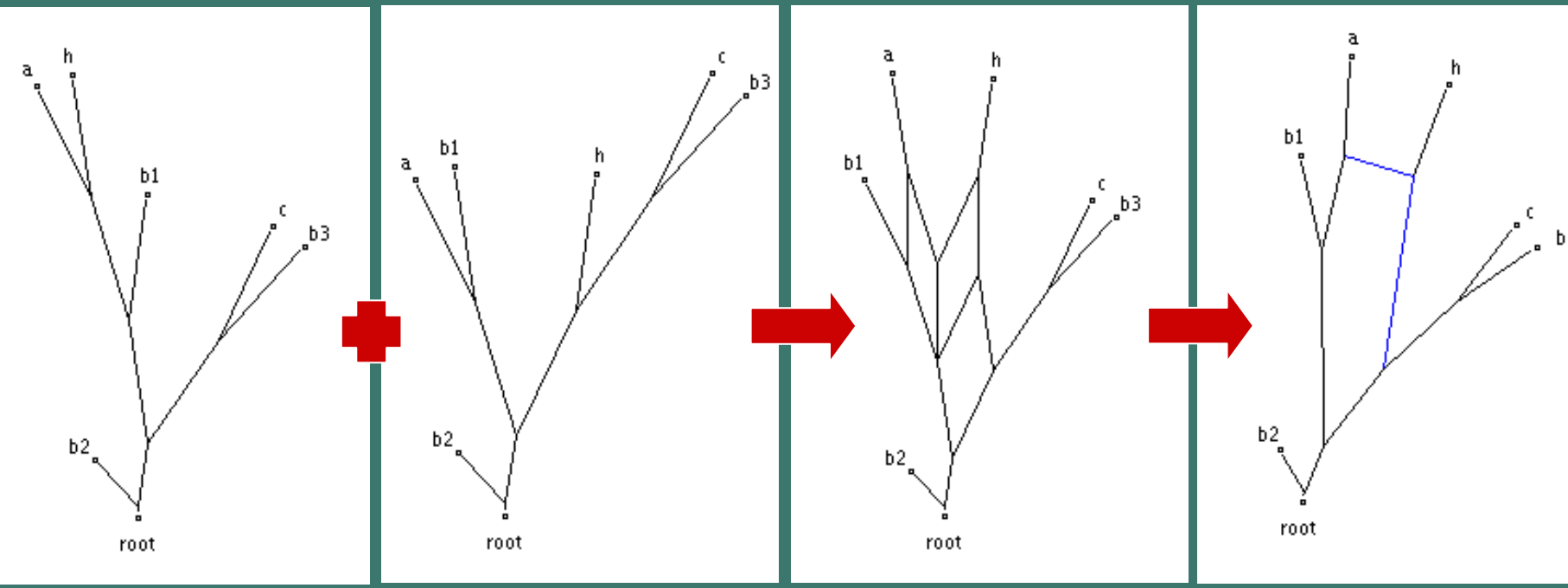
SPR-Based Algorithm

- Given two bifurcating trees, compute their SPR distance:
  - If = 0, return the tree
  - If = 1, return the reticulate network
  - Else, return fail

- Generalized to networks with multiple independent reticulations [Nakhleh *et al* 2004]
- Maximum agreement forest approach (Semple *et al* 2005)

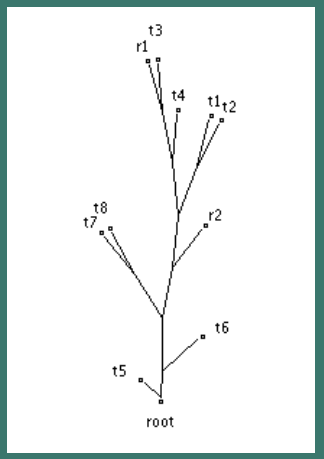
Splits-Based Approach

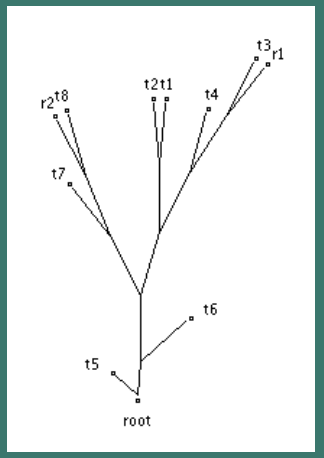
A new splits-based approach [Huson, Kloepper, Lockhart and Steel 2005]:

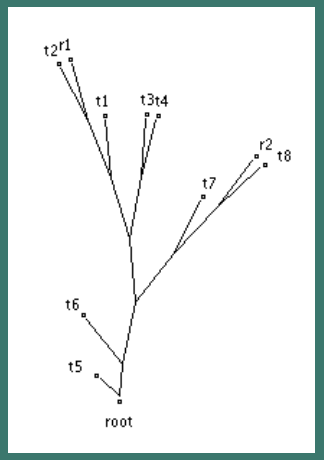


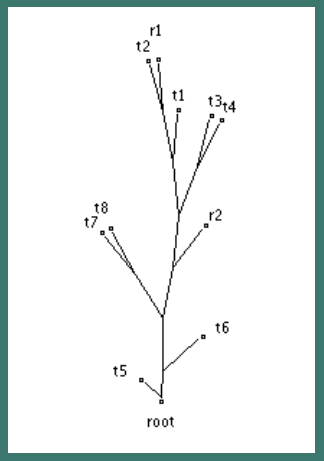
gene tree1 | gene tree2 | splits network of all splits | reticulate network

Multiple Independent Reticulations

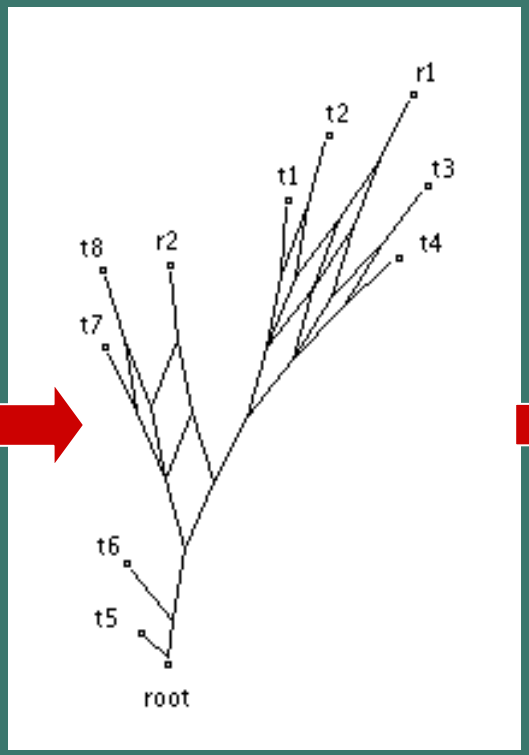




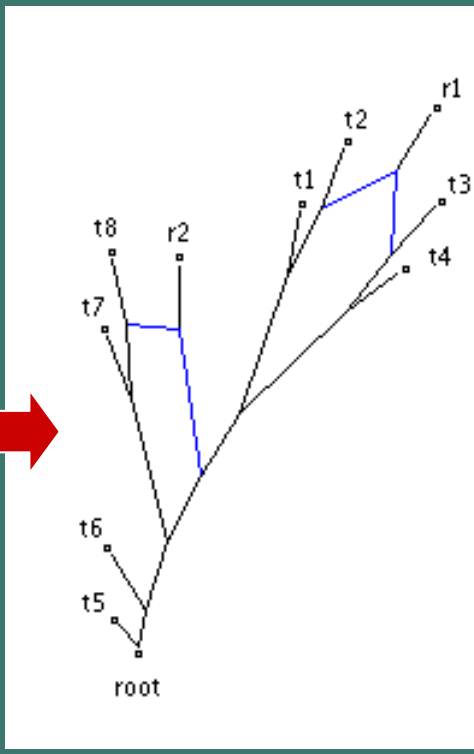




Two reticulations ⇒
four different gene trees

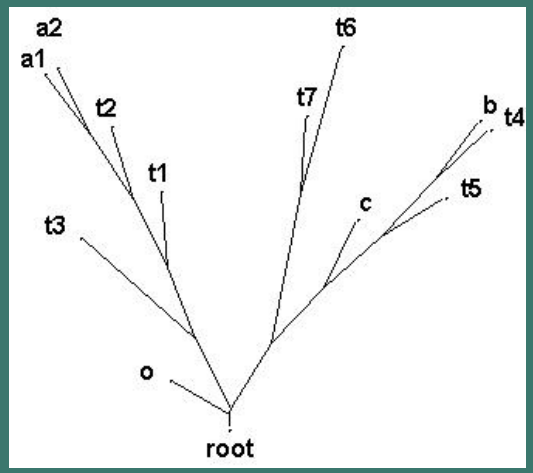


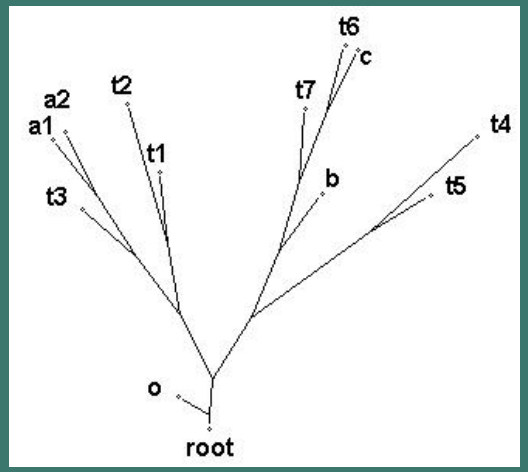
all splits



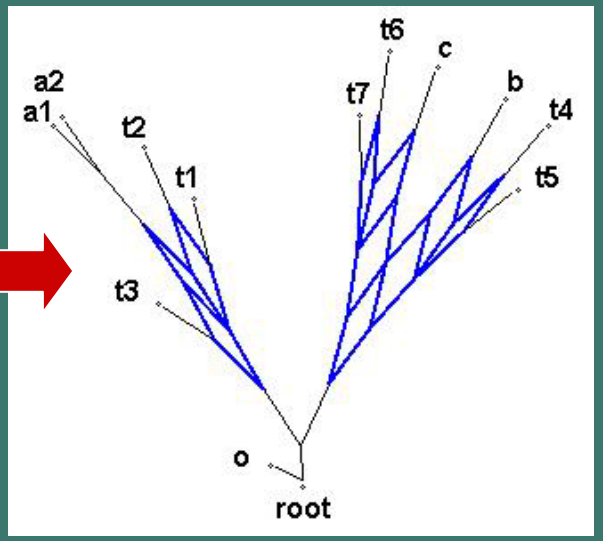
Reticulate network
that induces all
input trees

Multiple and Overlapping Reticulations

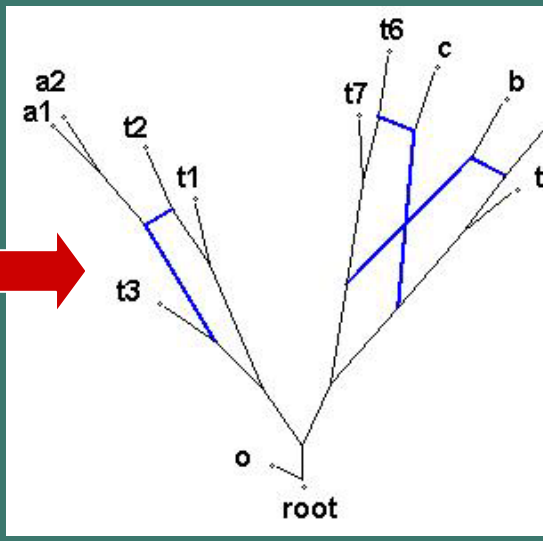




Input trees



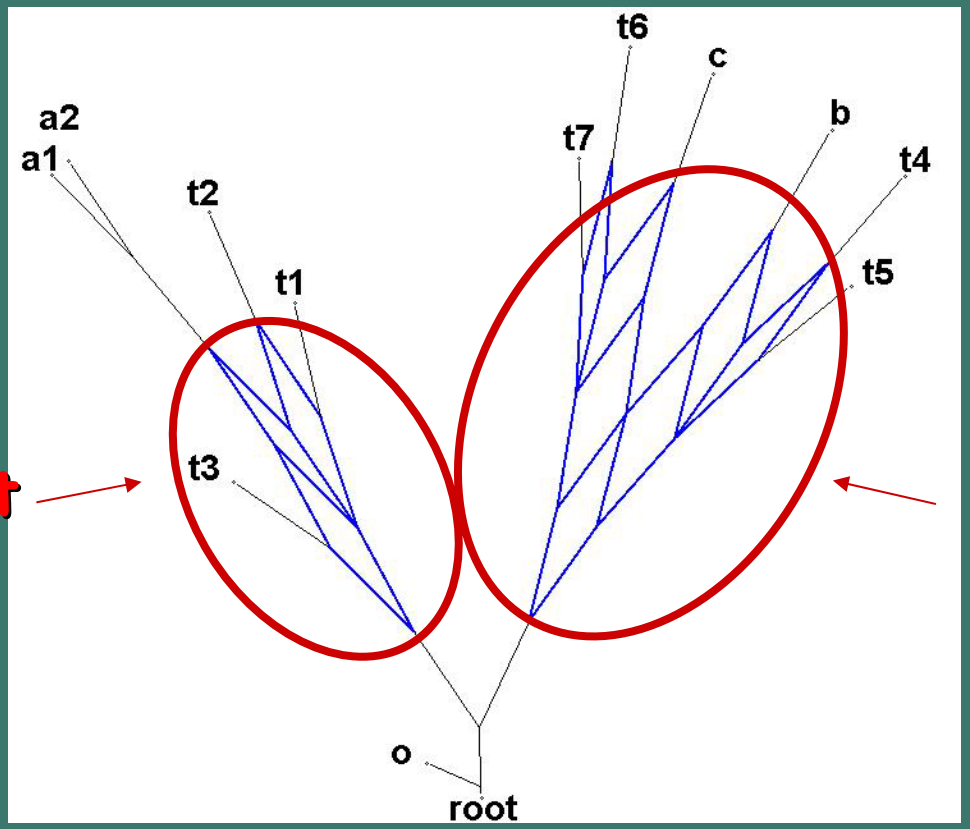
all splits



Reticulate network that induces all input trees

Decomposition Theorem

- Each incompatibility component can be considered independently:

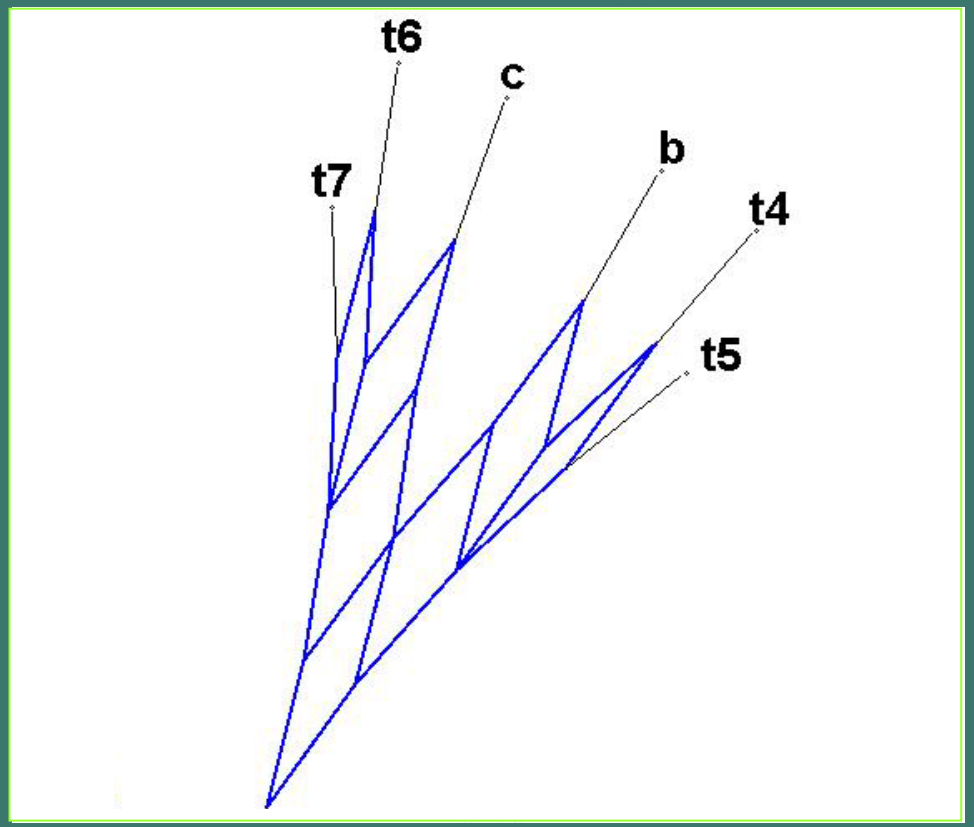


1. component

2. component

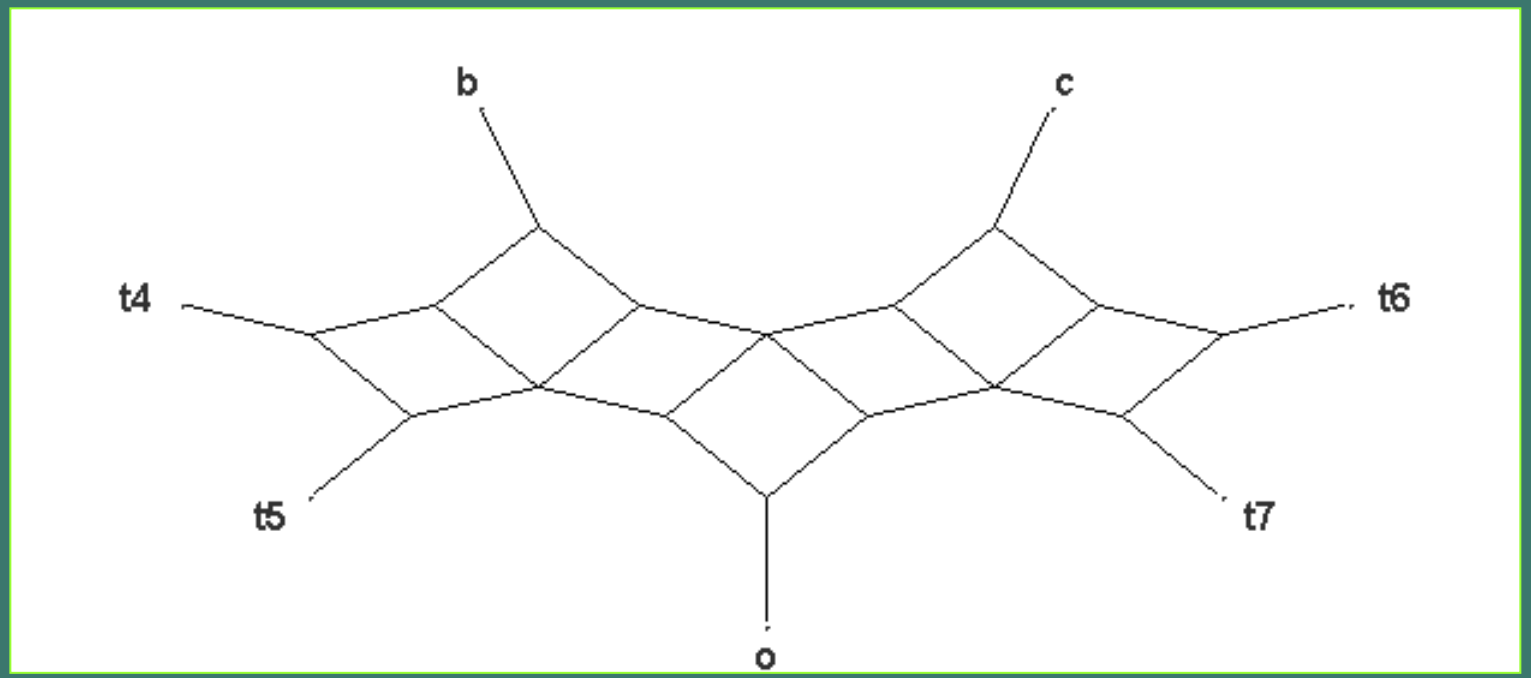
Decomposition Theorem

- Consider a component:



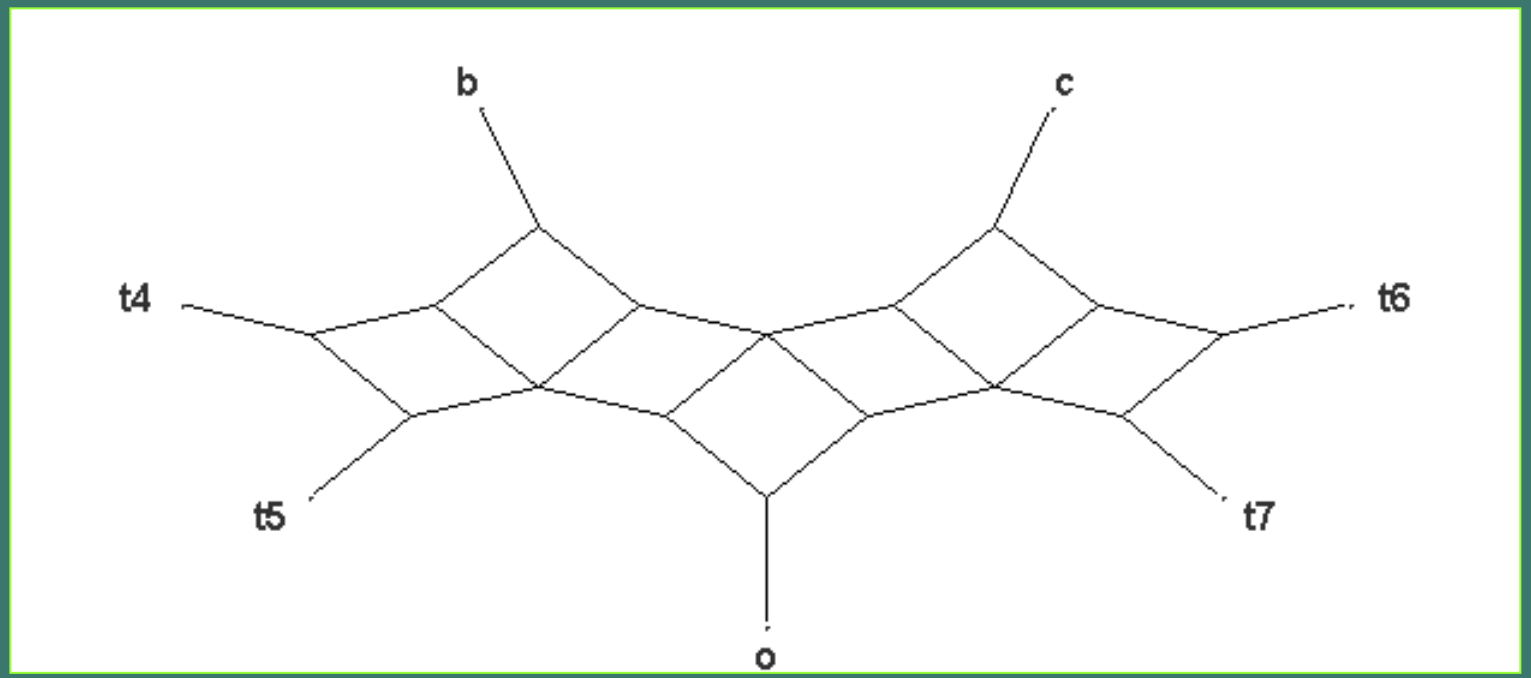
Algorithm

- Find decomposition $R \cup B$ as a set of reticulate taxa and backbone taxa



Algorithm

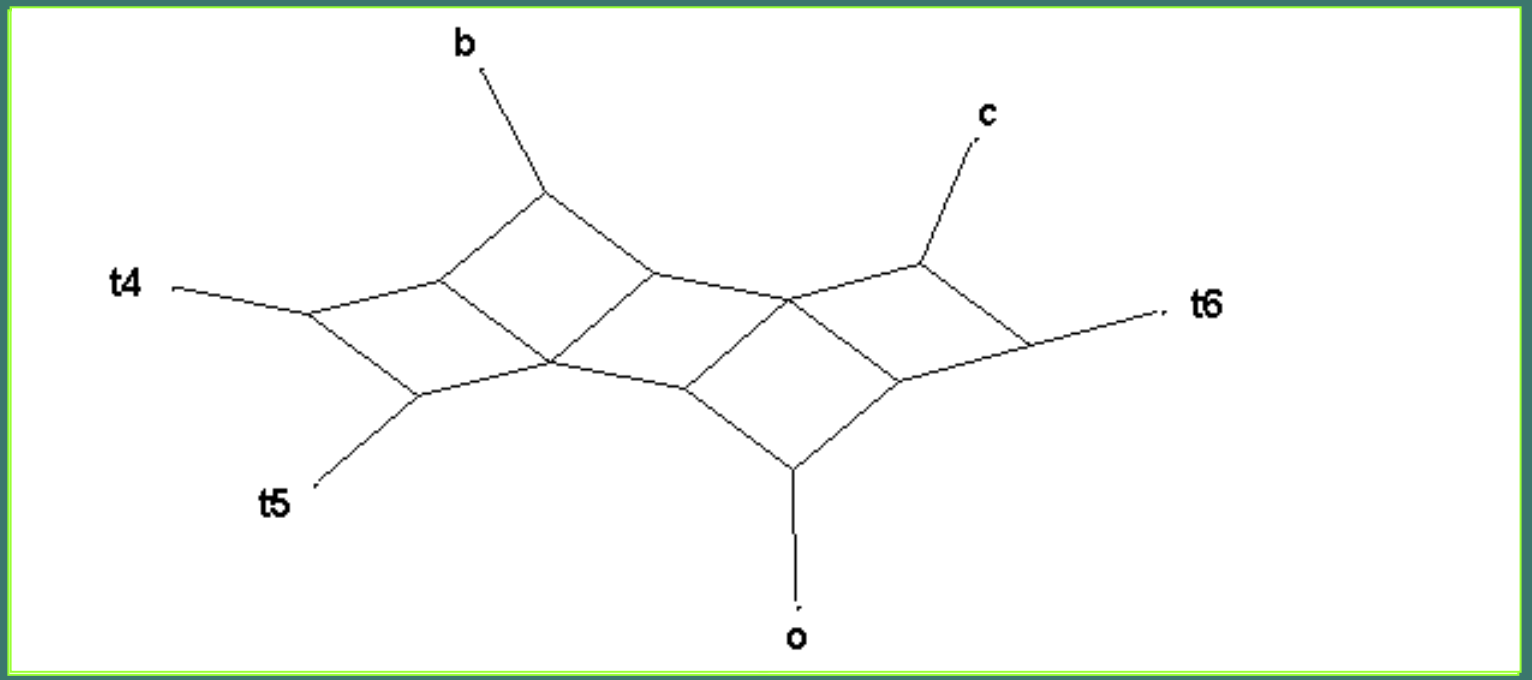
- Necessary condition: splits restricted to B must correspond to a tree



Algorithm

- Consider all choices for R of size **1**

  [Gusfield *et al.*, 2003, 2004]:

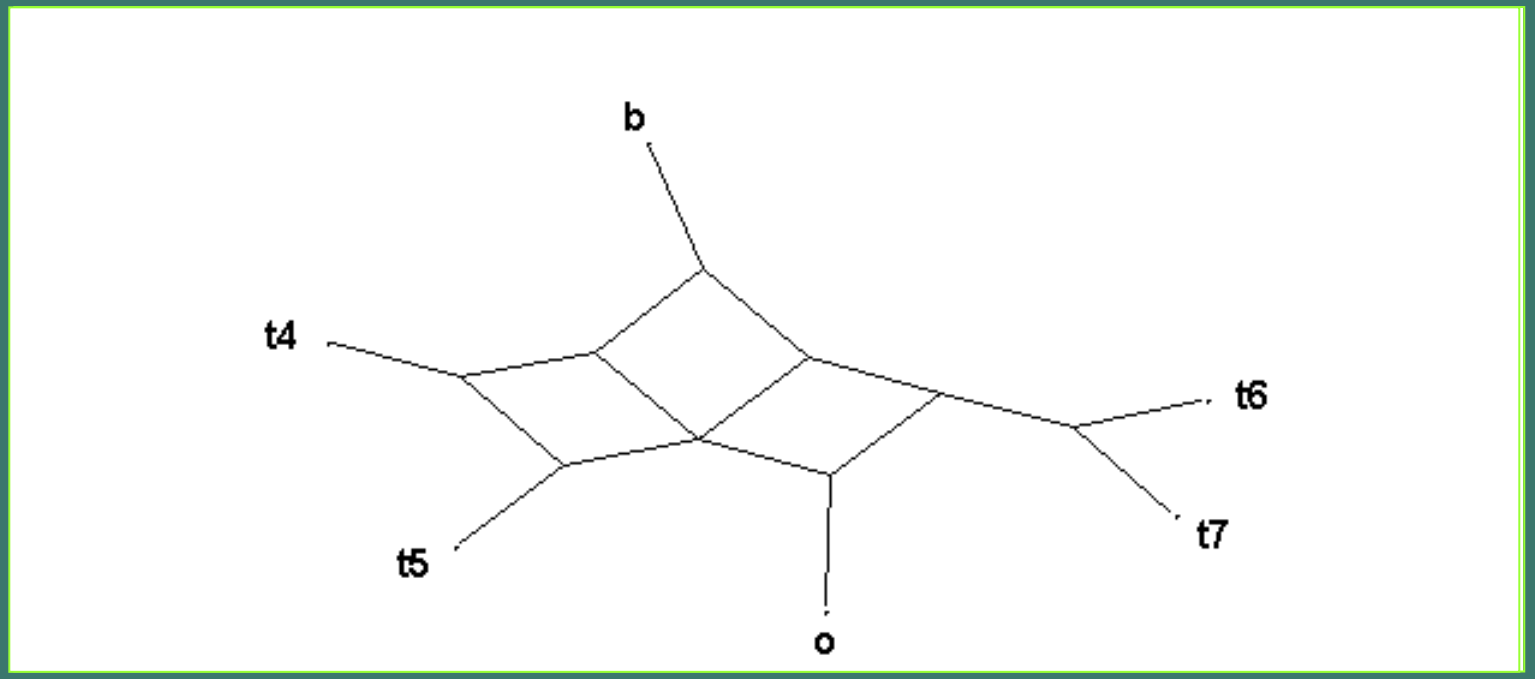


$R=\{t_7\}$ ☹...not a tree, R not good

Algorithm

- Consider all choices for R of size 1

  [Gusfield *et al.*, 2003, 2004]:



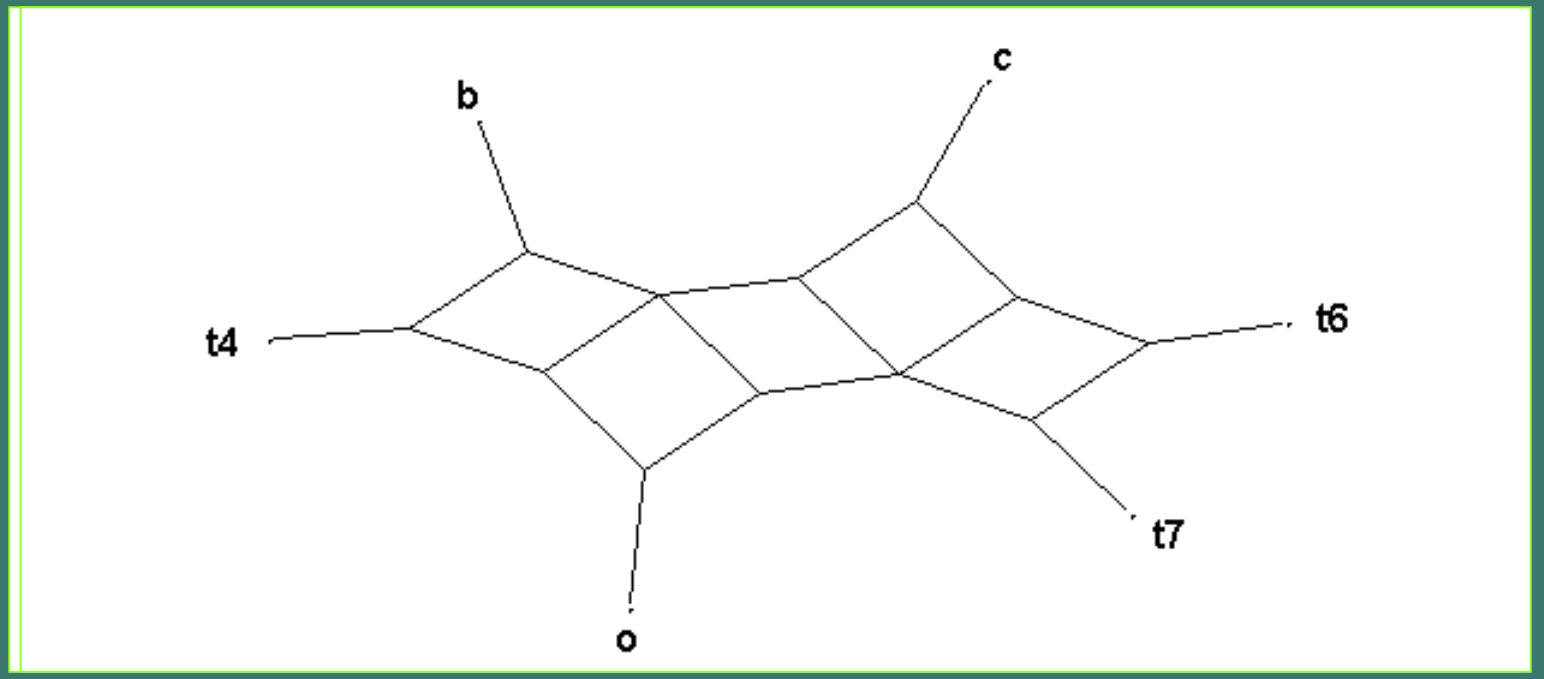
R={c}

☹...not a tree, R not good

Algorithm

- Consider all choices for R of size **1**

  [Gusfield *et al.*, 2003, 2004]:

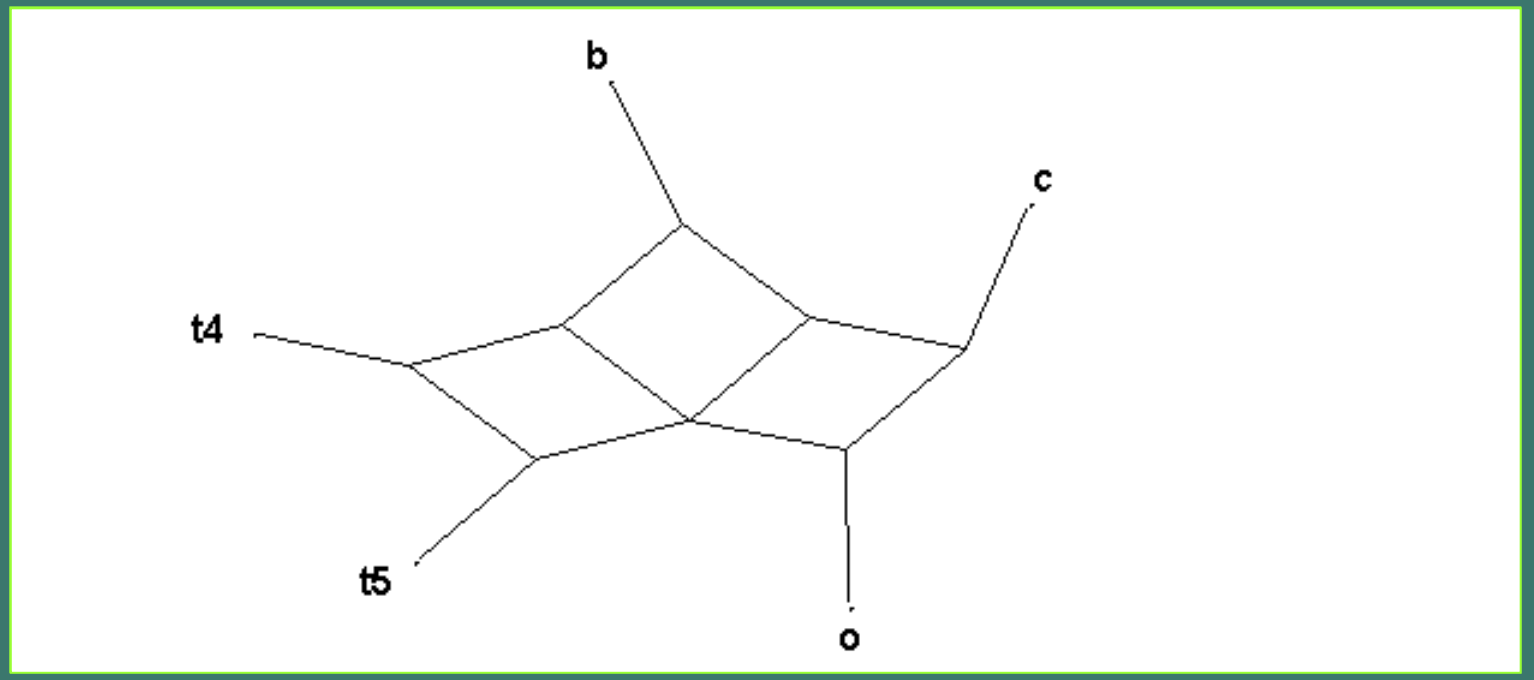


$R=\{t_5\}$ ☹...not a tree, R not good

Algorithm

- Consider all choices for R of size 2:

  [H., Kloepper, Lockhart and Steel, 2005]

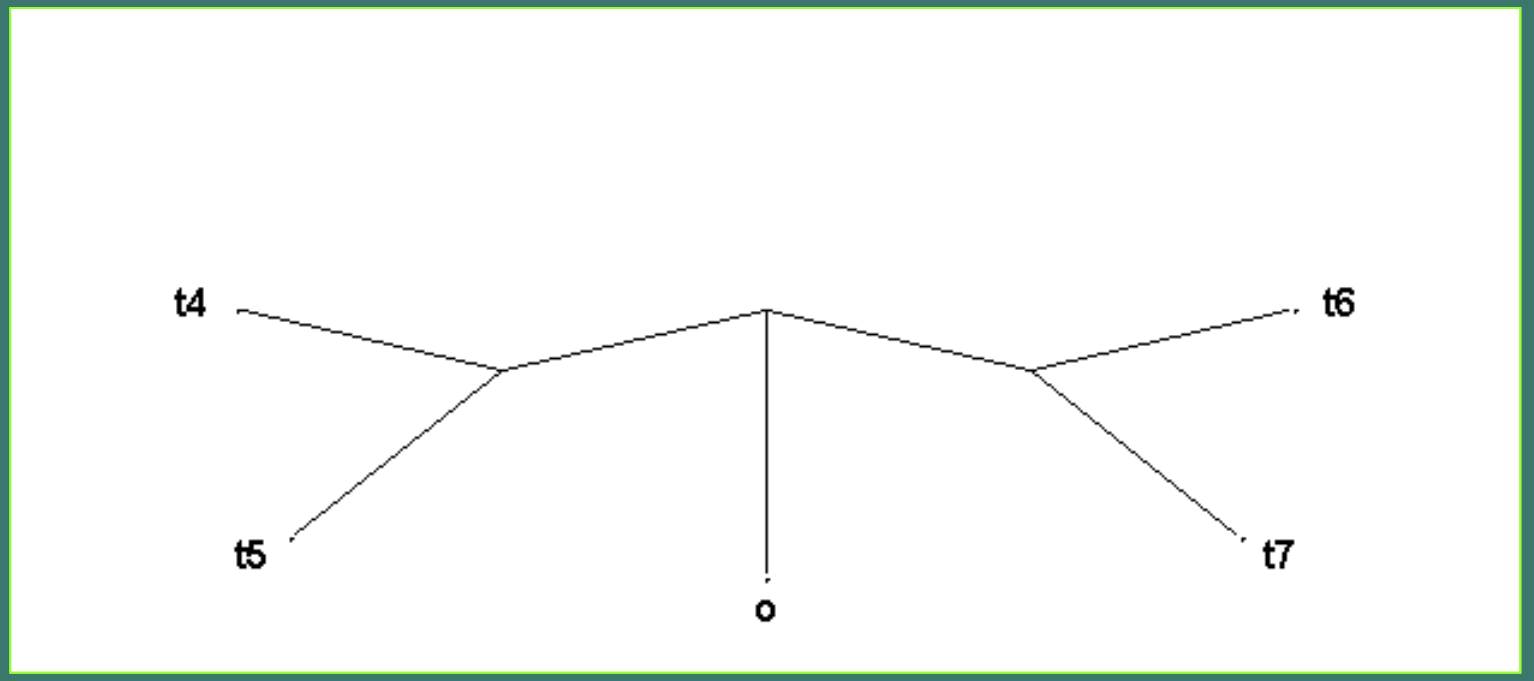


$R=\{t_6,t_7\}$ ☹...not a tree, R not good

Algorithm

- Consider all choices for R of size 2:

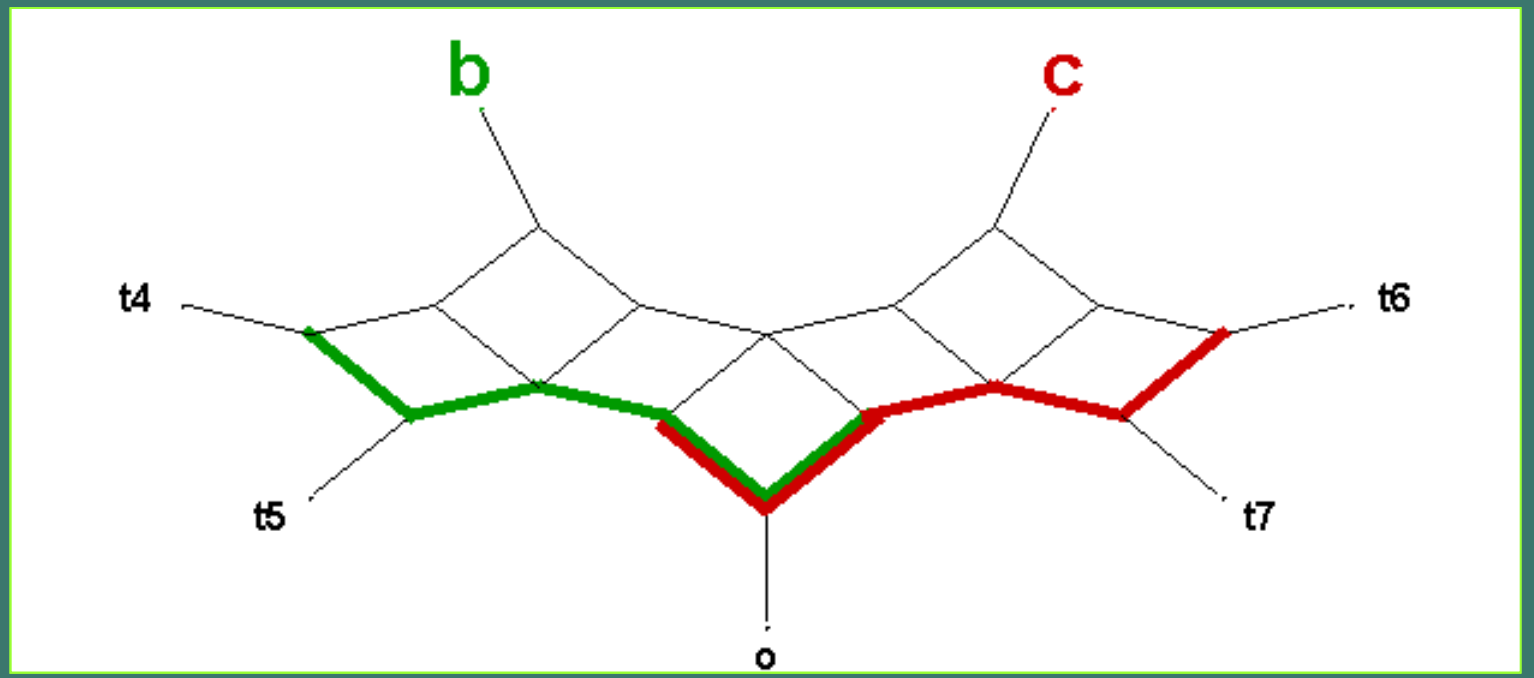
  [H., Kloepper, Lockhart and Steel, 2005]



R={b,c} ☺...is a **tree**, R is a candidate

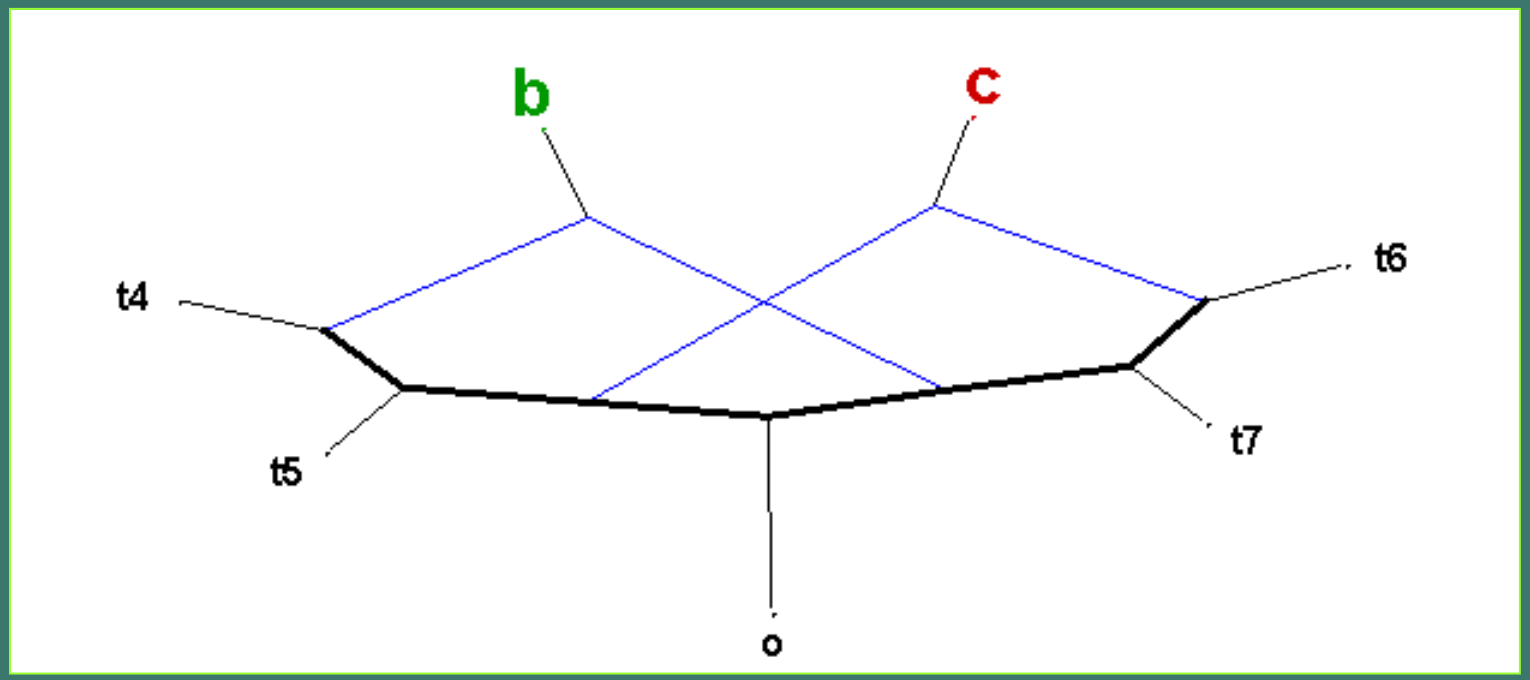
Check Candidate

- For R={b,c}, check that reticulation cycles overlap correctly along a path:



Network Construction

- Modify splits network to represent reticulations:

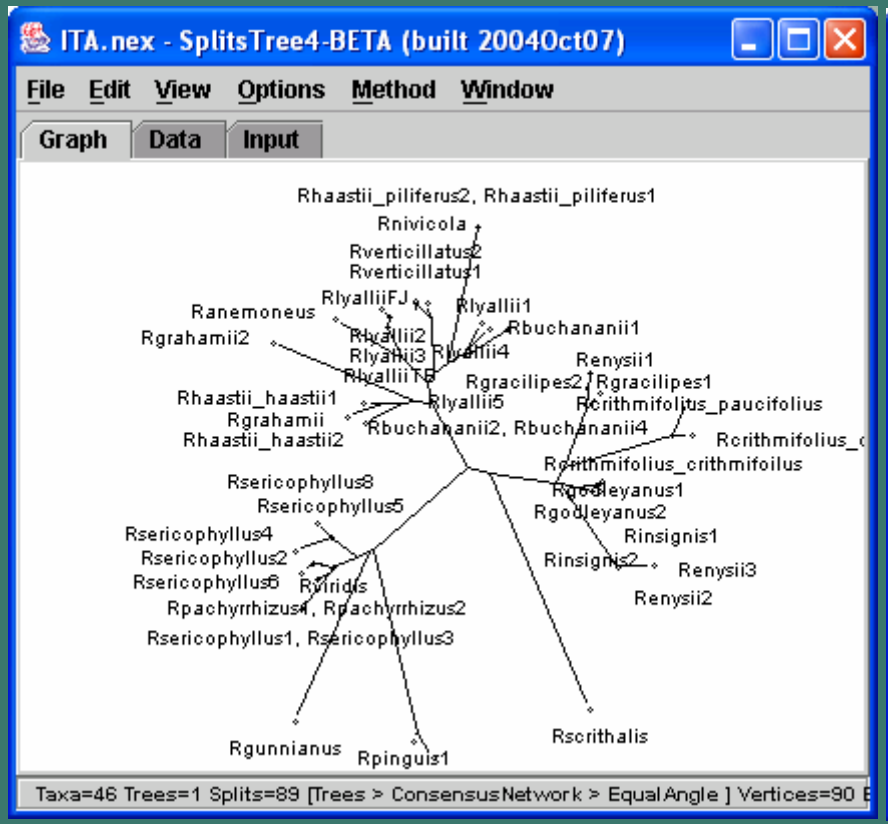


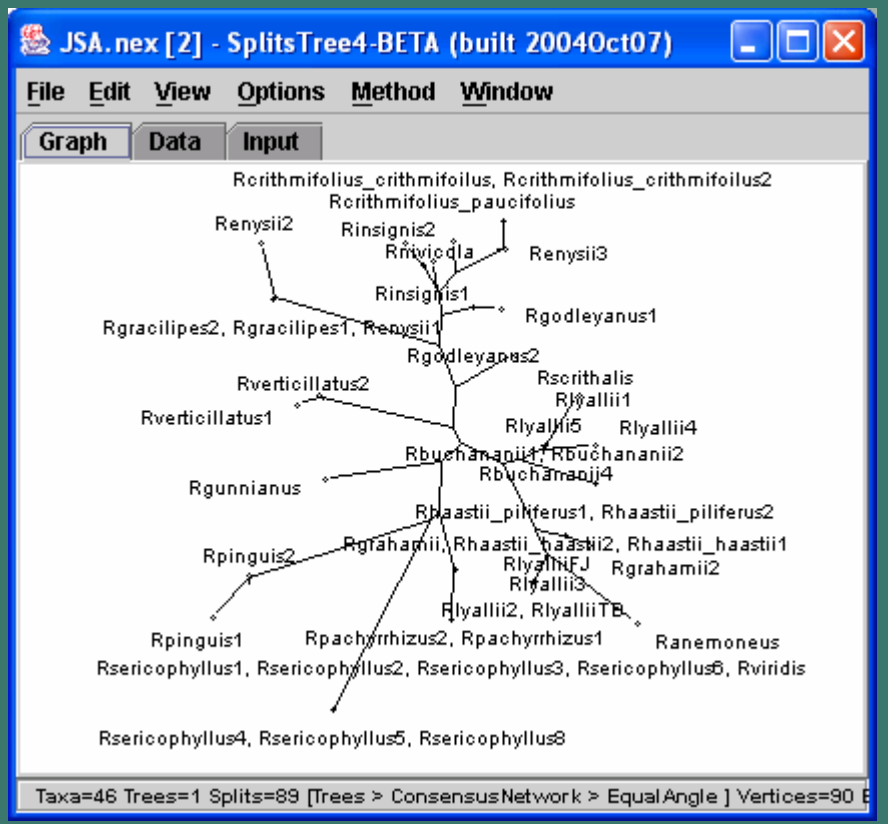
Splits-Based Algorithm

- Input:
  - Set of trees $\mathcal{T}$, not necessarily bifurcating, can be partial trees
  - Parameter k
- Output:
  - All reticulate networks N for which every incompatibility component can be explained by at most k overlapping reticulations
- Complexity: polynomial for fixed k

Application to Real Data

New Zealand *Ranunculus* (buttercup) species



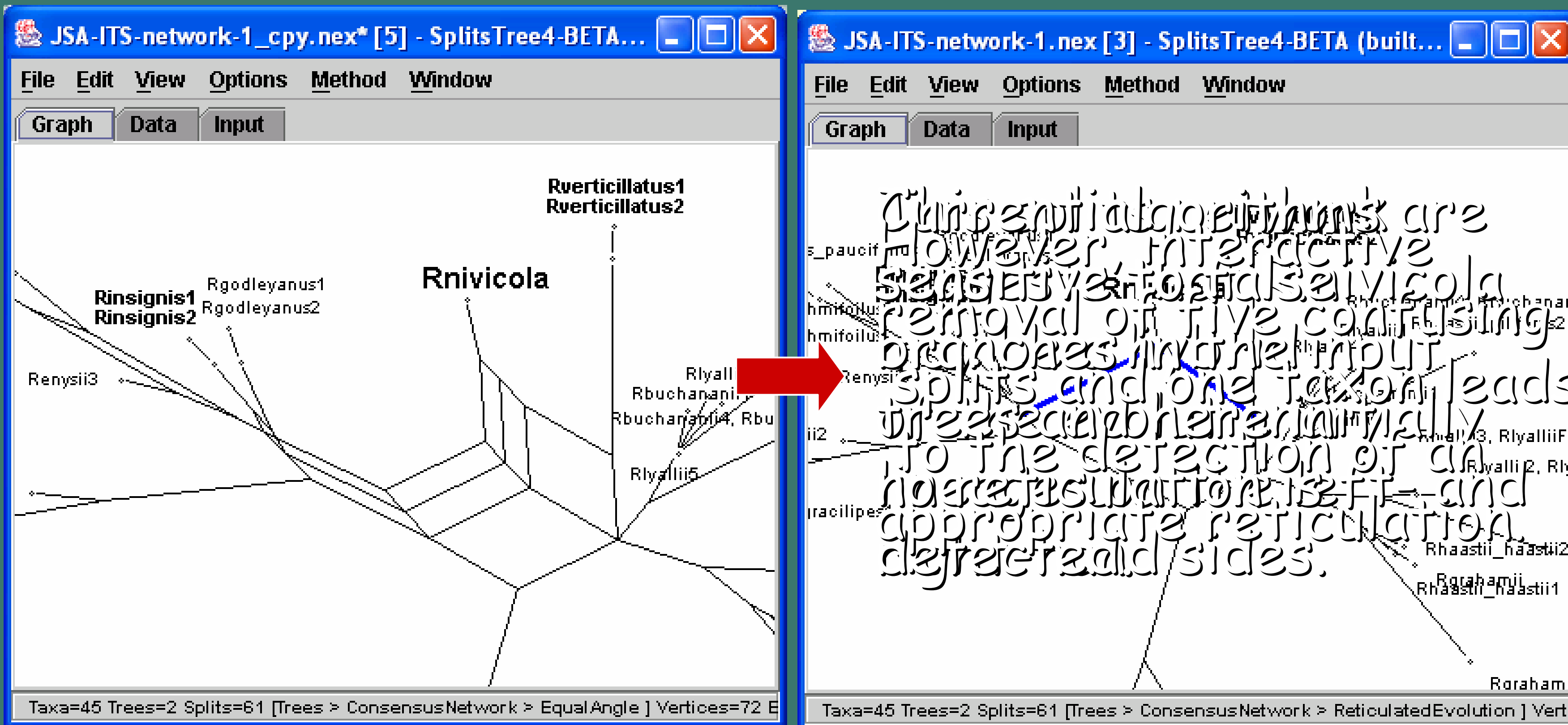


Nuclear ITS region

Chloroplast J_{SA} region

Application to Real Data

New Zealand *Ranunculus* (buttercup) species



Splits network for both genes

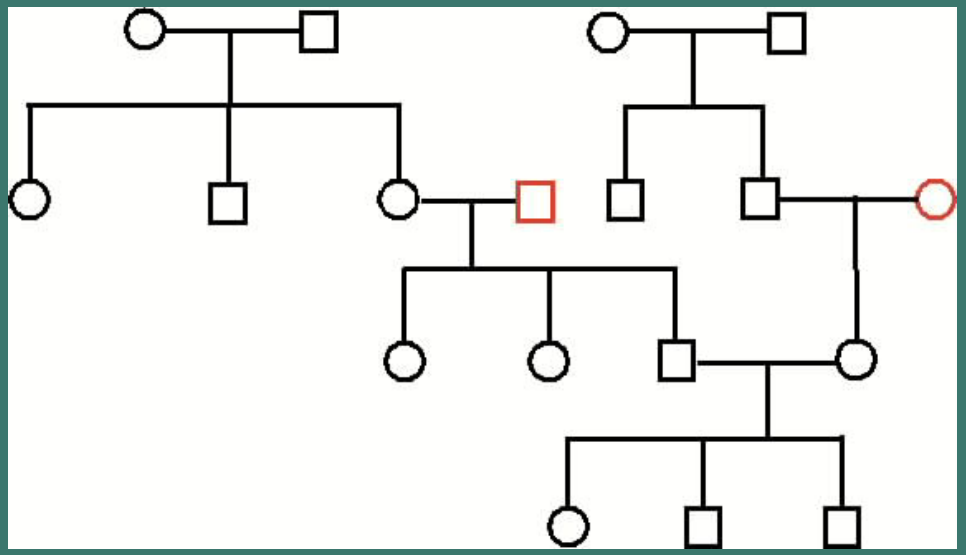
Reticulate network

Part V

1. Phylogenetic trees
2. Splits networks
3. Consensus networks
4. Hybridization and reticulate networks
5. Recombination networks

Recombination

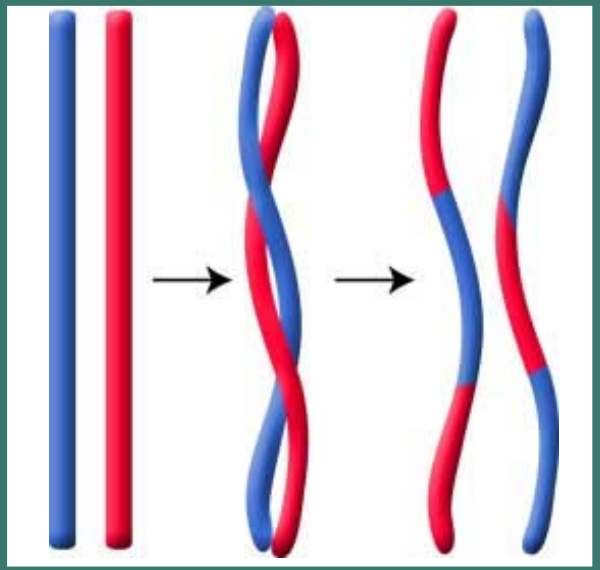
- Recombination is studied in population genetics [24, 20,16, 46, 47, 48] and there *ancestor recombination graphs* (ARGs) are used for statistical purposes.



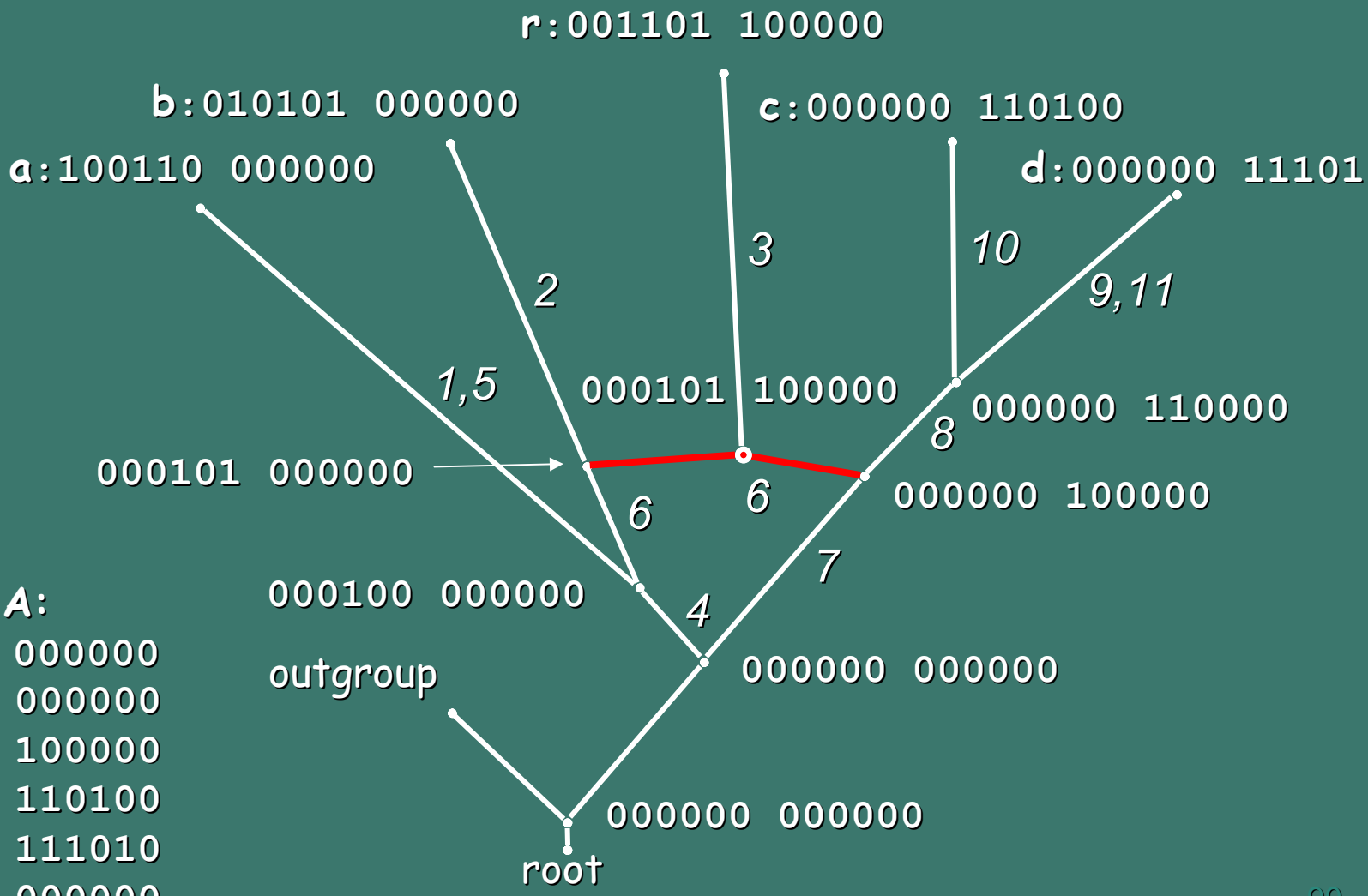
Chromosomal Recombination

- We will study the combinatorial aspects of chromosomal (meiotic) recombination and will consider *recombination networks* rather than ARGs.

- Simplifying assumptions:
  - all sequences have a common ancestor, and
  - any position can mutate at most once.



Example of a Recombination Network



Alignment A:

a:100110 000000
b:010101 000000
c:001101 100000
d:000000 110100
r:000000 111010
o:000000 000000

Recombination Network

Tree-based approach [Gusfield *et al.* 2003] for computing galled trees:

- For each component:
  - **Determine whether removing one taxon produces a perfect phylogeny**
  - If so, arrange taxa in gall
  - Return description of network

Recombination Network

Splits-based approach [Huson & Kloepper 2005] for computing overlapping networks:

- Determine a reticulate network as described above.
- Compute the labeling of nodes and edges.

Recombination Network

- [Lungso, Sun and Hein, to appear in WABI 2005]: Branch and bound approach to computing unrestricted recombination network

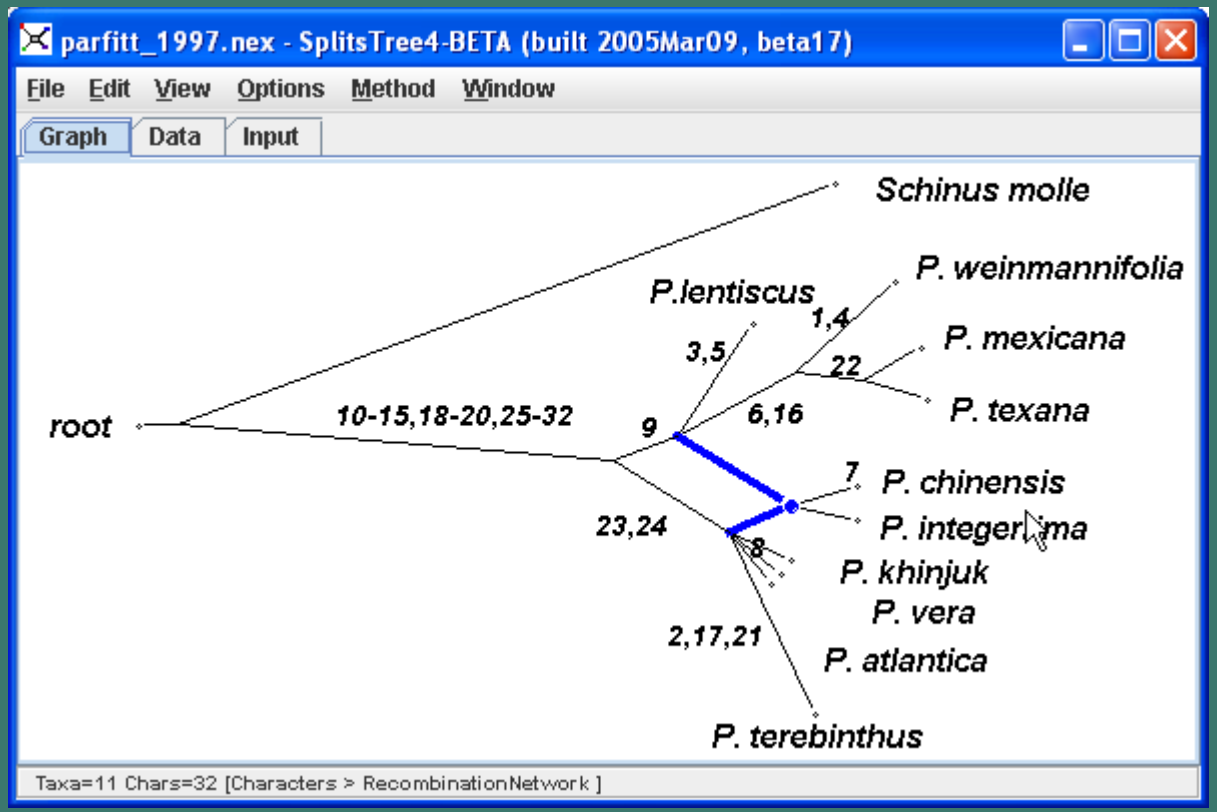
Example 1, Data

- Input: Presence (0) or absence (1) of a given restriction site in a 3.2kb region of variable chloroplast DNA in Pistacia [Parfitt & Badenes 1997]:

P.lentiscus	01110010100000001110000000010000
P.weinmannifolia	11001110100000010111000000010000
P.chinensis	01011000100000001110011000010000
P.integerrima	01011010100000001110011000010000
P.terebinthus	00011010000000001111101100010000
P.atlantica	01011011000000001110011000010000
P.mexicana	01011110100000010111010000010000
P.texana	01011110100000010111010000010000
P.khinjuk	01011010000000001110011000010000
P.vera	01011010000000001110011000010000
Schinus molle	01011010011111100000000011101111

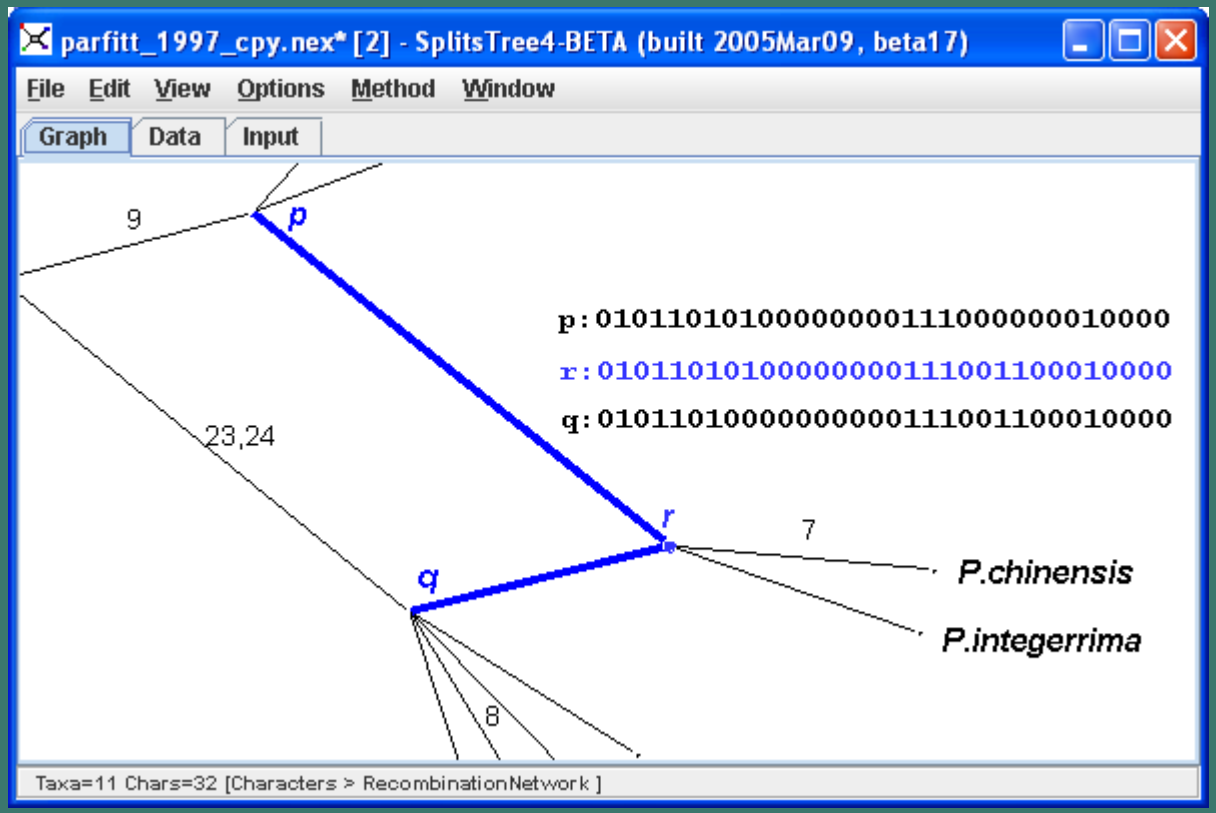
Example 1, Recombination Network

- Load this data in to SplitsTree4 and select RecombinationNetwork to obtain:



Example 1, Single Crossover

- Combinatorically, this can be explained using only one single-crossover recombination:



However: recombination of chloroplast is unlikely

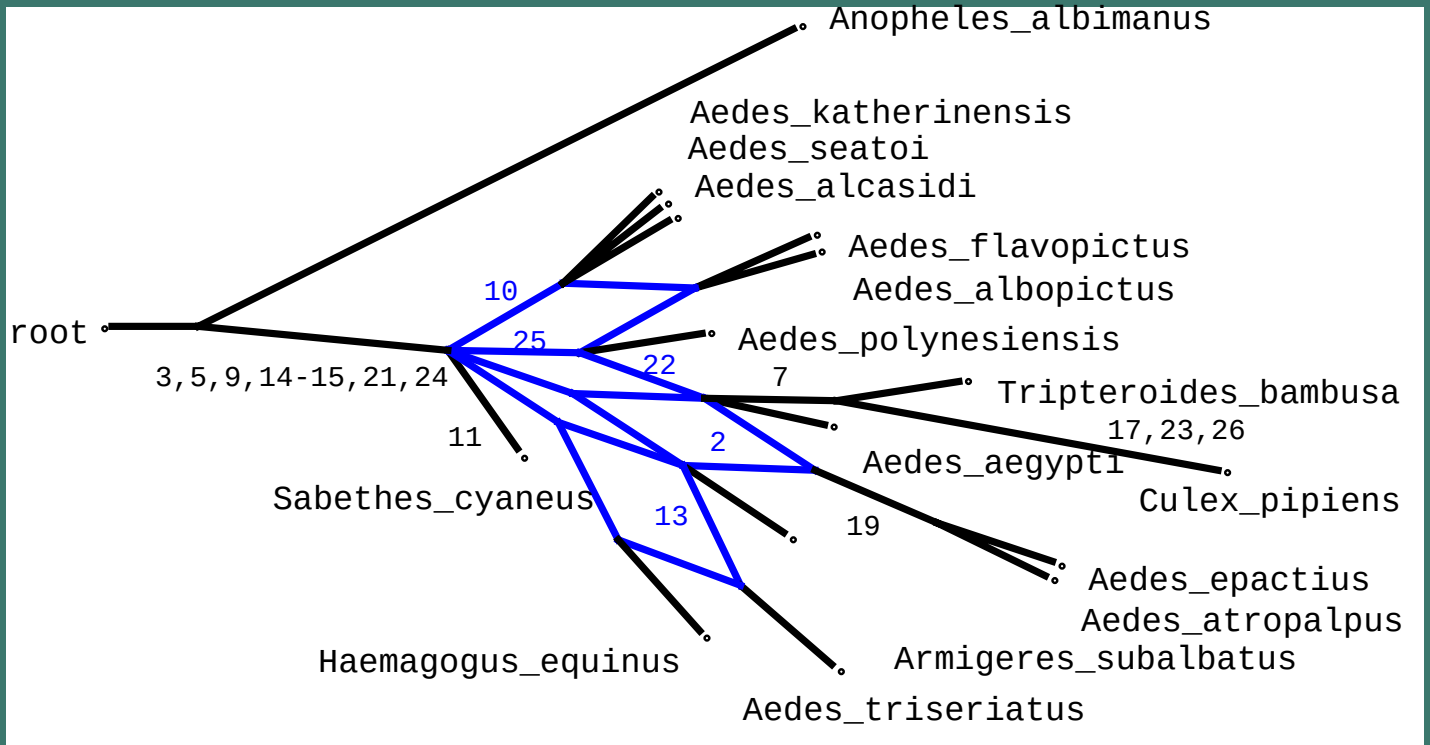
Example 2, Data

- Input: Restriction maps of the rDNA cistron (length ≈ 10kb) of twelve species of mosquitoes using eight 6bp recognition restriction enzymes [Kumar *et al*, 1998]:

Aedes albopictus	11110101010100010101010010
Aedes aegypti	11110101000100010101000010
Aedes seatoi	11110101010100010101010000
Aedes avopictus	11110101010100010101010010
Aedes alcasidi	11110101010100010101010000
Aedes katherinensis	11110101010100010101010000
Aedes polynesiensis	11110101000100010101010010
Aedes triseriatus	10110101000110010101000000
Aedes atropalpus	10110101000100010111000010
Aedes epactius	10110101000100010111000010
Haemagogus equinus	10110101000110010101010000
Armigeres subalbatus	10110101000100010101000000
Culex pipiens	11110111000100011101001011
Tripteroides bambusa	11110111000100010101000010
Sabethes cyaneus	11110101001100010101010000
Anopheles albimanus	11011101100101110101110100

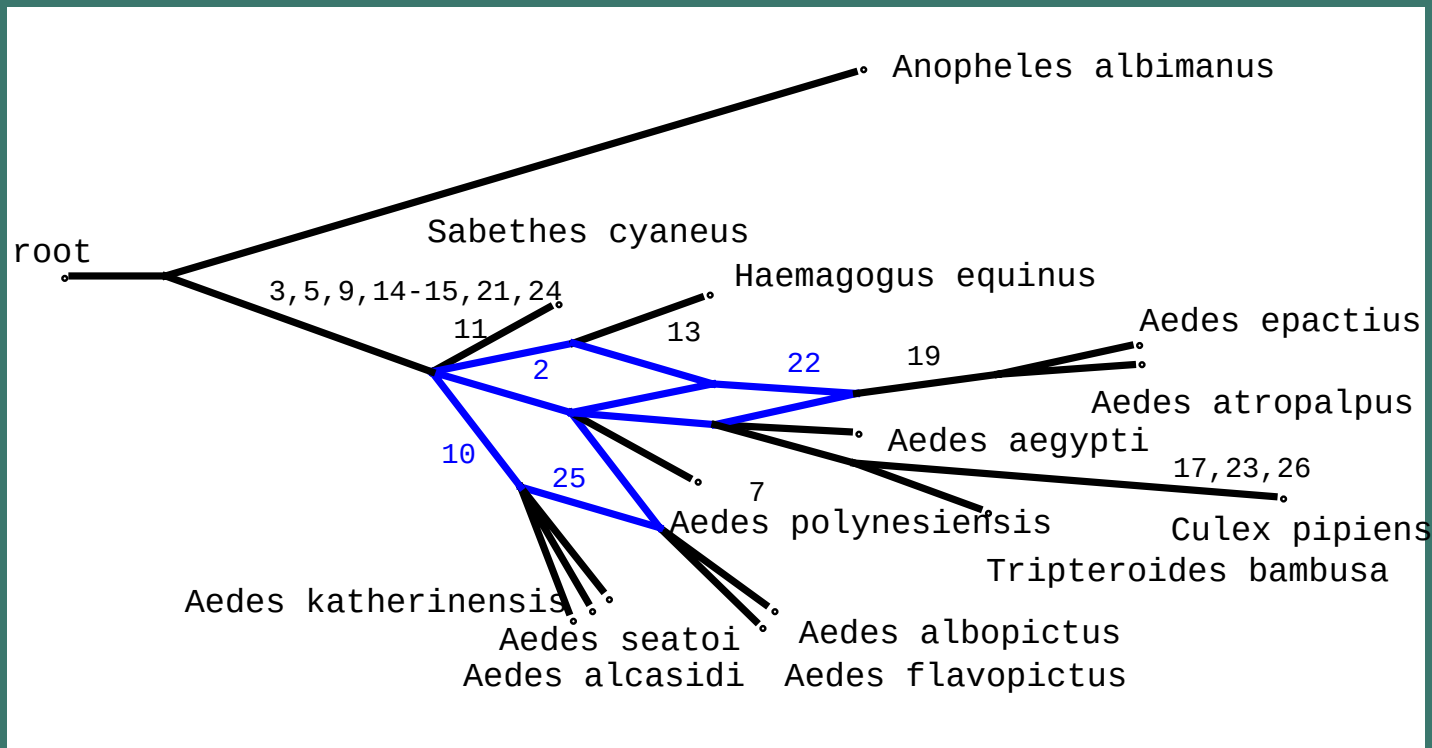
Example 2, Median Network

- This data set was analyzed using different tree-reconstruction methods with inconclusive results. The associated splits network (or *median network* [Bandelt *et al*, 1995] in this context), with edges labeled by the corresponding mutations:



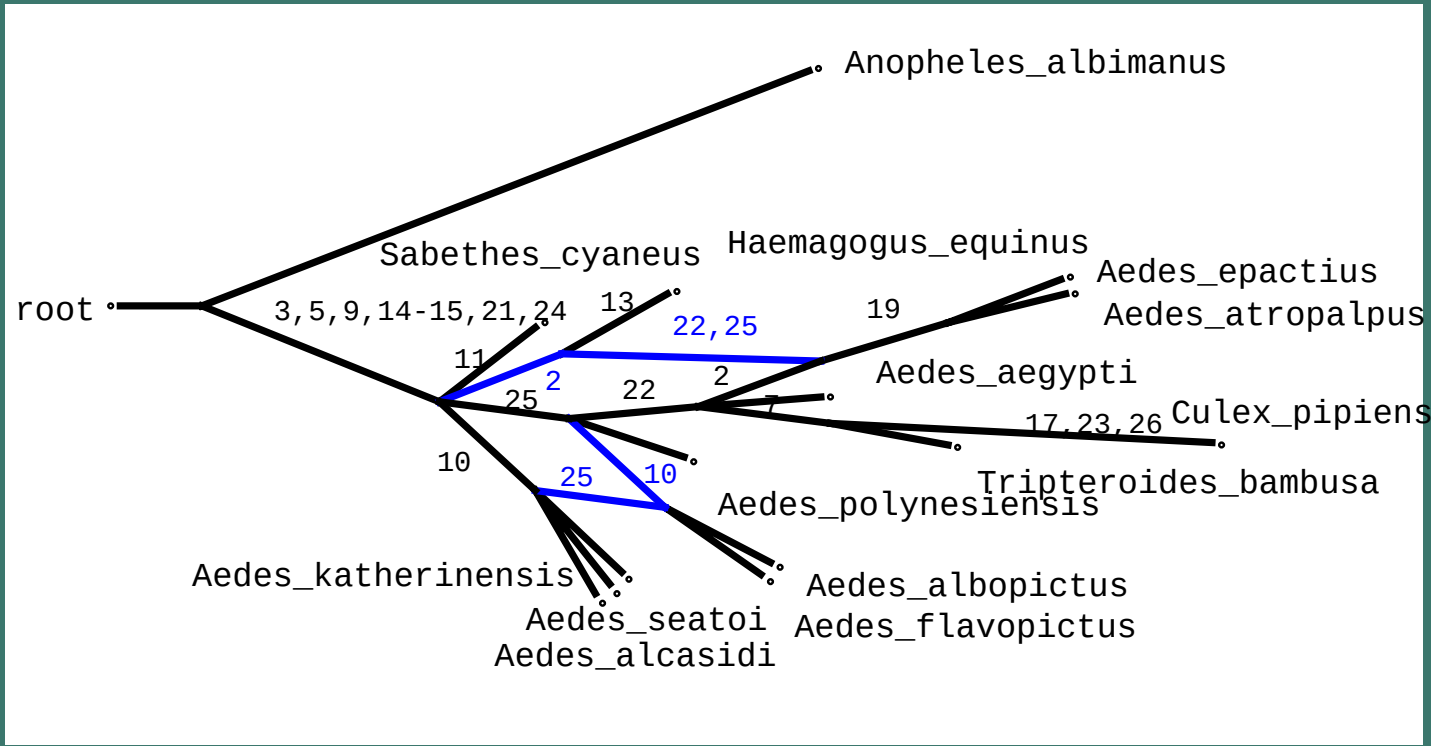
Example 2, Subset

- Recombination scenarios based on the complete data set look unconvincing. However, trial-and-error removal of two taxa *Aedes triseriatus* and *Armigeres subalbatus* gives rise to a simpler splits network:



Example 2, Recombination Network

- A possible recombination scenario is given by:



- Here, *Haemagogus equinus* appears to arise by a single-crossover recombination, and a second such recombination leads to *A.albopictus* and *A.avopictus.*

Example 3, Data

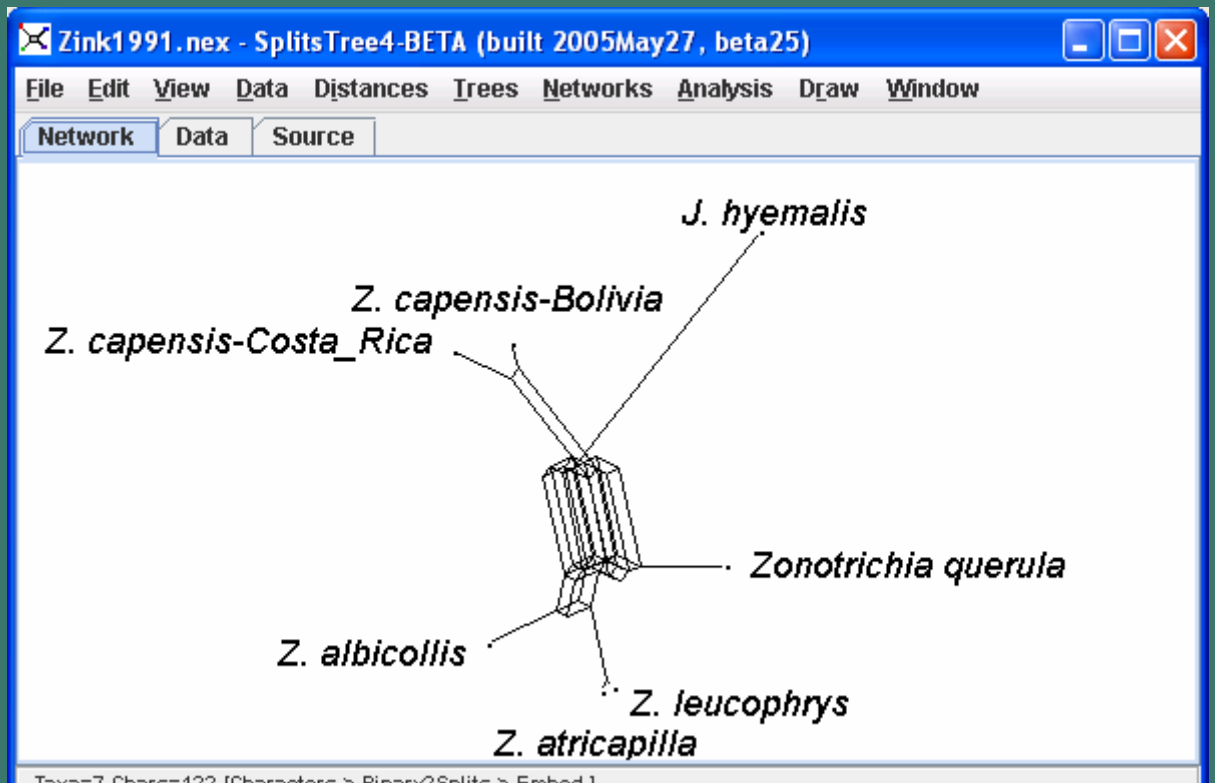
- 19 restriction endonucleases were used to analyze patterns of cleavage site variation in the mtDNA of Zonotrichia.
- 7 taxa, 122 characters
- [Zink *et al*, 1991]

```
notrichia_querula'      11100011111111000111100111111111100001111000111010101100011111001111111100010011111001111110111100110111
_atricapilla'           11100011111110000110000111111111000001111100111000101110011111001111111100010011111101111110111110110110
_leucophrys'            11100011111110000110000111111111000001111100111000101110011111001111111100010011111101111110111110110110
_albicollis'            11100011111110000110000101111111010000111010111000101111011111101111111100100111111001111100111100110110
_capensis--Bolivia'     01110011100100001000001111111000110110000011101010111101111100110111100101101111100011111001110110111010
_capensis--Costa_Rica'  11100111100101101000001111111000110110000011100010111101111100110111100101111111100011011001110110111010
_hyemalis'              11101011100100011000110011111000000111000111101111110011100101010101110011001111000001100101110011011010
```

- However: recombination of mtDNA unlikely

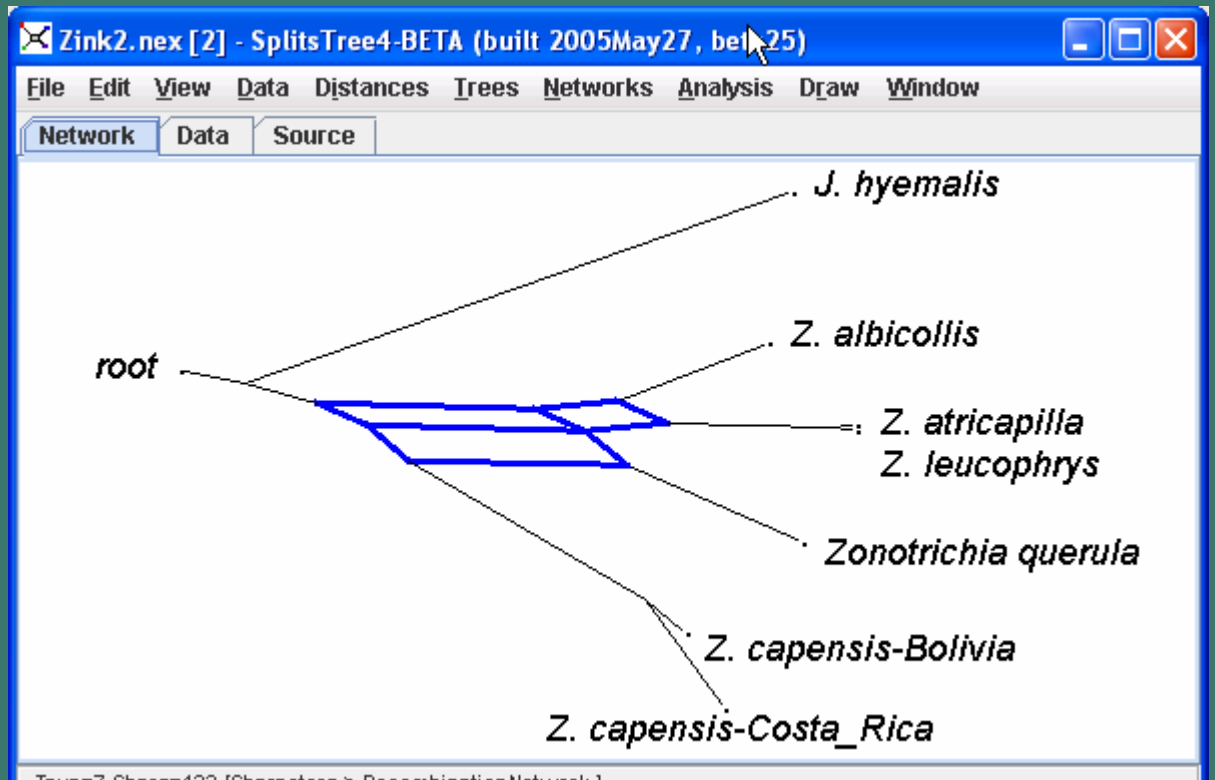
Example 3, Median Network

- The unrooted splits network for a dataset of restriction sites in the mtDNA of *Zonotrichia*:



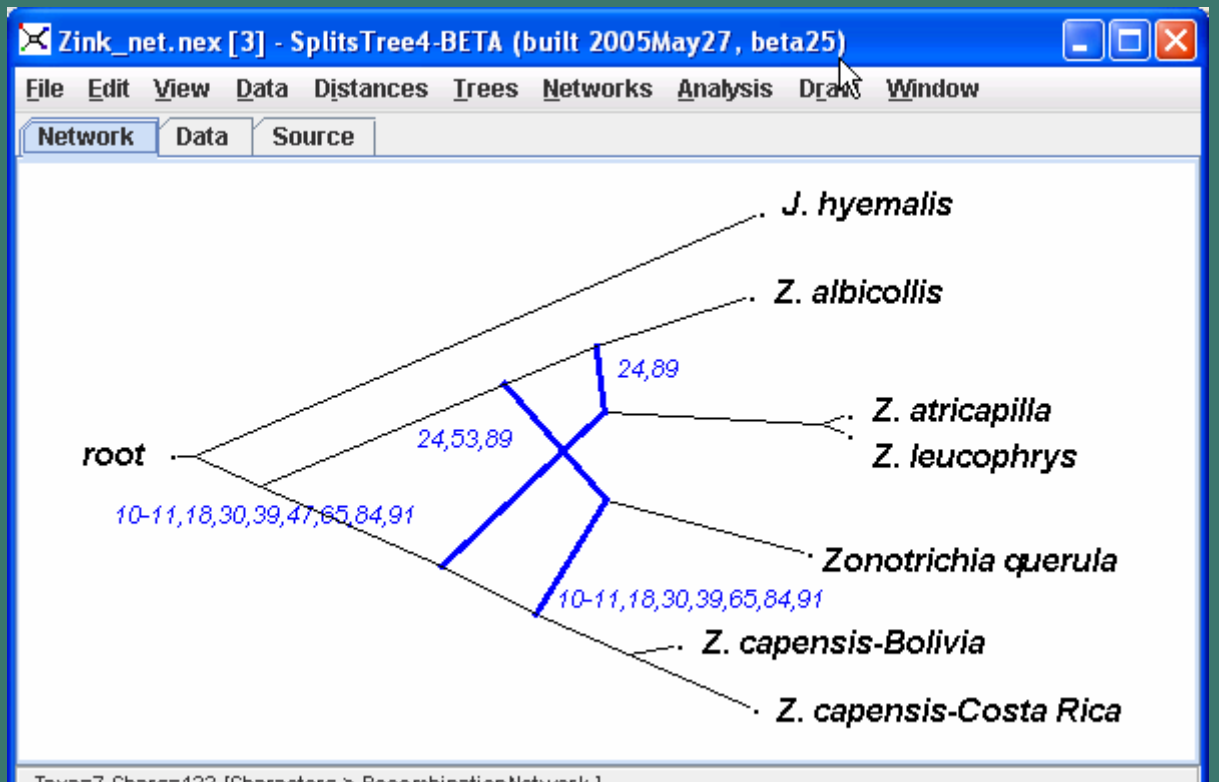
Example 3, Significant Differences

- The rooted splits network for the same data set, but suppressing all splits that are only supported by one site in the data:



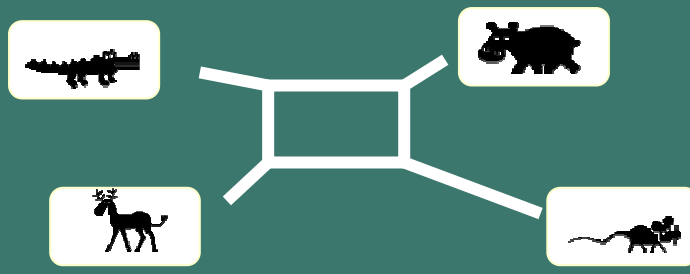
Recombination Network

- Possible recombination scenario involving two non-independent reticulations:



Summary

- Incompatible signals in gene trees can be usefully displayed using splits networks
- A reticulate network may be extracted by combinatorial analysis of individual components



- Implementations of many tree and network methods are available in **SplitsTree4**

3rd RECOMB Comparative Genomics Satellite Workshop

Trinity College Dublin

18th-20th September, 2005

Bringing together researchers from molecular biology, computer science and mathematics to discuss the latest developments in comparative genomics. This is the third workshop in this highly successful series.

Scientific organising committee

Aoife McLysaght, Trinity College Dublin
Daniel Huson, Tübingen University, Germany
Jens Lagergren, Stockholm Bioinformatics Centre and KTH
David Sankoff, University of Ottawa

http://www.gen.tcd.ie/recomb_cg