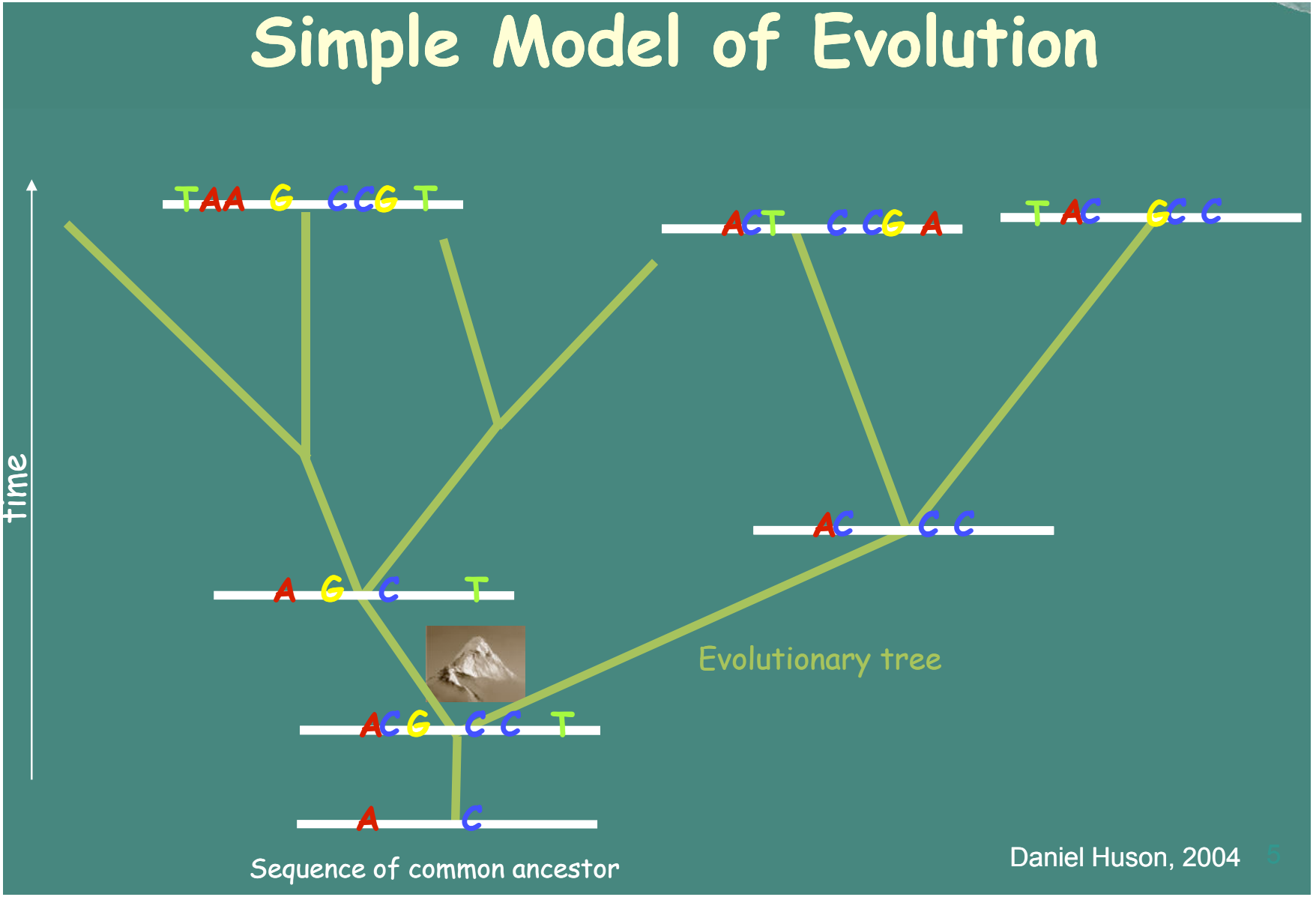
M2 - PHYL

Phylogenetic networks

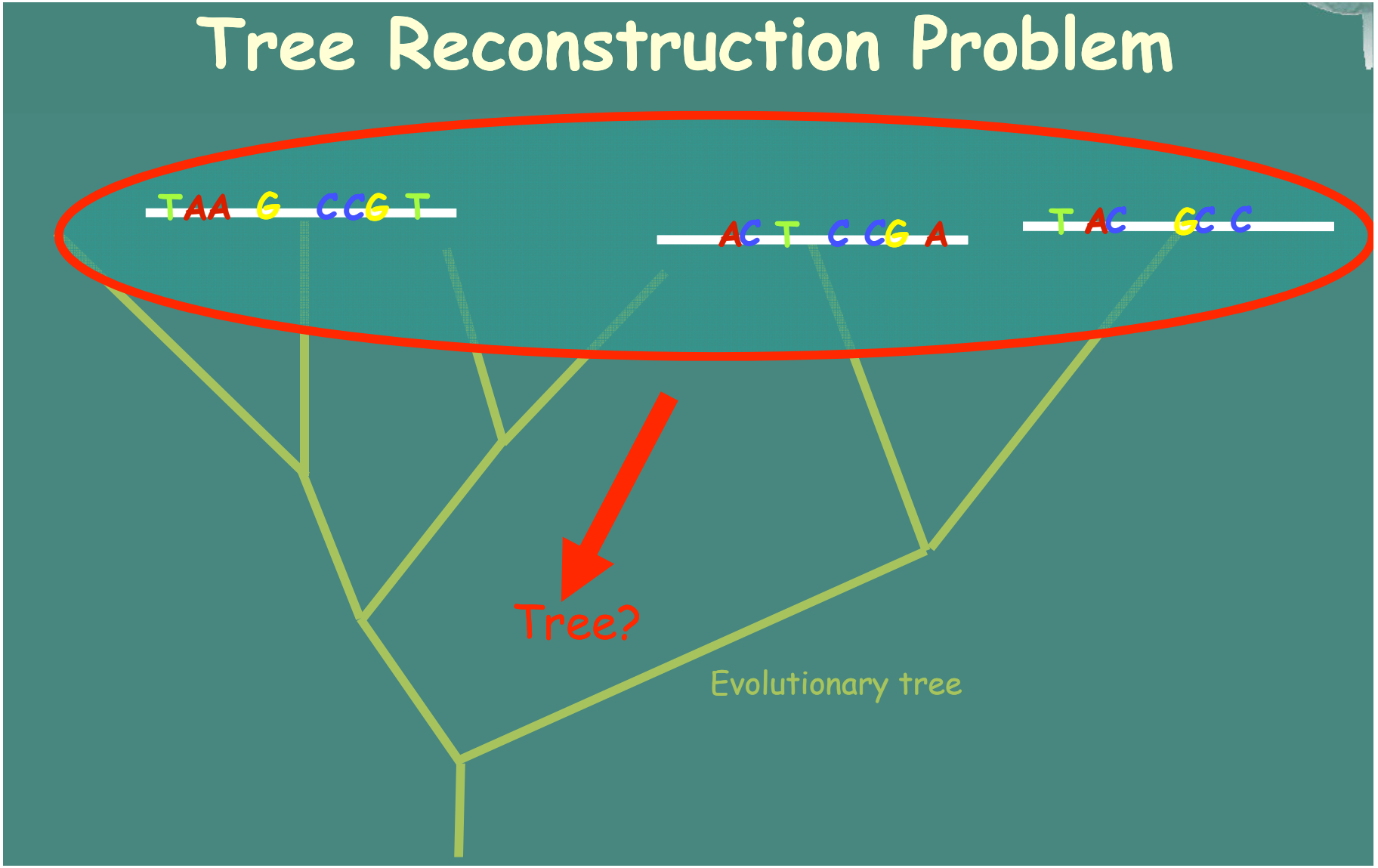
Alessandra Carbone
Université Pierre et Marie Curie

Suppose to know the tree….



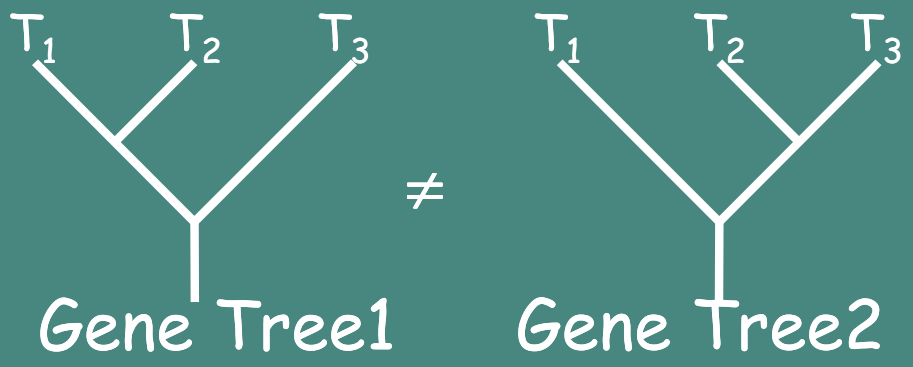
Sequences pick up mutations going along the branches of the tree

Tree Reconstruction Problem



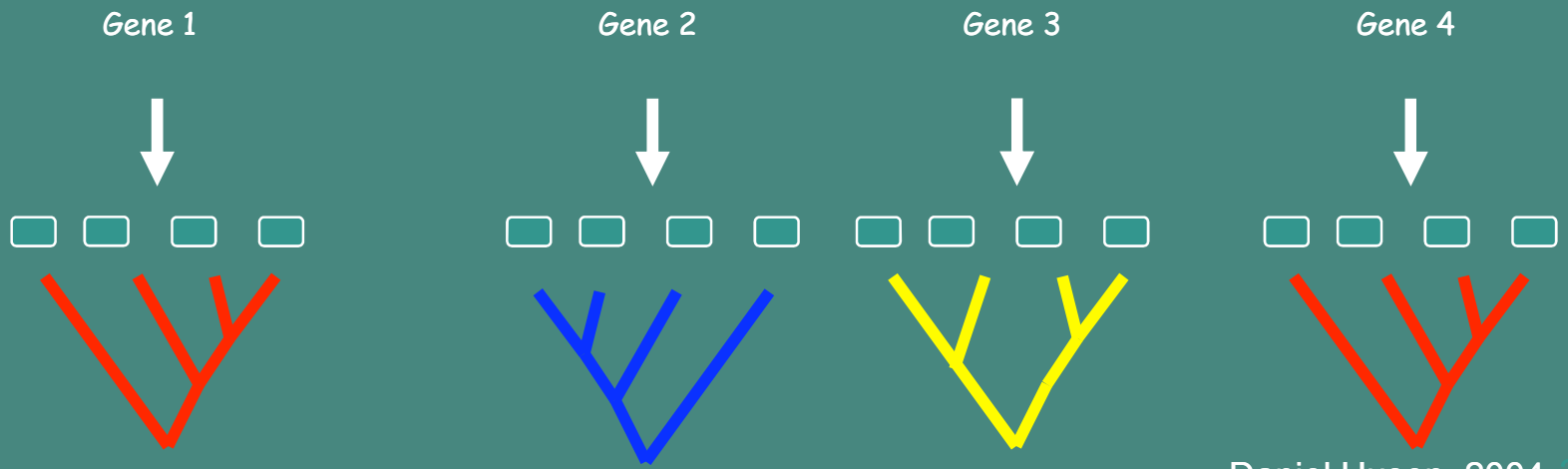
Fact: Gene Trees Disagree

- Ask two different genes what the phylogeny of a set of species is and you will get two different answers

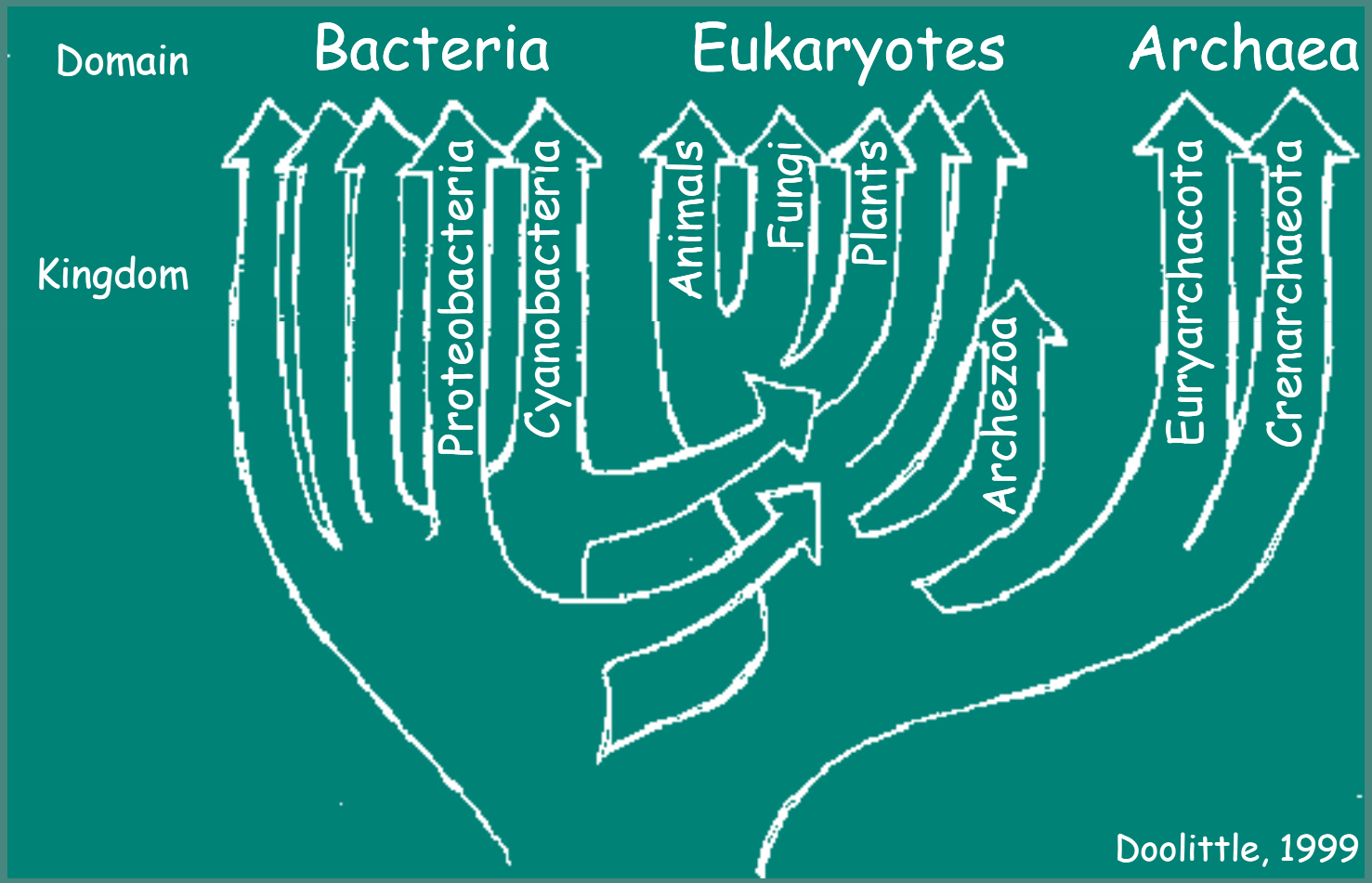


Gene Trees vs Species Trees

Differing gene trees give rise to "mosaic sequences"

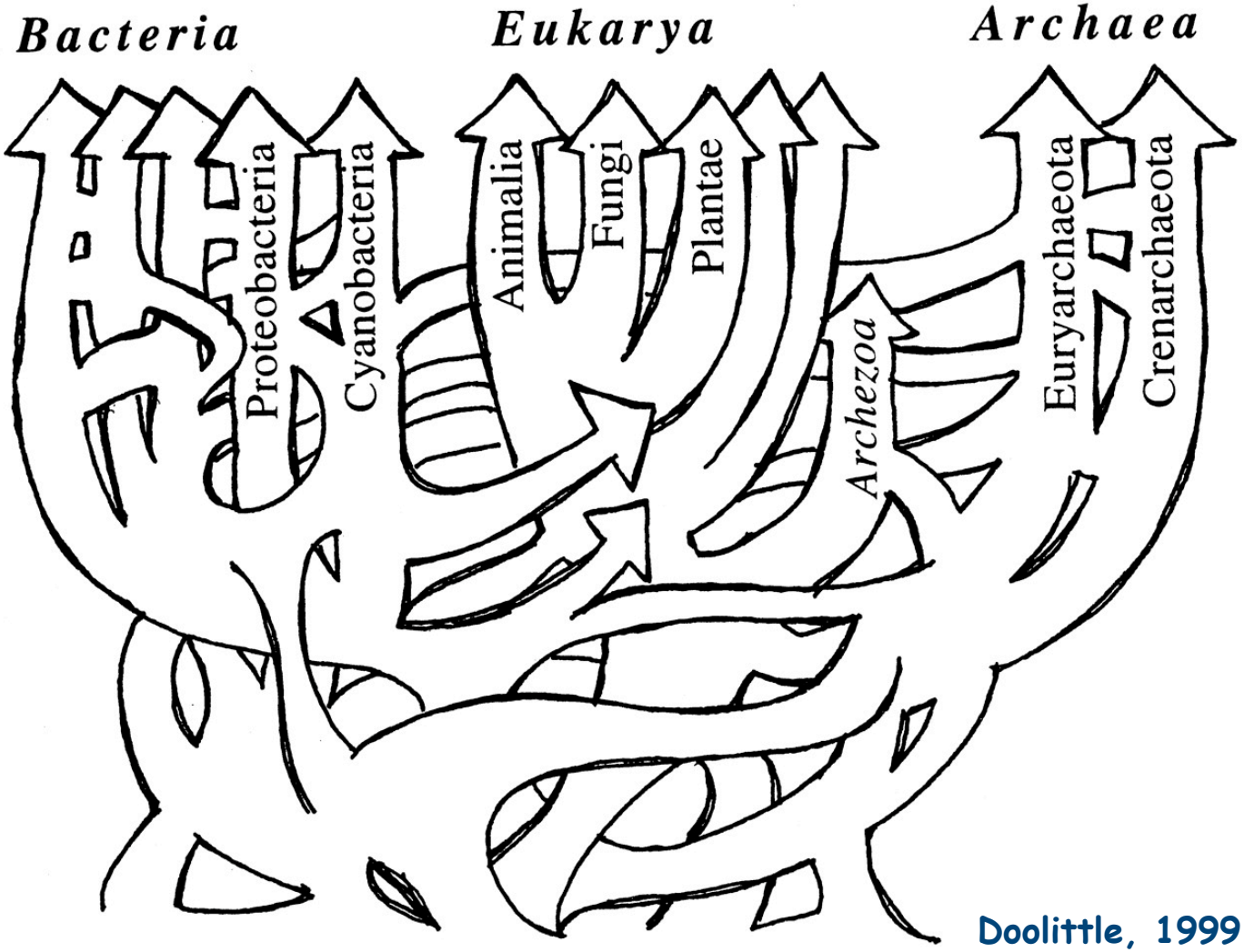


Current Classification of Life



Daniel Huson, 2004

Network of Life…



Doolittle, 1999

Definition of **phylogenetic network**

any graph that represents evolutionary relationships between a set of taxa that label some of its nodes (usually leaves)

taxa are represented by nodes

evolutionary relationships are represented by edges

DAG Directed Acyclic Graph

Definition of **phylogenetic network**

any graph that represents evolutionary relationships between a set of taxa that label some of its nodes (usually leaves)

taxa are represented by nodes

evolutionary relationships are represented by edges

DAG Directed Acyclic Graph

Different types of networks:
Unrooted or rooted
Abstract or explicit
Softwired or hardwired
Combined or transfer view

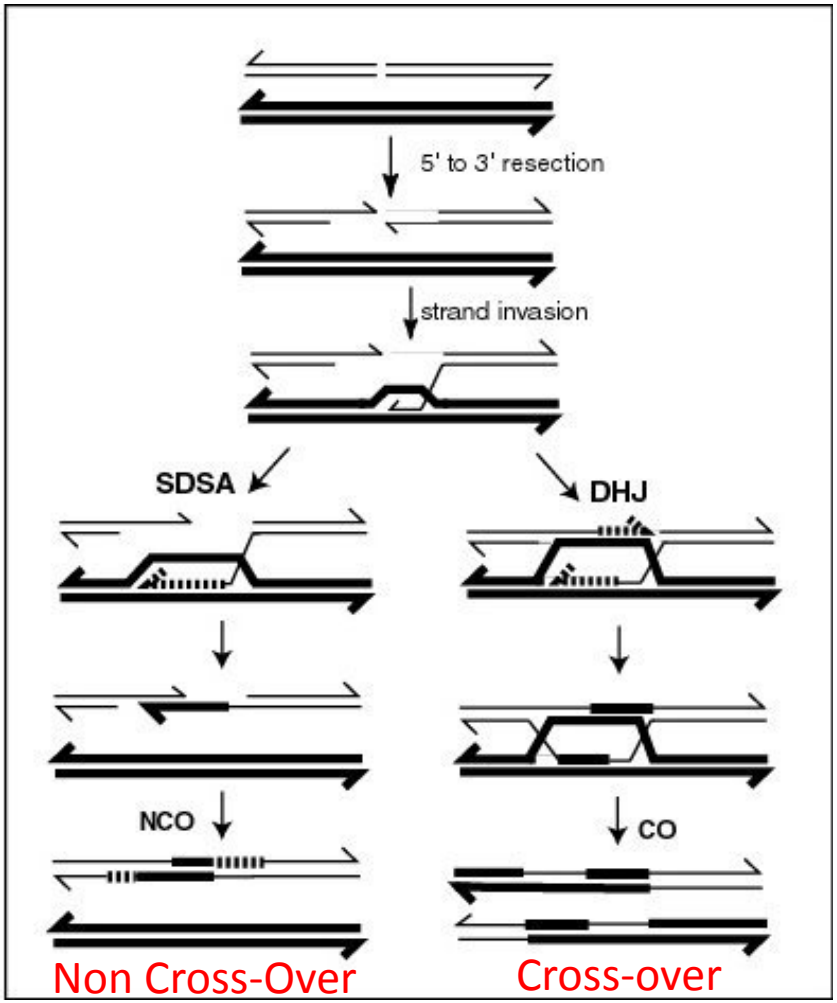
Underlying idea: to describe the evolution of life in a way that explicitly includes reticulate events.

Mechanisms of reticulate evolution

Hybridization (plants, frogs, fish)
Horizontal gene transfer (prokaryotes)
Recombination (populations)

Recombination

Meiotic recombination



Chromosomal cross-over

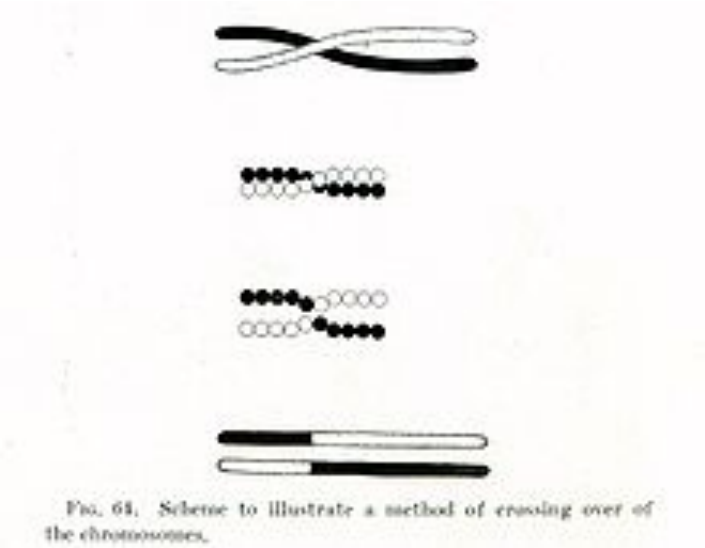
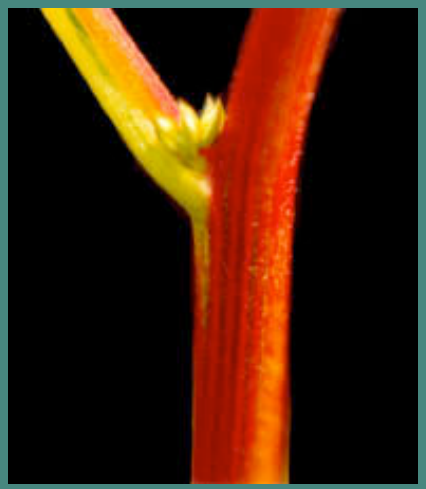


FIG. 61. Scheme to illustrate a method of crossing over of the chromosomes.

Genetic recombination during meiosis can lead to a novel set of genetic information that can be passed on to progeny

Hybridization

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



Water hemp

Hybrid

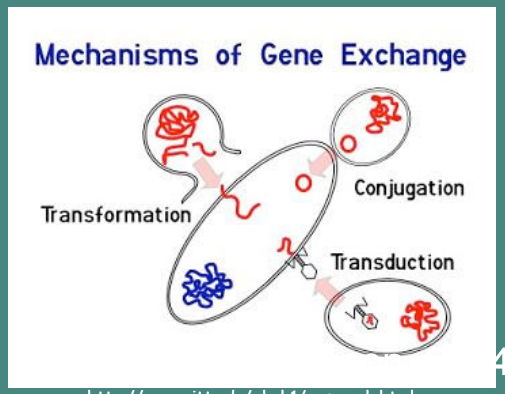
Pigs weed

Daniel Huson, 2004

Horizontal Gene Transfer

- Bacteria can become resistant to an antibiotic by having contact with other types of bacteria that are already resistant to the drug.

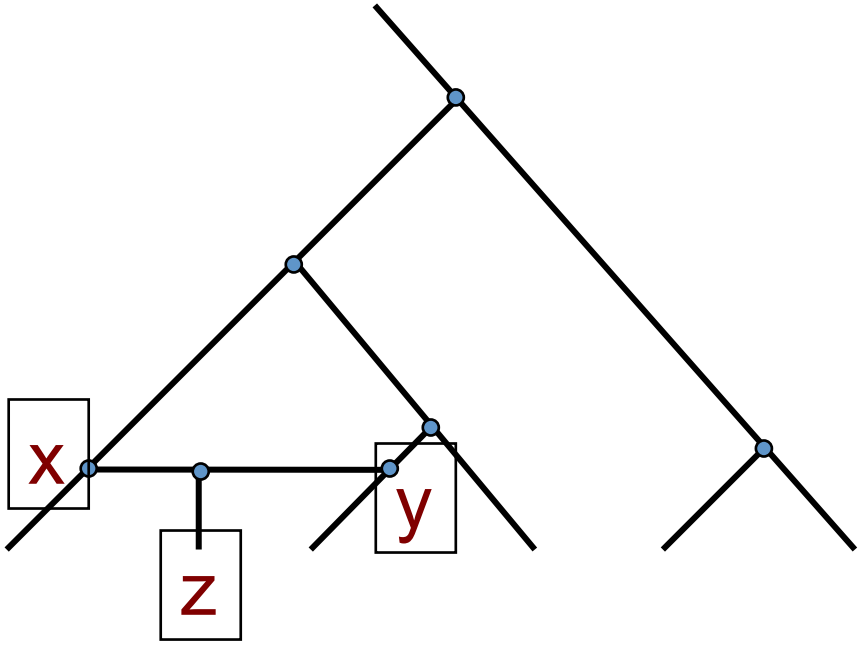
- This supports the idea that bacteria may commonly swap genes



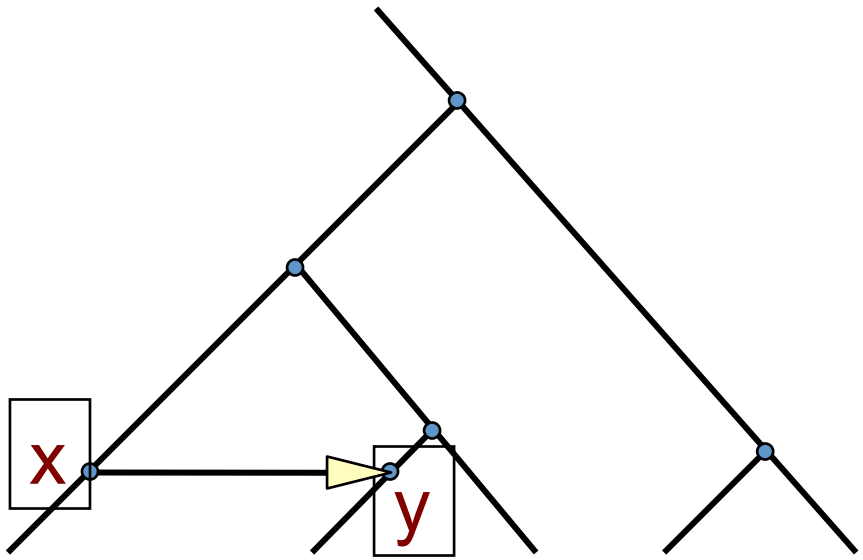
http://www.pitt.edu/~heh1/research.html

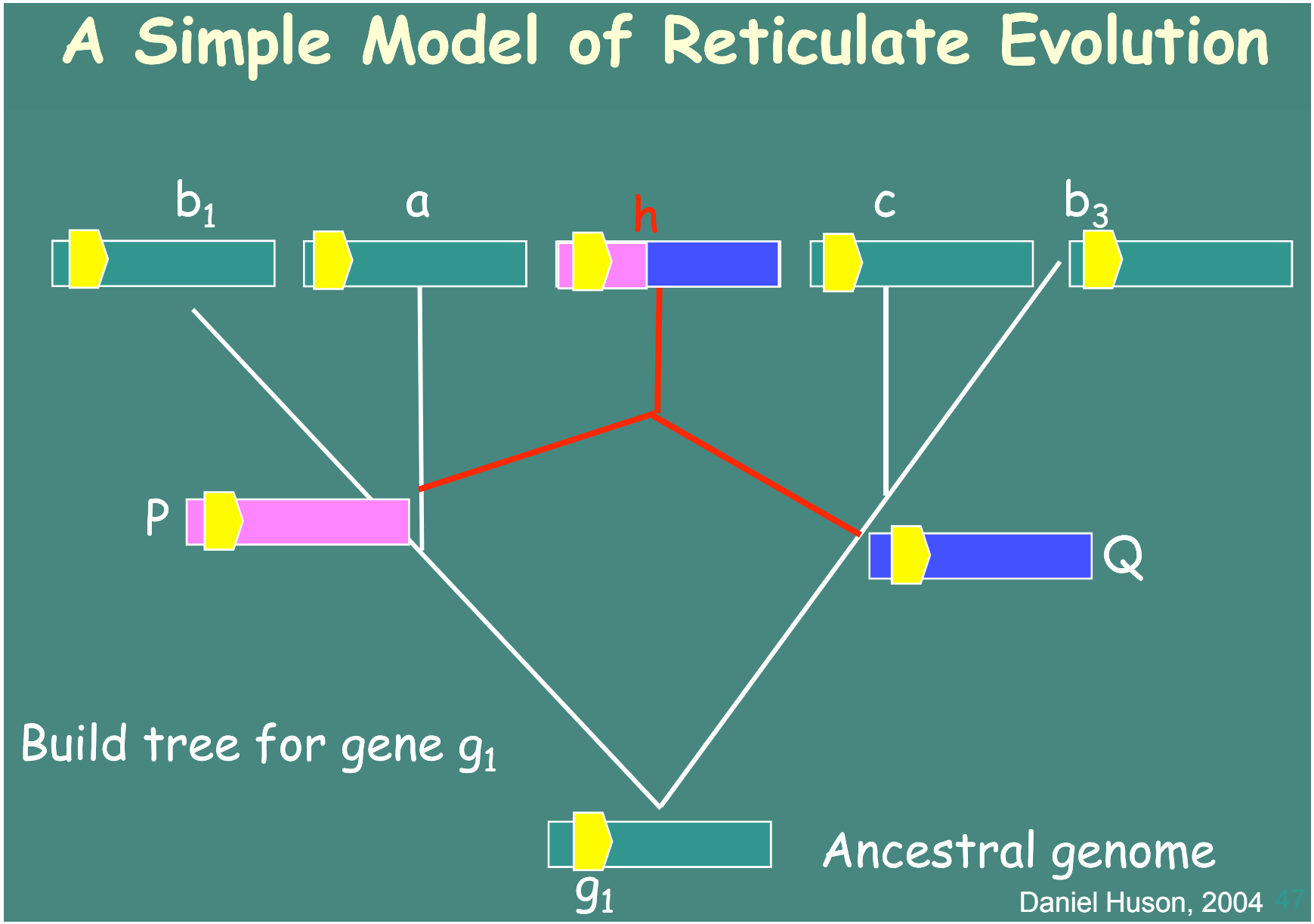
Types of reticulate evolution

- Hybrid speciation

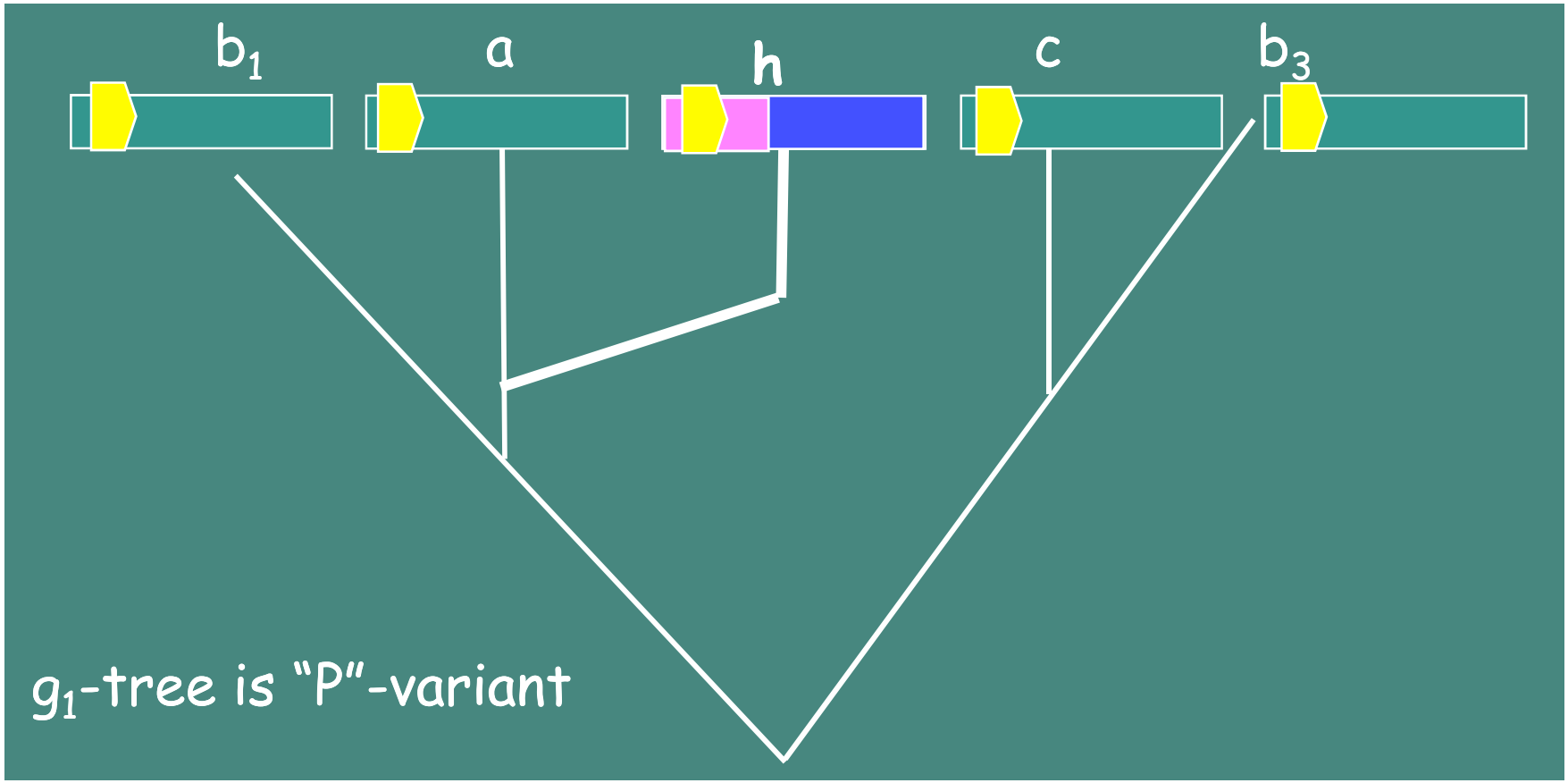


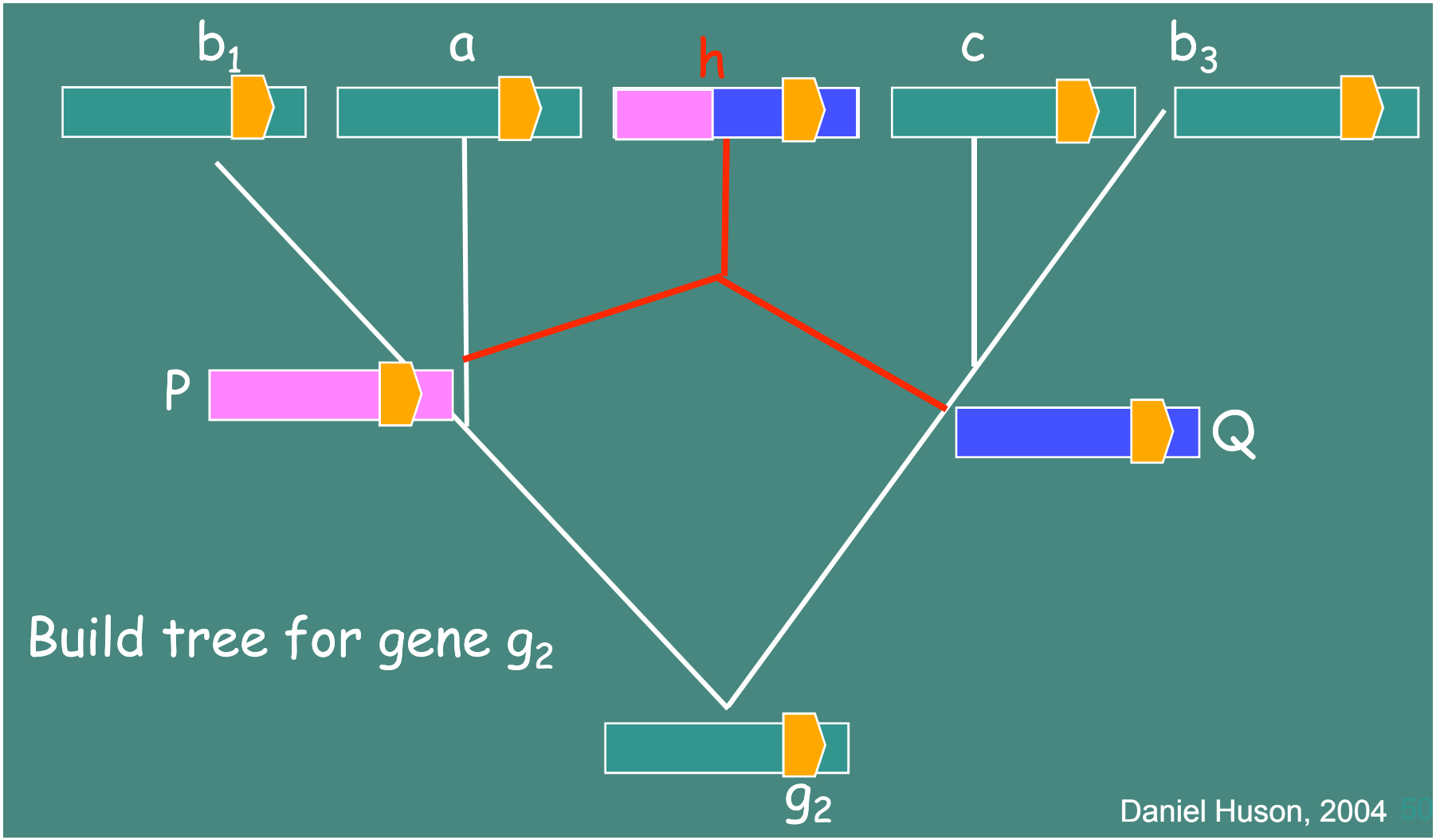
- Lateral gene transfer

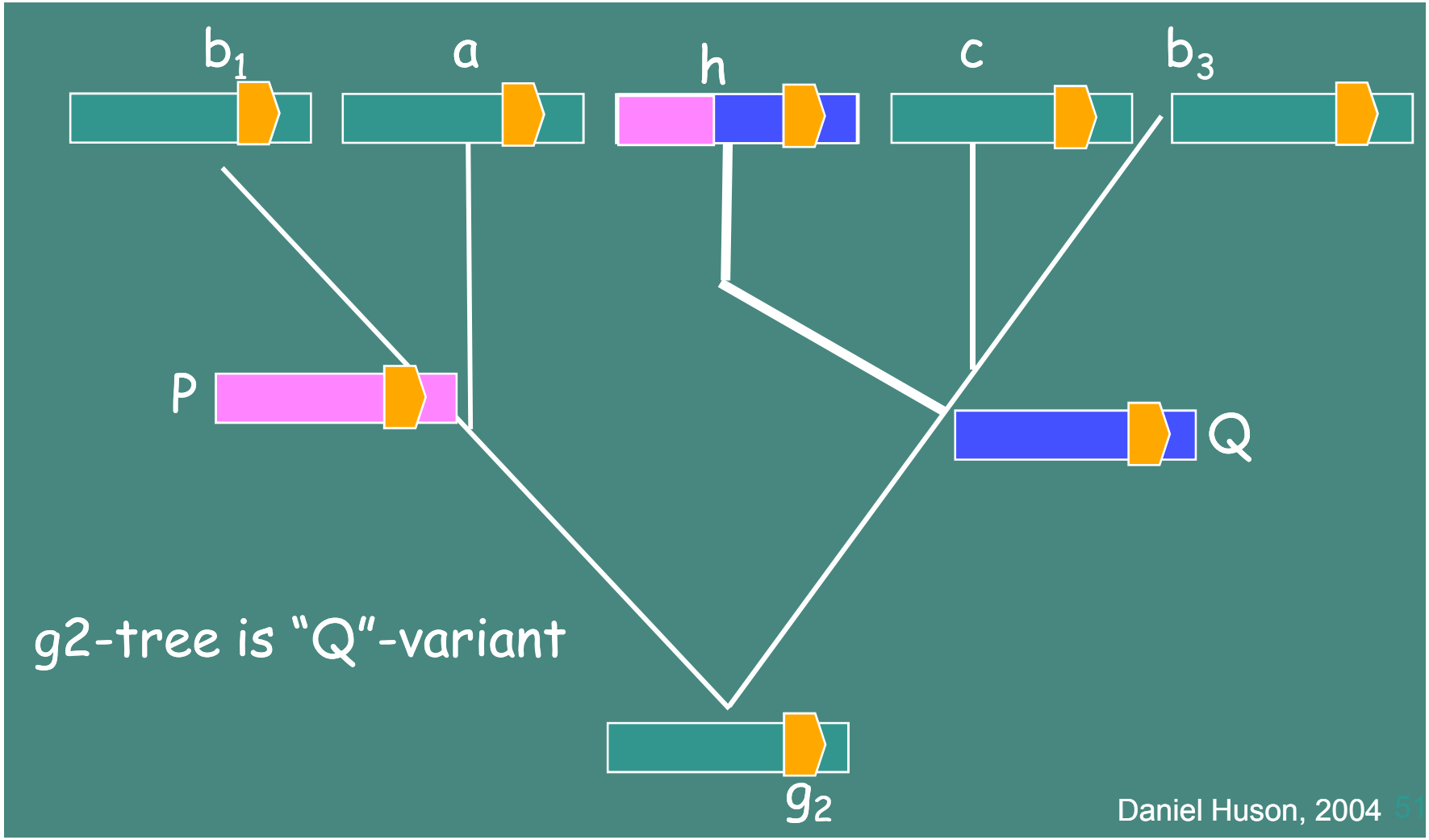


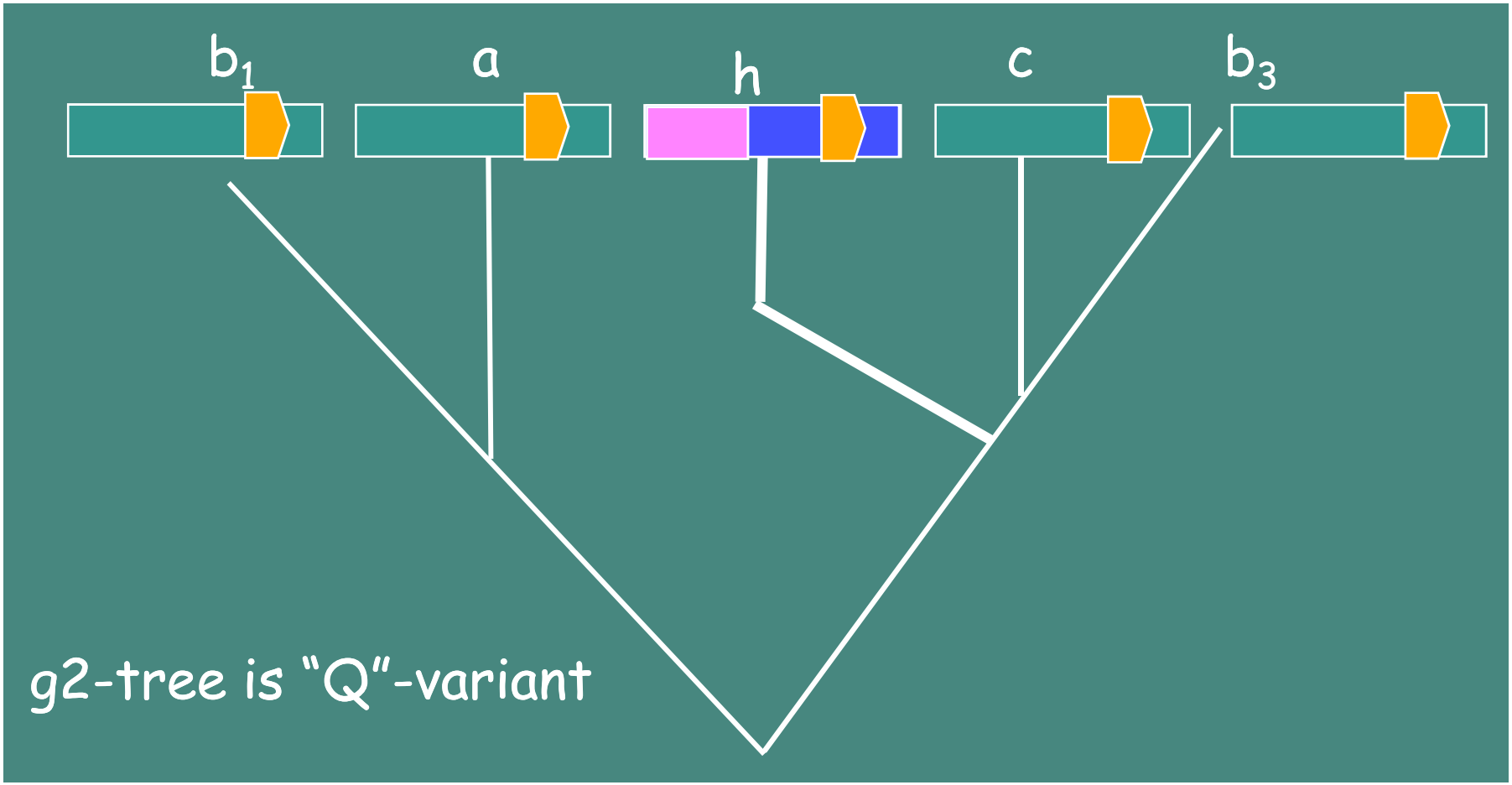


Where shall we insert h?


b1
a
h
c
b3
g1-tree is "P"-variant

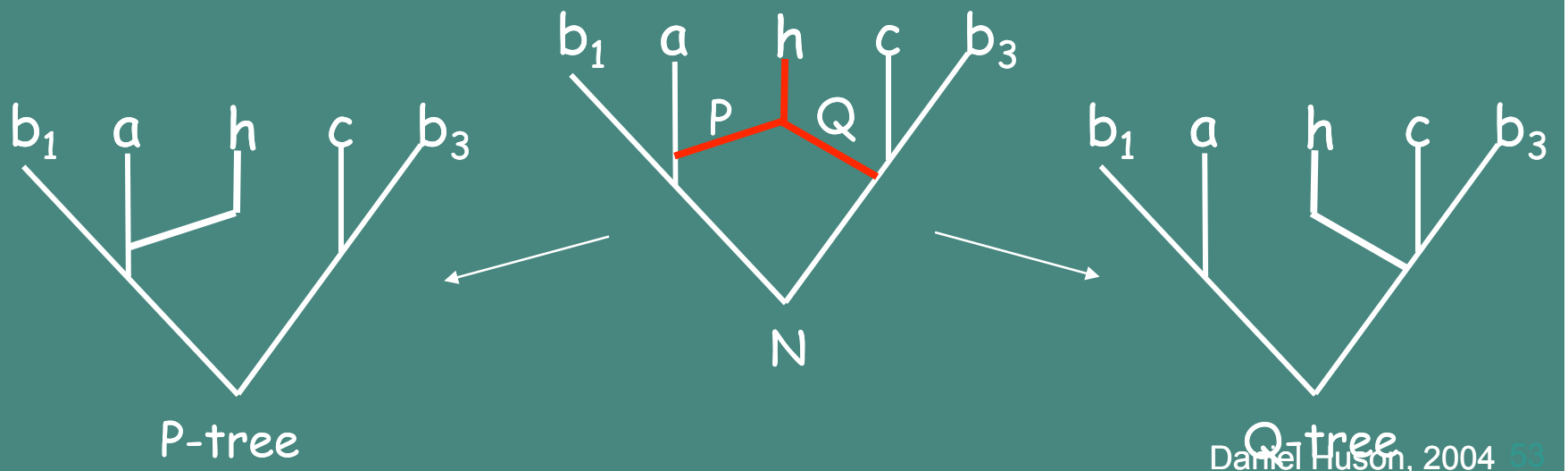

b1
a
h
c
b3
P
Q
Build tree for gene g2
g2


b1
a
h
c
b3
P
Q
g2-tree is "Q"-variant
g2


b1
a
h
c
b3
g2-tree is "Q"-variant

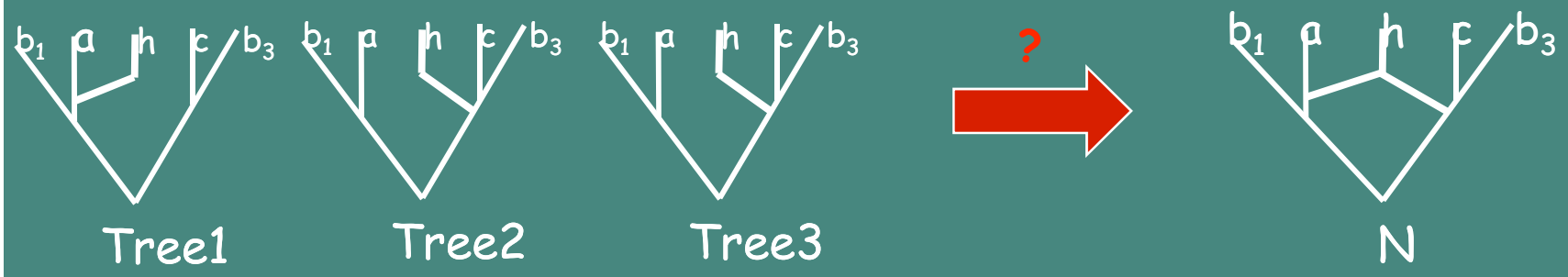
Reticulate Evolution and Induced Trees

- 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

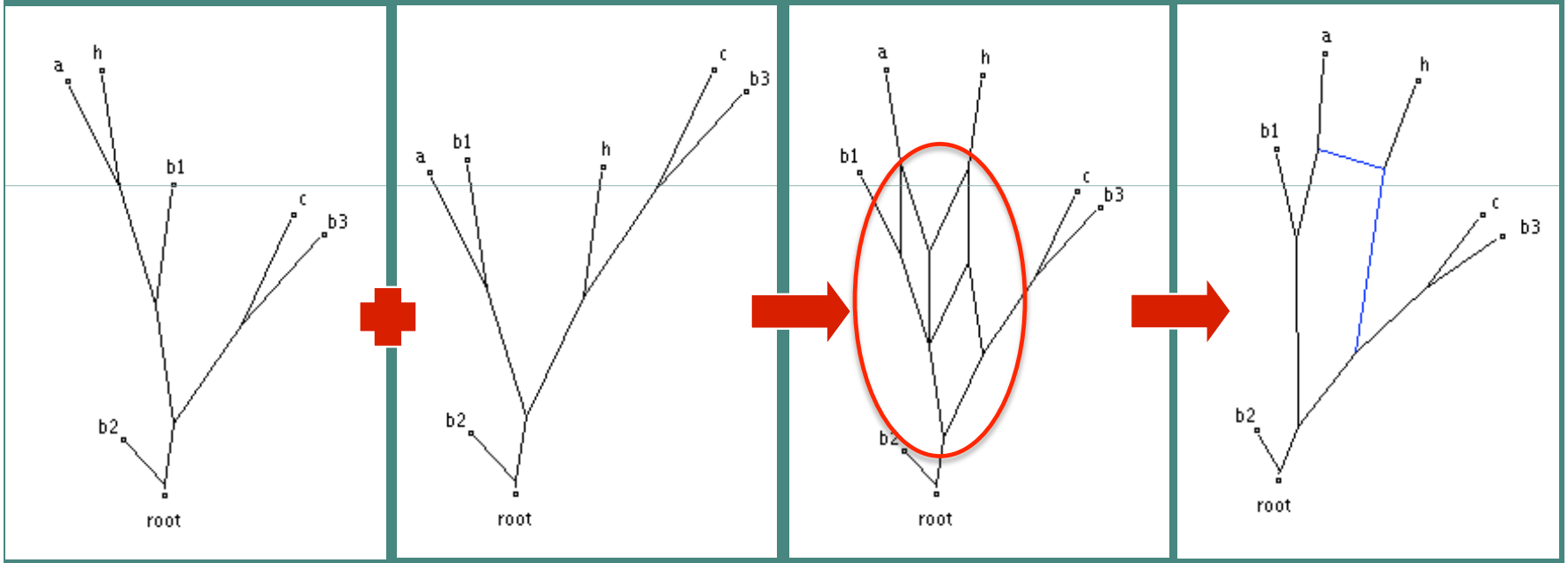


Reticulate Evolution Reconstruction Problem:

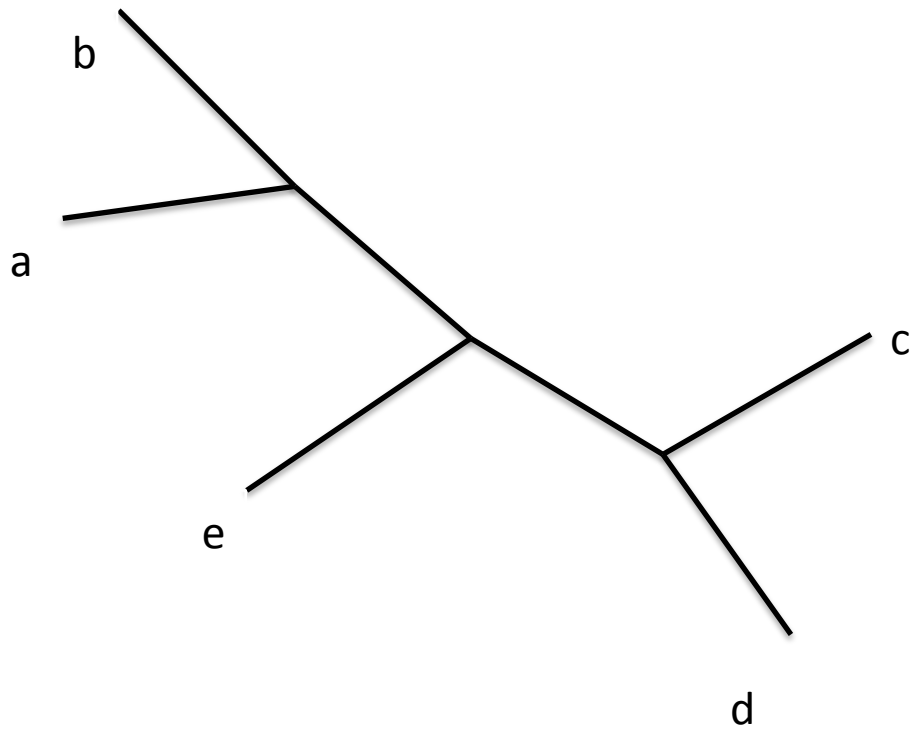
- Given a set of gene trees H sampled from an unknown reticulate network N.
- Reconstruct the network



From Gene Trees to Reticulate Network



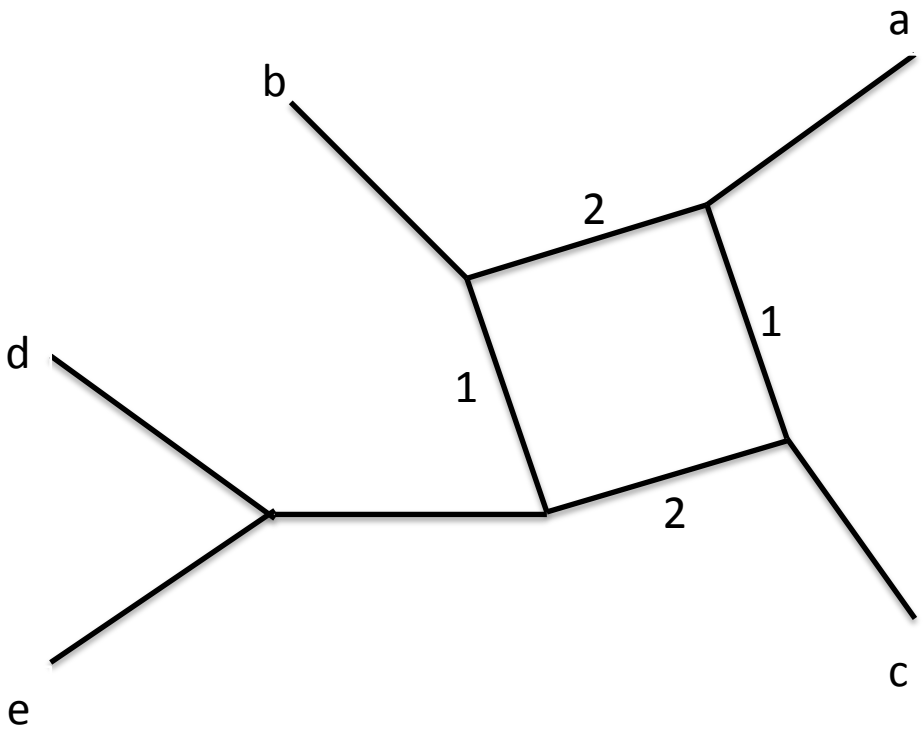
gene tree1 | gene tree2 | splits graph of all splits | reticulate network



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

All splits of the tree

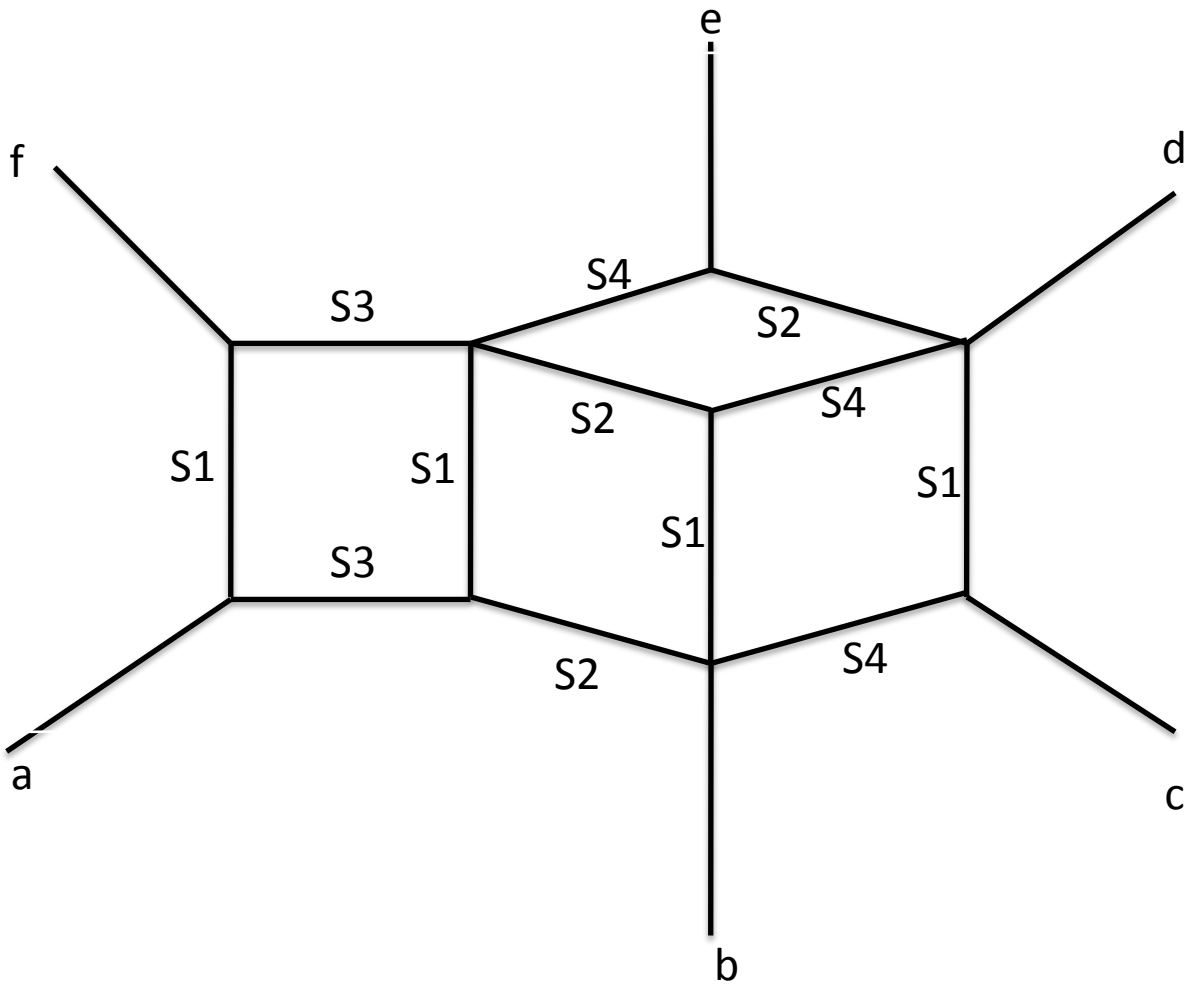
They are all the partitions of the leaves that can be defined by considering edges one at the time.



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

Splits represented by the two edges labeled 1 and labeled 2, respectively.

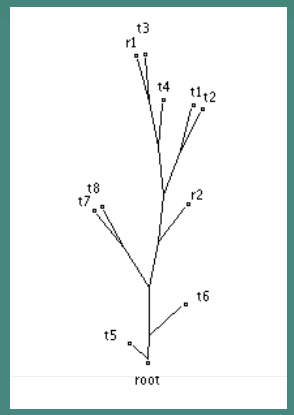
Splits of the network. They are all the partitions of the leaves that can be defined by considering ensembles of edges in the network one at the time.

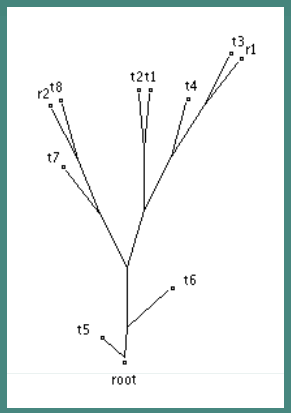


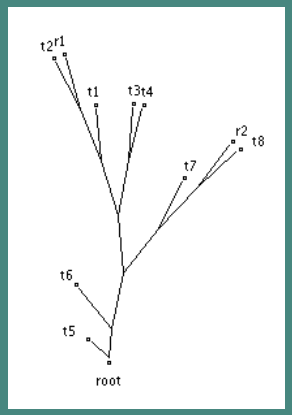
S1= {a,b,c}|{d,e,f}
S2= {a,e,f}|{b,c,d}
S3= {a,f}|{b,c,d,e}
S4= {a,b,f}|{c,d,e}

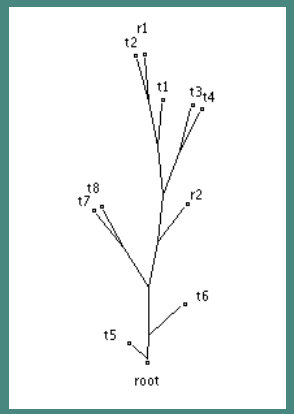
Splits of the network. They are all the partitions of the leaves that can be defined by considering ensembles of edges in the network one at the time.

Multiple Independent Reticulations

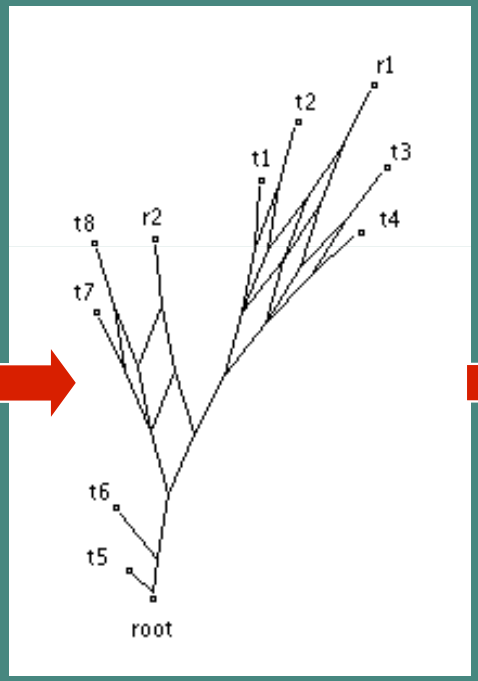




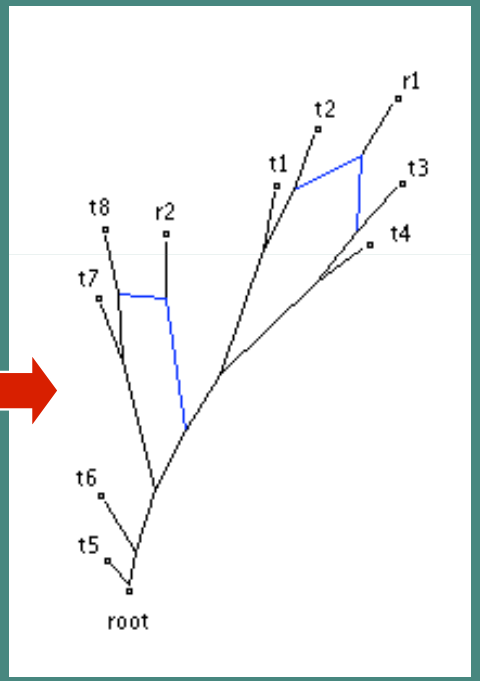




Two reticulations ⇒
four different gene trees



all splits



Reticulate network that induces all input trees

Optimal network reconstruction problem

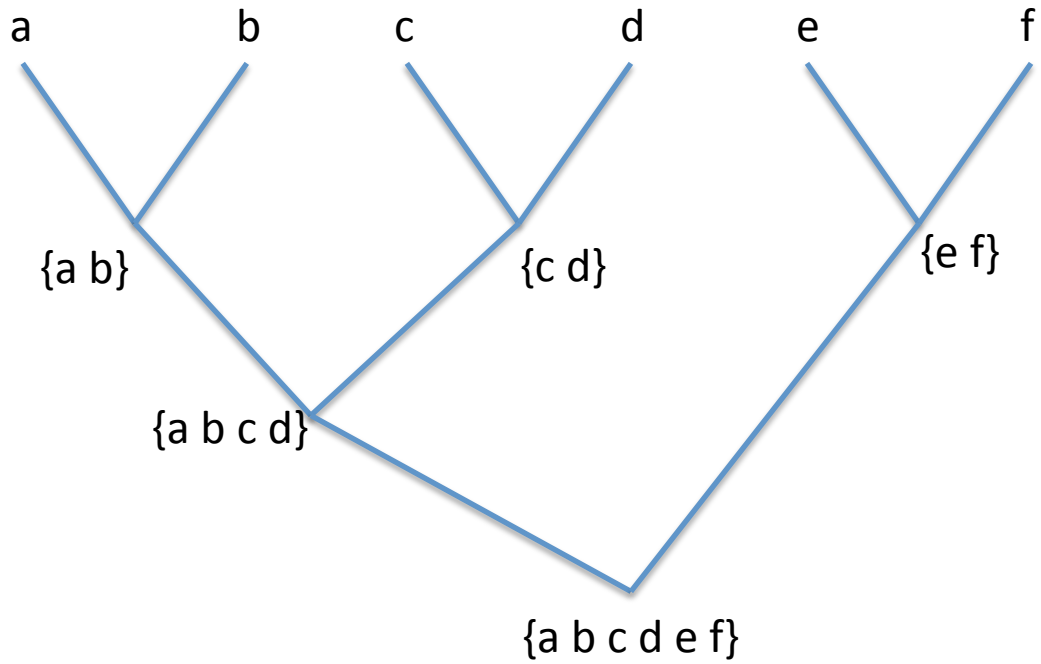
Given a set of gene trees sampled from a « species network », reconstruct the network. The idea is to start from two incongruent trees P and Q and explain their incongruency with some network.

Computational problem. Given a collection of input trees $T_1 \ldots T_k$, find a network N that contains all input trees and **minimizes** the number of « reticulation nodes ».

NP-hard problem (Bordewich and Semple, 2007)

FPT-algorithm for two trees
[Fixed Parameter Tractable – problems that can be solved in time $f(k)|x|^{O(1)}$, for some computable function f]

Trees and compatible clusters



If you have a phylogenetic tree, all nodes in the tree describe a cluster.
These clusters are nested.

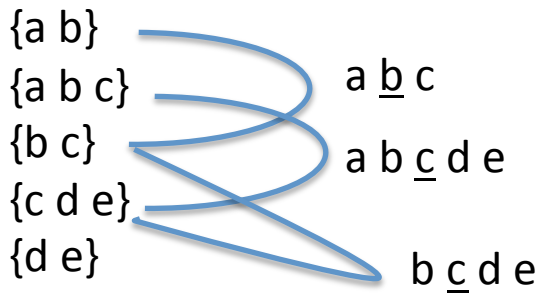
Two clusters C and C’ are compatible iff $\varnothing \in \{C\text{-}C', C \cap C', C'\text{-}C\}$

(that is, either they are disjoint or they have to be one the subset of the other)

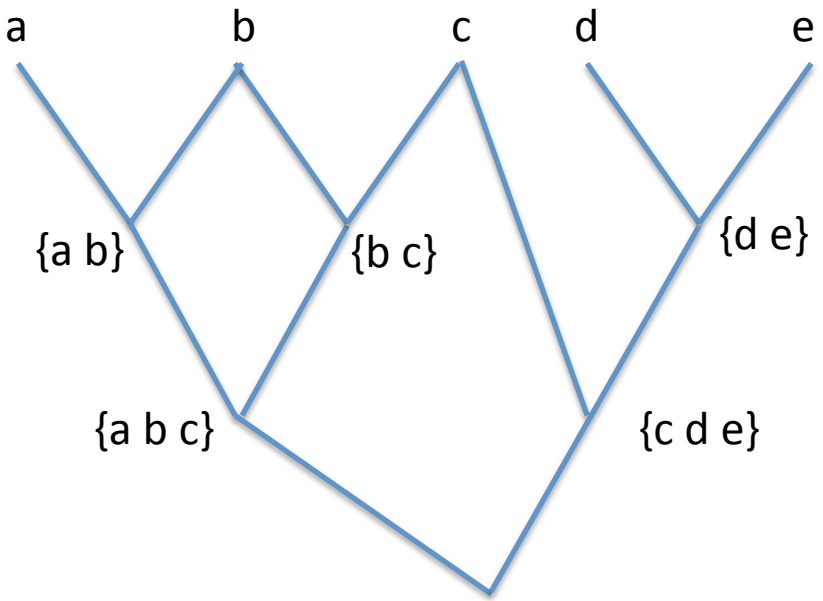
Lemma: the Hasse diagram of a set of clusters is a tree iff all pairs of clusters are compatible.

Incompatible clusters and the Hasse diagram

We have a set of clusters that are incompatible



There are three pairs of clusters that are incompatible with each other



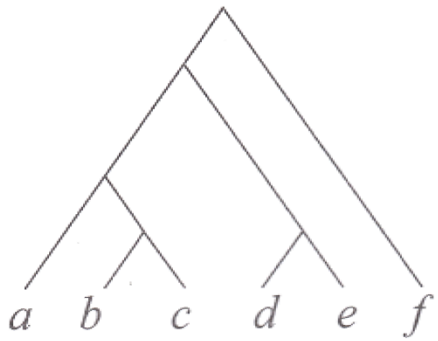
Not part of the input

How does a rooted phylogenetic network N on X represent a cluster?
where X is the set of taxa.

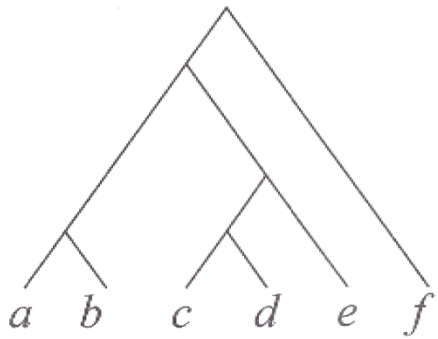
There are two possible answers:

A network N represents a given cluster C on X in the **hardwired sense**, if there exists a tree edge e in N such that the set of labels of leaves below e equals C.

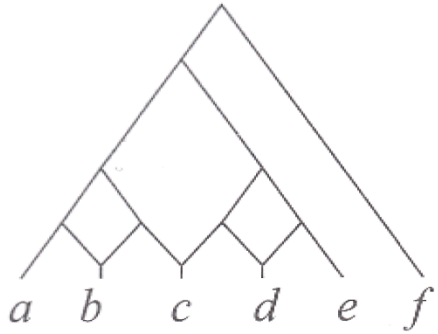
A network N represents a given cluster C on X in the **softwired sense**, if there exists a phylogenetic rooted tree T that is contained in N and represents C (in the hardwired sense).



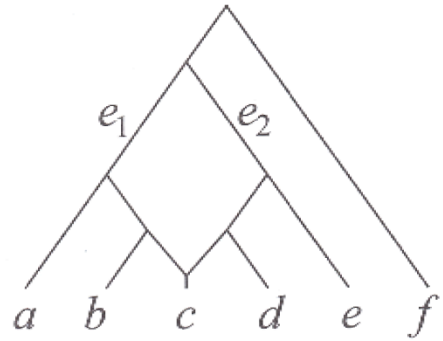
(a) Tree T_1



(b) Tree T_2



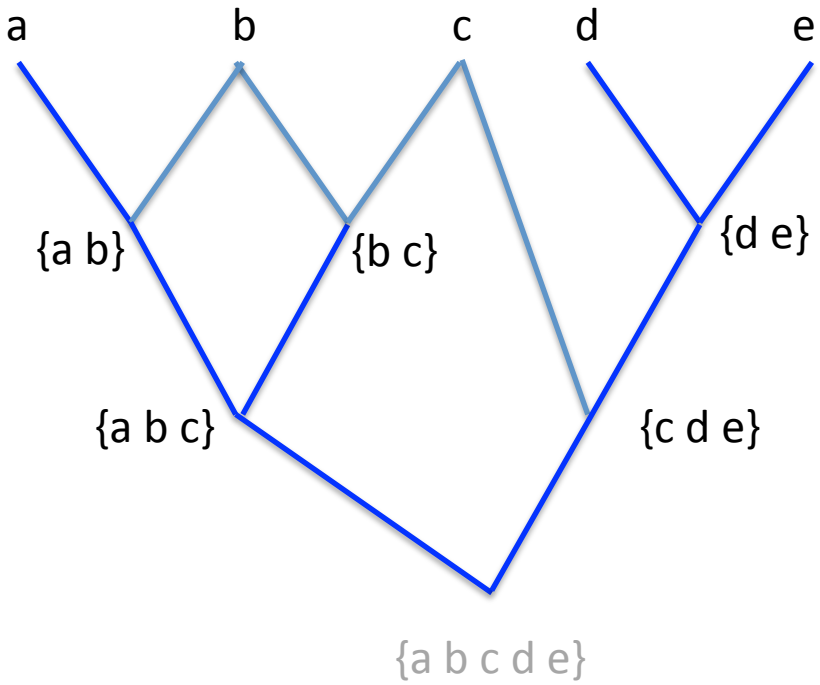
(c) Hardwired representation N_1



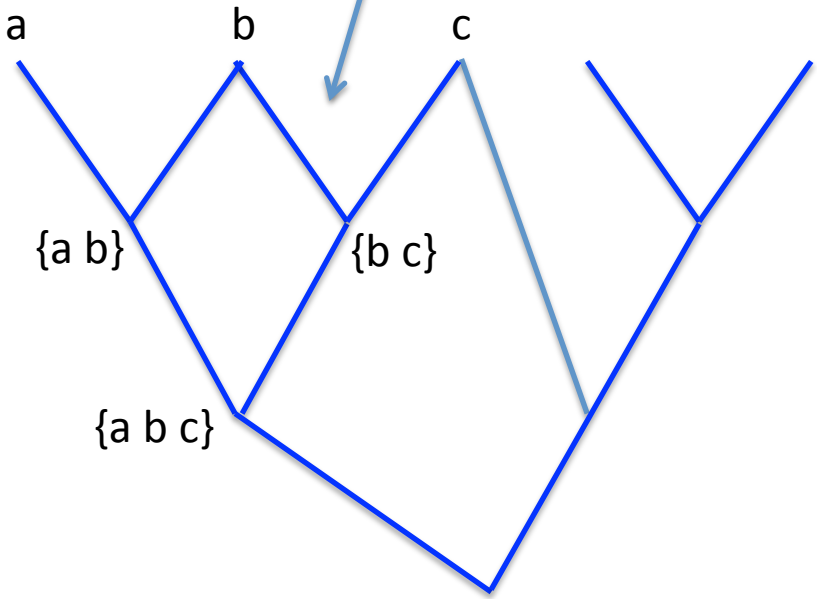
(d) Softwired representation N_2

Figure 6.14 In (a) and (b) we show two rooted phylogenetic trees T_1 and T_2. (c) A cluster network N_1 that represents all clusters contained in T_1 and T_2 (in the hardwired sense). (d) A second rooted phylogenetic network N_2 representing the same set of clusters in the softwired sense. In this particular example, both input trees are also represented by both networks.

Hasse diagram and hardwired networks

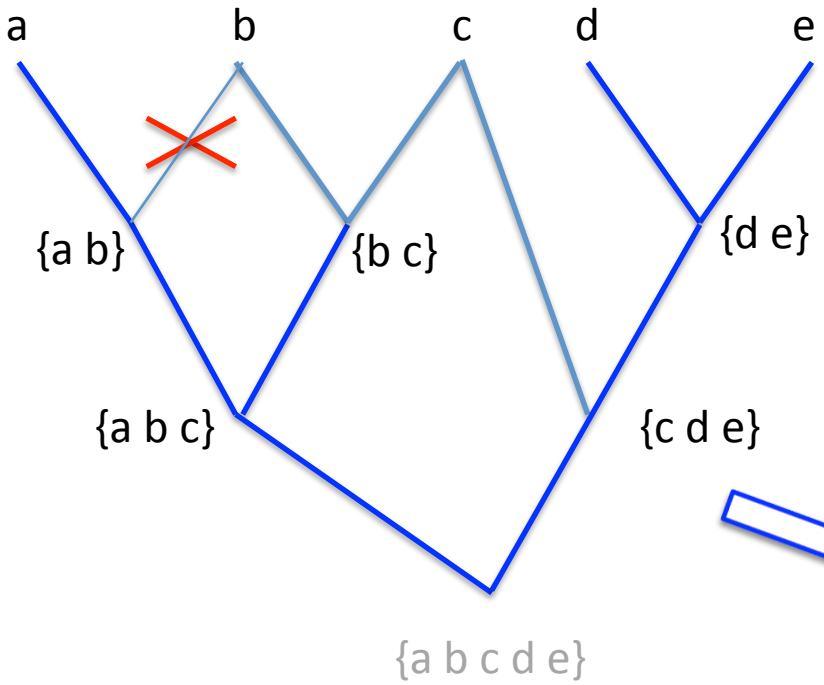


This is not a tree:
this is not the kind of networks that biologists are looking for.



If we want to look what is labeling the node {a b c} we need to go down reading the subtrees: **hardwired network**
(the label of a node is the union of the labels of its children)

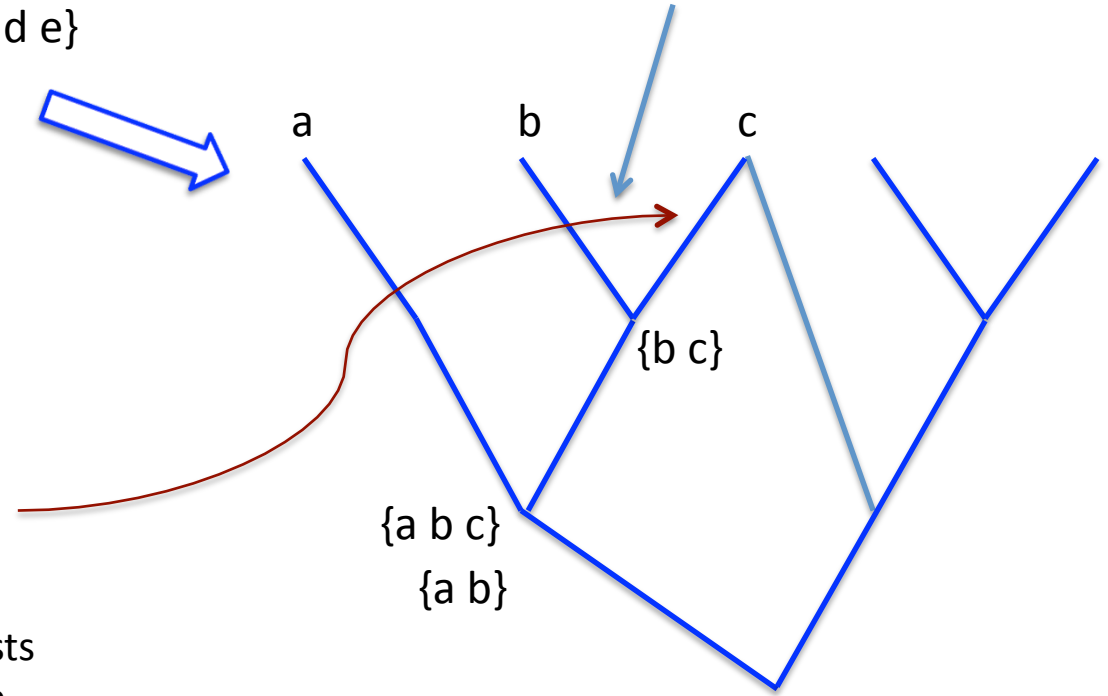
Softwired networks



Modify the network slightly! Take away an edge.

Check whether all clusters are represented by the tree. You have to move cluster {a b} up.

This network does not represent the original clusters in the hardwire sense!



Softwired network: this edge is turned off by {a b}

This is the kind of networks that biologists look for, to describe genes that might be recovered by one parent or by the other

Property:

Let N be a rooted phylogenetic network on X. The set of clusters represented by a hardwired network N is contained in the set of clusters represented by the network in the softwire sense.

How do we find an optimal softwired network?

The idea of the algorithm (Huson, 2010) is
-List the clusters.
-Write down all incompatibilities
-Find the minimal set of taxa that hits all (incompatible) triplets. This should provide the reticulation nodes.

Conjecture: the minimum number of taxa needed to remove all incompatibilities equals the minimum number of reticulate nodes in a softwired network.

Check whether this conjecture has been proven!

Interested in phylogenetic networks

How to represent incompatible signals, reticulate events?

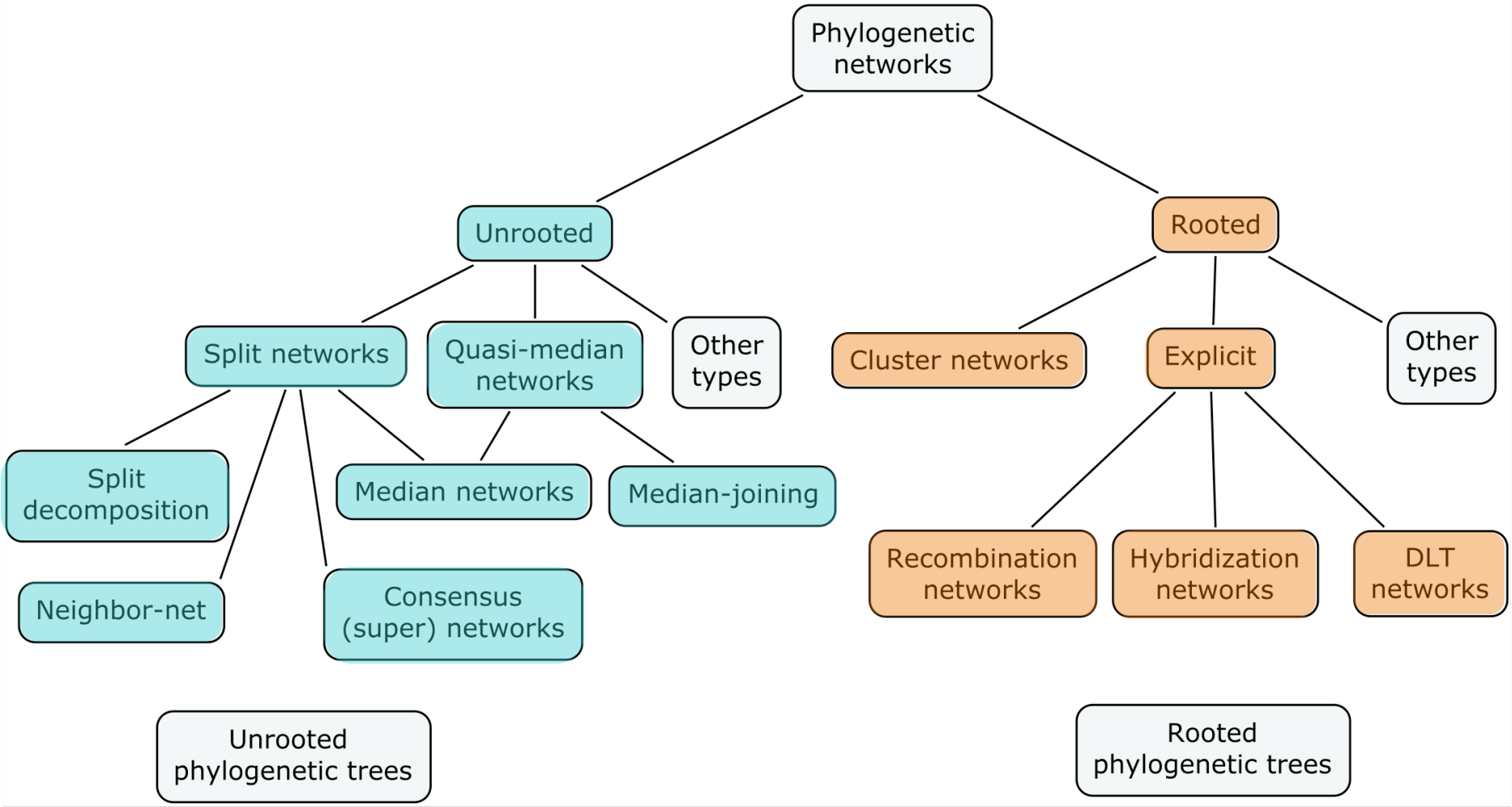
Unrooted networks

- Split networks

Rooted networks

- abstract networks
- explicit networks

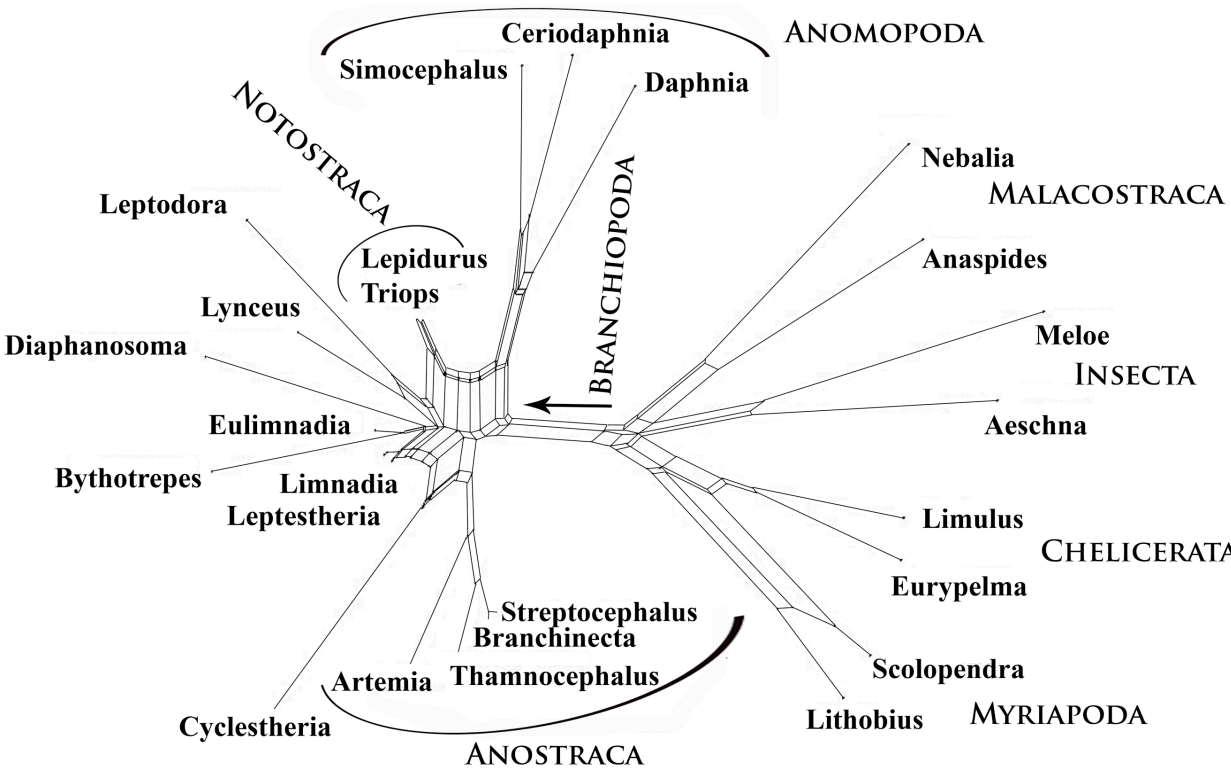
A tree of terms..



SplitsTree4

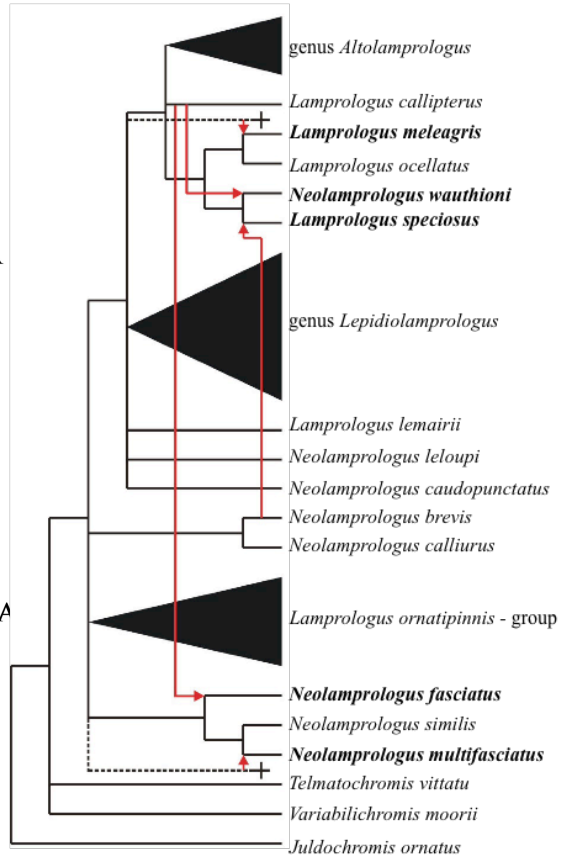
- Integrates a wide range of phylogenetic network and phylogenetic tree methods, inference tools, data management utilities, and validation methods.
- Included methods for inferring split networks:
  - *From character data*. Median networks, parsimony splits, spectral analysis
  - *From distance matrices*. Split decomposition and neighbor-net
  - *From sets of trees.* Consensus networks and supernetworks.
- Also constructs other types of phylogenetic networks, eg recombination and hybridization networks
- User friendly?!

Split network:



Waegele and Meyer (2007)

Hybridization network:

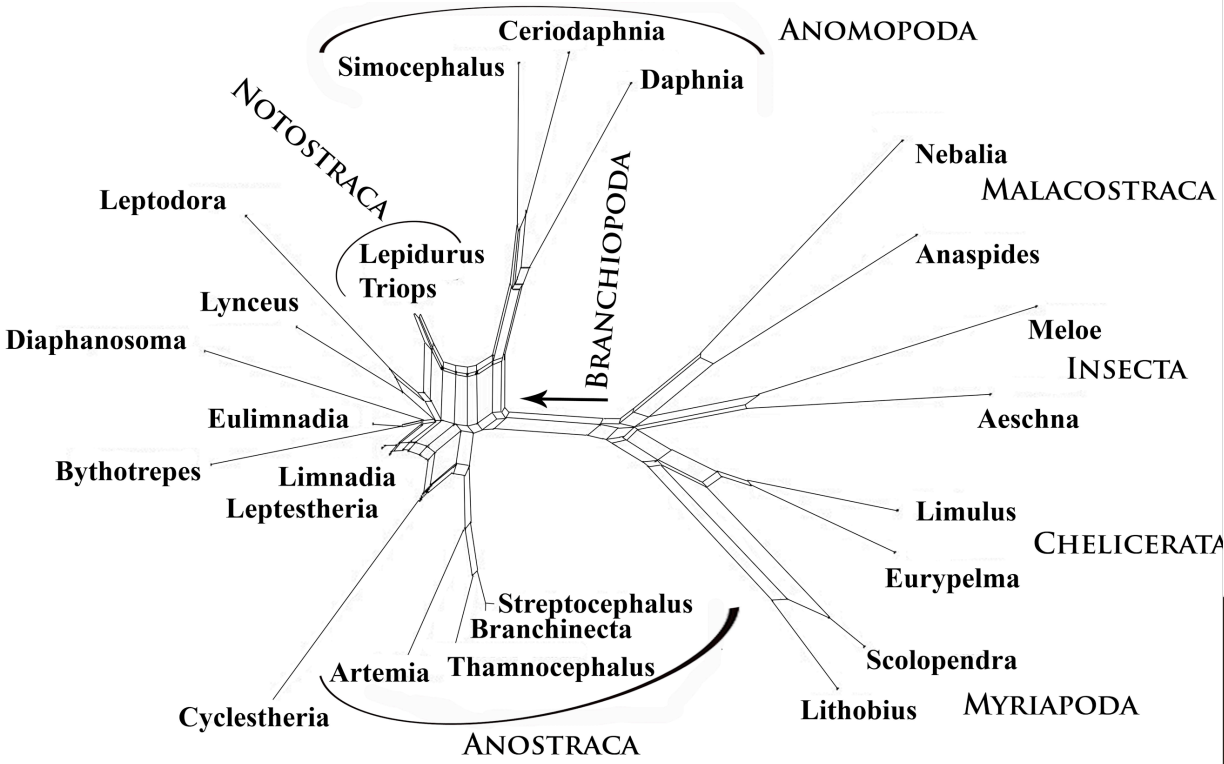


Koblmuller et al (2007)

No clear direction of time

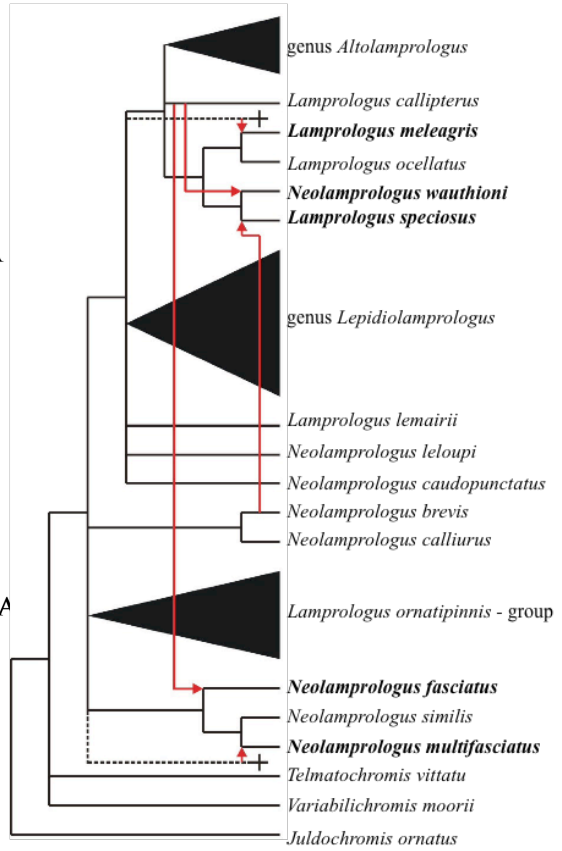
Has direction of time

Split network:



Shows conflicting placement of taxa

Hybridization network:

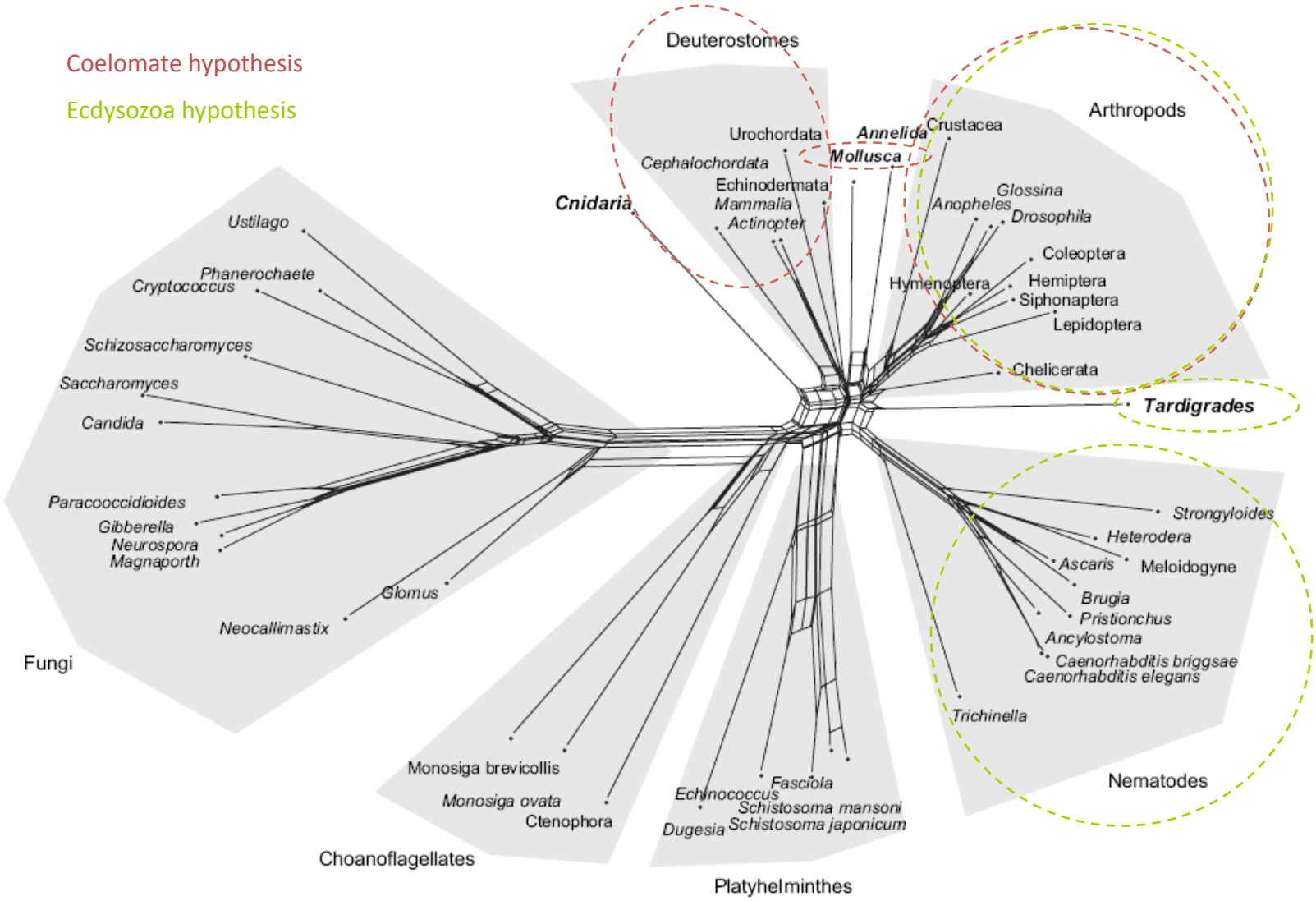


Shows putative hybridization history

Conclusion

- Split networks are useful for visualization.
- However they are not useful for making conclusive phylogenetic analysis.
- SplitsTree4 encompasses many tools, but are they really that useful?

Example 2:
Animal Phylogeny

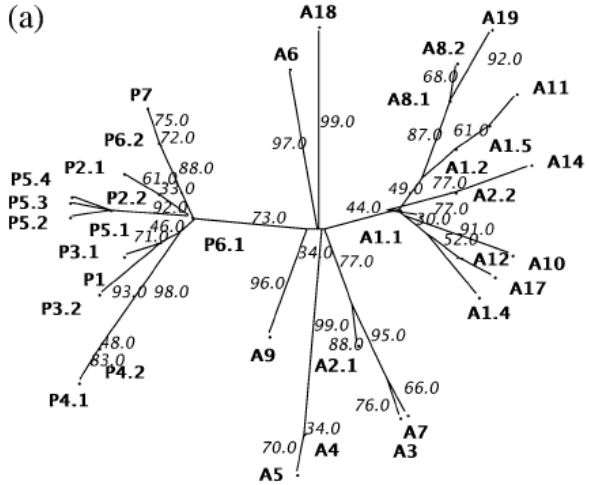


More examples..

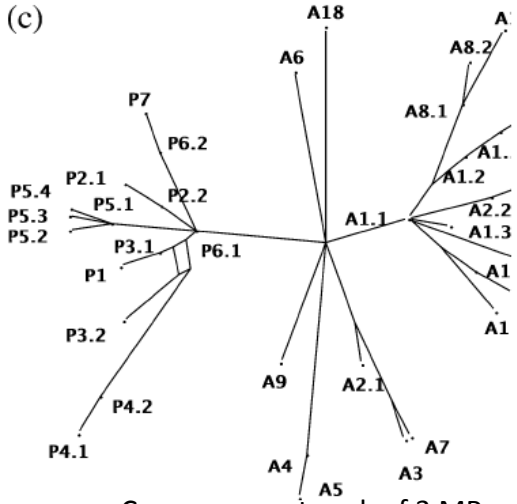
Dusky dolphins

60 variables (sites of DNA)

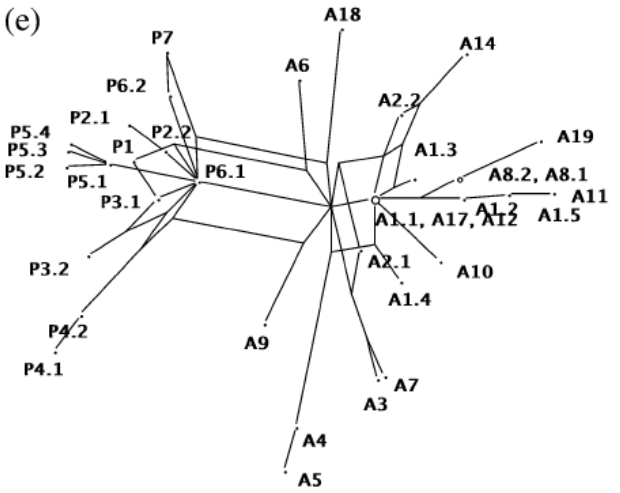
35 haplotypes



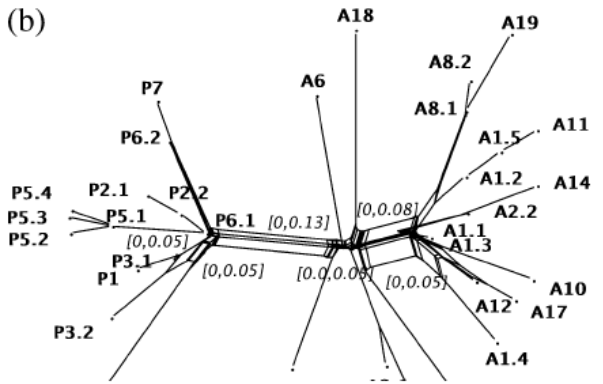
Neighbor-joining tree with bootstrap values



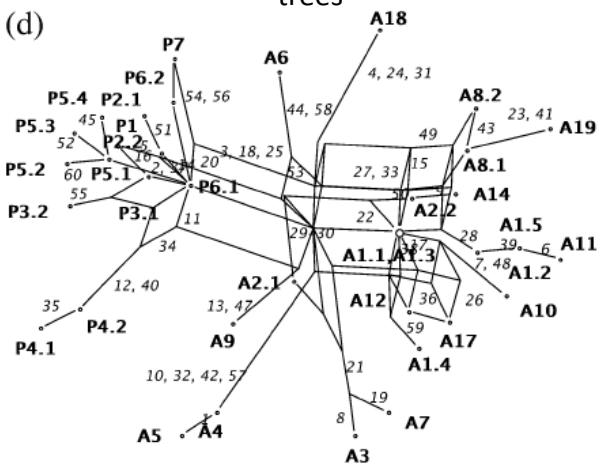
Consensus network of 3 MP trees



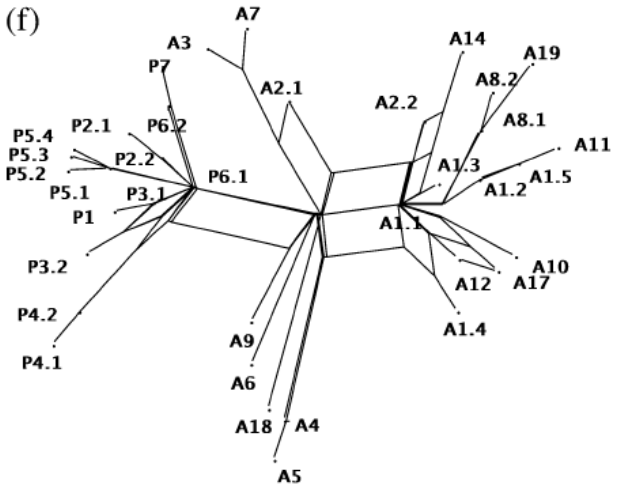
Split decomposition network



95% confidence network



Median network



Neighbor-net network