Phylogenetic Networks

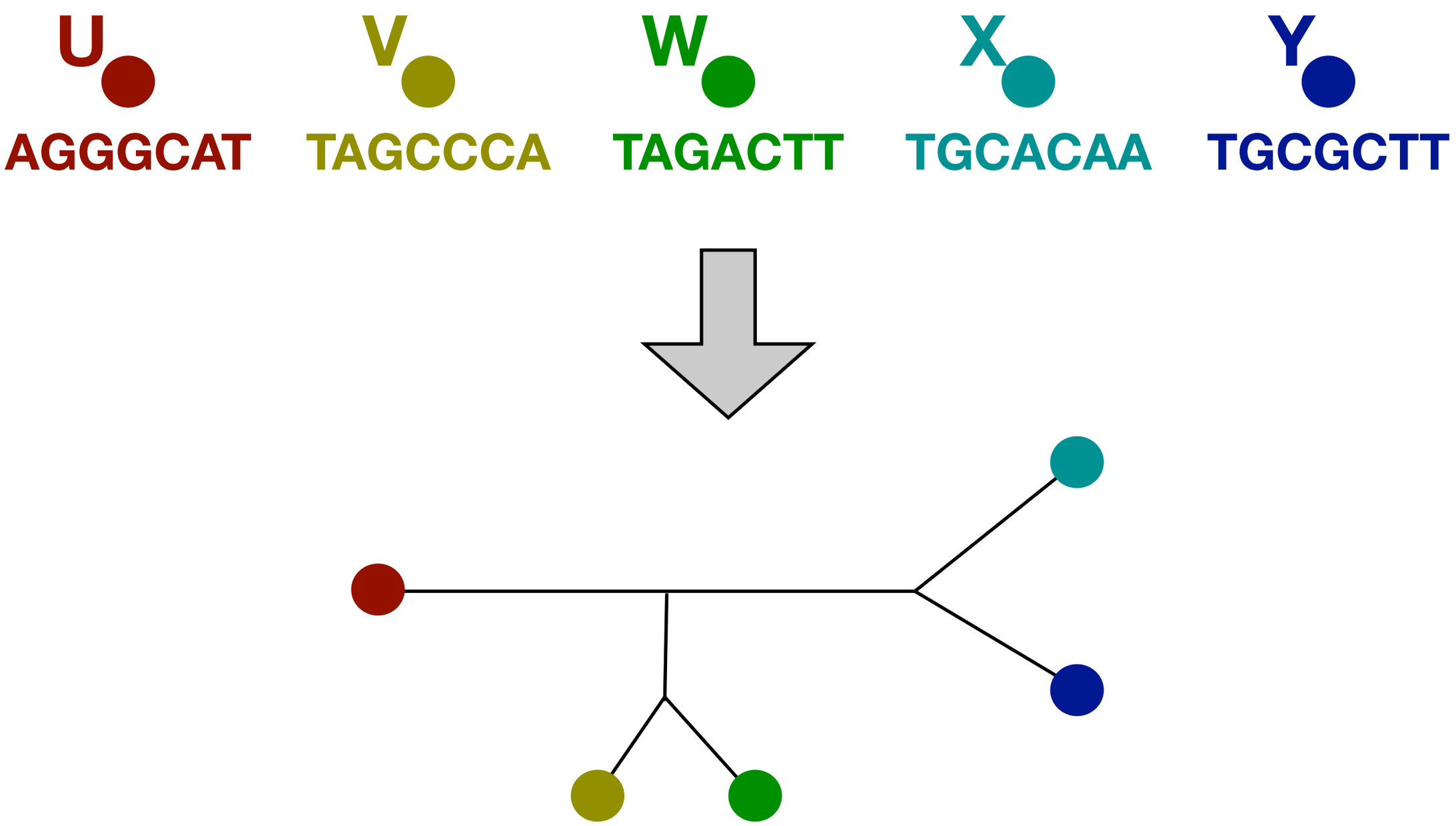
Luay Nakhleh
Department of Computer Science
Rice University

RICE
Unconventional Wisdom

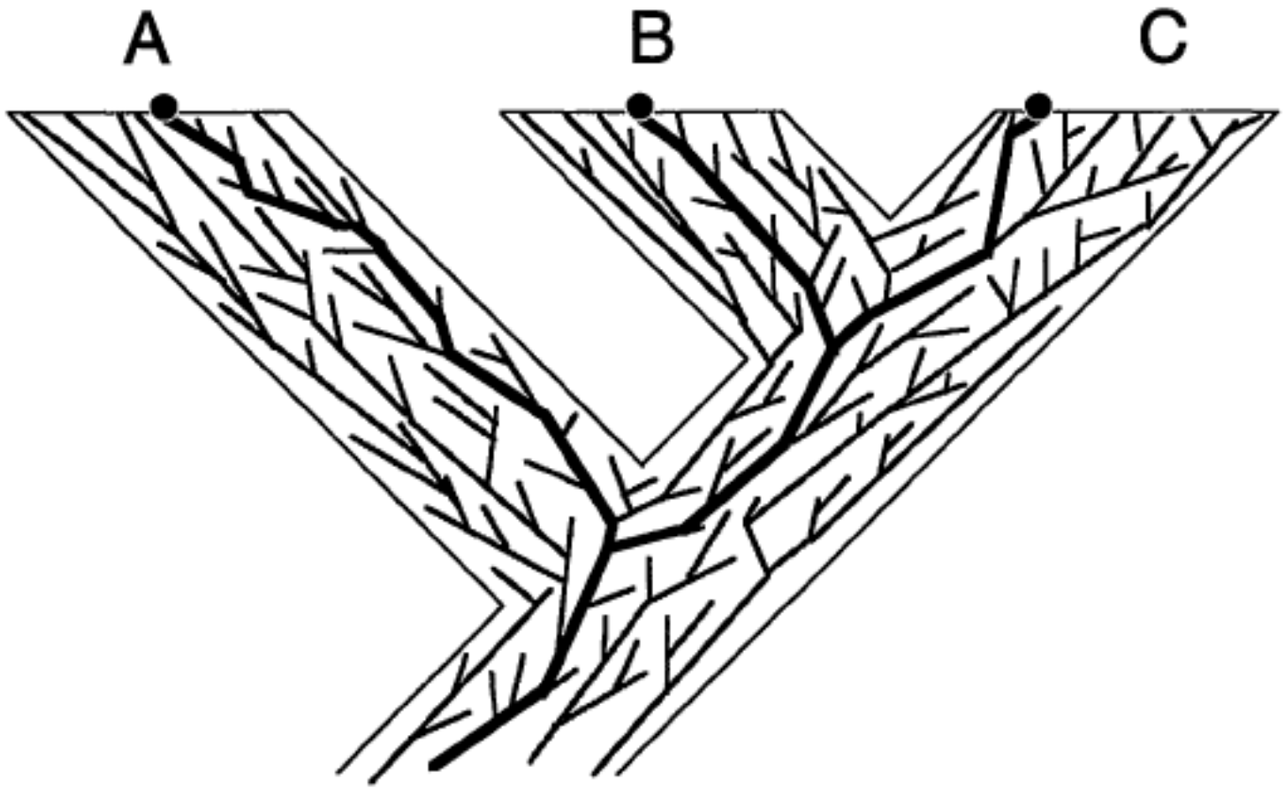
Bioinformatics for Biologists
University of Houston
9 November 2011

The Phylogeny Reconstruction Problem

The Phylogeny Reconstruction Problem

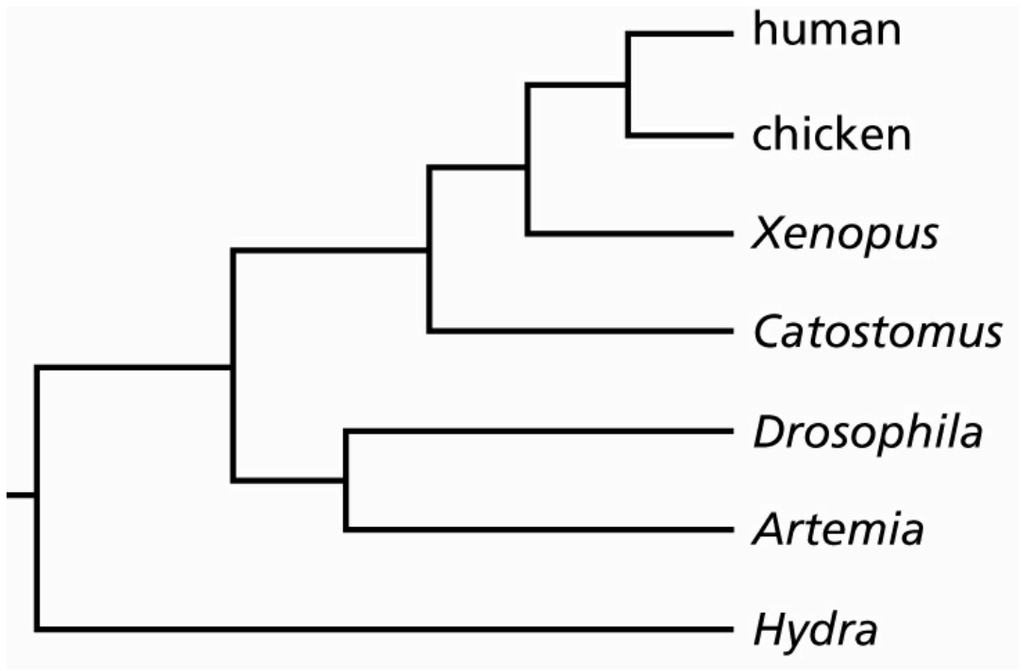


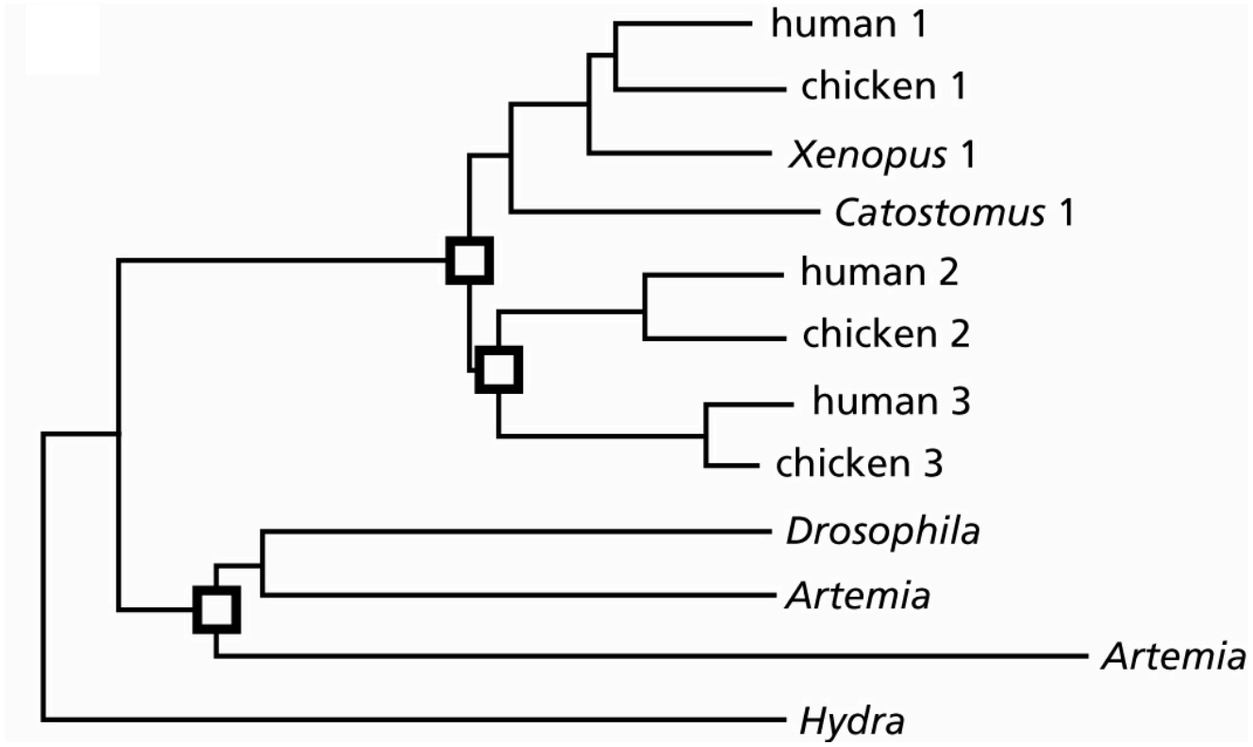
Gene Trees in Species Trees



[Source: W.P. Maddison, Syst. Biol. 46(3):523-536,1997.]

What Tree is Being Reconstructed?

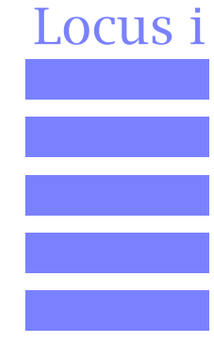




The Pre-Genomic Era

The Pre-Genomic Era

A
B
C
D
E

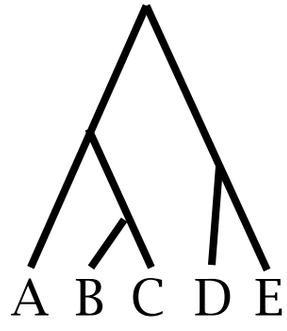


Gene Tree

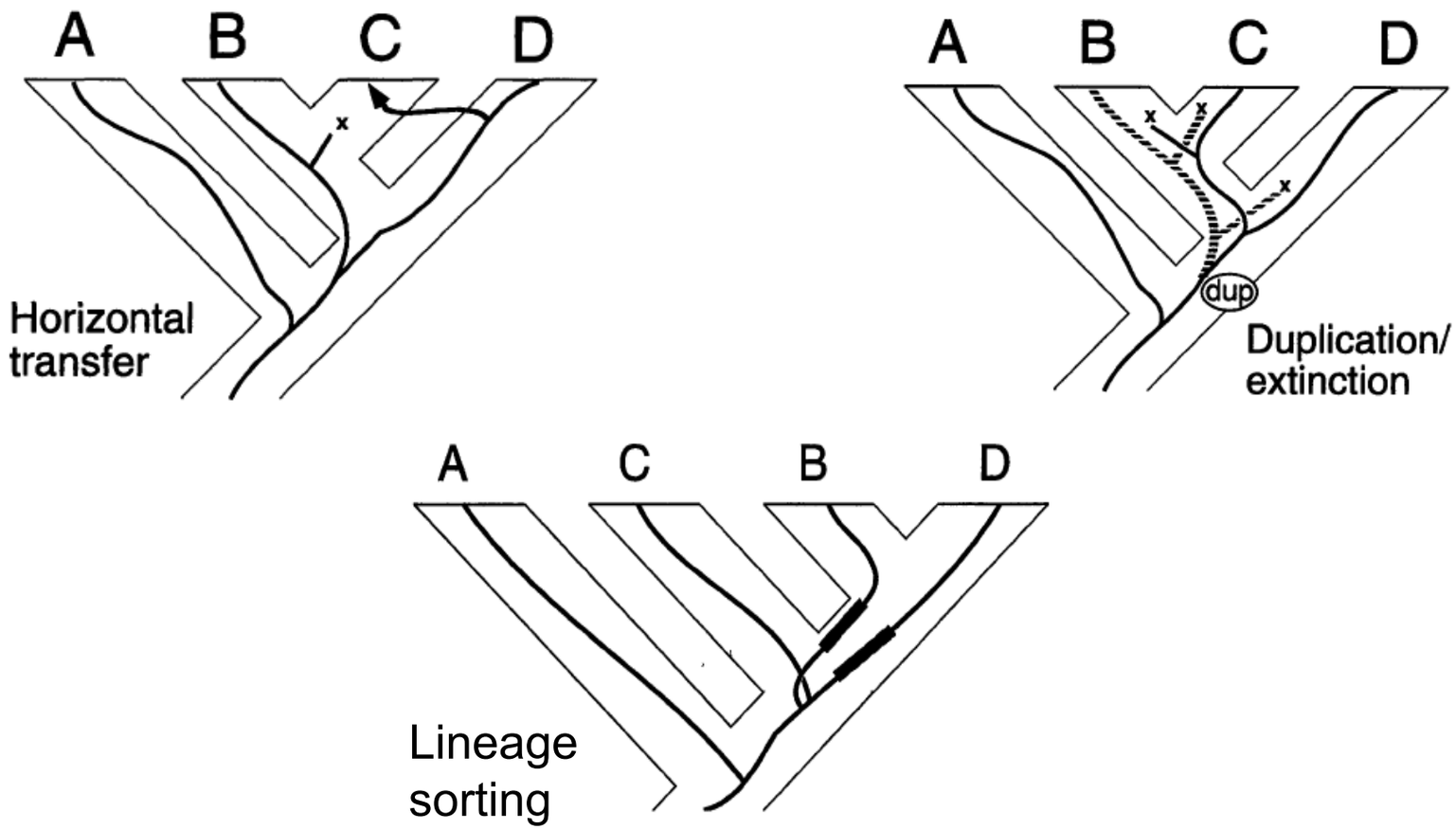
The Pre-Genomic Era

Gene Tree

Species Phylogeny

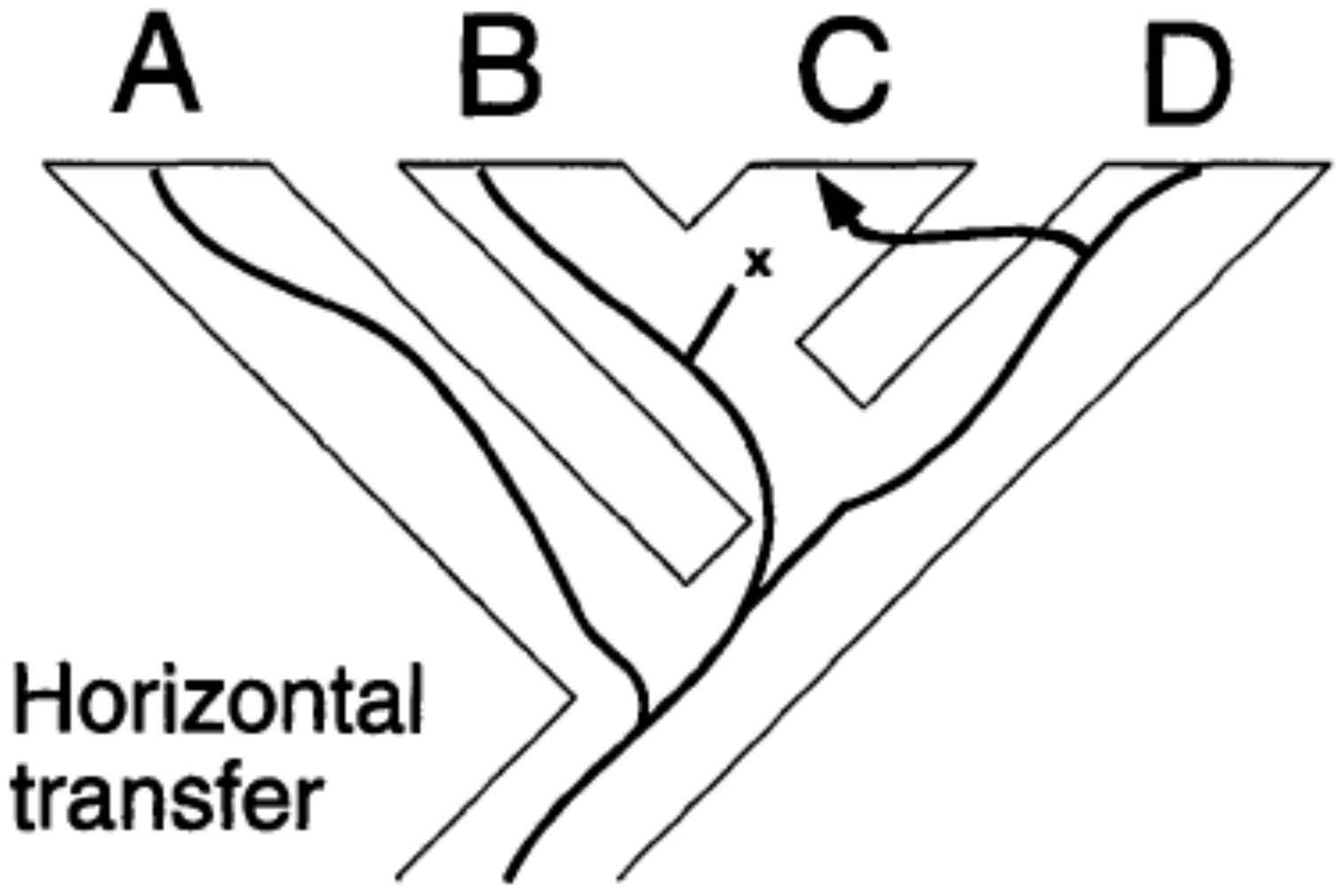


However...

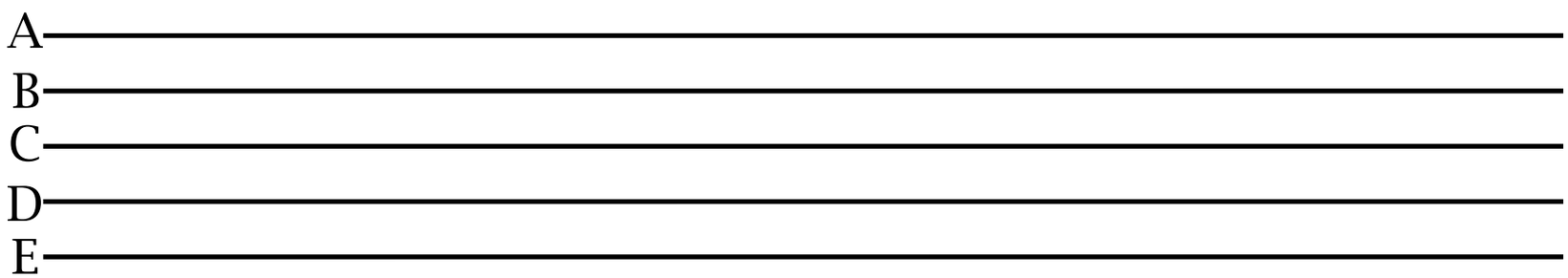


[Source: W.P. Maddison, Syst. Biol. 46(3):523-536,1997.]

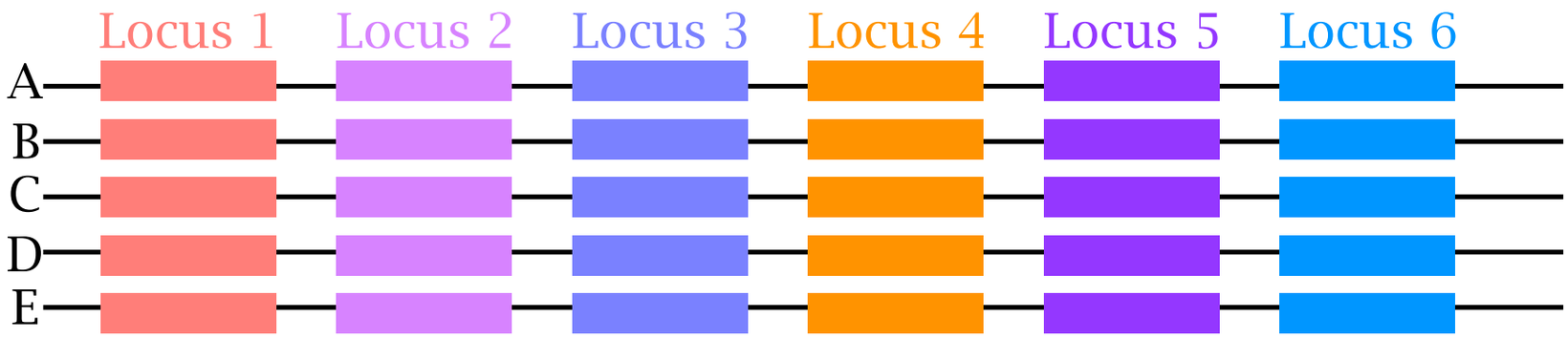
In This Lecture



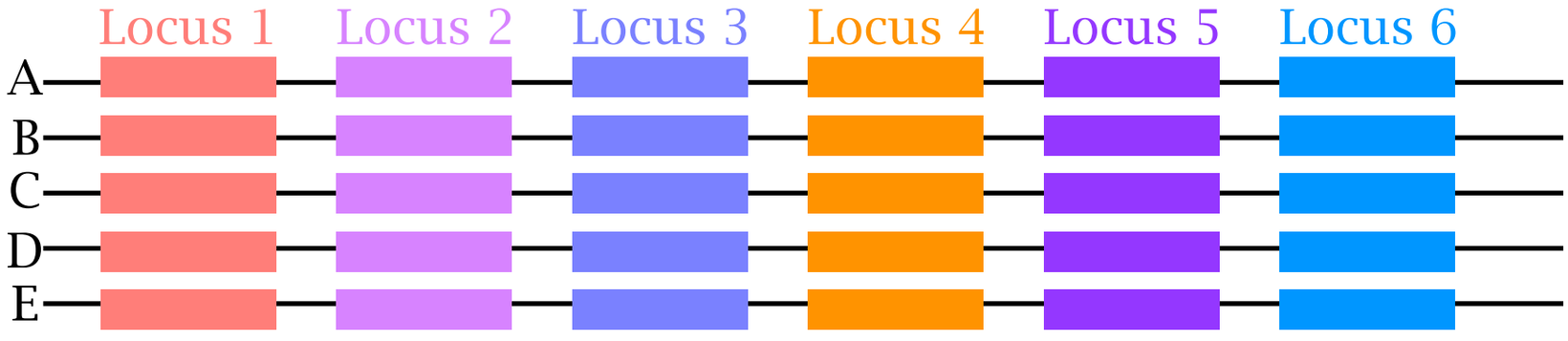
The post-genomic era



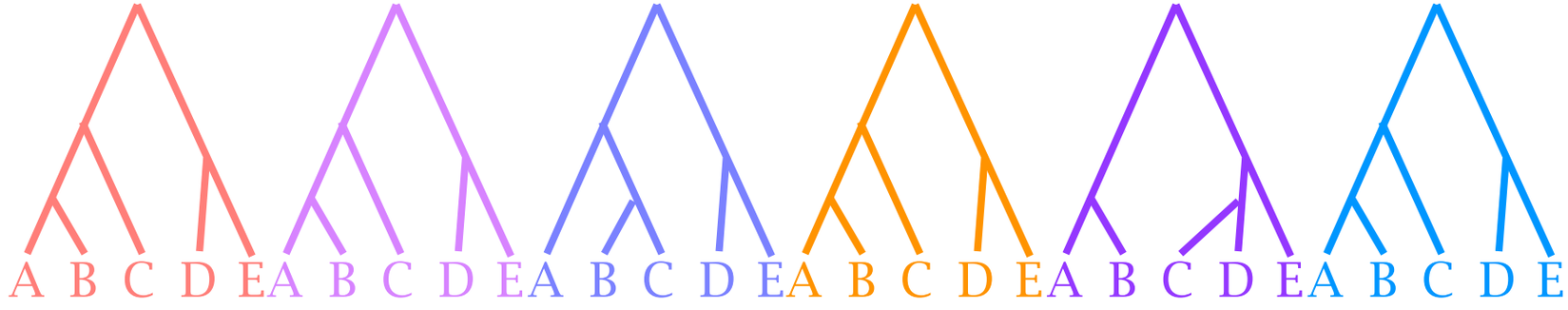
The post-genomic era



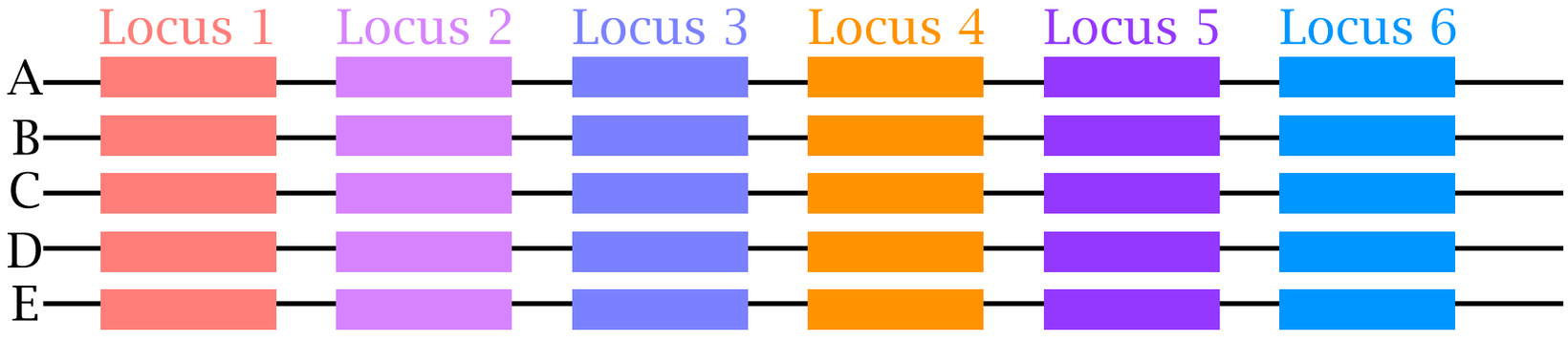
The post-genomic era



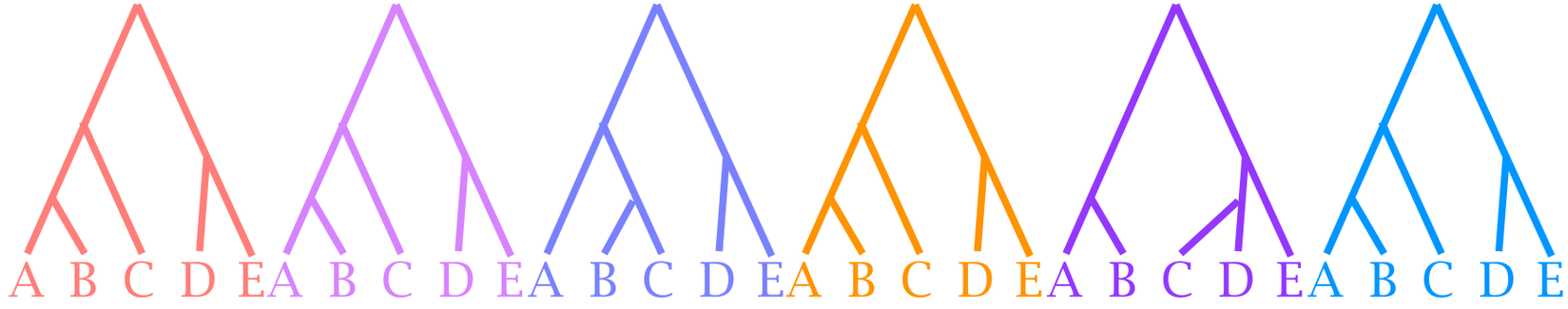
Gene Trees



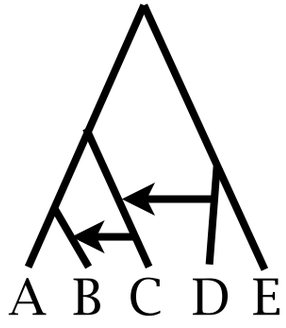
The post-genomic era



Gene Trees



Species Phylogeny

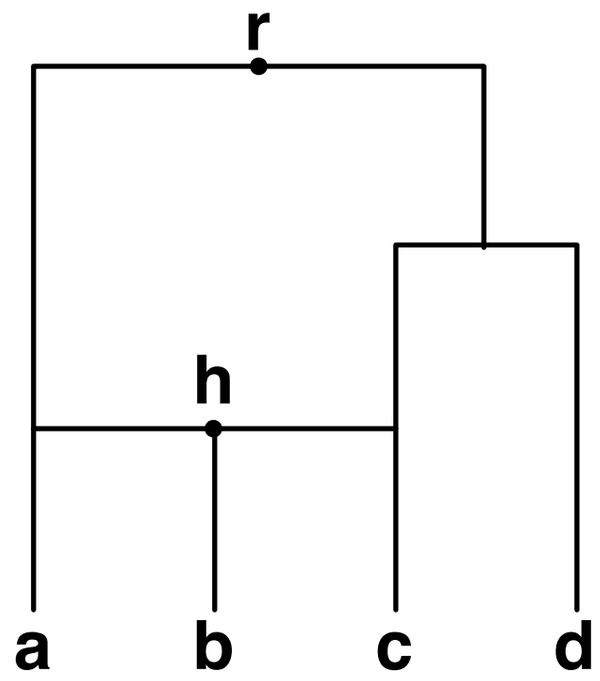


Outline of the Talk

- The phylogenetic network model
- From trees to networks
- From sequences to networks
- Should we build a network
- Summary

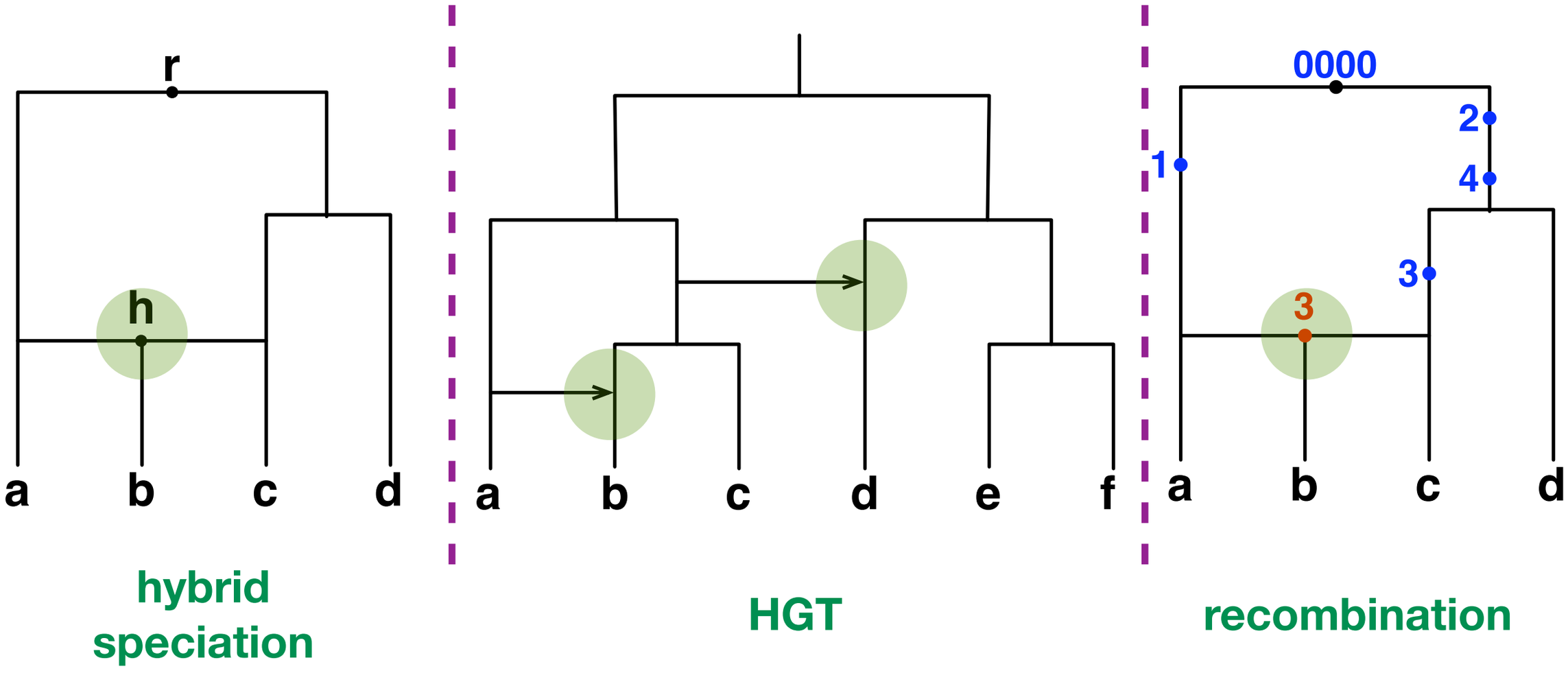
Phylogenetic Networks

- When HGT occurs, the evolutionary history reconstructed from the genomic sequences is more appropriately represented as a phylogenetic network



Phylogenetic Networks

- Phylogenetic networks generalize trees and allow for modeling vertical and non-vertical evolution in a variety of scenarios

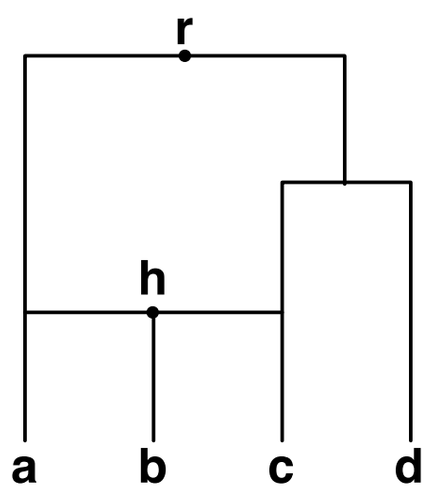


Phylogenetic Networks

A *phylogenetic network* N on set $\mathcal{X}$ of taxa is an ordered pair (G, f), where

- $G = (V, E)$ is a directed, acyclic graph (DAG) with $V = \{r\} \cup V_L \cup V_T \cup V_N$, where
  - $indeg(r) = 0$ (r is the *root* of N);
  - $\forall v \in V_L, indeg(v) = 1$ and $outdeg(v) = 0$ (V_L are the *leaves* of N);
  - $\forall v \in V_T, indeg(v) = 1$ and $outdeg(v) \geq 2$ (V_T are the *tree nodes* of N); and,
  - $\forall v \in V_N, indeg(v) = 2$ and $outdeg(v) = 1$ (V_N are the *reticulation nodes* of N),

  and $E \subseteq V \times V$ are the network's edges (we distinguish between *reticulation edges*, edges whose heads are reticulation nodes, and *tree edges*, edges whose heads are tree nodes.
- $f : V_L \to \mathcal{X}$ is the *leaf-labeling* function, which is a bijection from V_L to $\mathcal{X}$.



From Trees to Networks

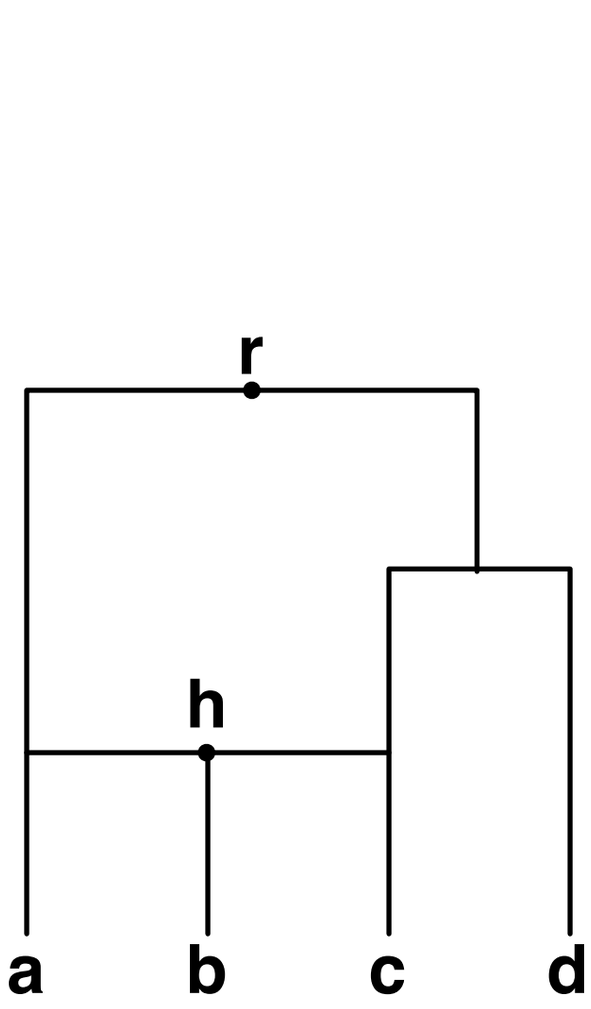
Central Observation

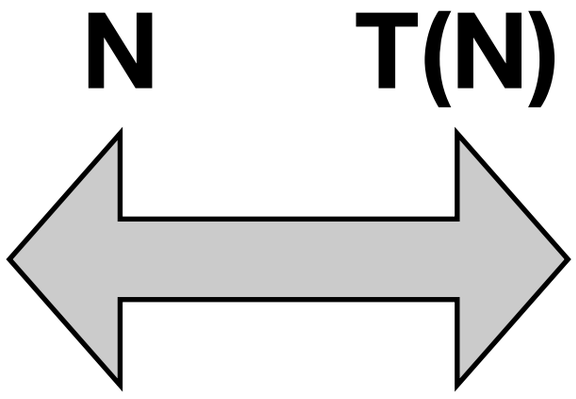
- At the lowest level of “atomicity”: every nucleotide in a genome has evolved down a tree

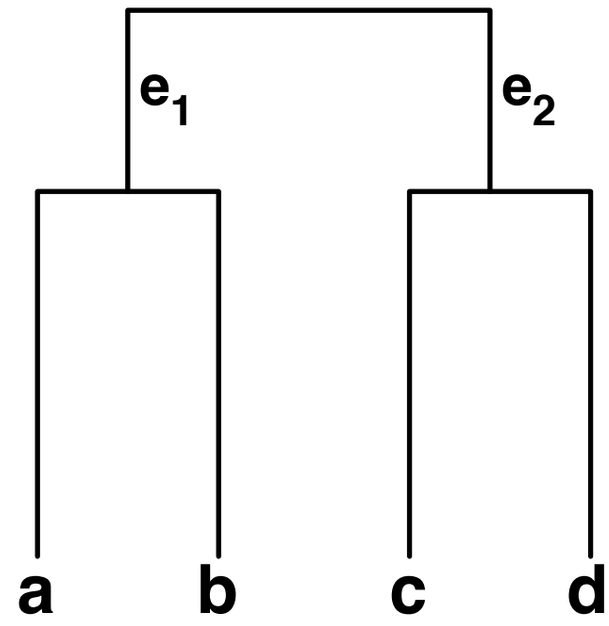
- More generally: barring recombination, the evolutionary history of an individual gene is treelike

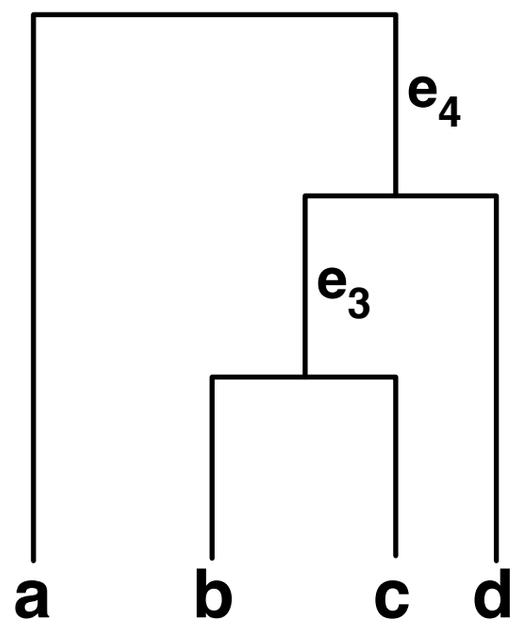
- Hence, a phylogenetic network is viewed as the reconciliation of the gene trees

Trees and Networks

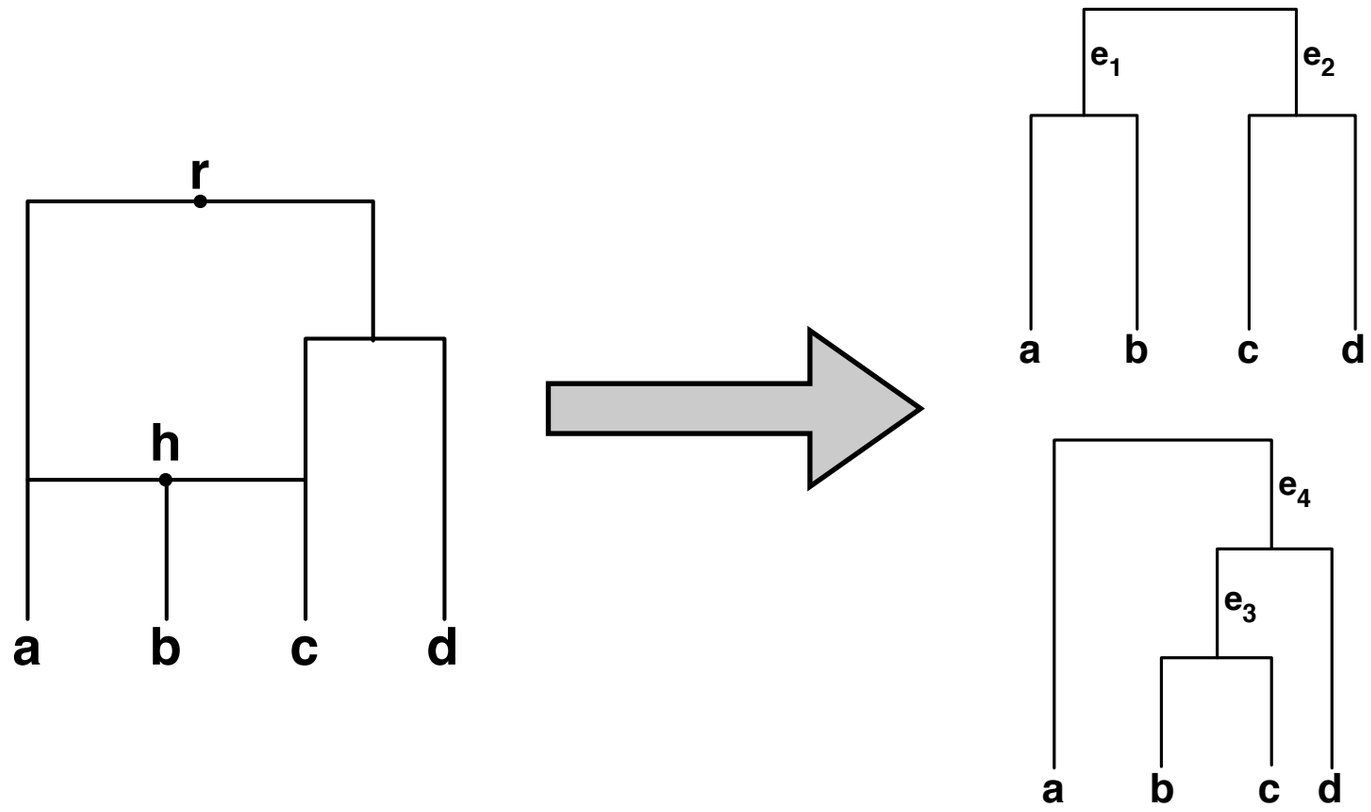








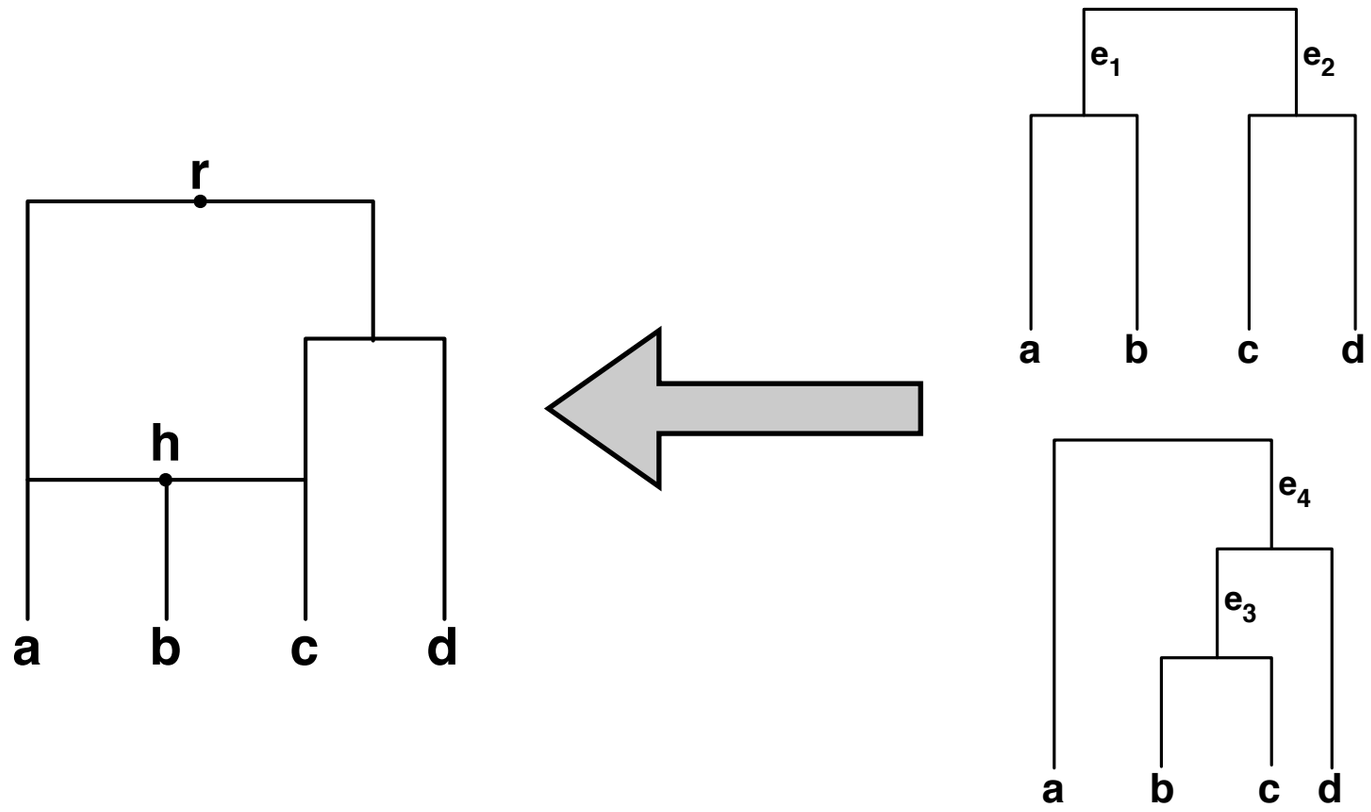
From a Network to Its Constituent Trees



- Tells about the different gene genealogies and sequence evolution (more later)

- Given a network, it is easy to compute the set of induced trees

From a Set of Trees to Their Containing Network

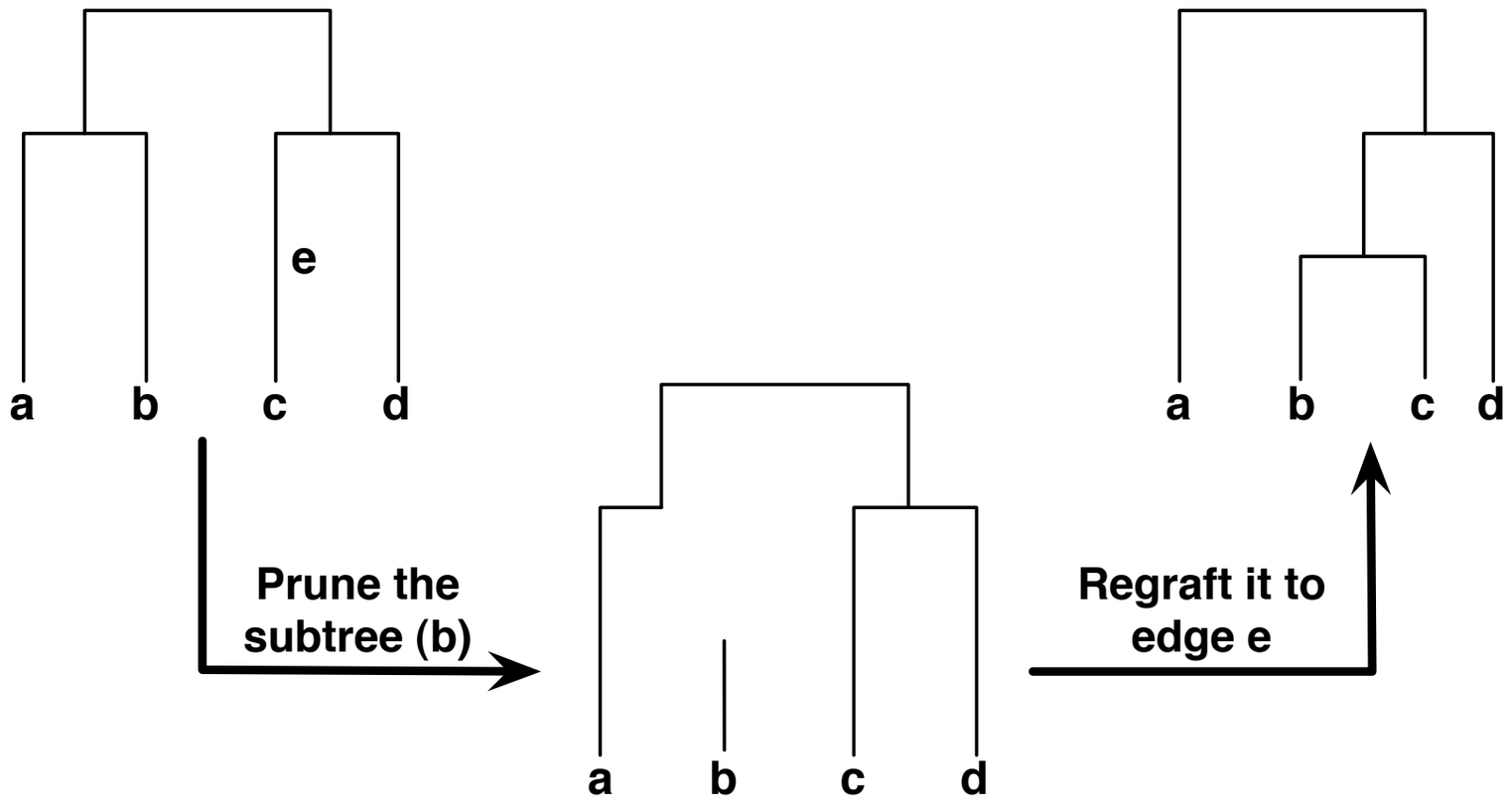


- Amounts to reconstructing the evolutionary history (of genomes, species, etc.)

- Given a set of trees, it is very hard (in general) to compute a “good” network that contains them

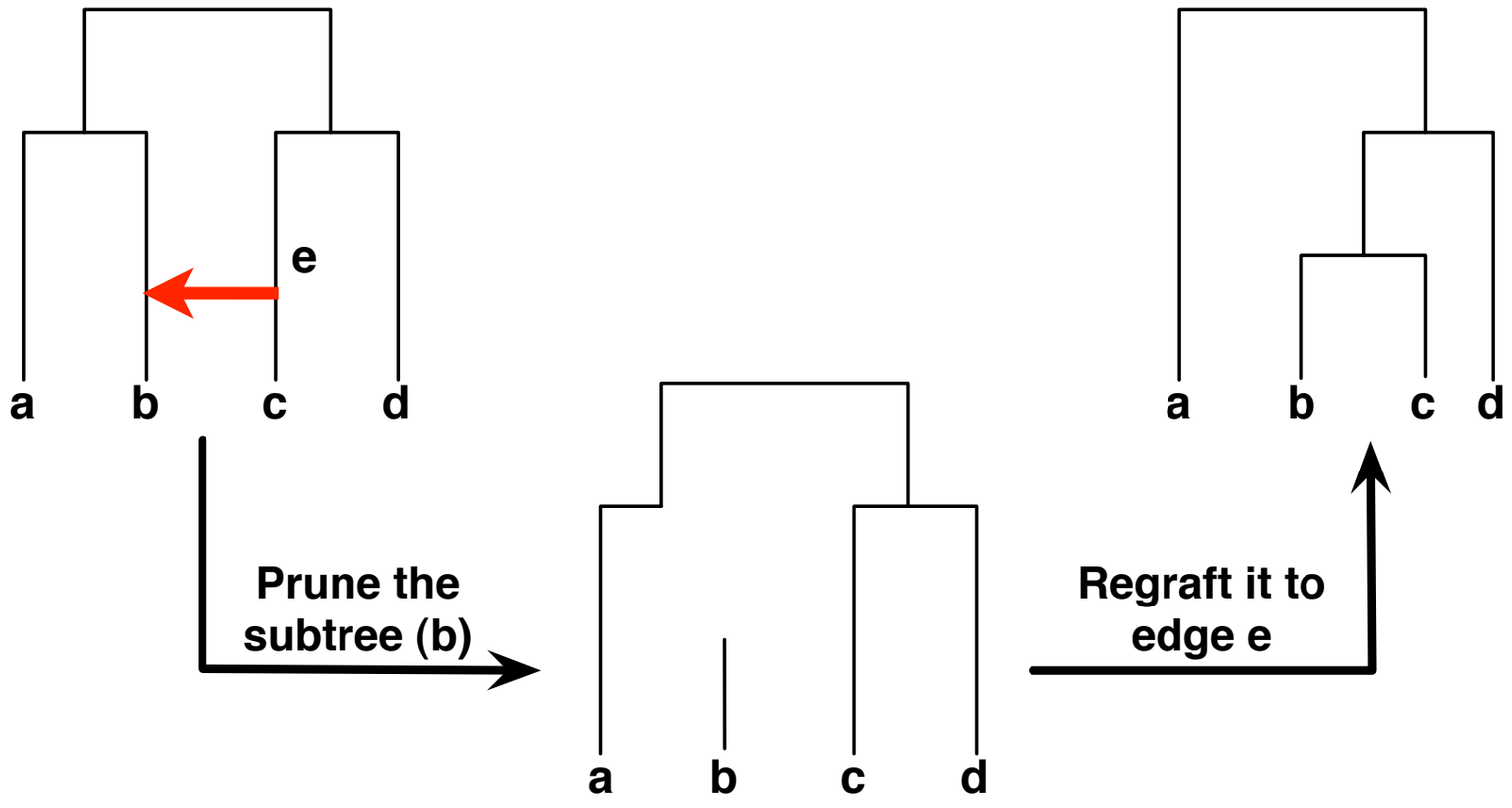
From a Set of Trees to Their Containing Network

- The Subtree Prune and Regraft (SPR) operation mimics the effect of a reticulation event



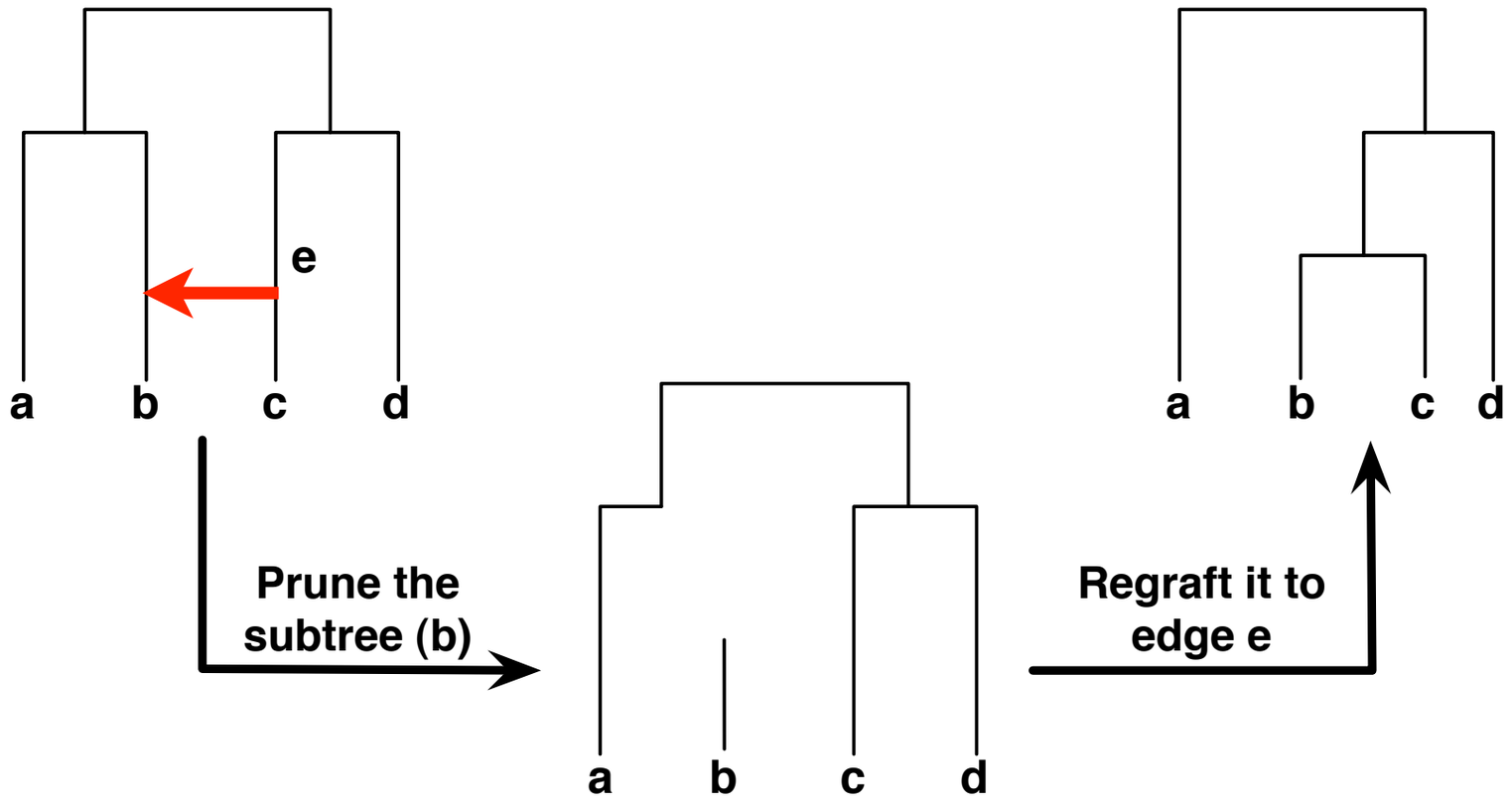
From a Set of Trees to Their Containing Network

- The Subtree Prune and Regraft (SPR) operation mimics the effect of a reticulation event



From a Set of Trees to Their Containing Network

- The Subtree Prune and Regraft (SPR) operation mimics the effect of a reticulation event



- The SPR distance (the minimum number of SPR moves required to transform one tree into another) is taken as a proxy for the (minimum) number of reticulation events

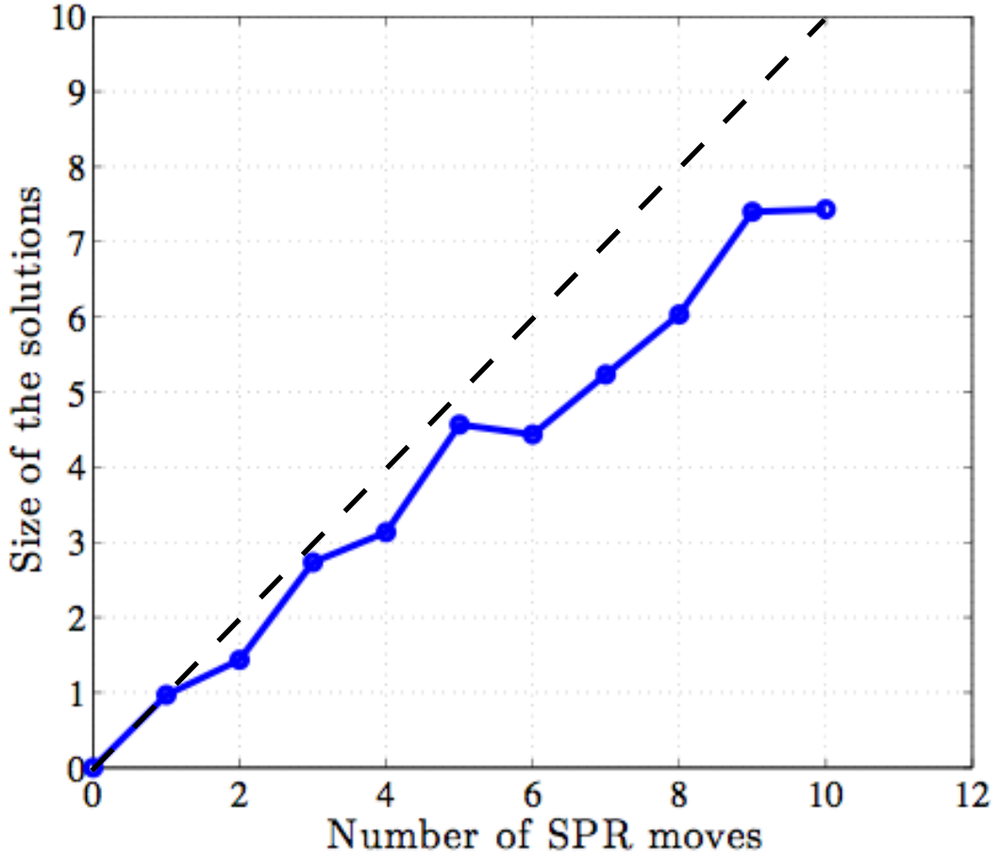
Programs for Computing (exactly or heuristically) the SPR Distance

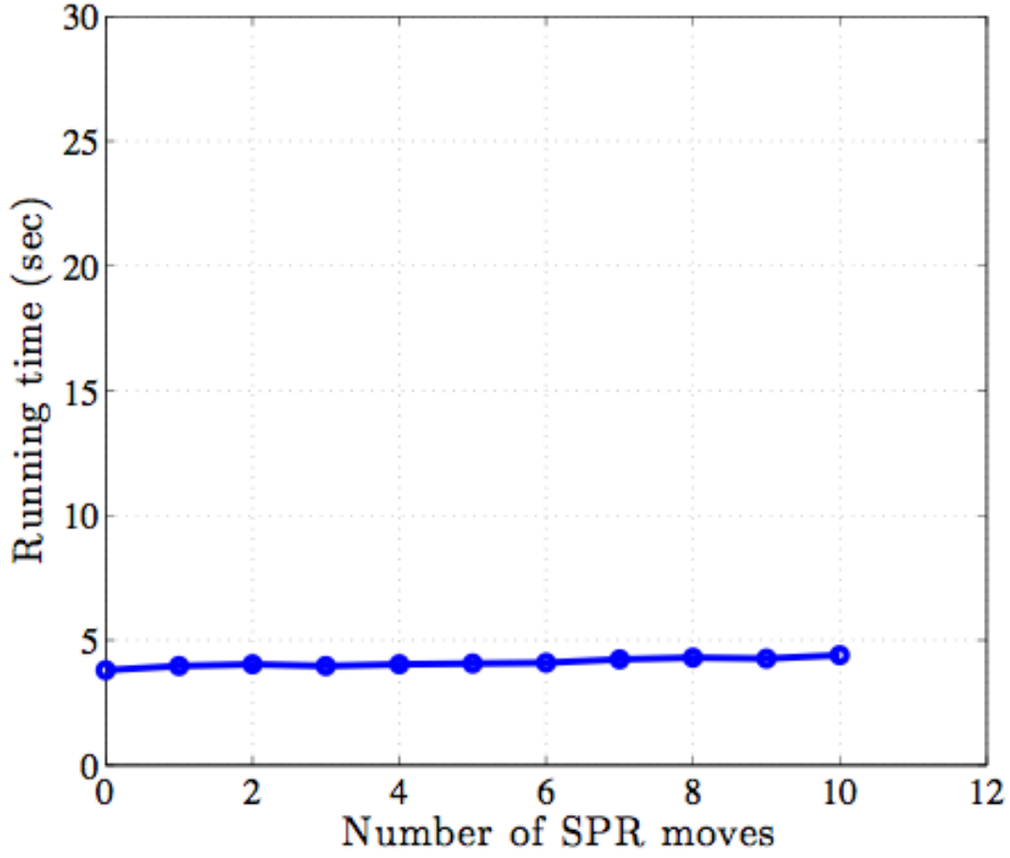
- EEEP: Beiko and Hamilton.
- HorizStory: MacLeod, Charlebois, Doolittle, and Bapteste.
- HorizTrans: Hallett and Lagergren.
- RIATA-HGT: Nakhleh, Ruths, and Wang.
- SPRDist: Wu.
- TNT: Goloboff.
- ...

The SPR Distance

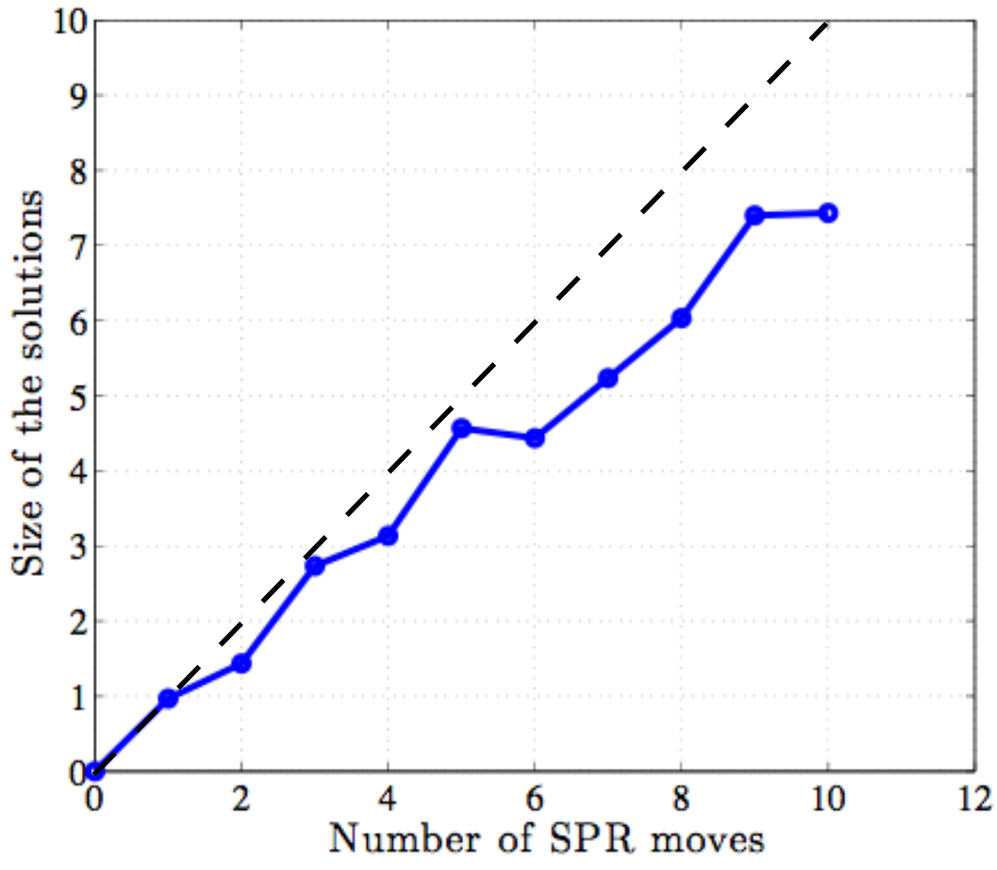
- Very hard to compute (NP-hard)

- Several fast heuristics exist, with very good performance in practice, including our own RIATA-HGT

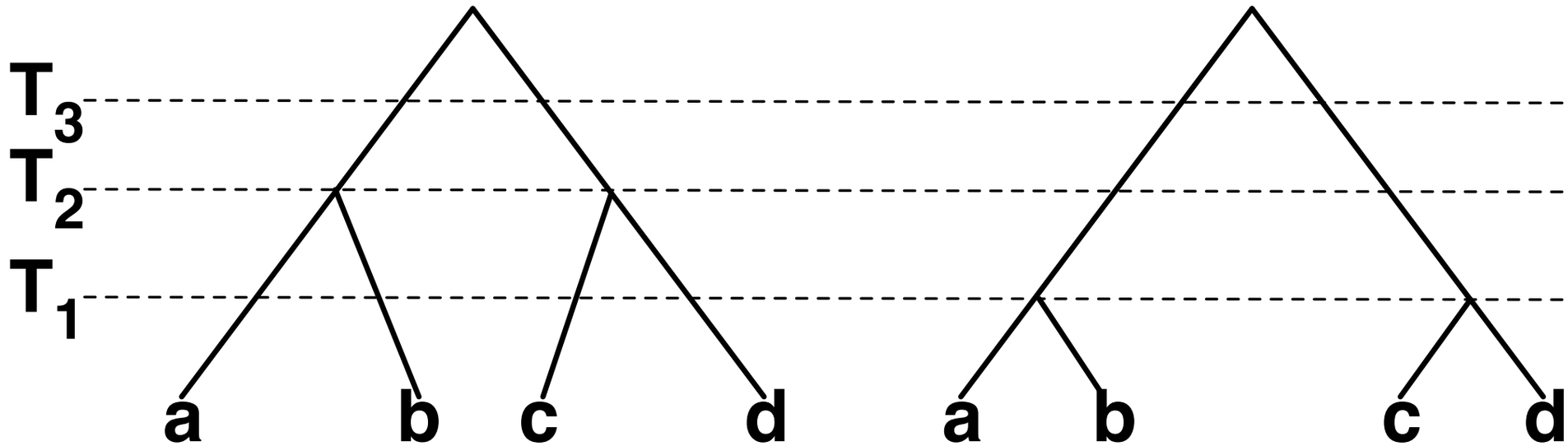




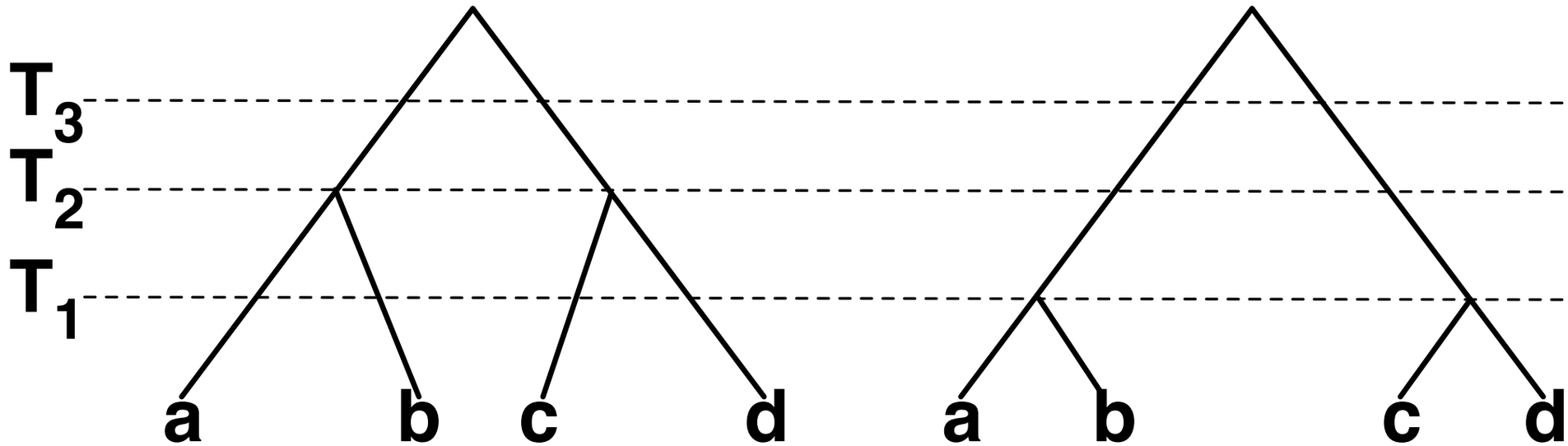
Issues with the SPR Distance:
(1) Underestimation

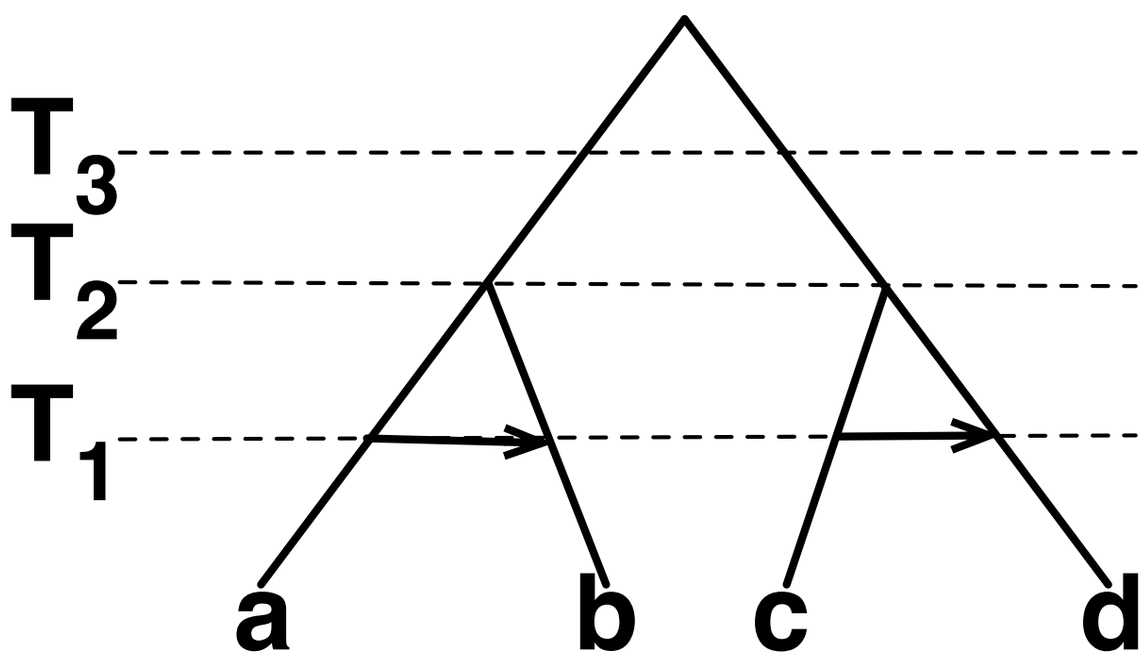


Issues with the SPR Distance:
(2) Ordered Trees

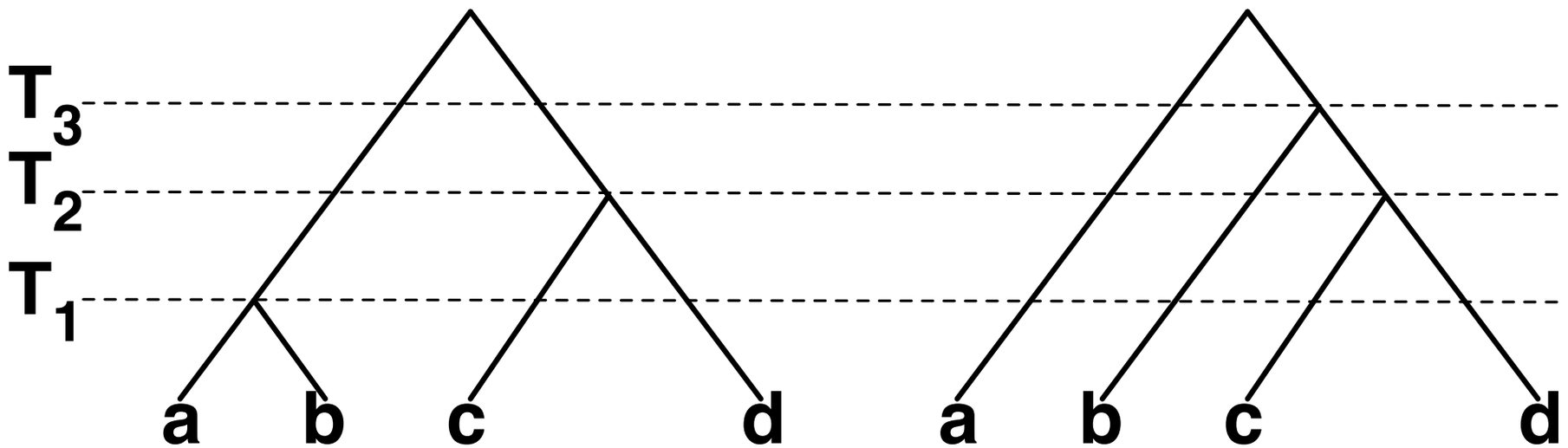


Issues with the SPR Distance:
(2) Ordered Trees

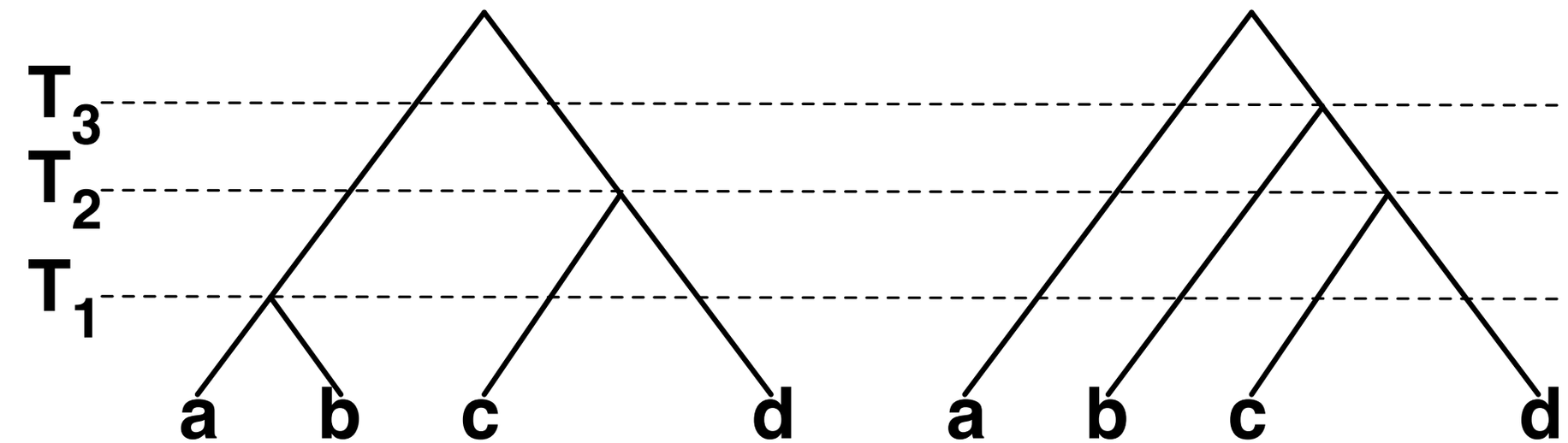


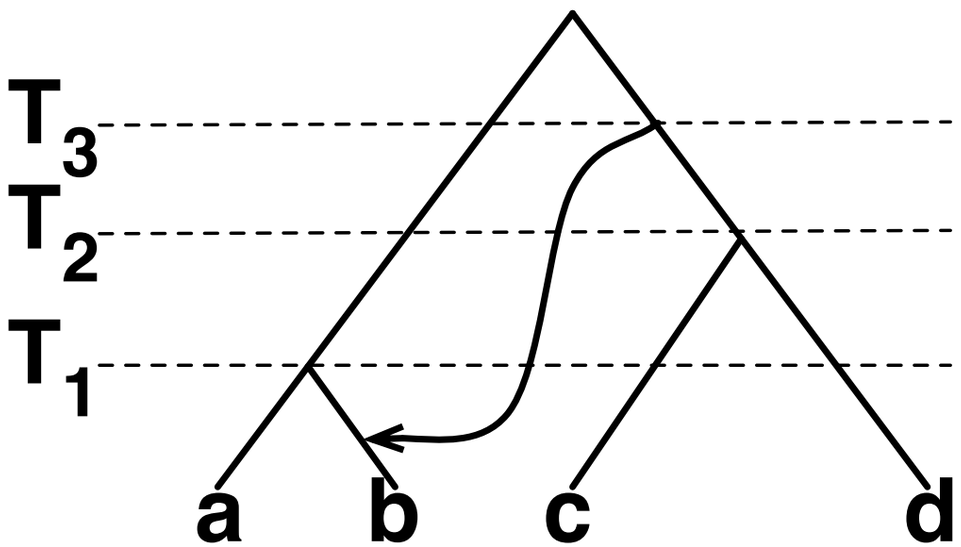


Issues with the SPR Distance:
(3) Time-inconsistent Moves

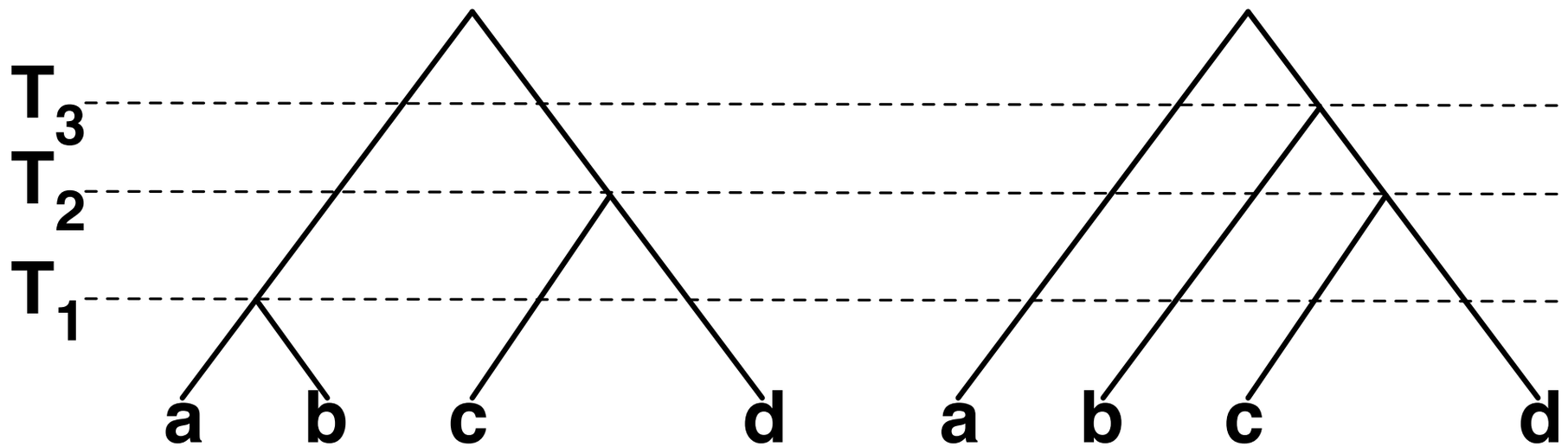


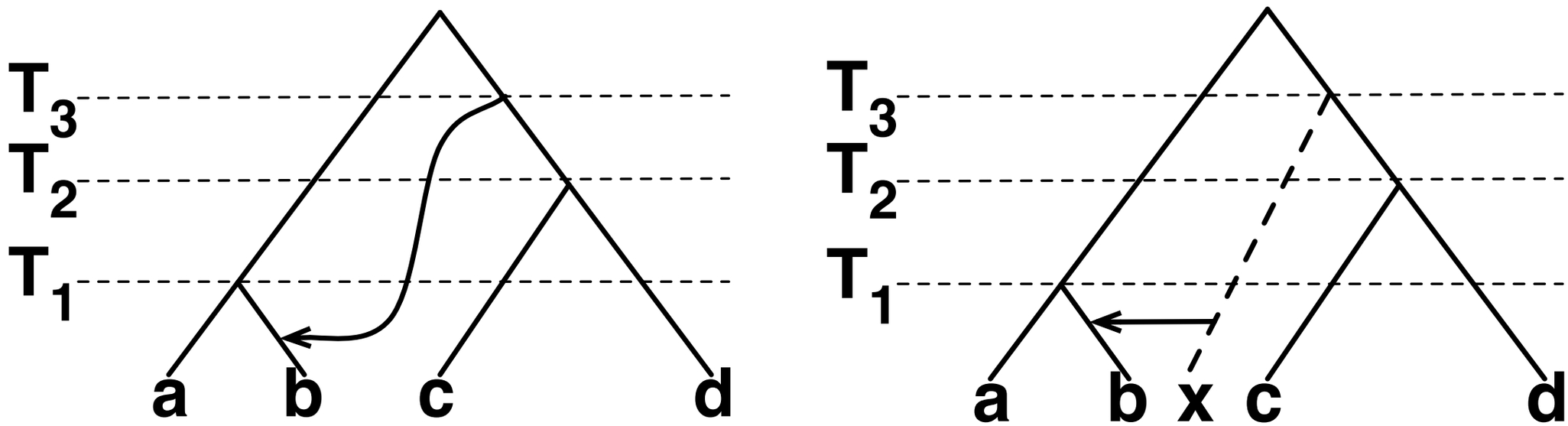
Issues with the SPR Distance:
(3) Time-inconsistent Moves



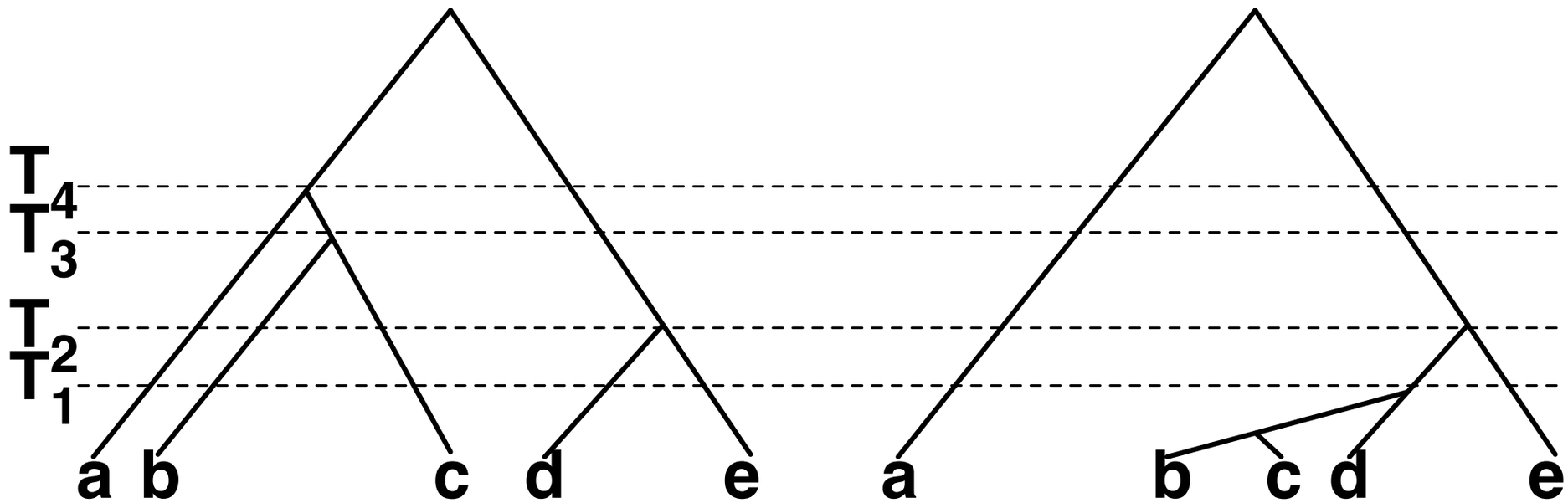


Issues with the SPR Distance:
(3) Time-inconsistent Moves

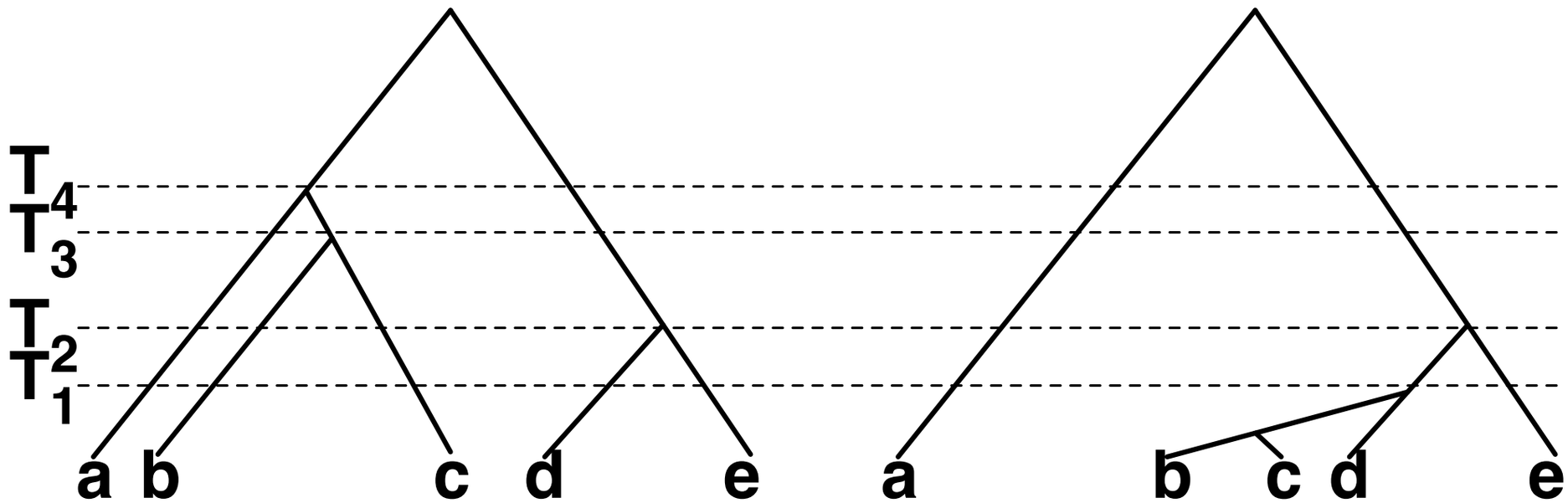


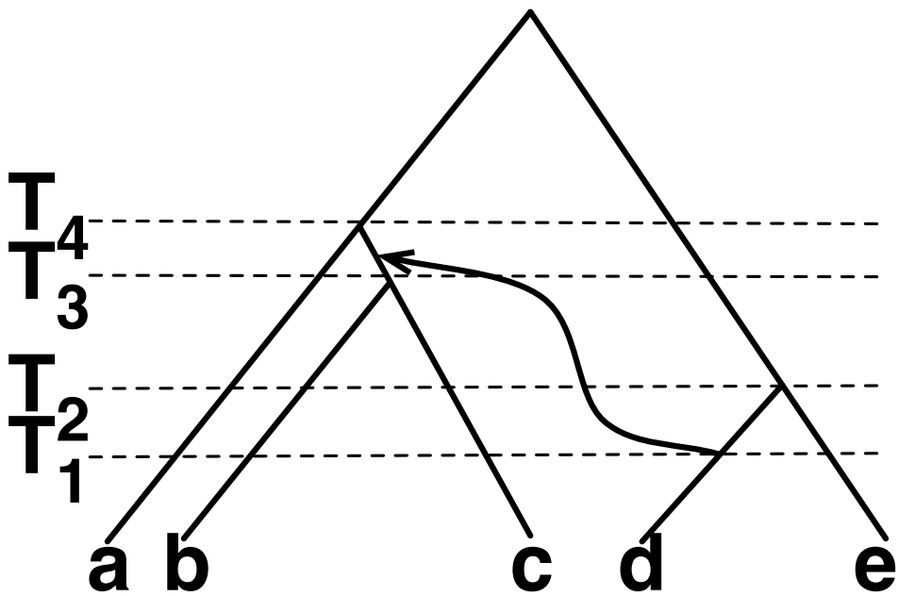


Issues with the SPR Distance:
(3) Time-inconsistent Moves

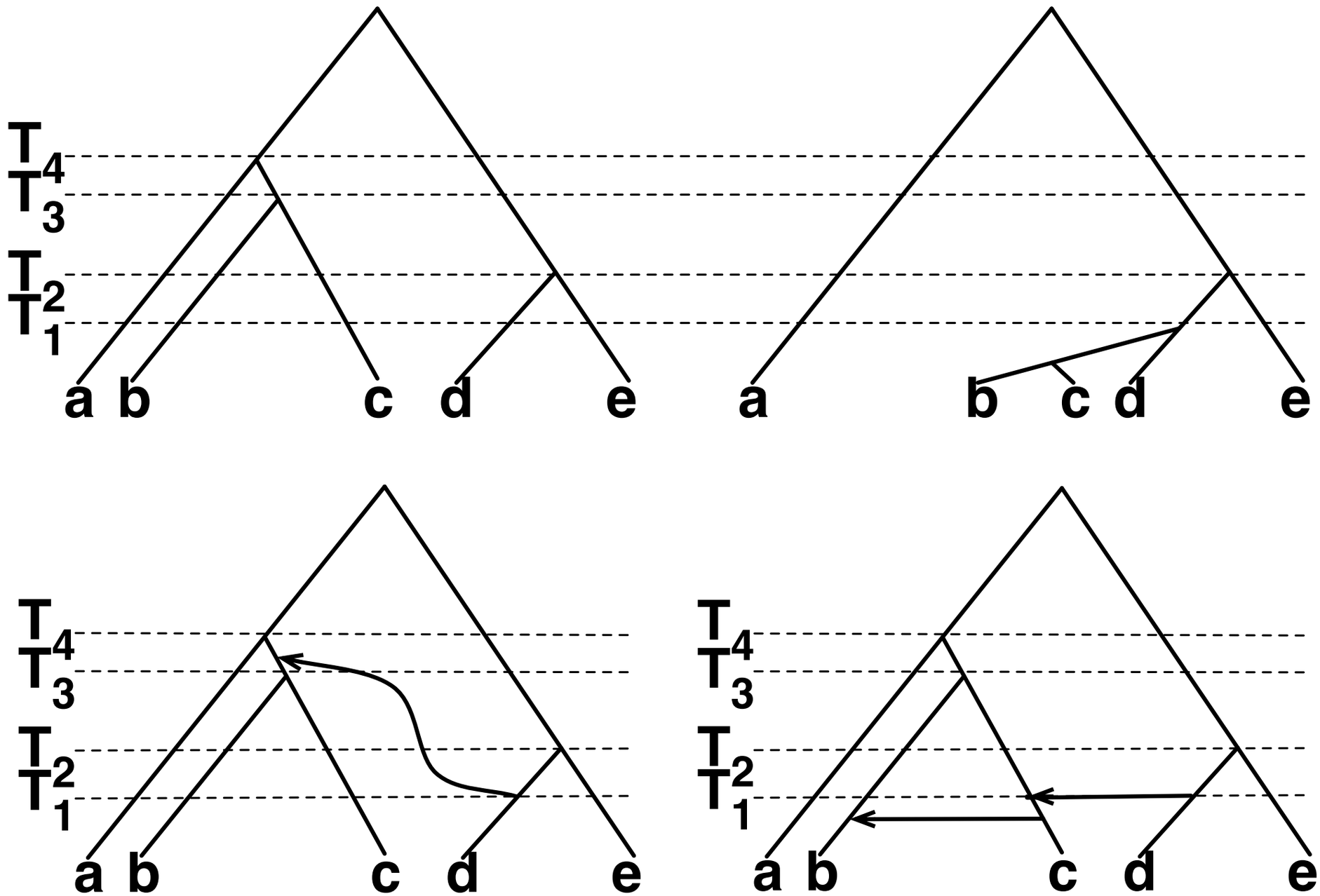


Issues with the SPR Distance:
(3) Time-inconsistent Moves





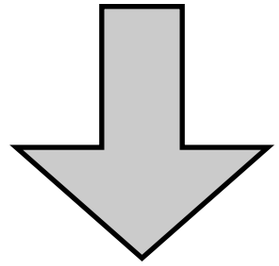
Issues with the SPR Distance:
(3) Time-inconsistent Moves



Issues with the SPR Distance:
(4) Multiple Trees

- Recall:

Gene Trees



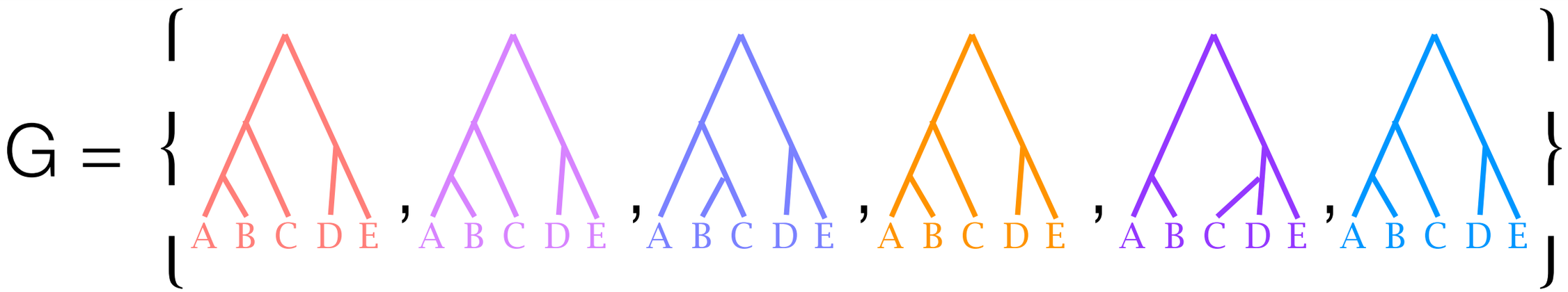
Species
Phylogeny

Issues with the SPR Distance:
(4) Multiple Trees

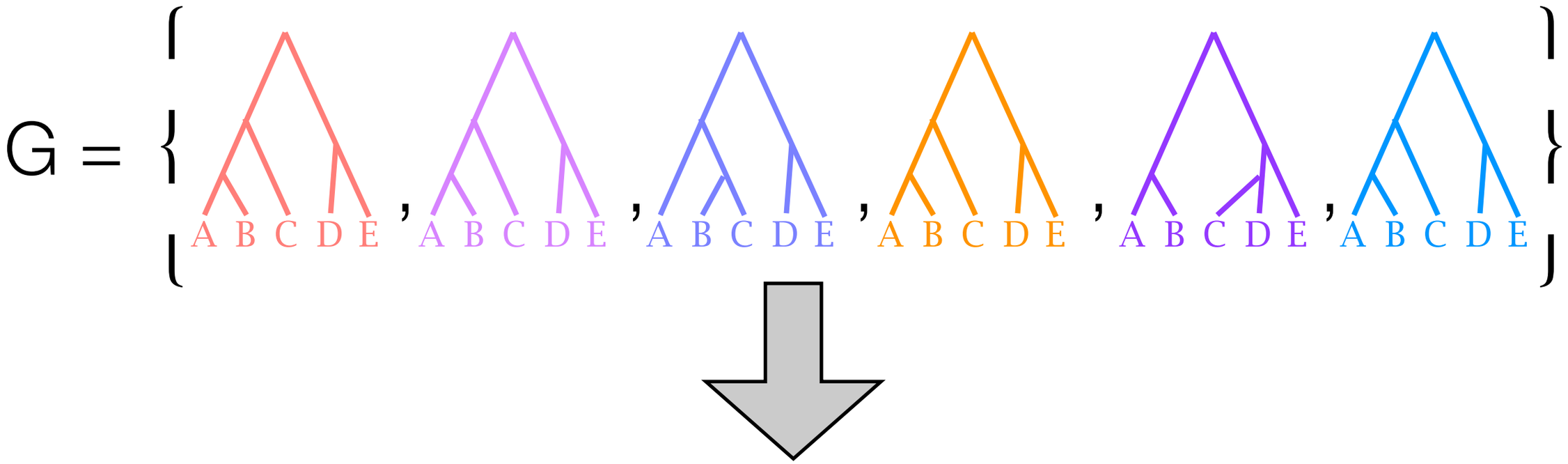
- However, SPR is defined on a pair of trees.

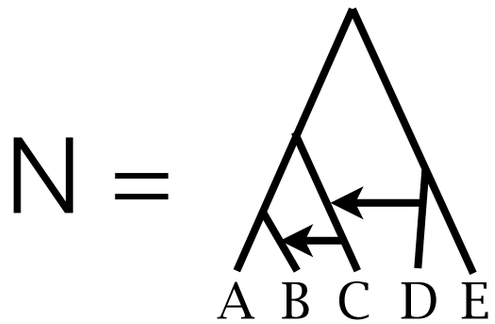
- The problem now becomes: Given an input set G of gene trees, find a phylogenetic network N with the minimum number of reticulation nodes such that G⊆T(N).

Issues with the SPR Distance:
(4) Multiple Trees

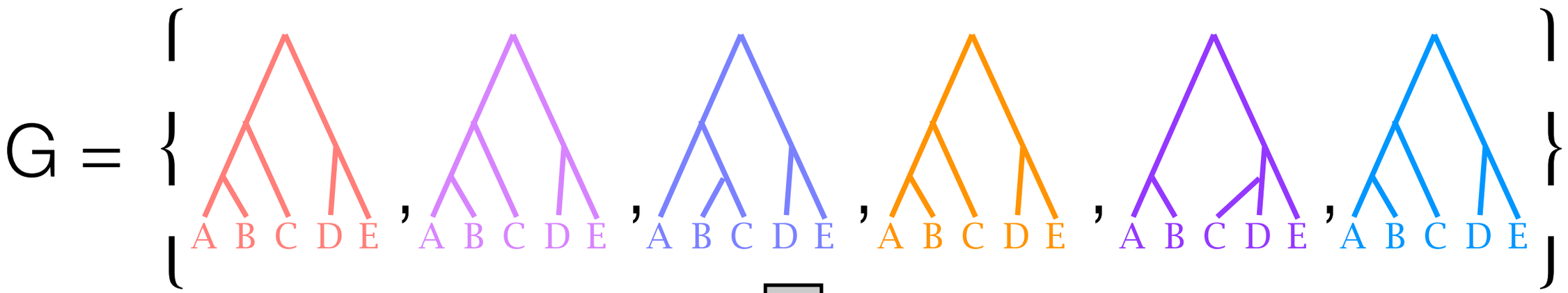


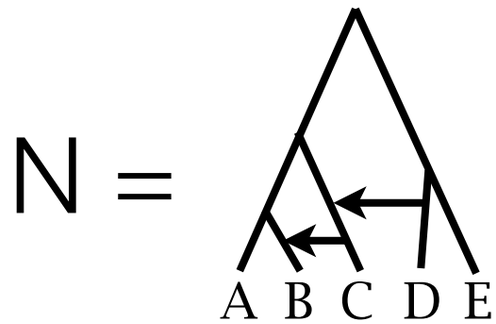
Issues with the SPR Distance:
(4) Multiple Trees

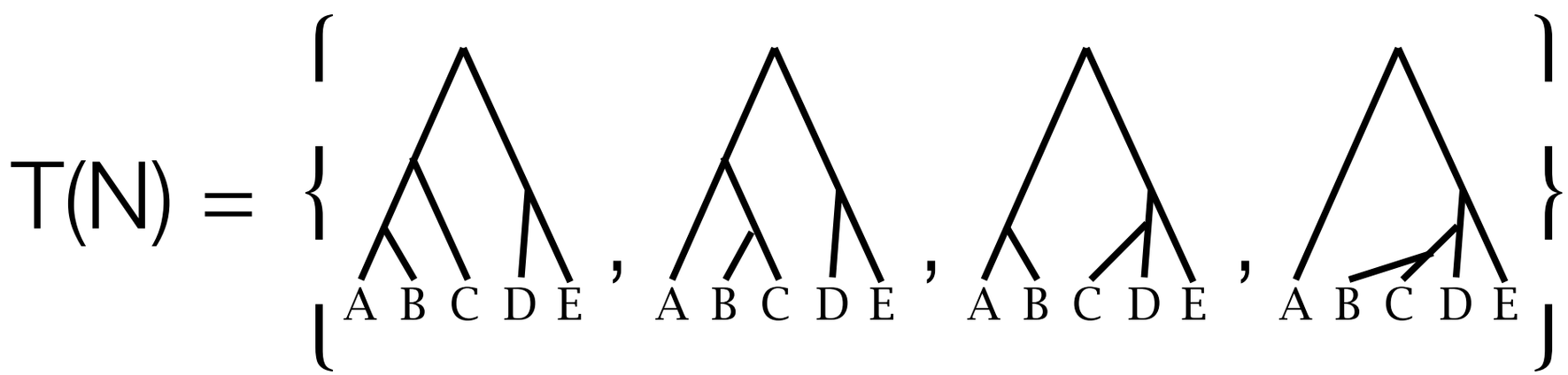




Issues with the SPR Distance:
(4) Multiple Trees







Issues with the SPR Distance:
(4) Multiple Trees

- Programs that allow for multiple trees in the input:
    - CASS: van Iersel and Kelk.
    - MURPAR: Park, Jin, and Nakhleh
    - PIRN: Wu.

Issues with the SPR Distance:
(5) Unknown Species Tree

- To guarantee that an SPR move reflects an HGT event, it must be computed on a gene tree with respect to the species tree.

- In practice, the species tree may not be known.

- Heuristics:

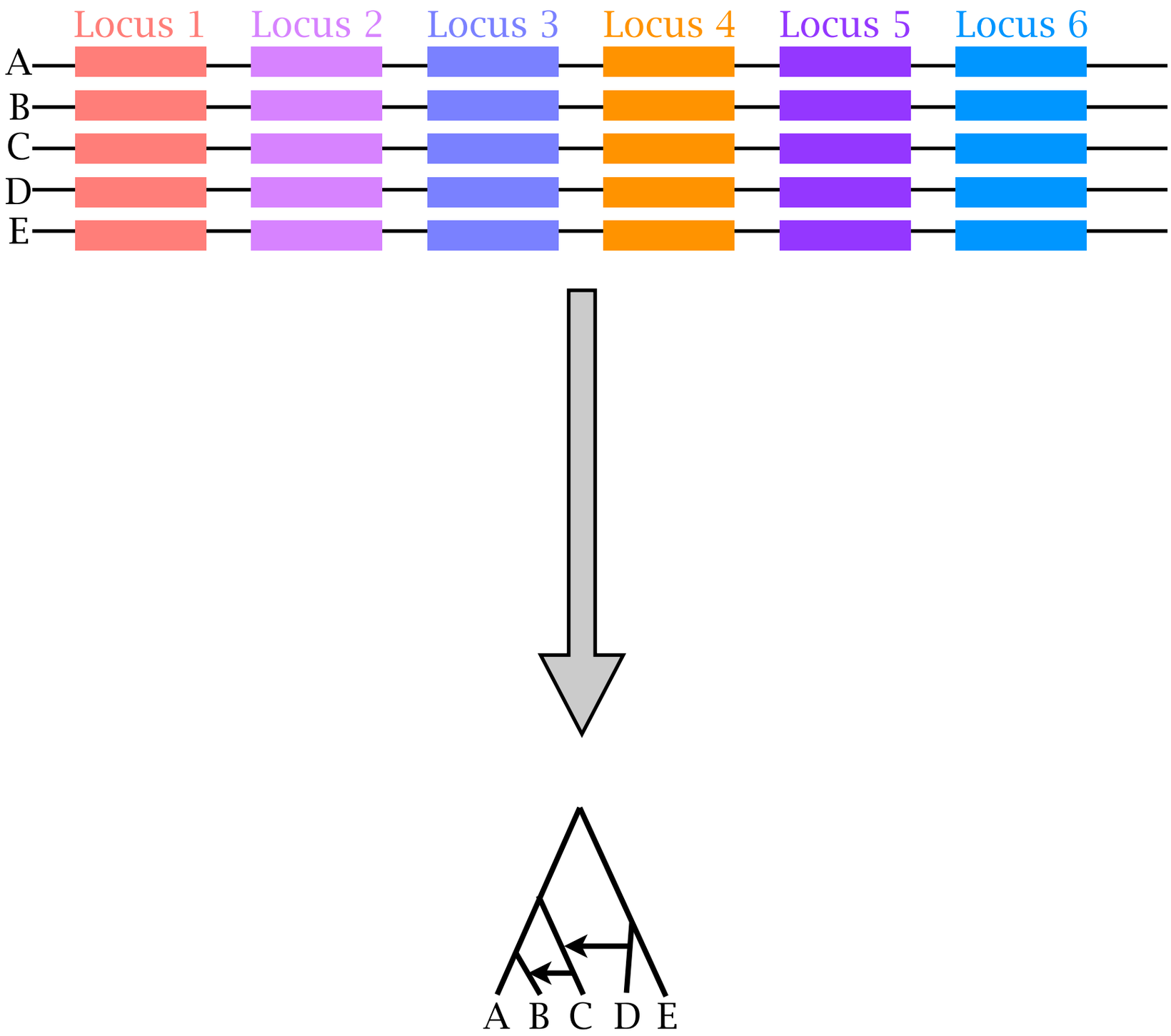
  - Take the consensus of all gene trees to be the candidate for species tree (Warning: May necessitate dealing with non-binary trees).

  - Take the gene tree with the highest frequency to be the candidate for species tree (May be problematic under certain settings).

  - Try each of the gene trees as a species tree candidate, infer networks, and choose the one that is optimal over all choices of gene trees.

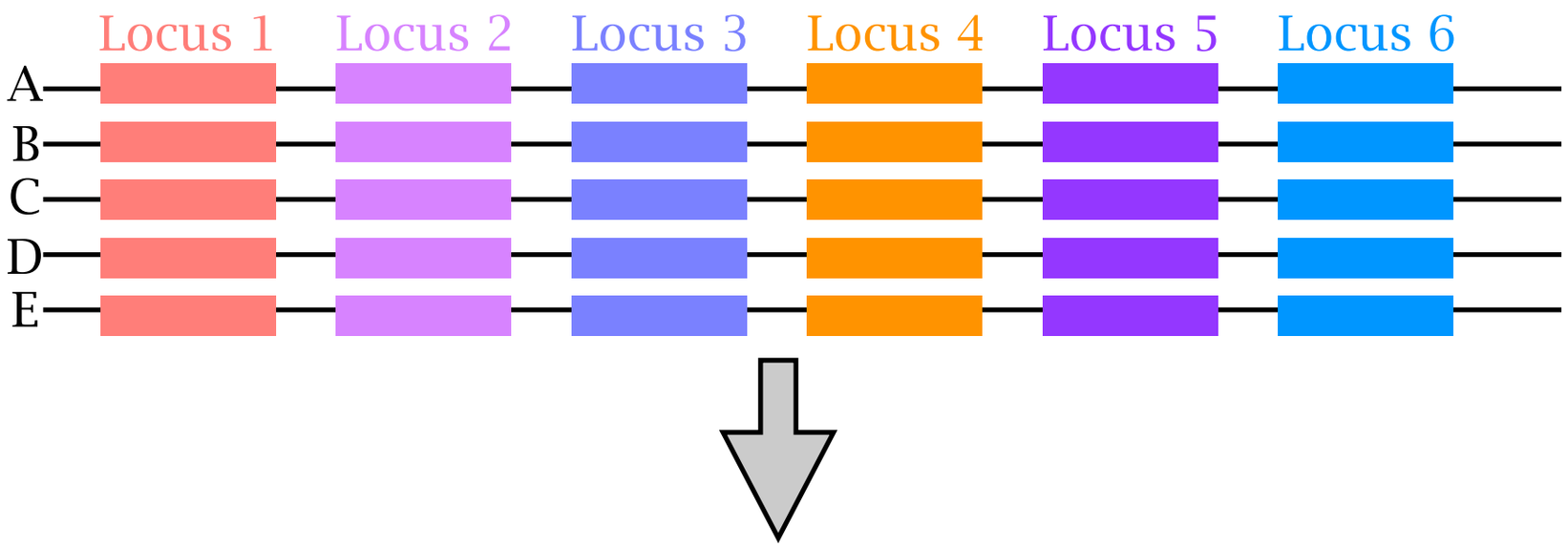
From Sequences to Networks

Recall: The actual phylogenetic network reconstruction problem is...



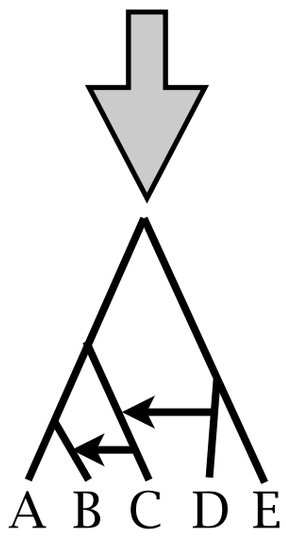
Species Phylogeny

The approach we have shown thus far is...



Gene Trees

Species Phylogeny



Optimization Criteria in Phylogenetics

- Maximum parsimony
- Character compatibility
- Maximum likelihood
- ...

Optimization Criteria in Phylogenetics

- Maximum parsimony
- Character compatibility
- Maximum likelihood
- ...

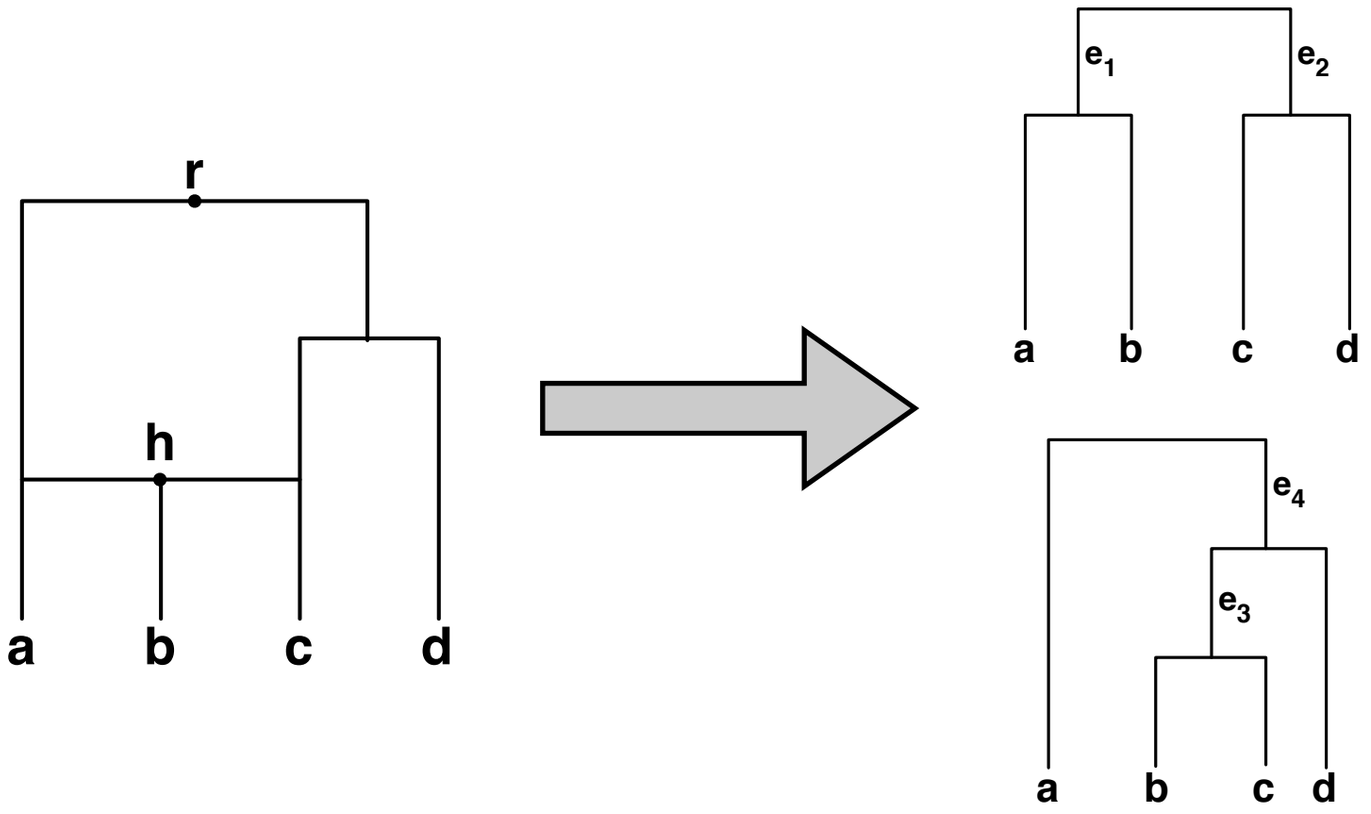
Question: How do we generalize these to network?

Generalizing Optimization Criteria to Networks

- Back to the central observation...

Generalizing Optimization Criteria to Networks

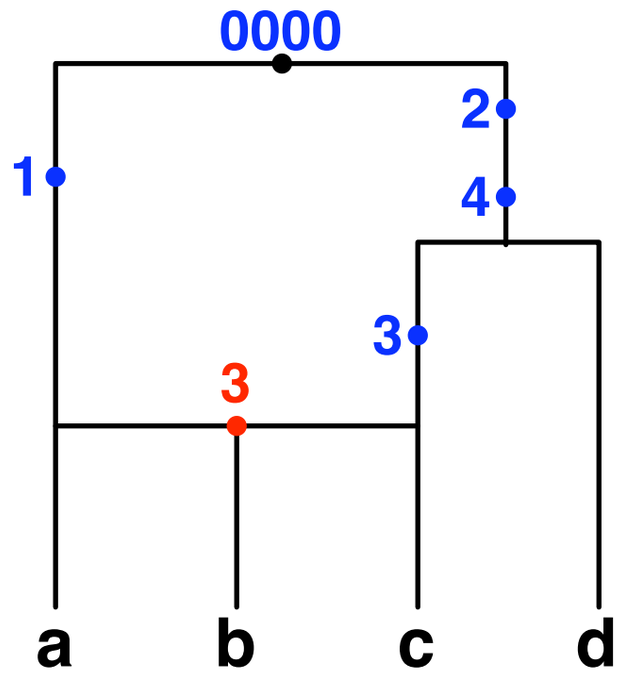
- Back to the central observation...



Generalizing Optimization Criteria to Networks

- The evolution of a site (or, more practically, a block of sites) on a network is best represented by one of the trees inside the network

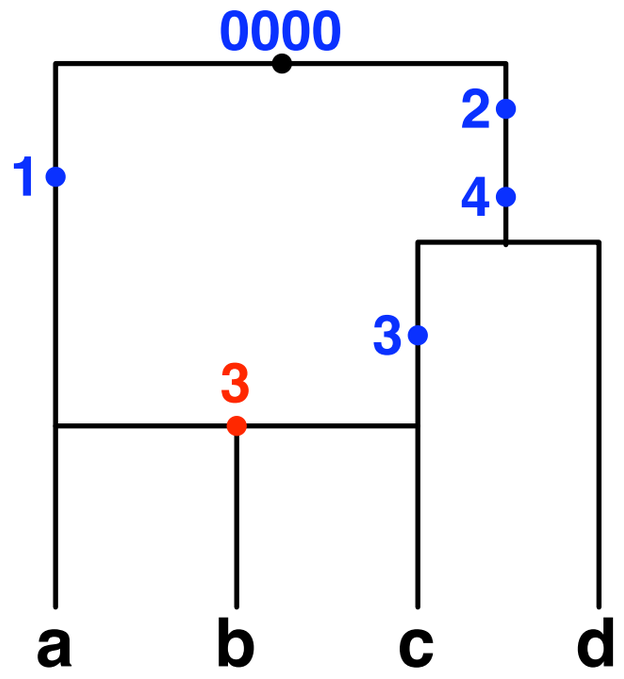
	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1

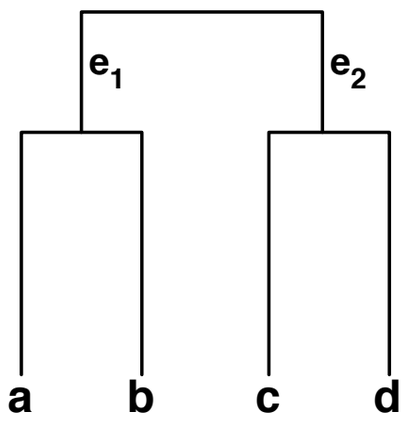


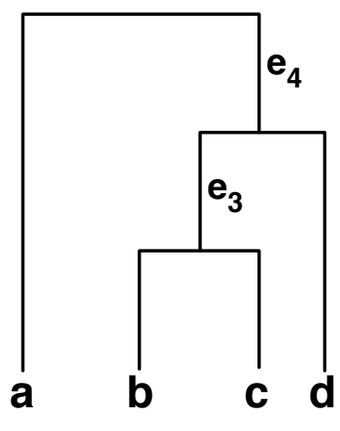
Generalizing Optimization Criteria to Networks

- The evolution of a site (or, more practically, a block of sites) on a network is best represented by one of the trees inside the network

	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1



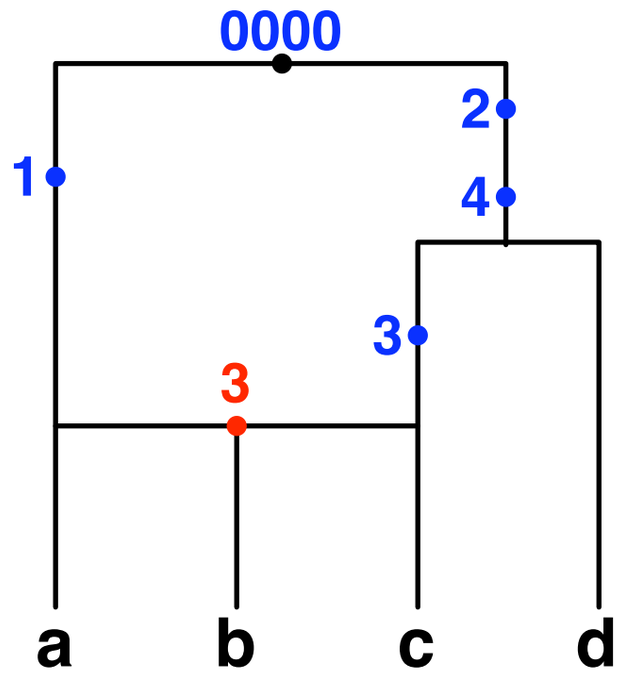


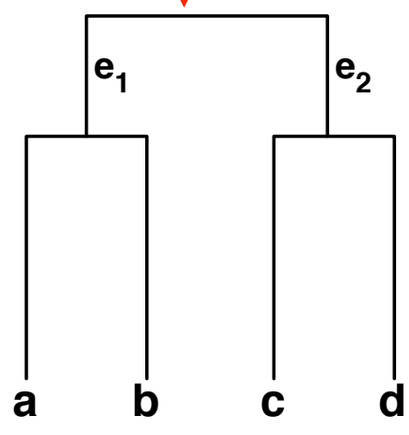


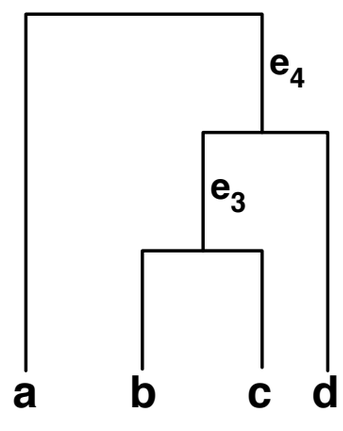
Generalizing Optimization Criteria to Networks

- The evolution of a site (or, more practically, a block of sites) on a network is best represented by one of the trees inside the network

	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1



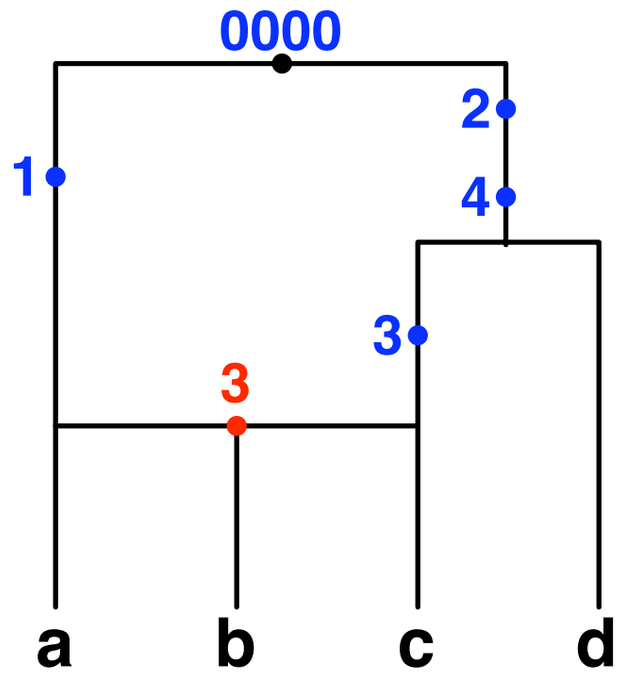


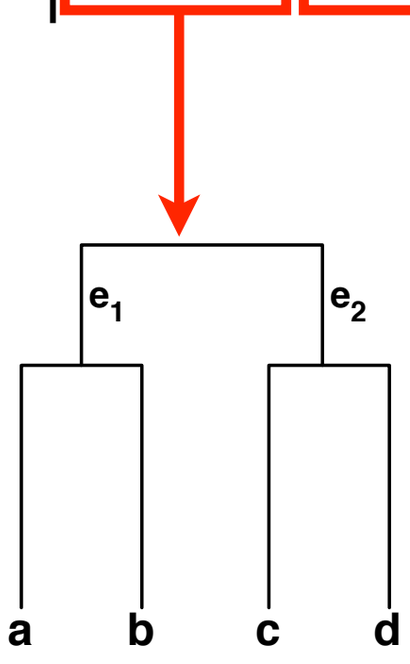


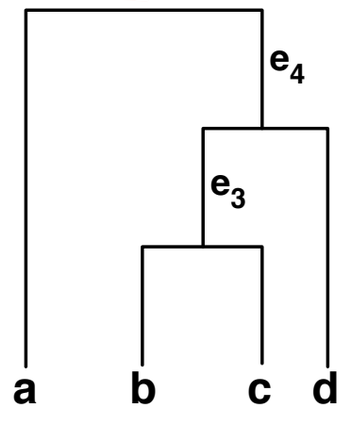
Generalizing Optimization Criteria to Networks

- The evolution of a site (or, more practically, a block of sites) on a network is best represented by one of the trees inside the network

	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1





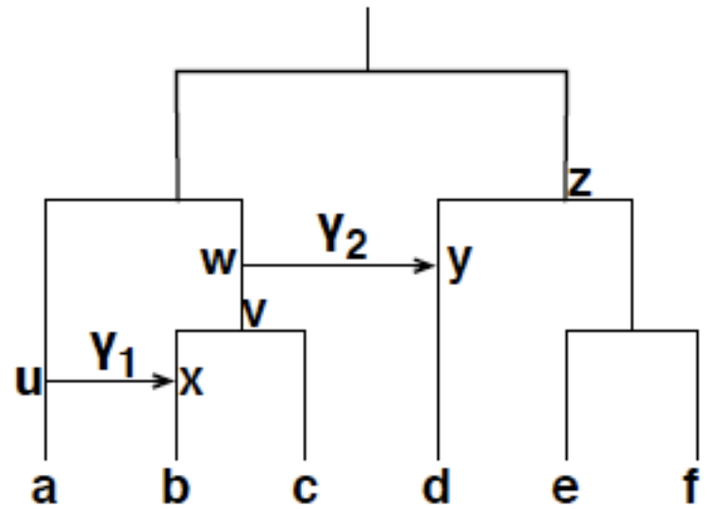


Maximum Parsimony on Phylogenetic Networks

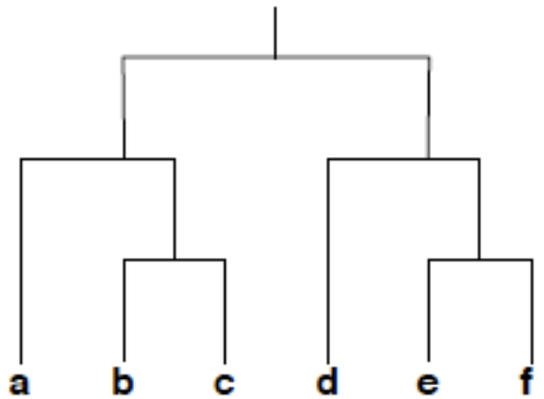
$$PS(N,S) = \sum_{S_i \in S} \left[\min_{T \in T(N)} PS(T, S_i) \right]$$

$$N^* = \text{argmin}_N PS(N,S)$$

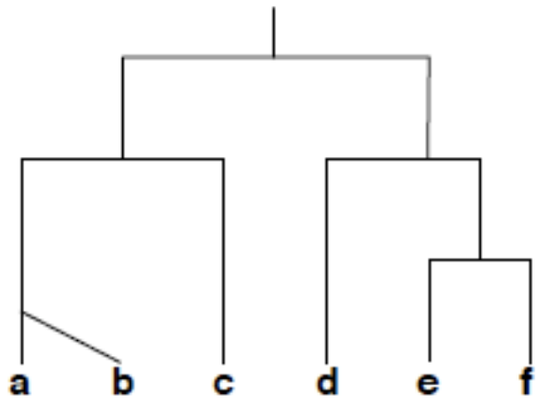
Maximum Likelihood on Phylogenetic Networks



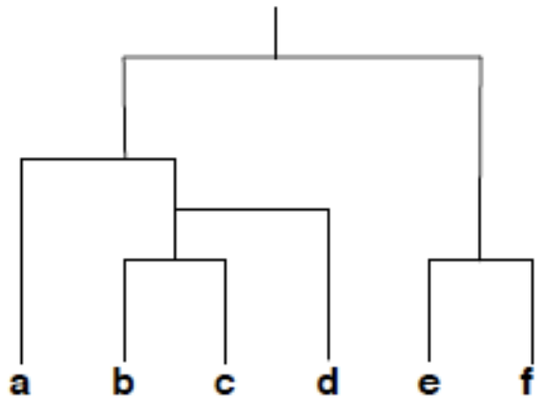
N



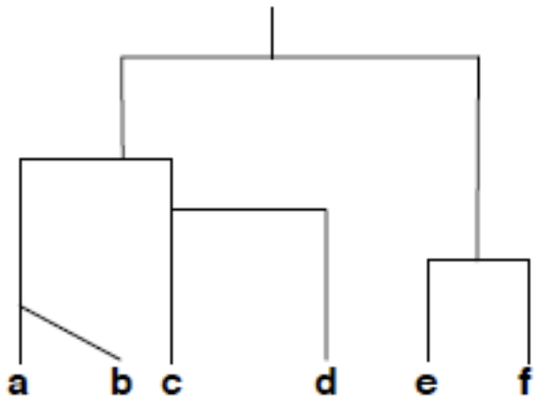
T_1



T_2

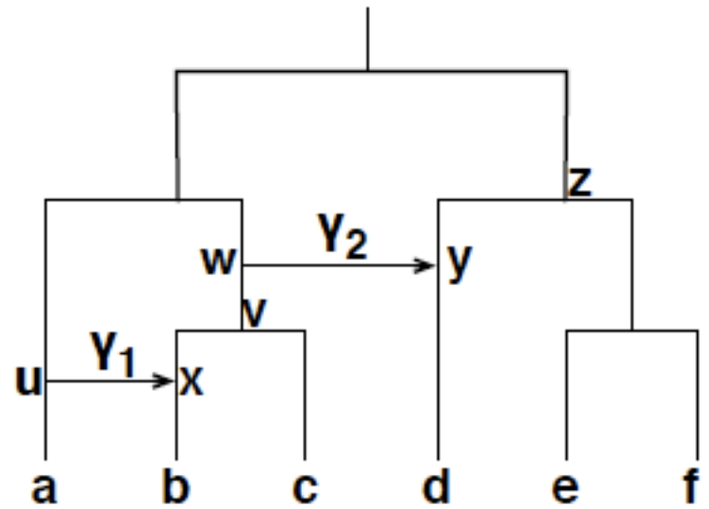


T_3

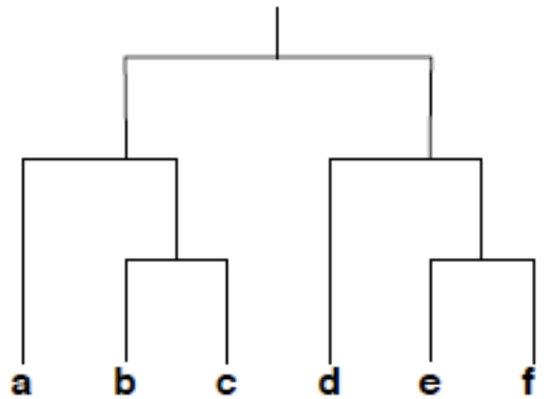


T_4

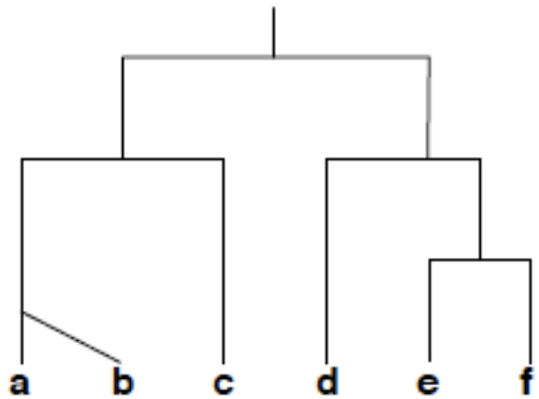
Maximum Likelihood on Phylogenetic Networks



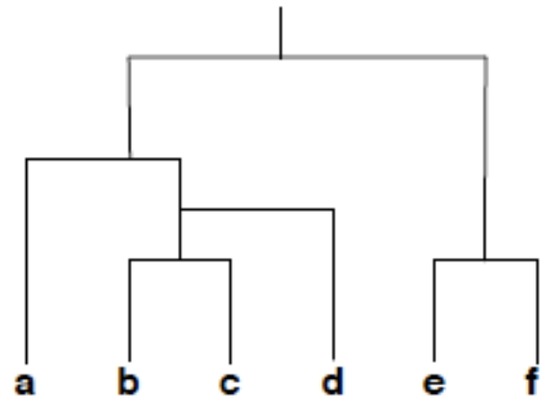
N



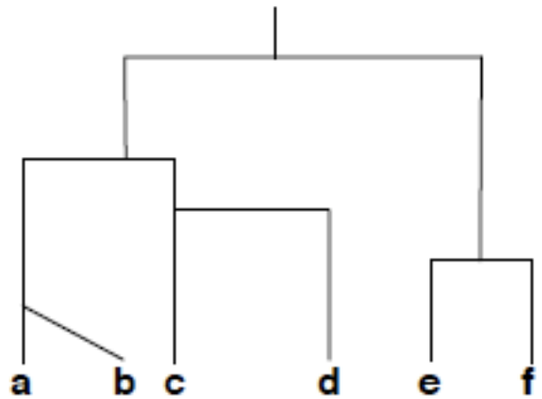
T_1



T_2



T_3



T_4

$P(T_1|N,\Gamma) = (1-\gamma_1)(1-\gamma_2)$ $P(T_2|N,\Gamma) = \gamma_1(1-\gamma_2)$ $P(T_3|N,\Gamma) = (1-\gamma_1)\gamma_2$ $P(T_4|N,\Gamma) = \gamma_1\gamma_2$

Maximum Likelihood on Phylogenetic Networks

$$L(N, \Gamma, \lambda; S) = P(S|N, \Gamma, \lambda) = \prod_{S_i \in S} \left[\sum_{T \in T(N)} [\mathbf{P}(S_i|T, \lambda) \cdot \mathbf{P}(T|N, \Gamma)] \right]$$

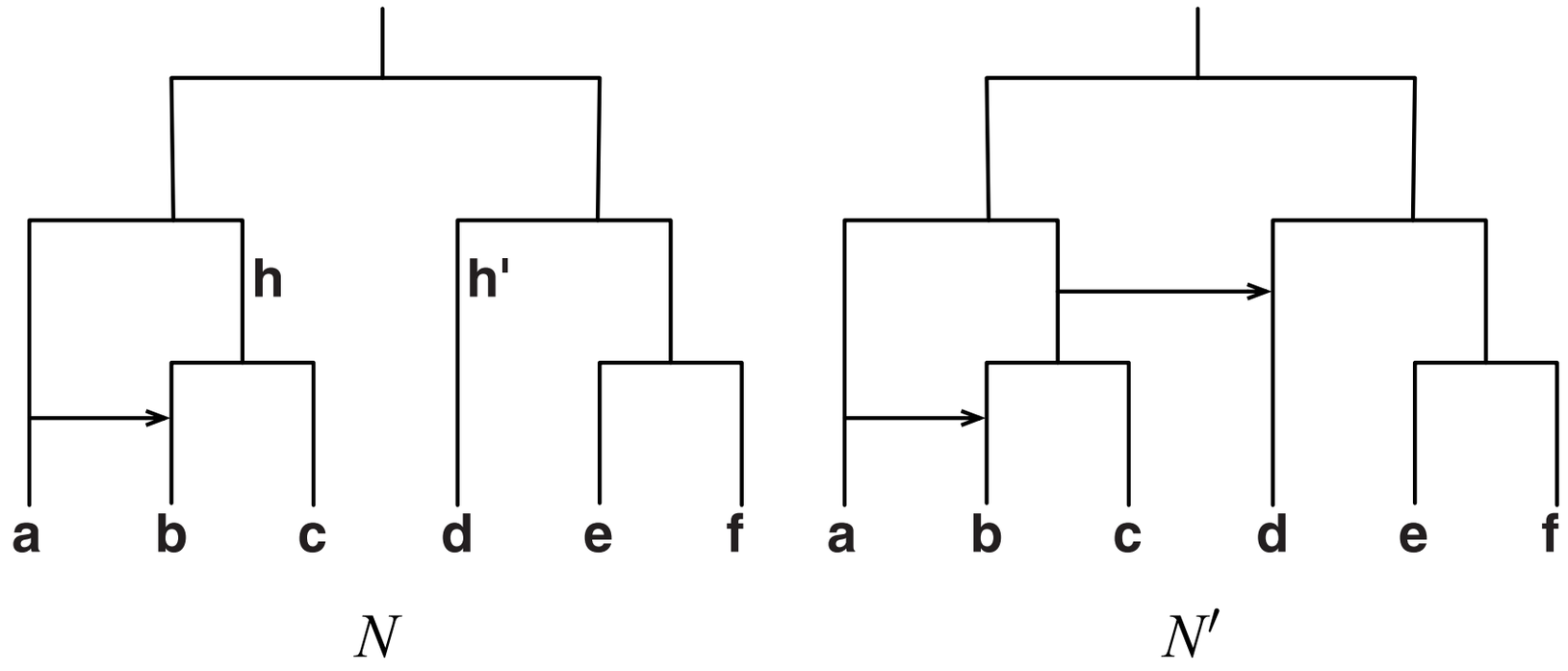
$$(N^*, \Gamma^*, \lambda^*) = \text{argmax}_{(N,\Gamma,\lambda)} L(N, \Gamma, \lambda; S)$$

Issues With Sequence-based Inference: (1) Computational Complexity

- The problems are NP-hard, even when the network is given.
- The network space is much larger than the tree space.
- No techniques currently exist for searching the network space (the equivalent of SPR, TBR, and NNI in searching the tree space).

Issues With Sequence-based Inference: (2) Overfitting

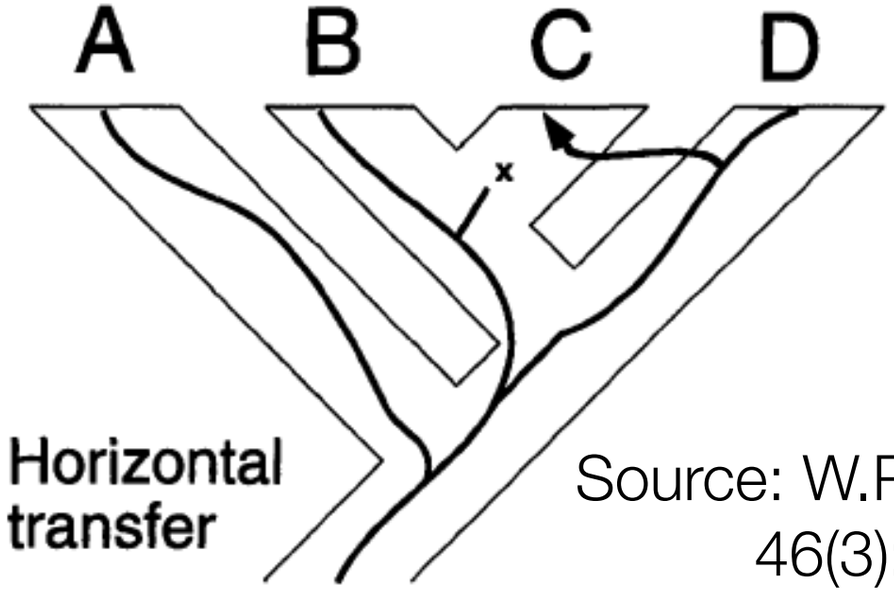
- The more [HGTs], the merrier! That is, adding more HGTs to the network can either improve the fit of the data or keep it unchanged, but never makes it worse.



- Have to control for complexity of the model

To Network, or Not to Network, That Is the Question

Recall



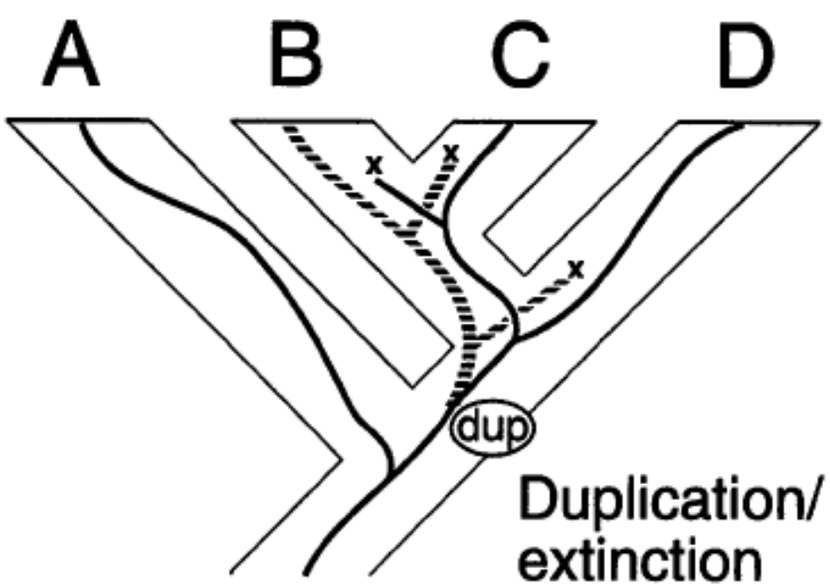
Source: W.P. Maddison, Syst. Biol., 46(3): 523-536, 1997.

But...

- Horizontal gene transfer is only one possible cause

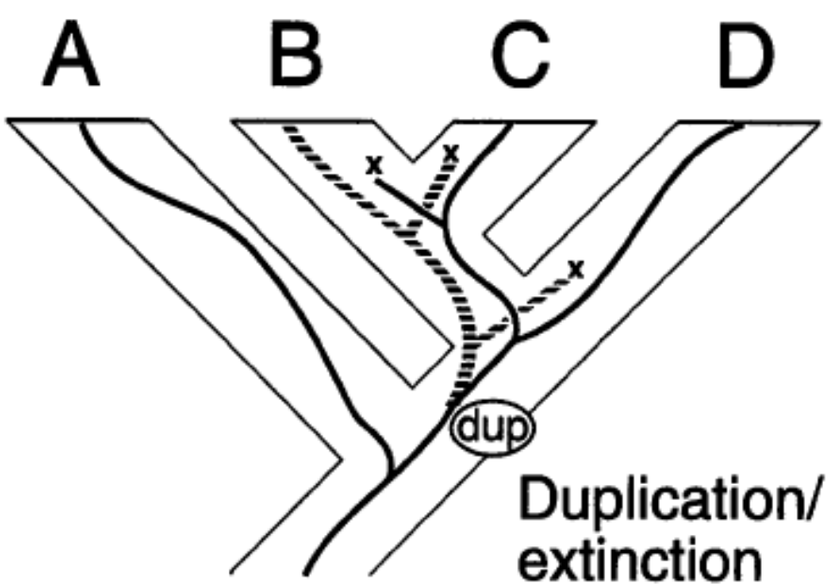
But...

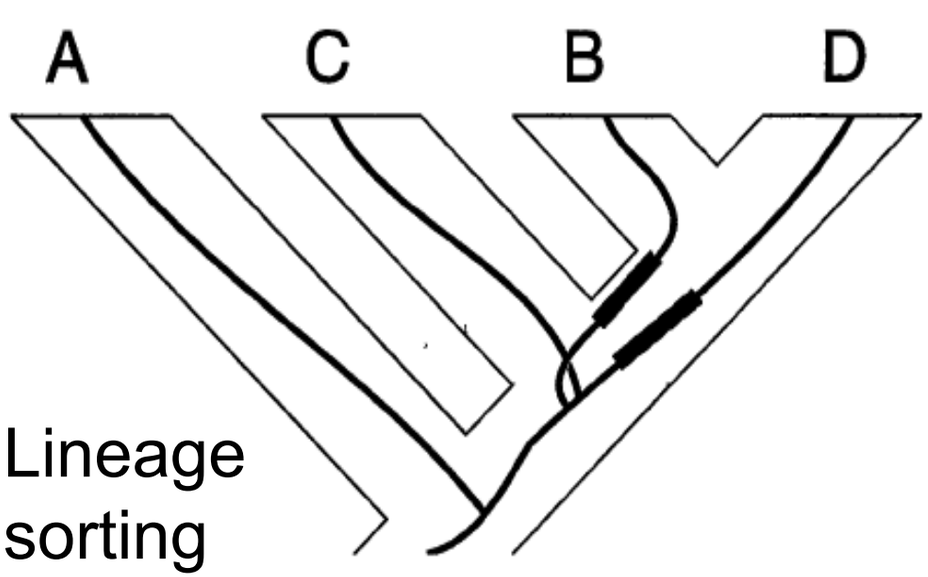
- Horizontal gene transfer is only one possible cause



But...

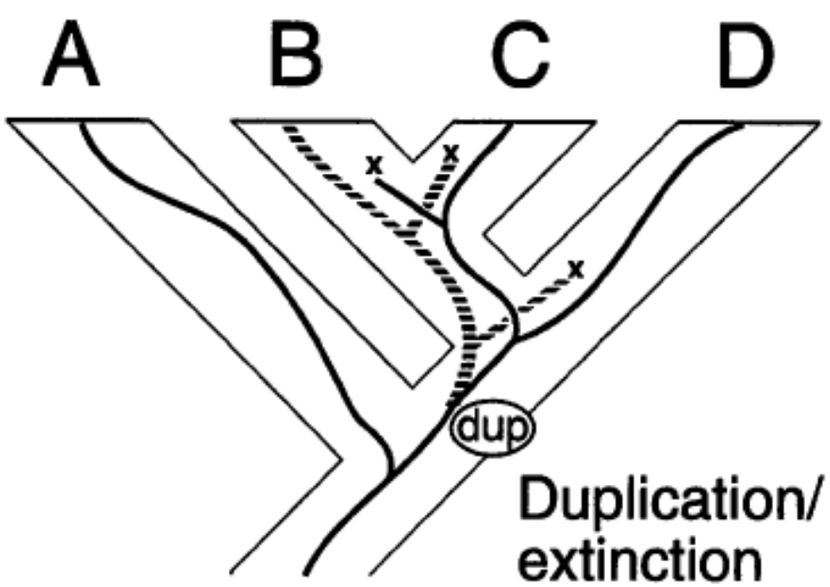
- Horizontal gene transfer is only one possible cause

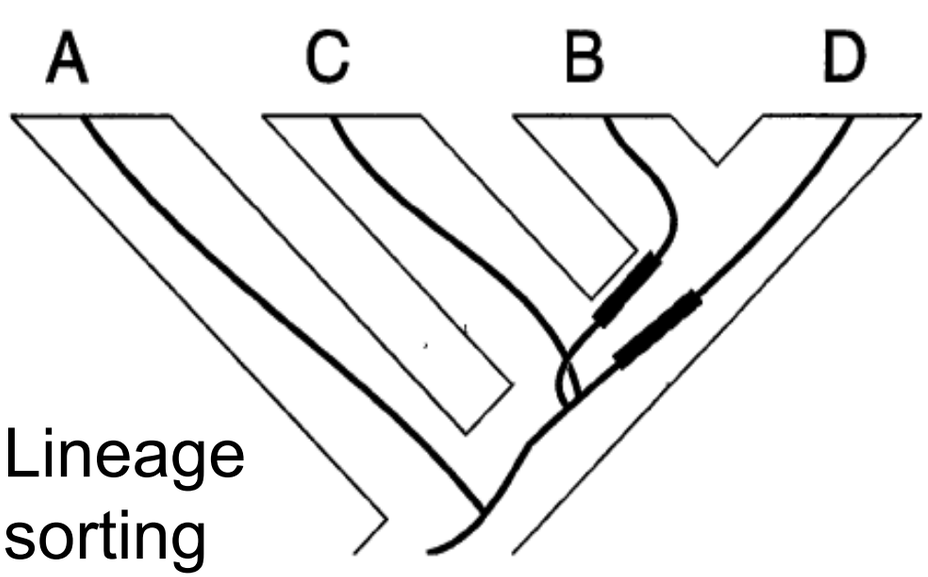




But...

- Horizontal gene transfer is only one possible cause





In these cases, the gene trees should *not* be reconciled into a phylogenetic network, but rather reconciled *within the branches* of the species tree

The Main Question

- Given a collection of gene trees, determine (rather than *assume*) the cause(s) of incongruence, and reconcile the trees accordingly

- Gives rise to the need for a stochastic framework that explains the observed patterns of gene trees

- A natural candidate is *the coalescent*, which allows for computing gene tree probabilities, among other things

- However, it needs to be augmented to allow for events such horizontal gene transfer, gene duplication/loss, ...

- Work is emerging in this area.

Summary

- Phylogenetic networks generalize trees to allow for modeling of non-treelike (reticulate) evolutionary histories

- The SPR operation and distance are the most commonly used tools for estimating reticulation from pairs of trees, yet they suffer from several issues

- Optimization criteria can be generalized in a straightforward manner to networks by considering the trees inside the network

- Incongruence is not necessarily a reflection of reticulate evolution; stochastic frameworks for determining the cause of incongruence are necessary; the coalescent is a natural candidate

Thank You!

http://www.cs.rice.edu/~nakhleh/