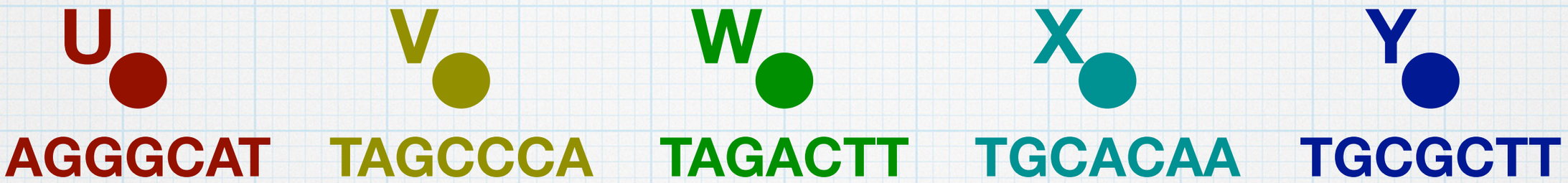
Phylogenomics:
Trees Within Trees and Trees Within Networks

COMP 571 - Spring 2015
Luay Nakhleh, Rice University

Recall



Recall

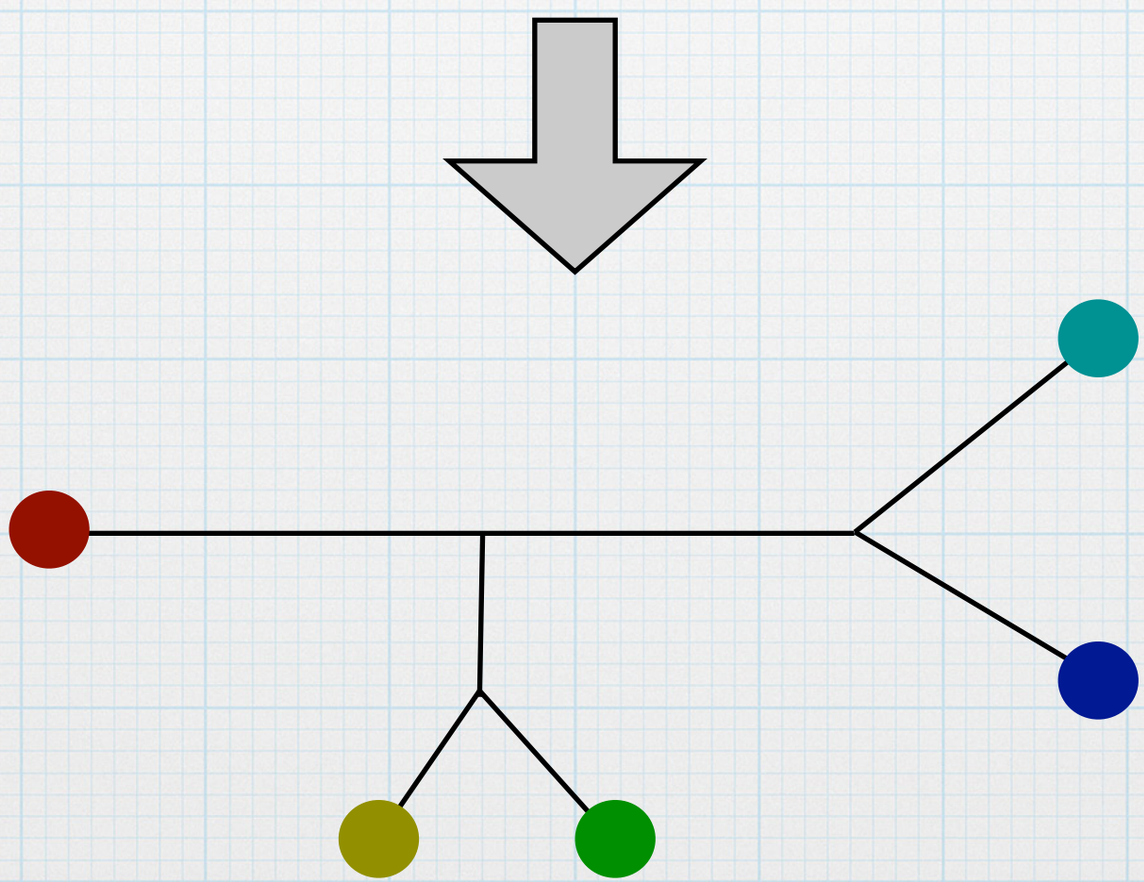
U AGGGCAT

V TAGCCCA

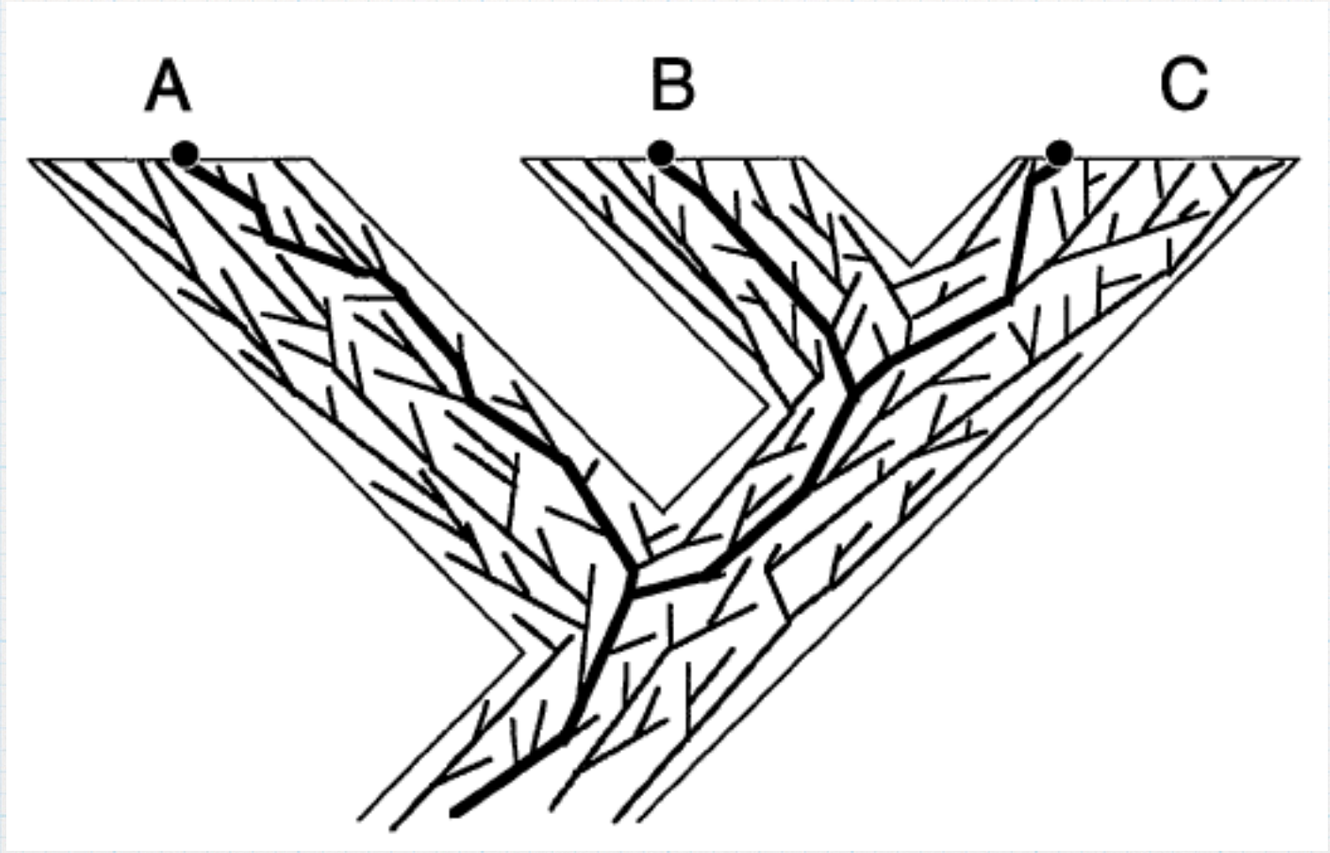
W TAGACTT

X TGCACAA

Y TGCGCTT

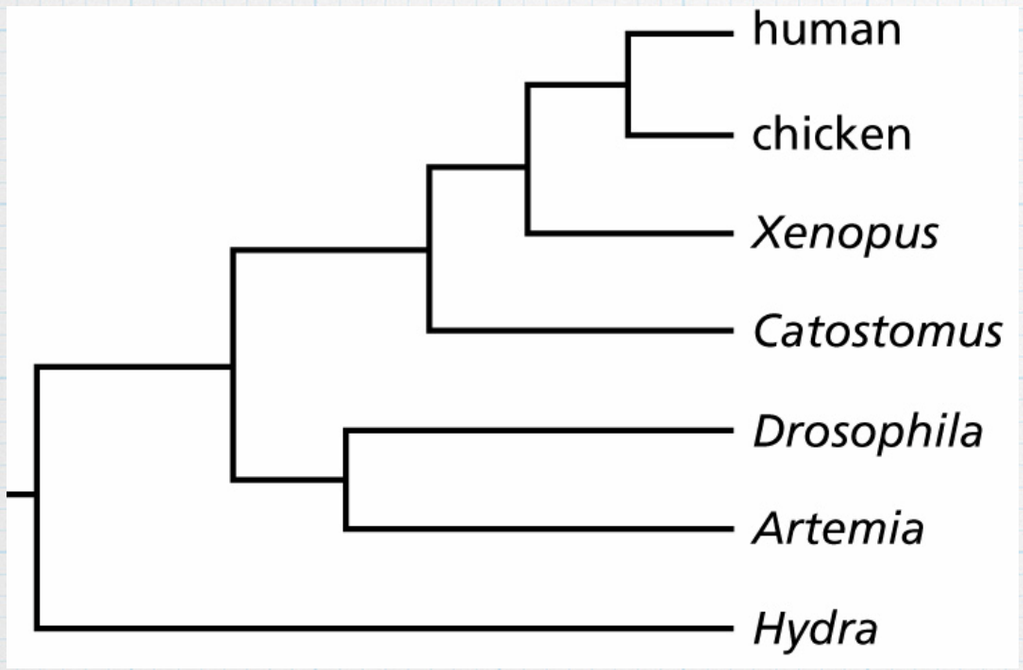


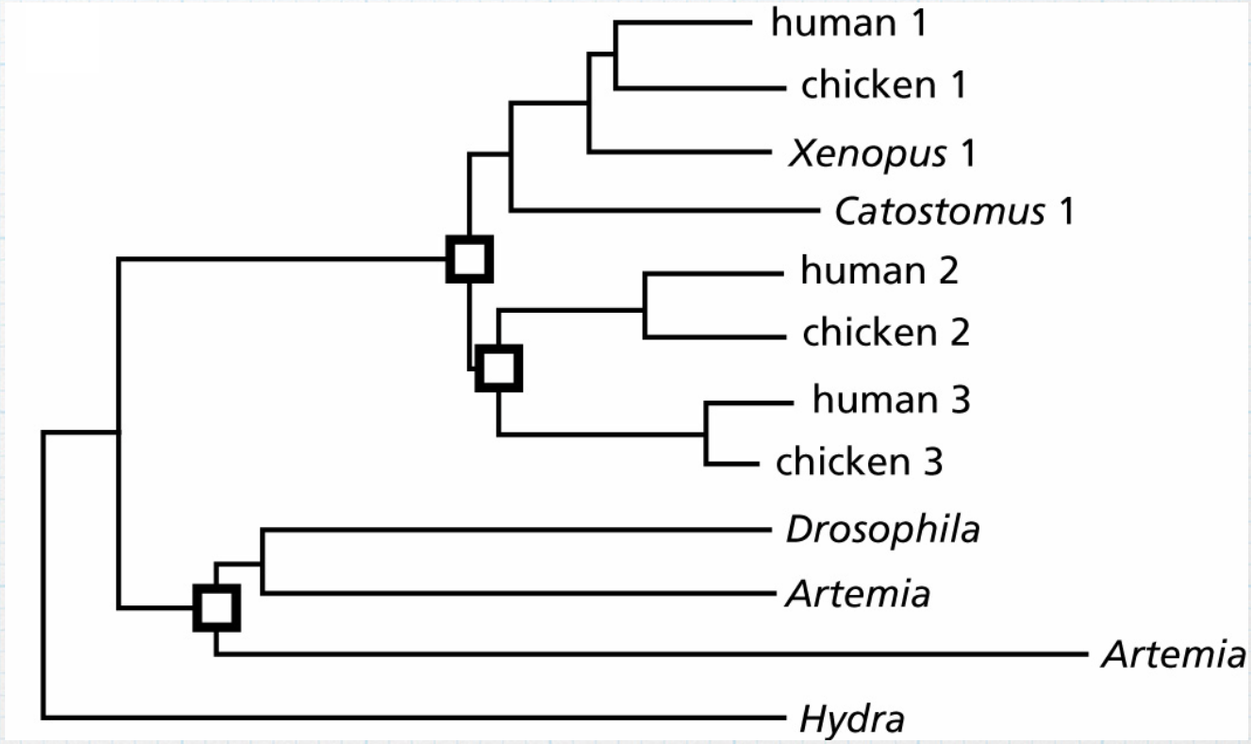
Gene Trees in Species Trees



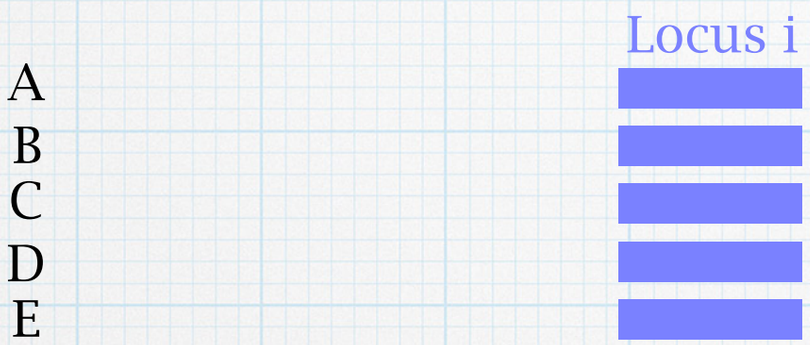
[Source: W.P. Maddison, Syst. Biol. 46(3):523-536,1997.]

What Tree Is Being Inferred?



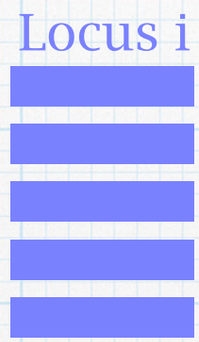


The Pre-genomic Era



The Pre-genomic Era

A
B
C
D
E

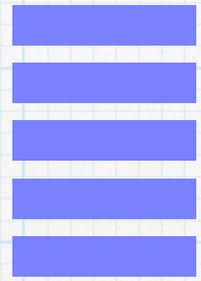


Gene Tree

The Pre-genomic Era

A
B
C
D
E

Locus i



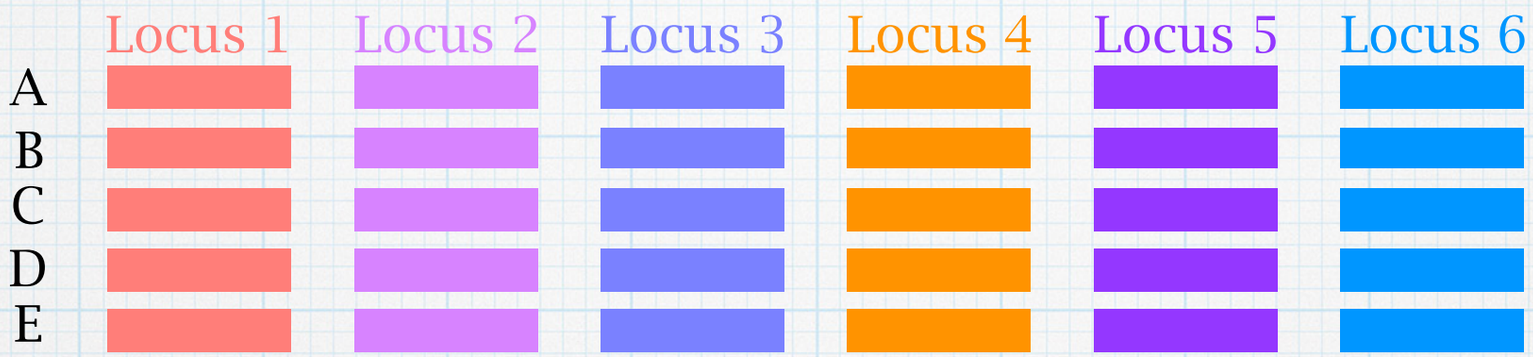
Gene Tree

Species
Phylogeny

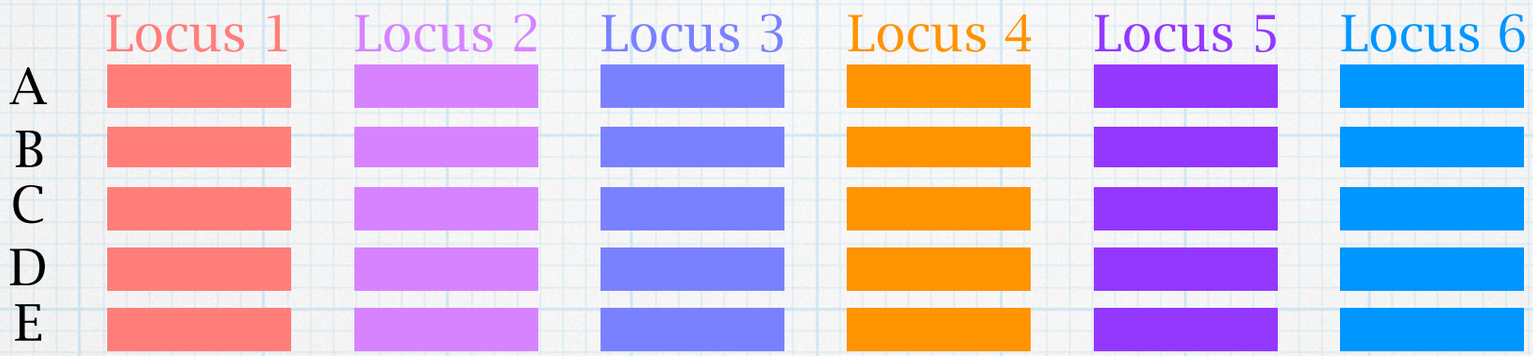
The Post-genomic Era

A
B
C
D
E

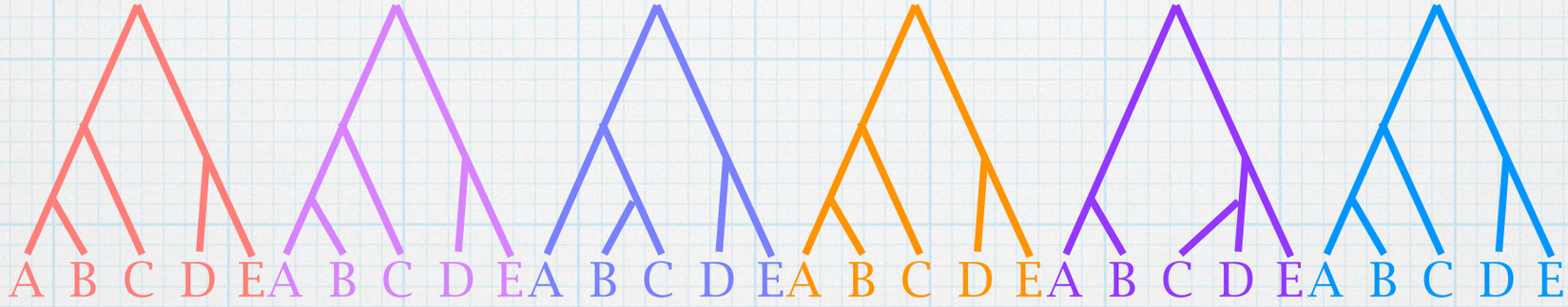
The Post-genomic Era



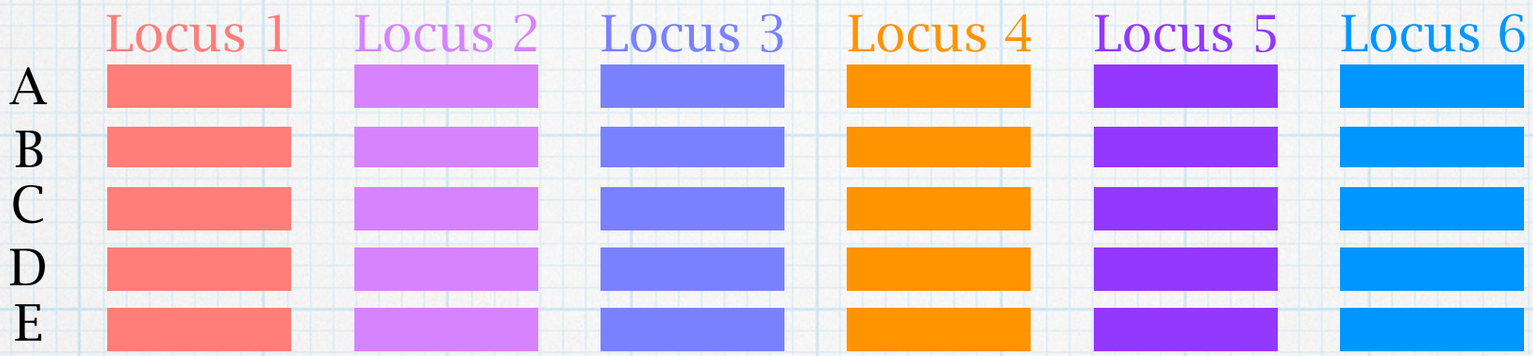
The Post-genomic Era



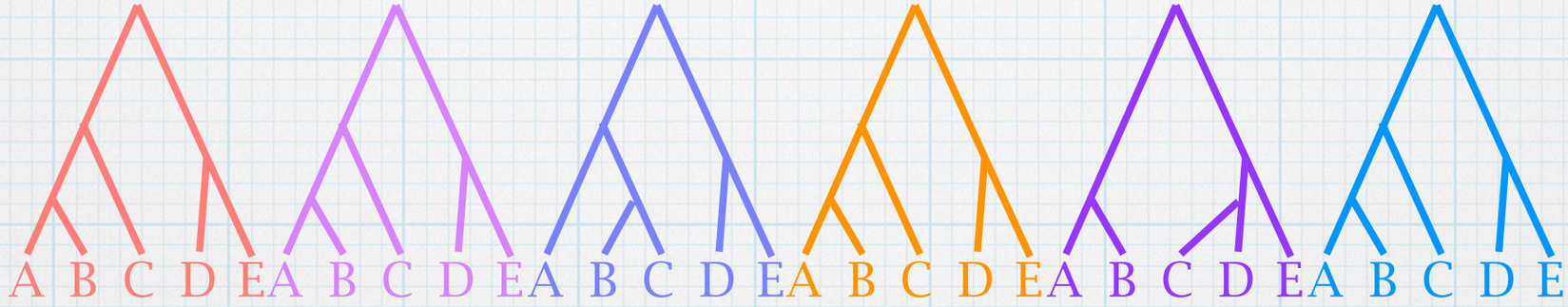
Gene Trees



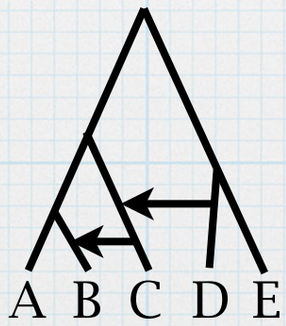
The Post-genomic Era



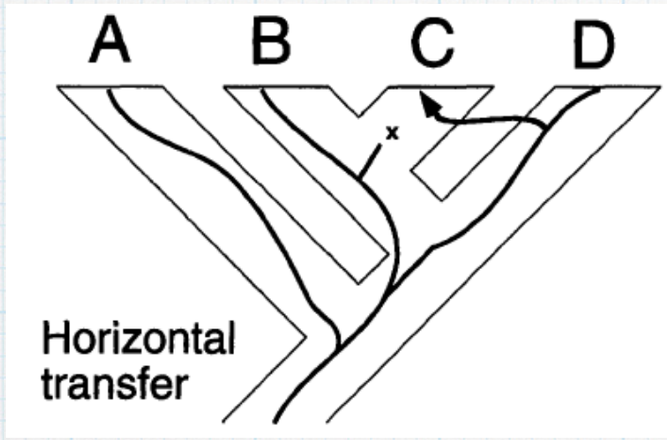
Gene Trees

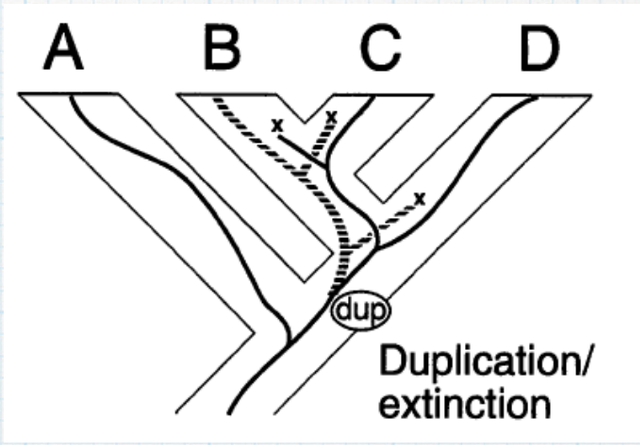


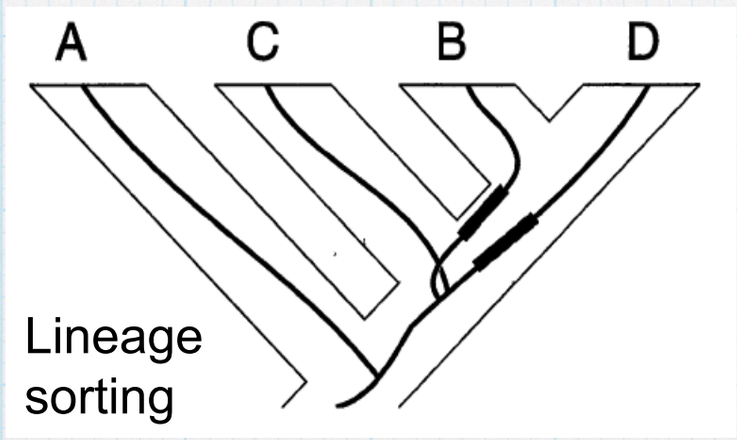
Species Phylogeny



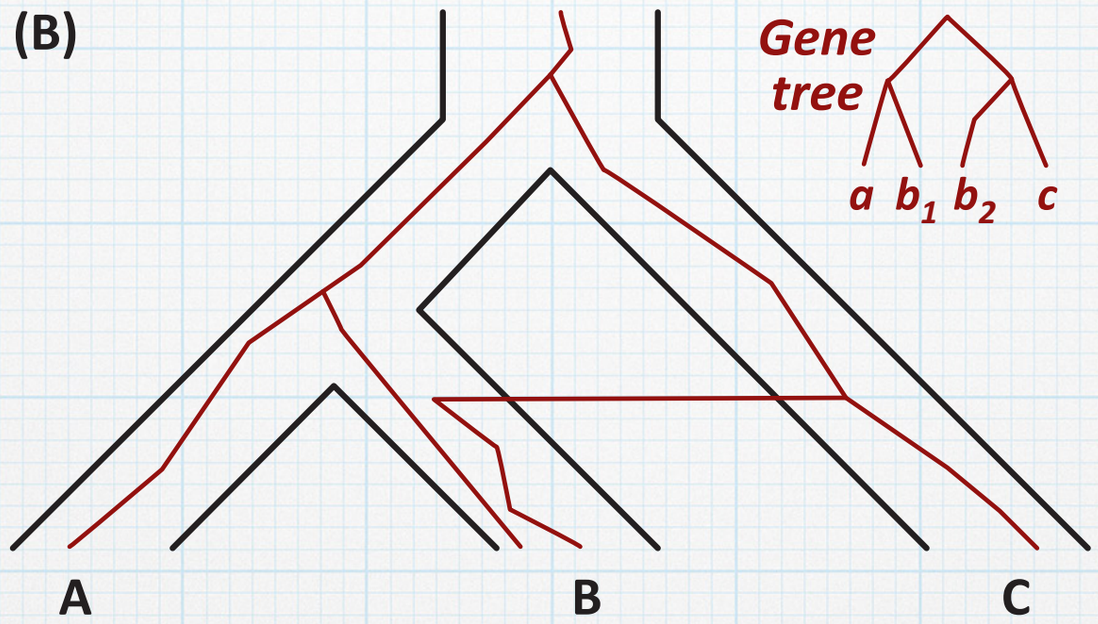
What Causes Incongruence?







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

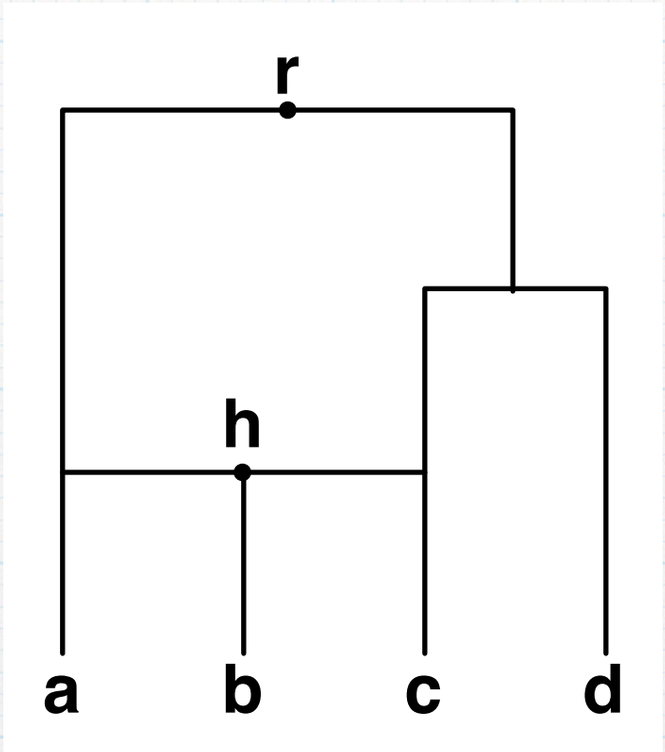


Reticulate Evolution

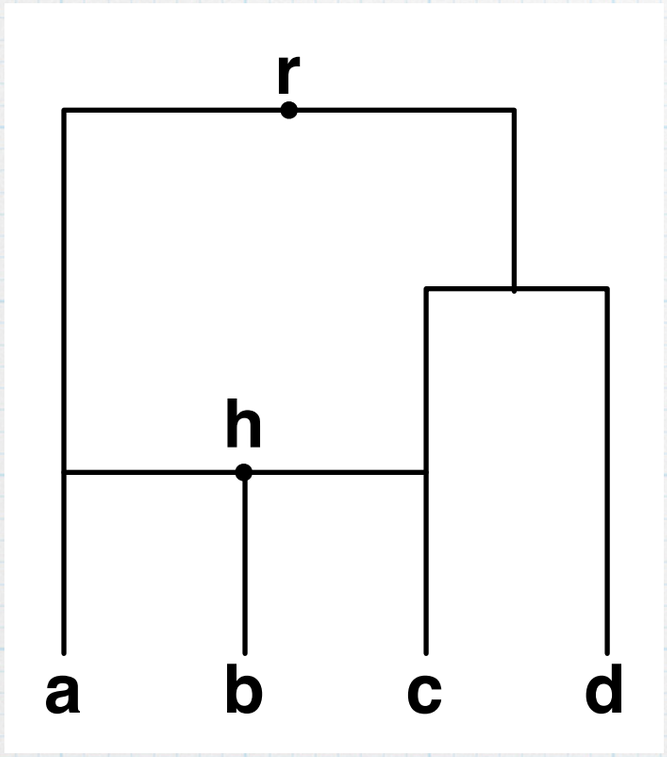
Phylogenetic Networks

* When reticulate evolutionary events (e.g., horizontal gene transfer or hybridization) occur, the evolutionary history of the genomes is best represented by a phylogenetic network.

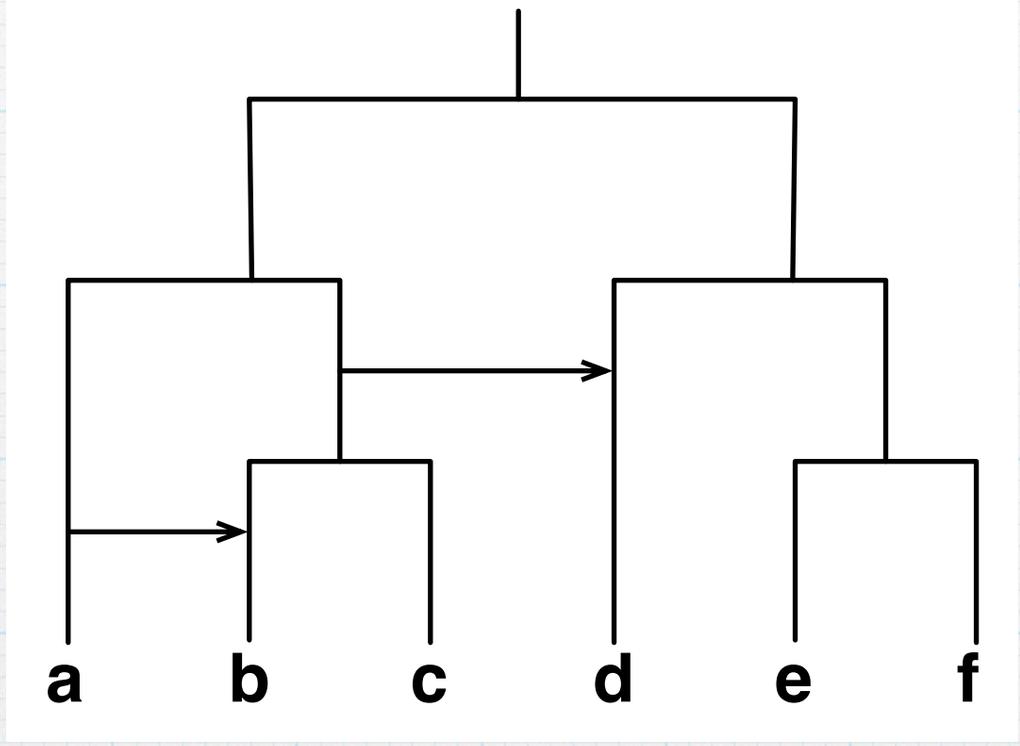
Phylogenetic Networks



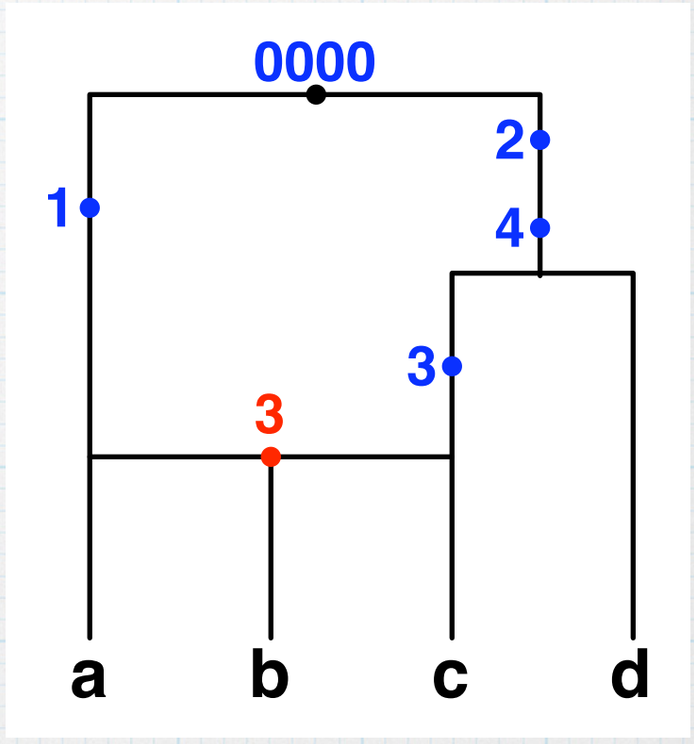
Phylogenetic Networks



hybrid speciation



HGT



recombination

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 $\mathcal{X}$.

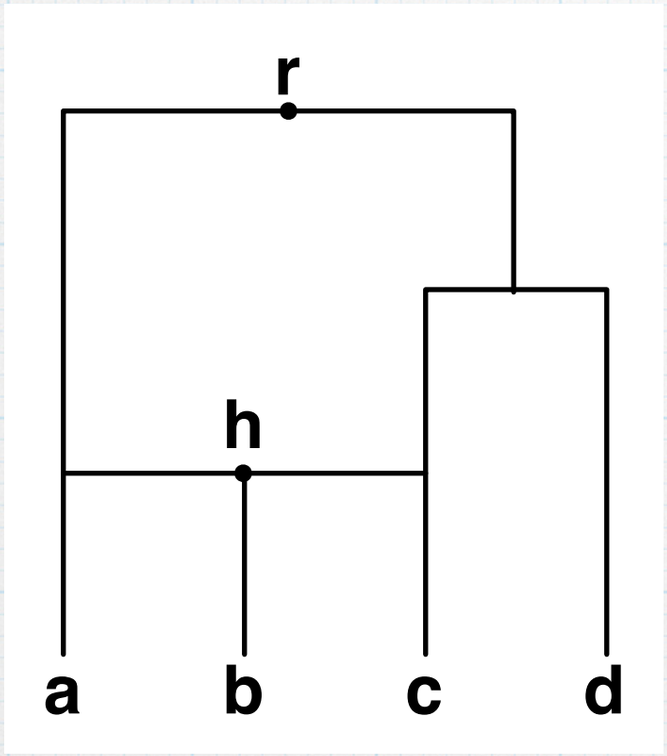
Trees Within Networks

* At the lowest level of atomicity, every nucleotide in a genome has evolved down a tree.
* More generally: every non-recombining region in the genome has evolved down a tree.

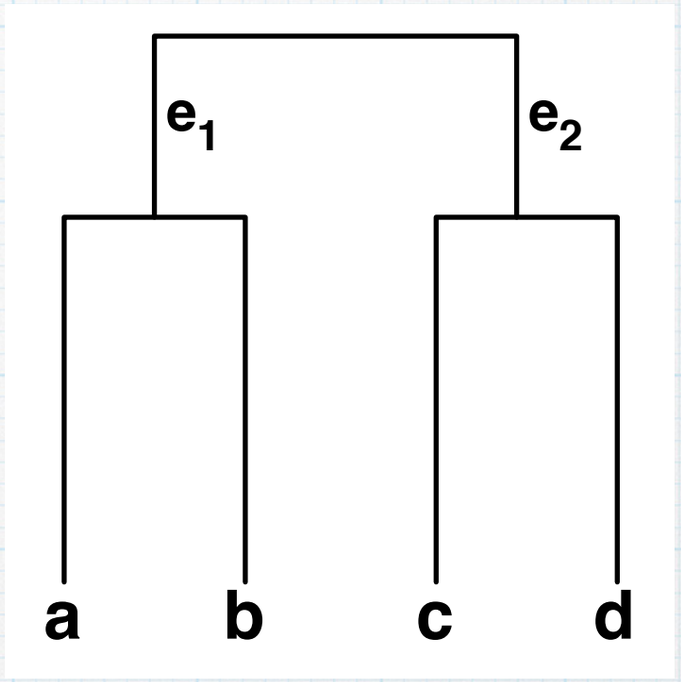
Trees Within Networks

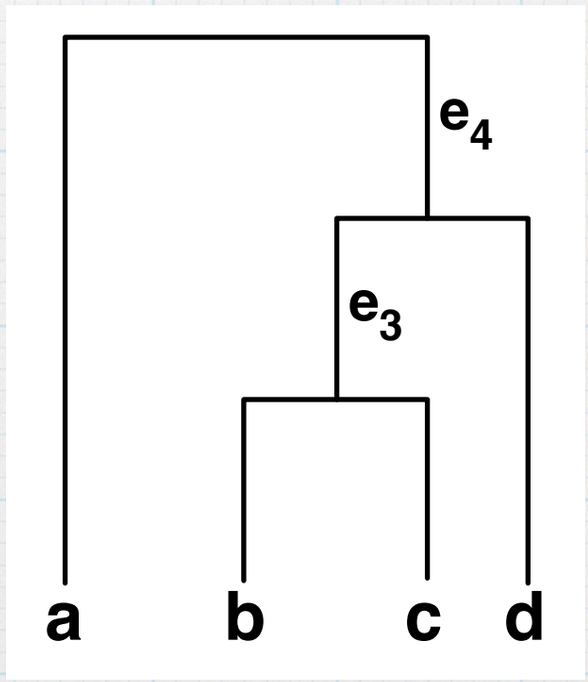
* Therefore, a phylogenetic network can be viewed as a DAG that embeds a collection of (gene) trees.

Trees Within Networks

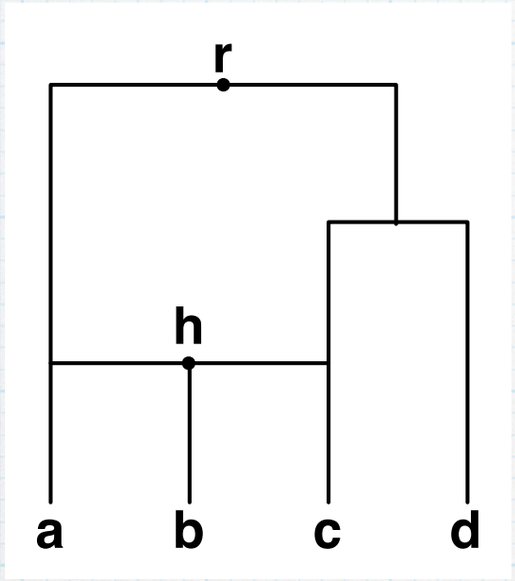


N T(N)

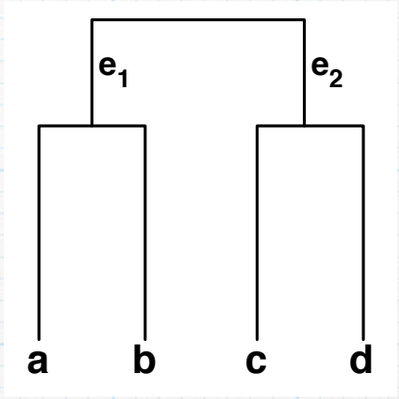


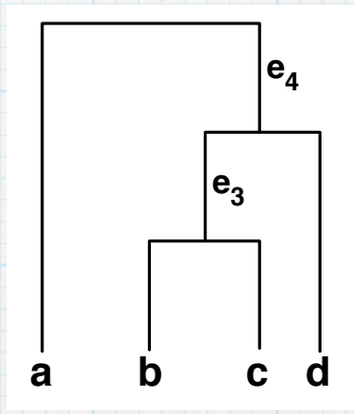


Trees Within Networks: From a network to its trees



N





T(N)

Trees Within Networks: From a network to its trees

* Given a network N with k reticulation nodes, the set T(N) contains $O(2^k)$ trees.
* Generating the trees is straightforward, but enumerating all of them is expensive for large values of k.

Trees Within Networks: From trees to a network

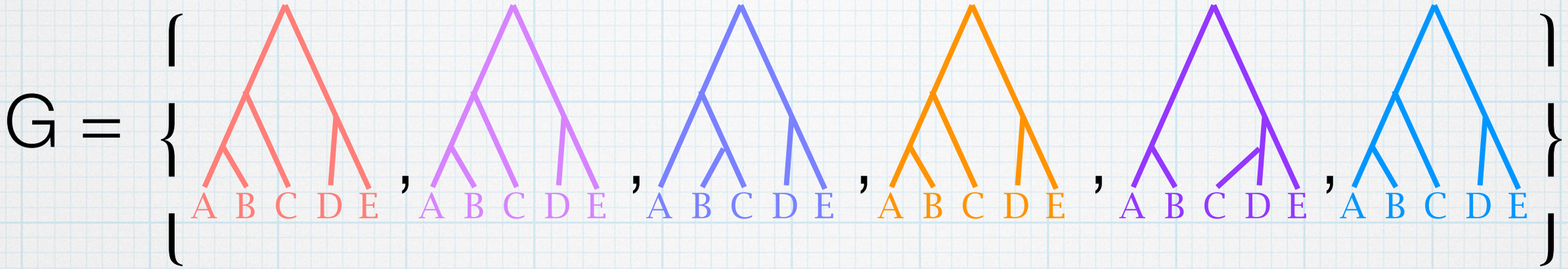
* The inference problem:
  * Input: Collection of gene trees $\mathcal{G}$.
  * Output: A network N that embeds all the trees in $\mathcal{G}$ (that is, $\mathcal{G} \subseteq \mathcal{T}(N)$)

Trees Within Networks: From trees to a network

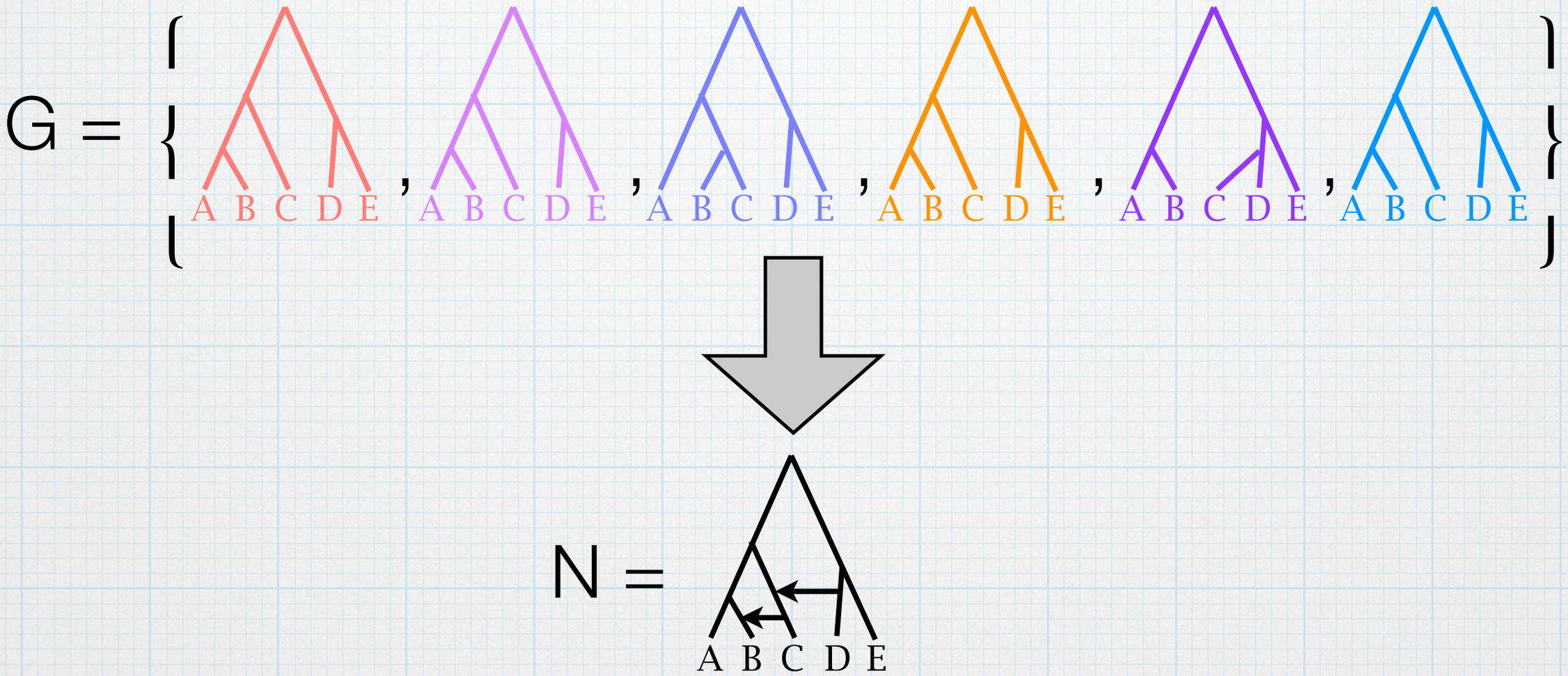
* The inference problem:
  * Input: Collection of gene trees G.
  * Output: A network N that embeds all the trees in G (that is, $G \subseteq T(N)$)

Maximum parsimony formulation:
Find N with smallest number of reticulation nodes such that $G \subseteq T(N)$.

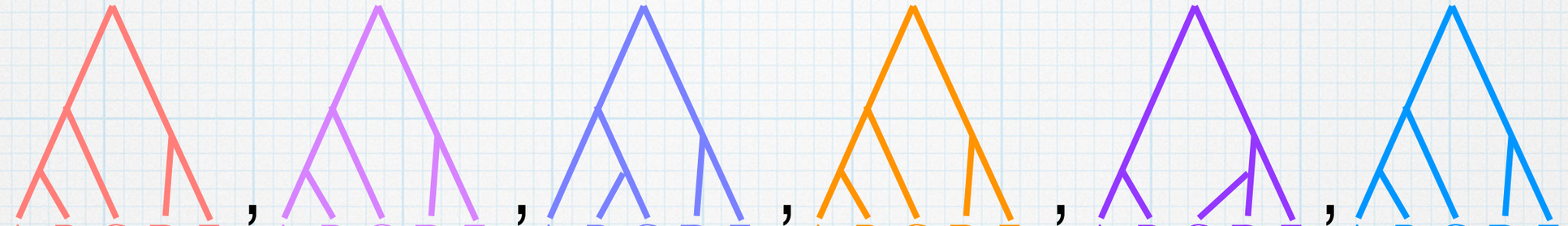
Trees Within Networks:
From trees to a network

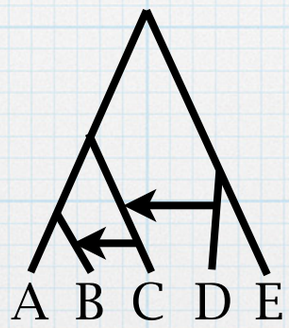


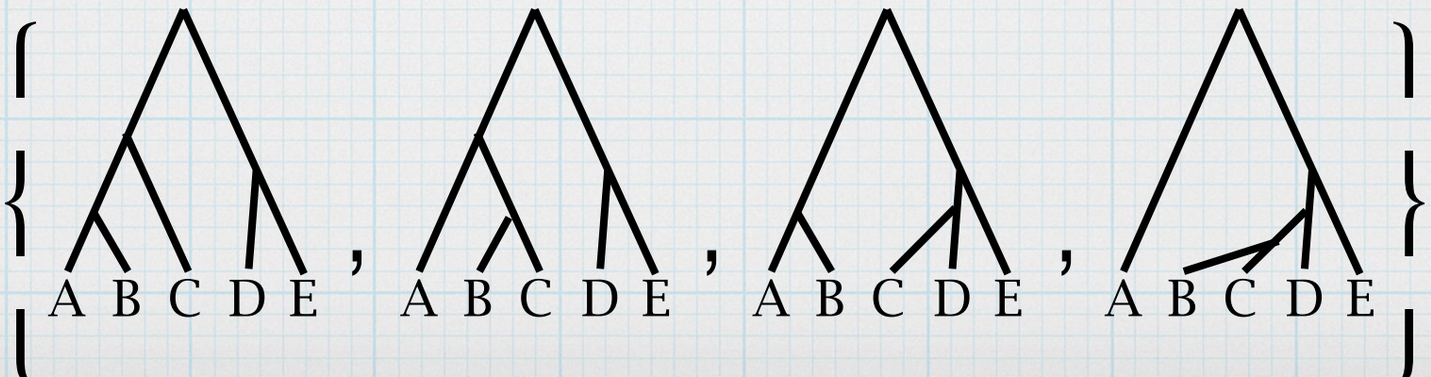
Trees Within Networks:
From trees to a network



Trees Within Networks:
From trees to a network

$G =$ {  }

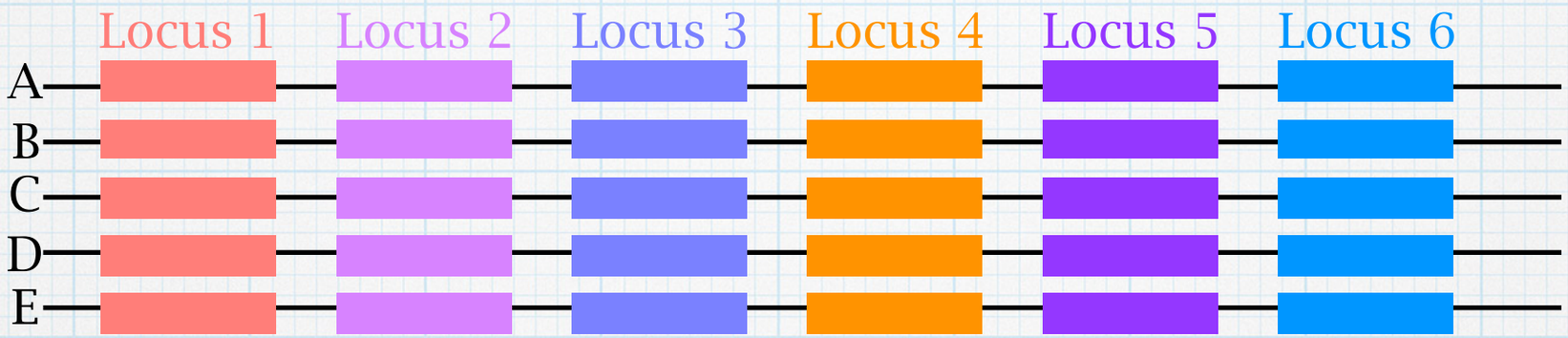
$N =$ 

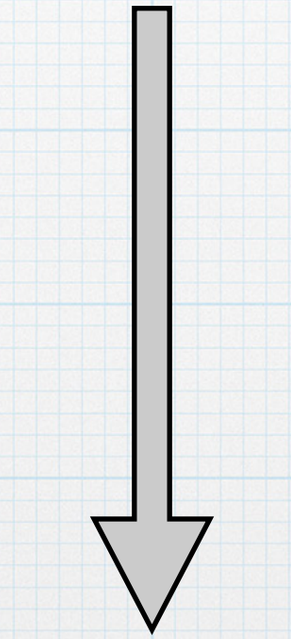
$T(N) =$ {  }

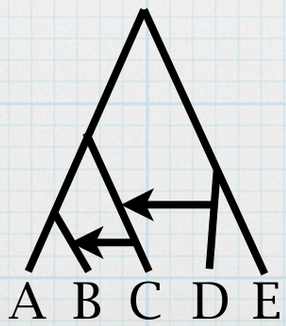
Trees Within Networks: From trees to a network

* The inference problem is NP-hard.
* Many heuristics and (worst-case exponential) exact algorithms exist.

The General Inference Problem

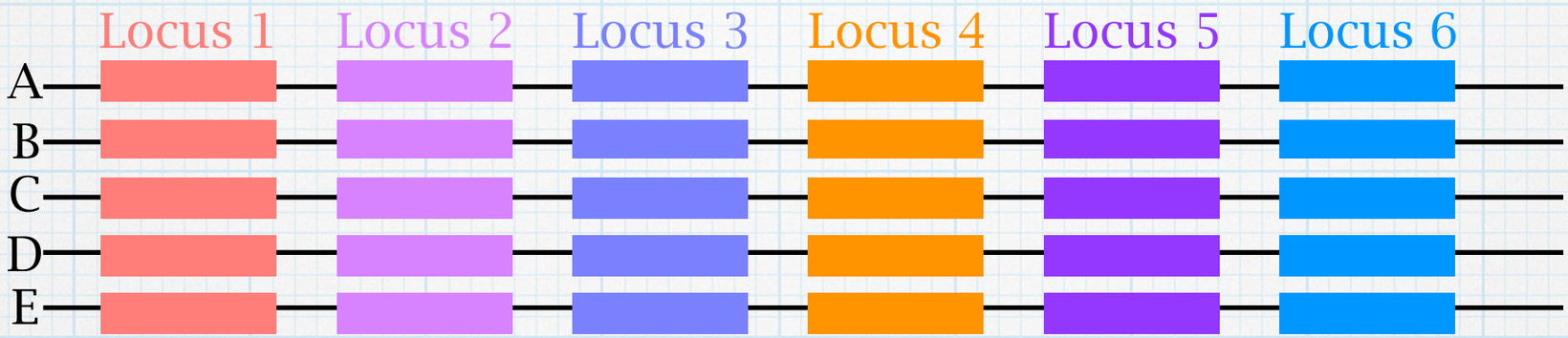




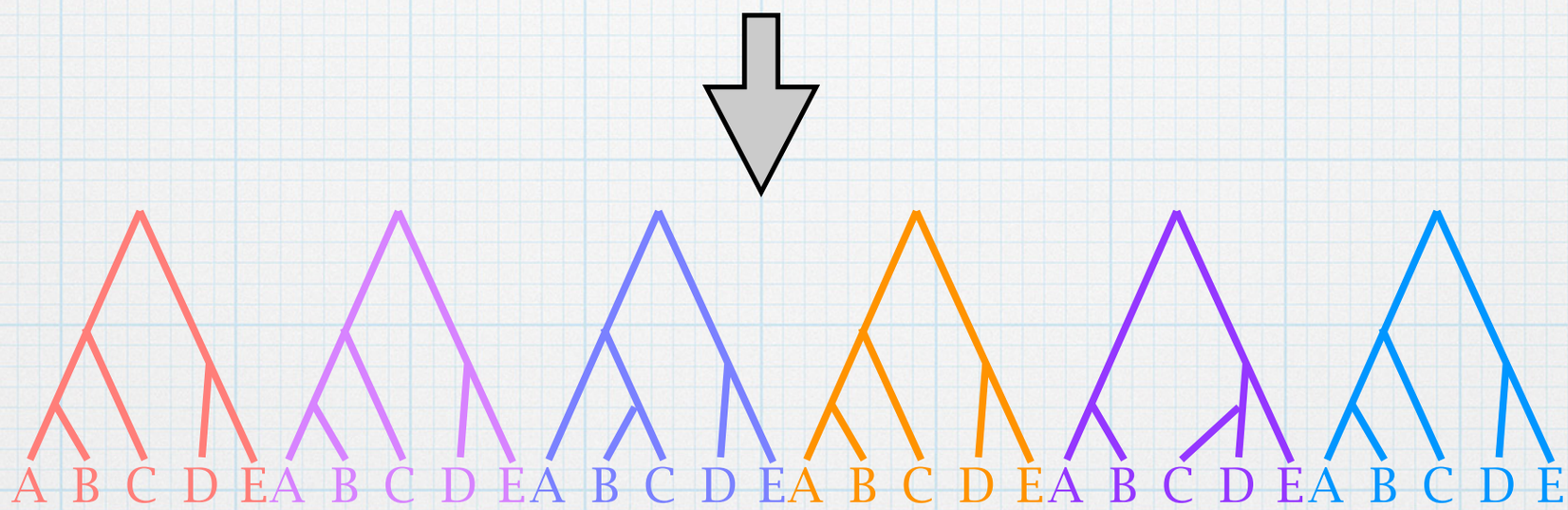


Species
Phylogeny

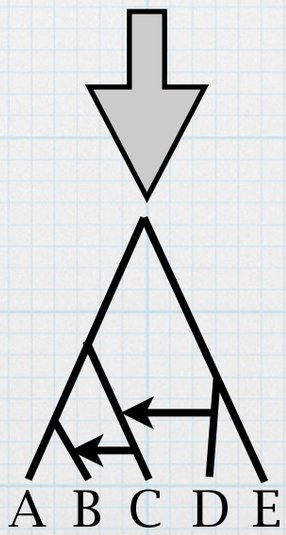
What we've seen so far...



Gene Trees



Species Phylogeny



The General Inference Problem

* We can extend the sequence-based inference methods to networks...

Maximum Parsimony 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 Phylogenetic Networks

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

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

Phylogenetic Networks and Model Selection

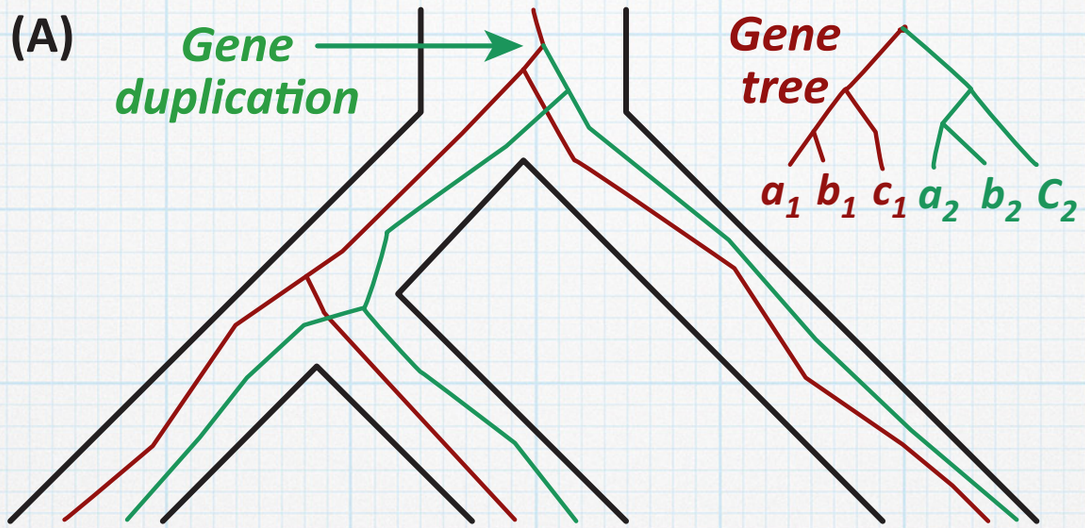
* Consider a phylogenetic network N' that contains phylogenetic network N.
* Then, $T(N) \subseteq T(N')$.
* N' would fit the data at least as well as N.
* Solution: To fit the data, make a network that is as complex as possible.

Phylogenetic Networks and Model Selection

* Consider a phylogenetic network N' that contains phylogenetic network N.
* Then, $T(N) \subseteq T(N')$.
* N' would fit the data at least as well as N.
* Solution: To fit the data, make a network that is as complex as possible.

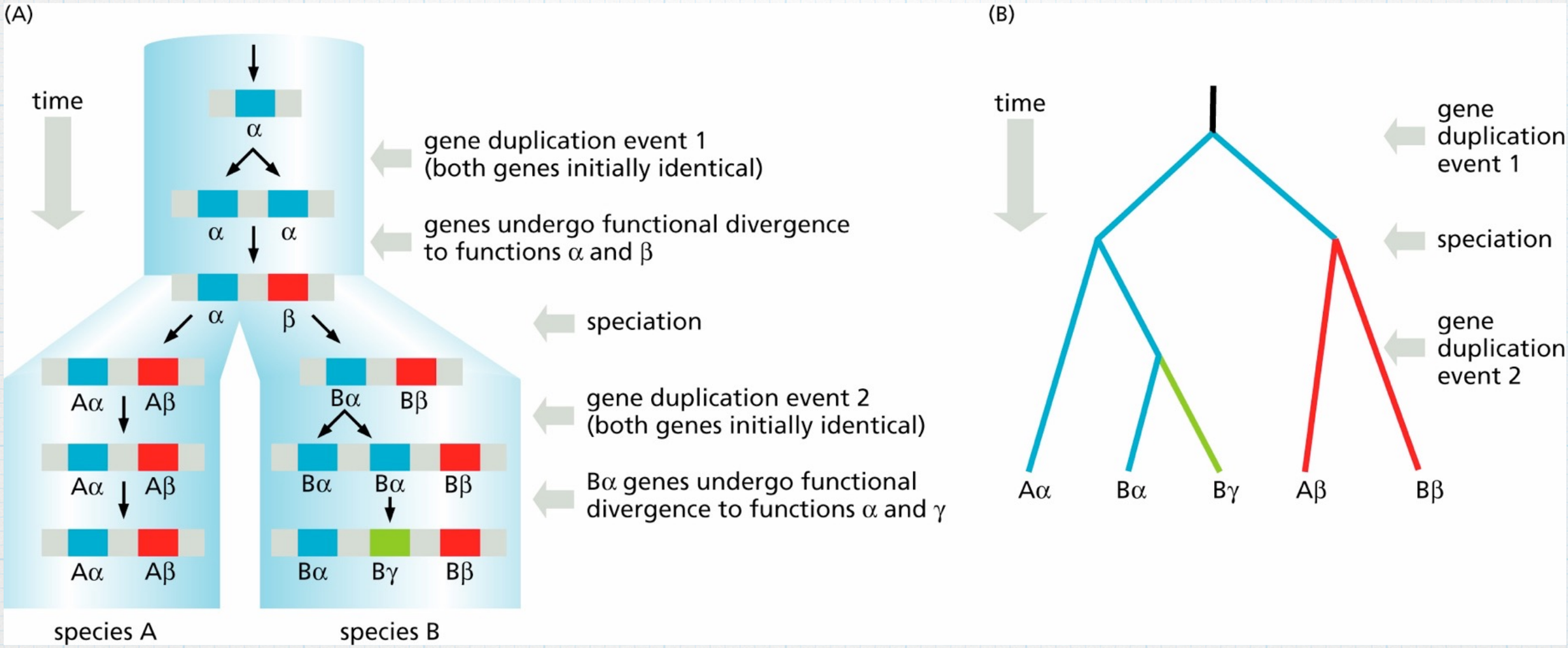
Very bad solution!

One needs to account for model complexity carefully.



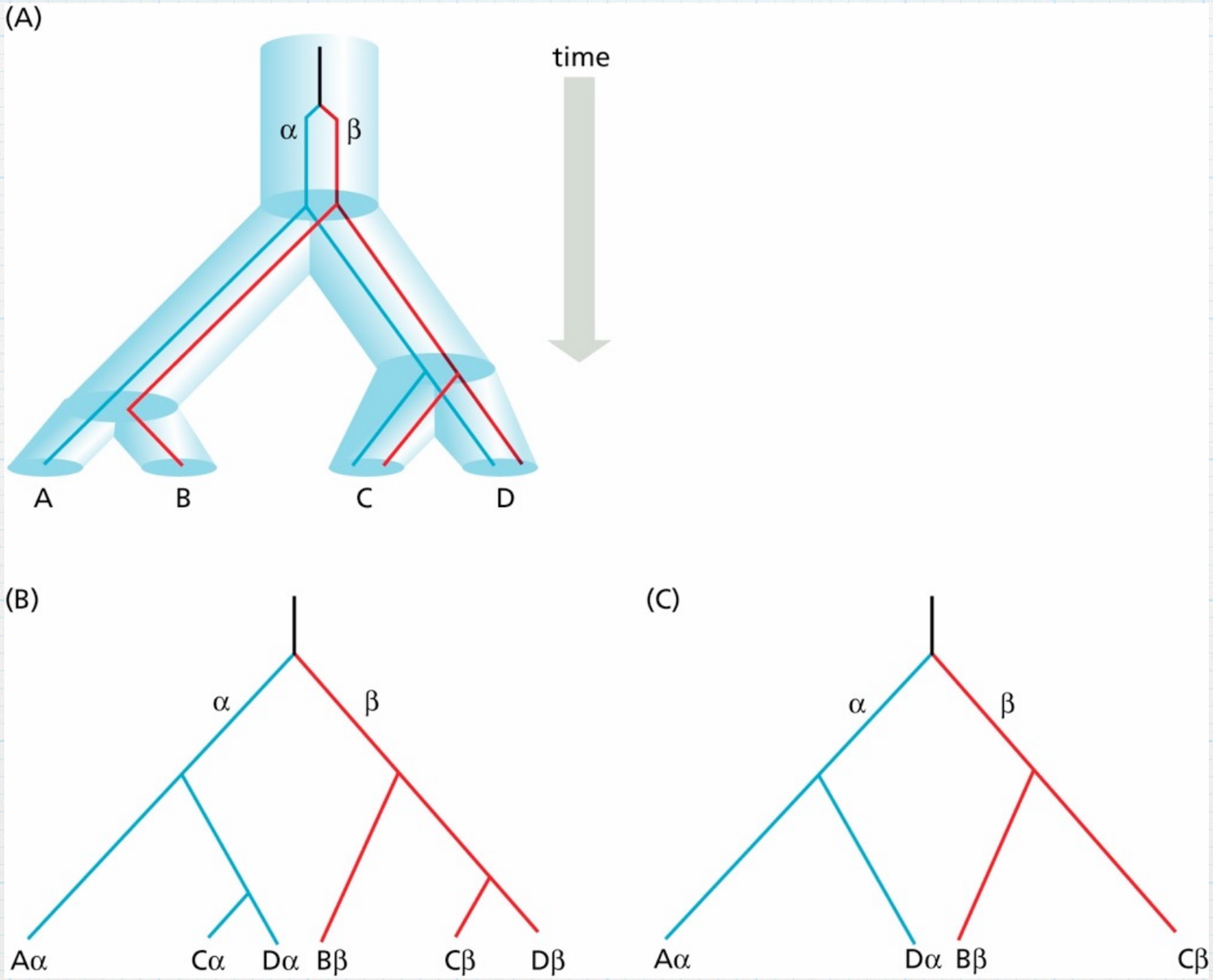
Gene Duplication/Loss

Gene Duplication



[Source: Understanding Bioinformatics]

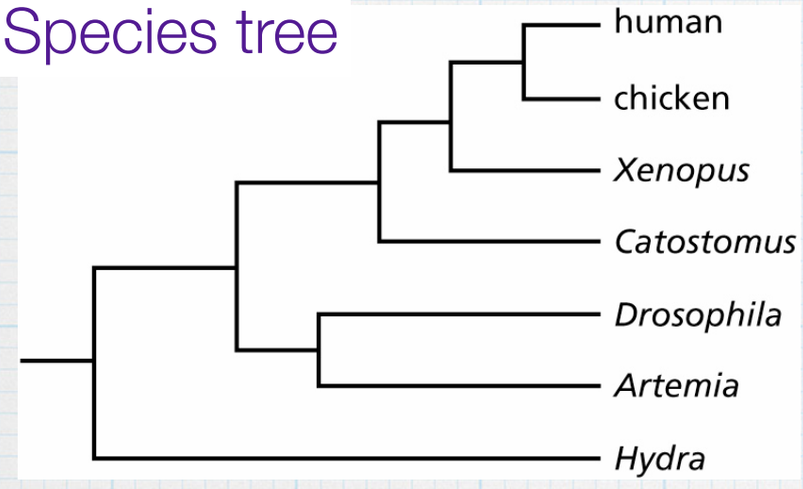
Gene Duplication/Loss



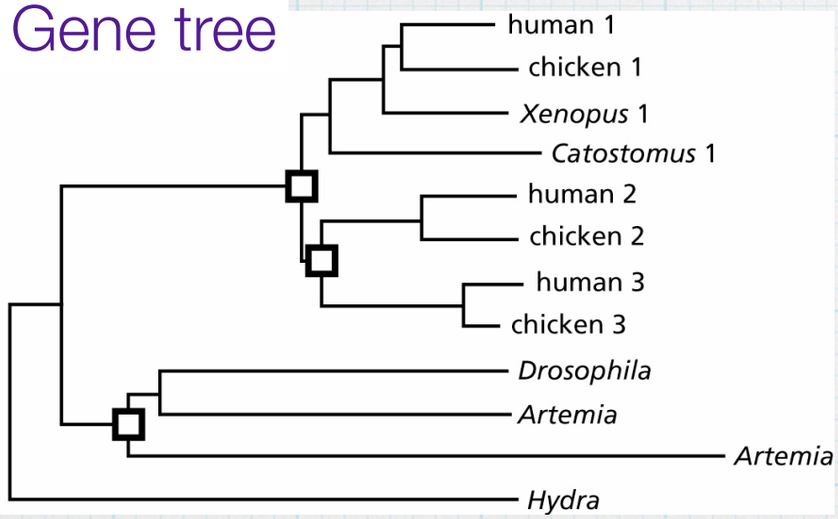
[Source: Understanding Bioinformatics]

Species/Gene Tree Reconciliation

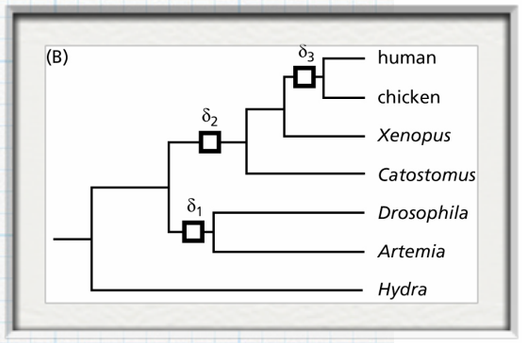
Species tree



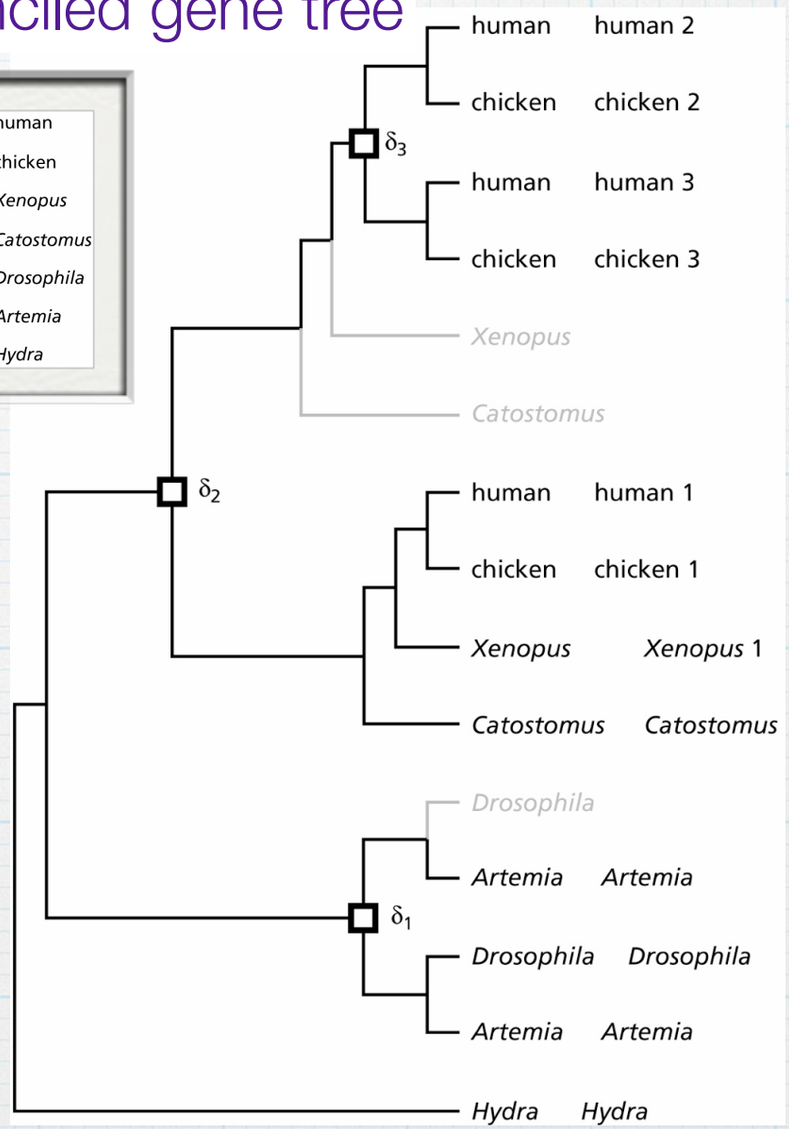
Gene tree



Reconcile



Reconciled gene tree

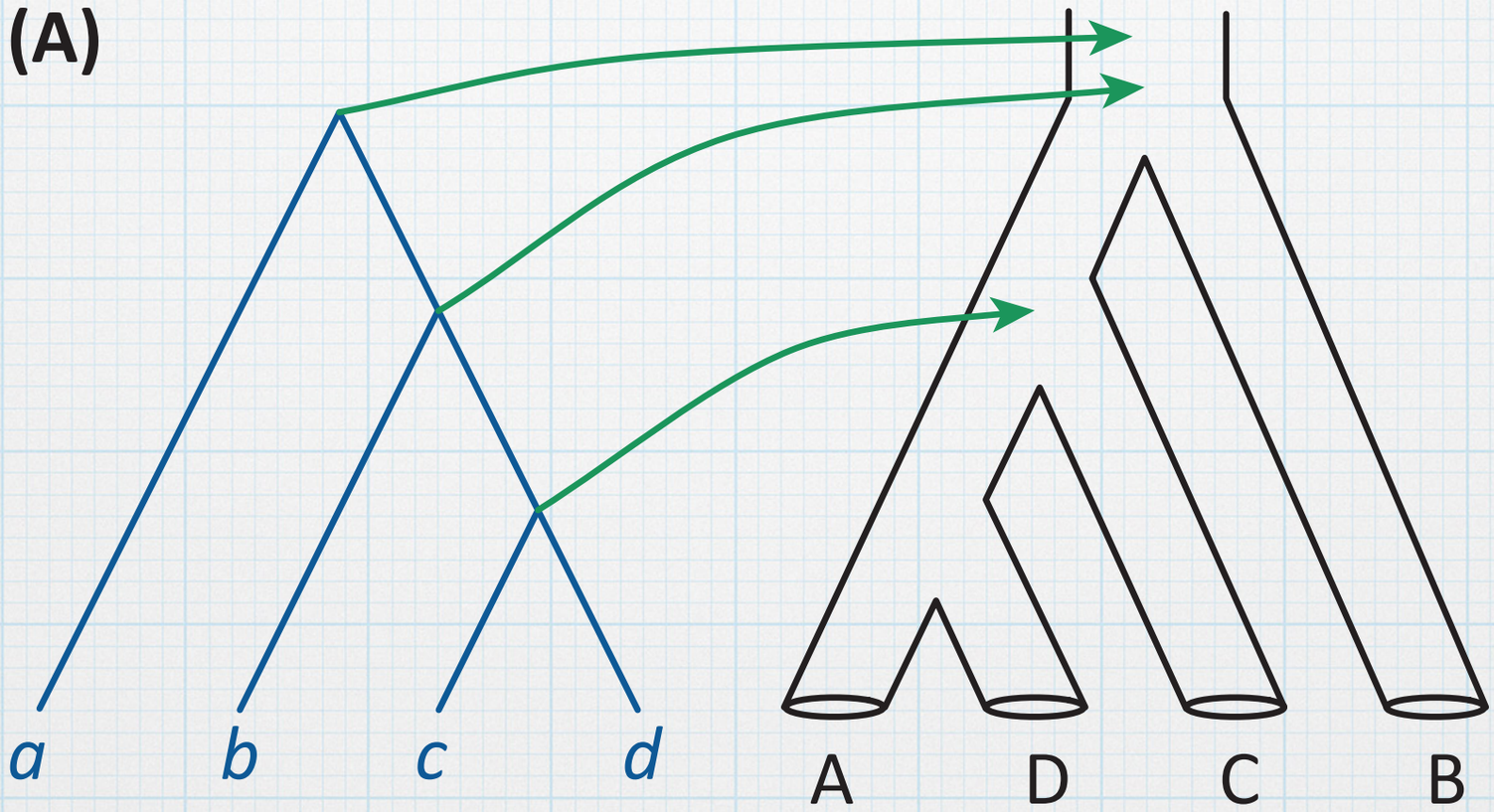


[Source: Understanding Bioinformatics]

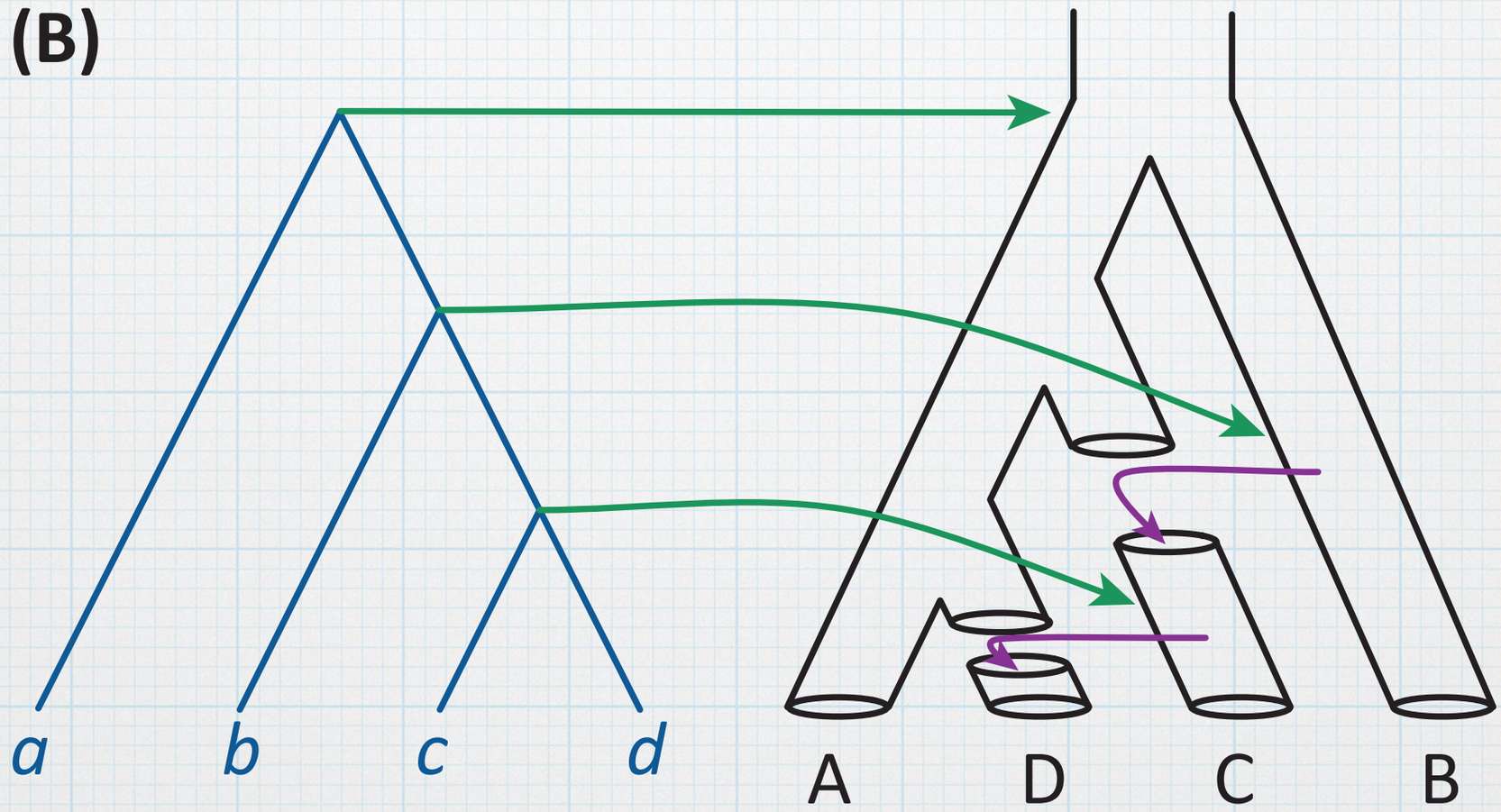
Species/Gene Tree Reconciliation

* The lca (least common ancestor) mapping:
  * Map every node x in the gene tree to the lca of L(x) in the species tree, where L(x) is the set of all leaf labels under x.

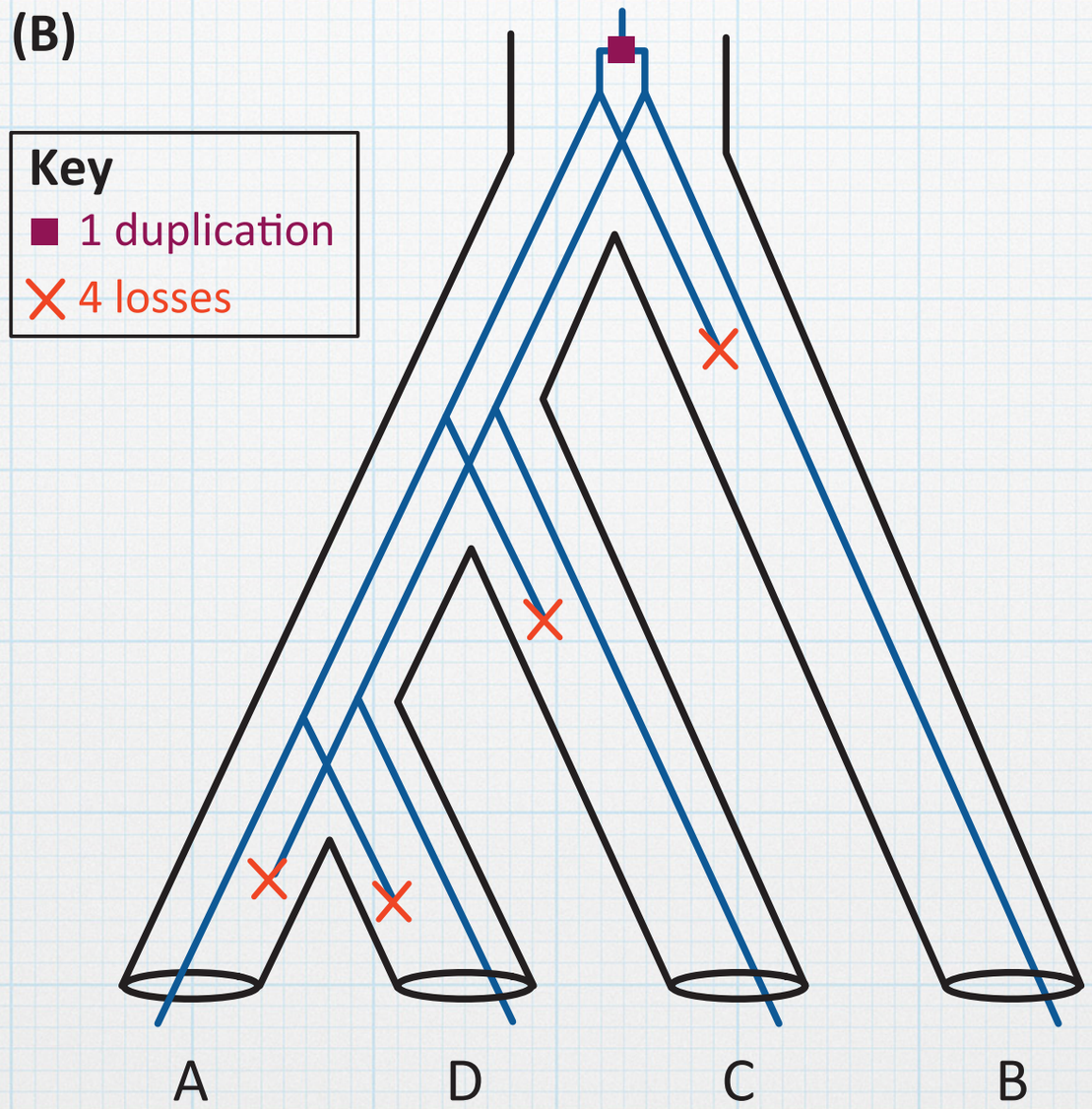
Species/Gene Tree Reconciliation



A similar technique can be used to reconcile a gene tree with a species tree to detect HGT...



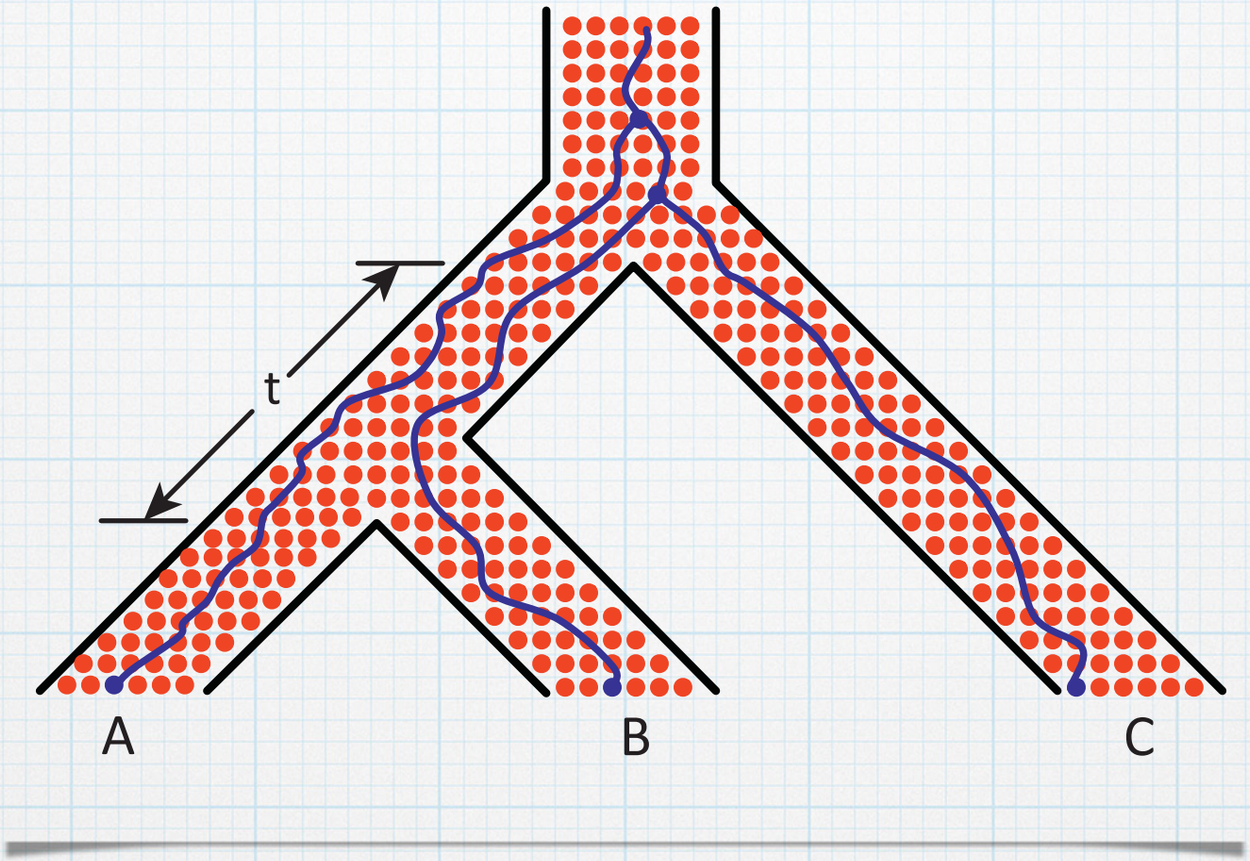
Species/Gene Tree Reconciliation



The reconciliation identifies dup/loss events.

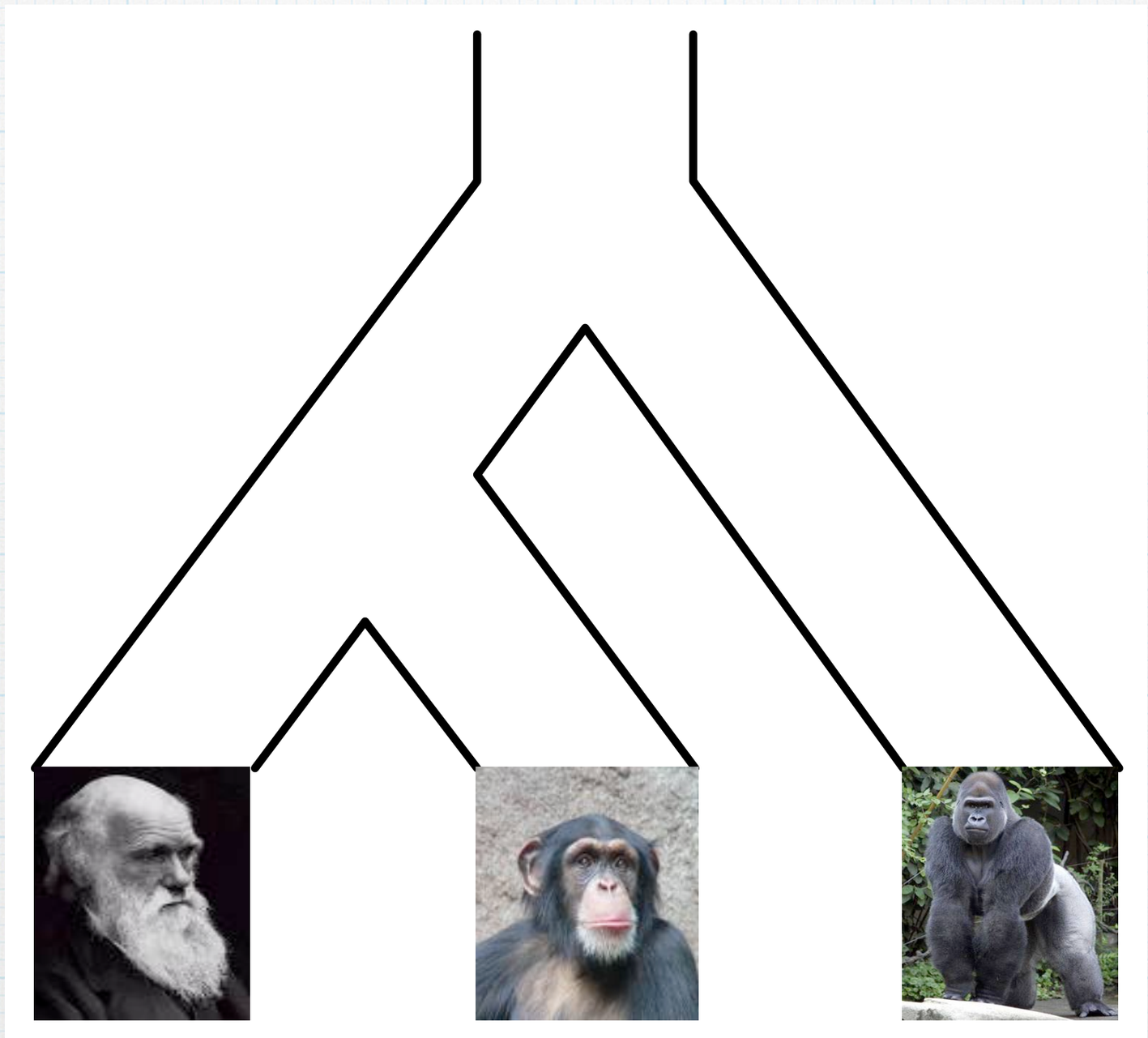
The Inference Problem

* Input: Collection of gene trees G.
* Output: Species tree T that minimizes the (weighted) number of duplication and loss events resulting from the reconciliation of all trees in G with T.

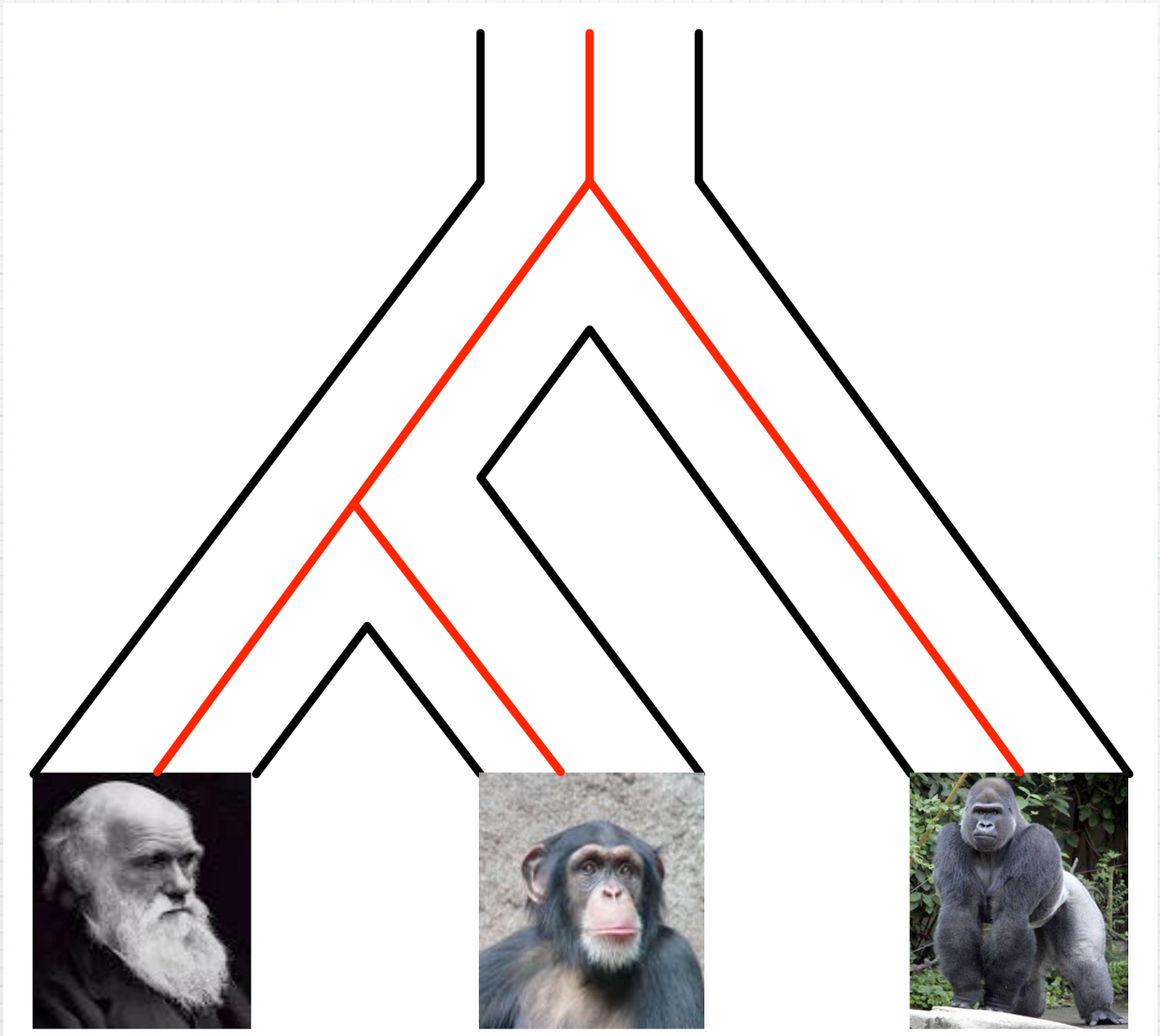


Incomplete Lineage Sorting

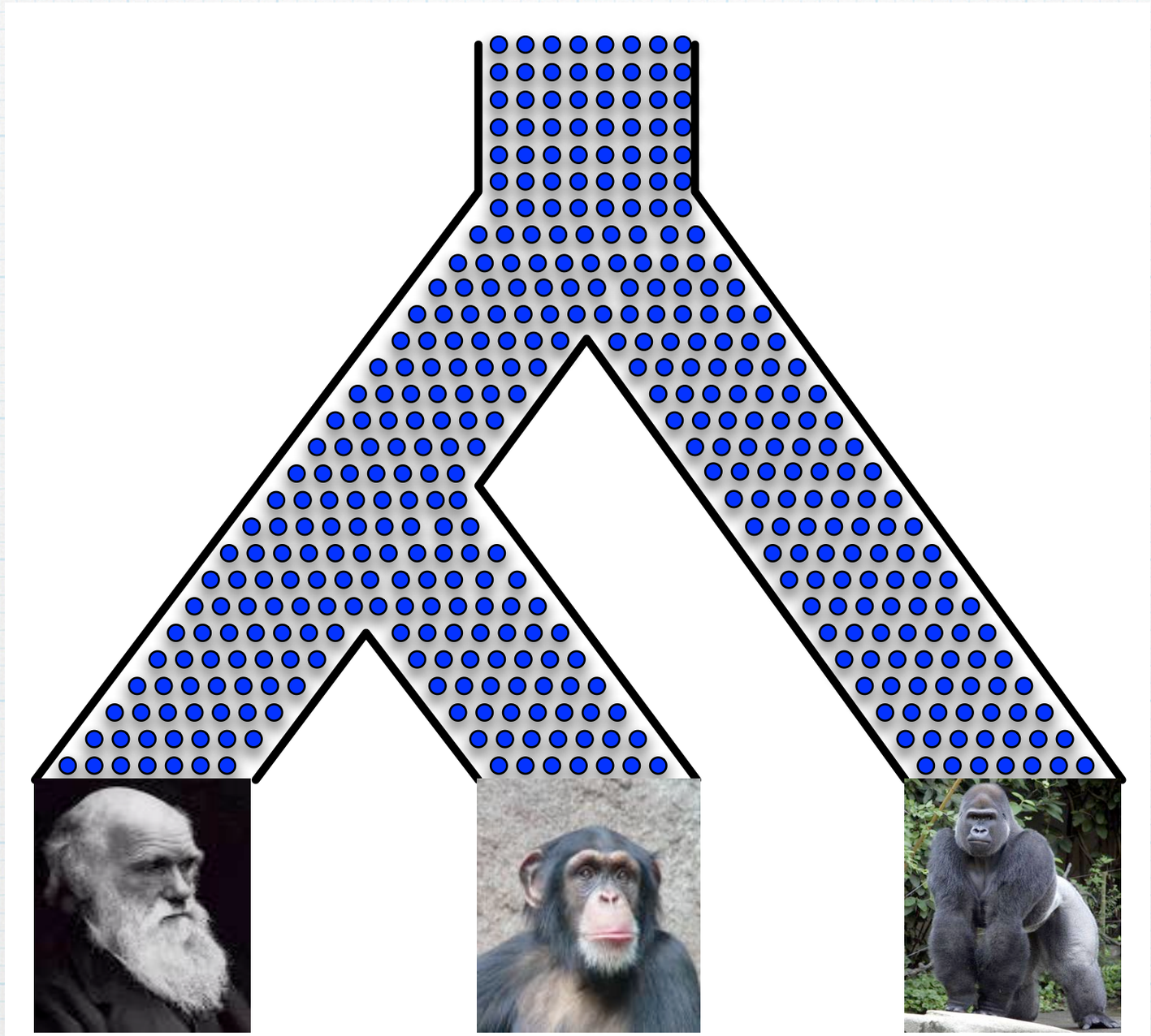
Species Trees



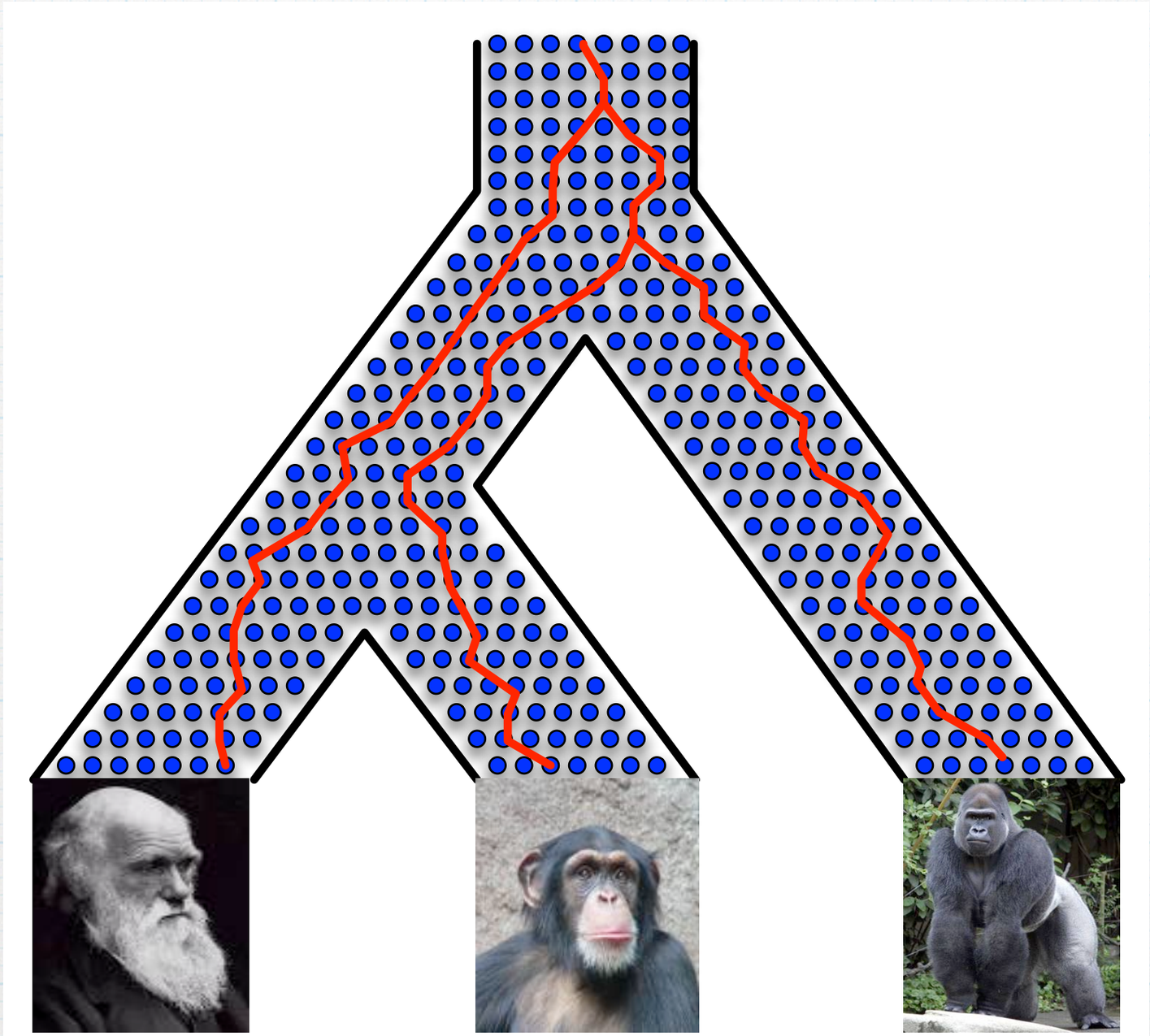
Species Trees and Gene Trees



Species Through the Population Lens

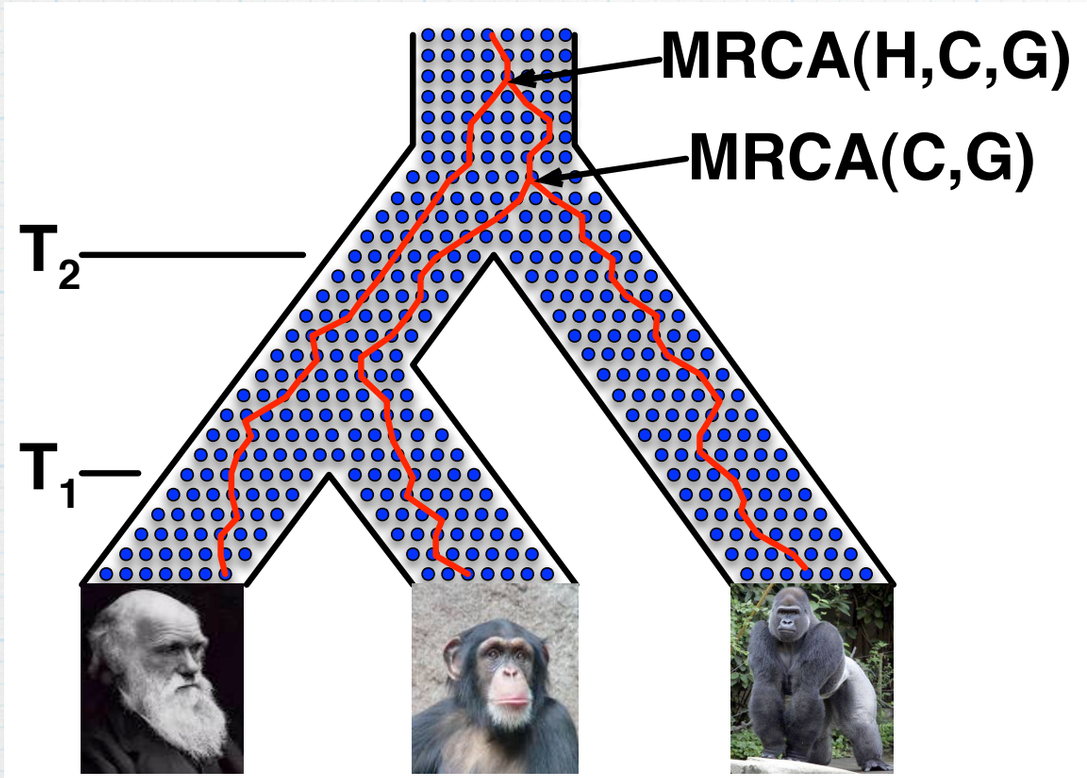


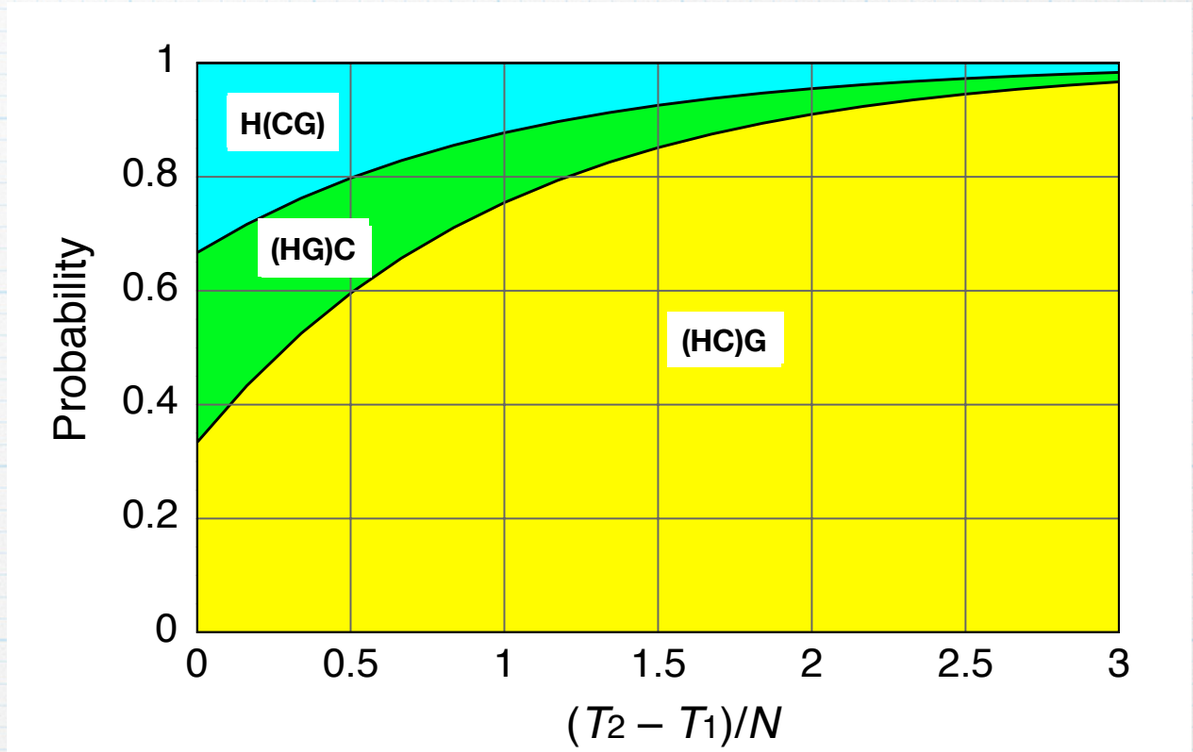
Species Through the Population Lens



Species Through the Population Lens

Species Through the Population Lens





$$\mathbf{P}[((H,C),G)] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}$$

$$\mathbf{P}[((H,G),C)] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

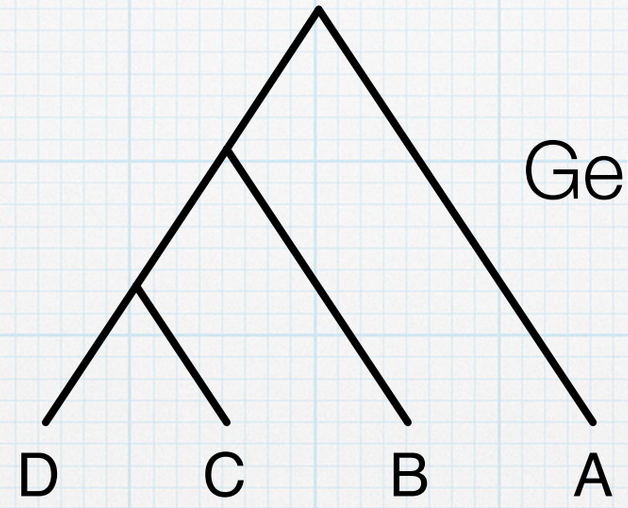
$$\mathbf{P}[(H,(C,G))] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

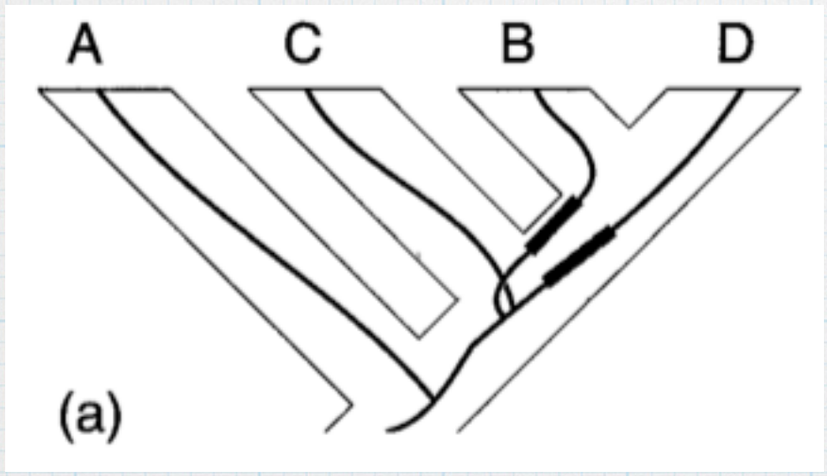
The Inference Problem

* Input: Collection S of sequence alignments, one per locus
* Output: Species tree Ψ and its branch lengths δ that maximizes

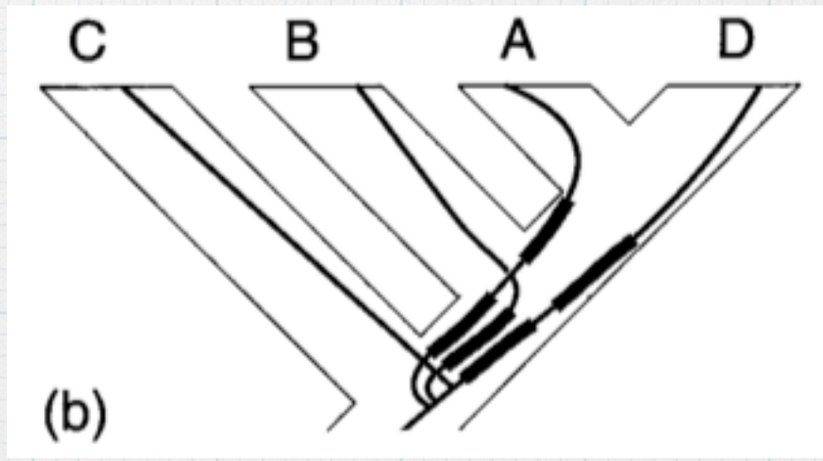
$$\prod_{s \in S} \sum_{g} \int_{\lambda} [\mathbf{P}(s|g,\lambda) \cdot \mathbf{P}(g|\Psi,\delta)]$$

A Maximum Parsimony Formulation



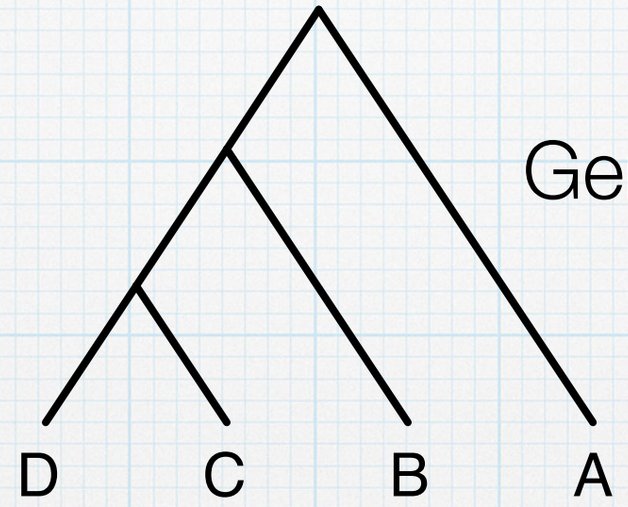


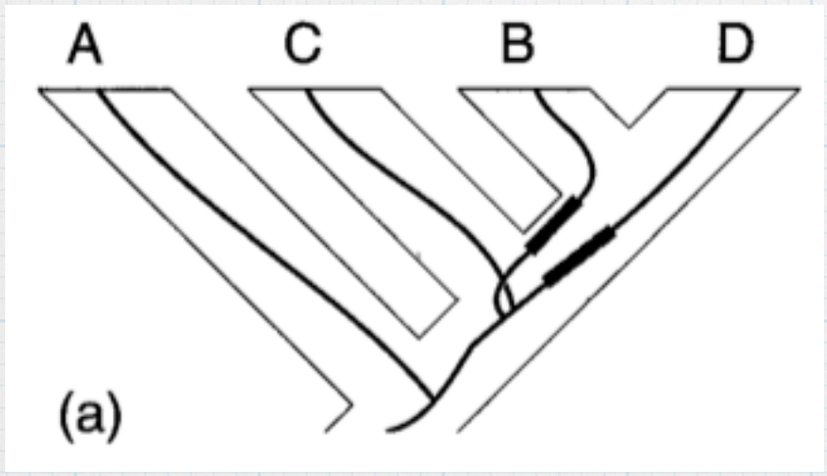
1 extra lineage

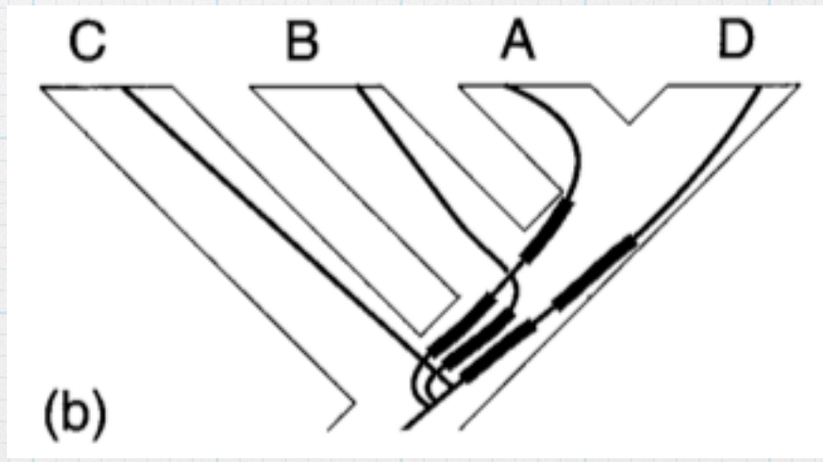


3 extra lineages

A Maximum Parsimony Formulation







1 extra lineage

3 extra lineages

lca mappings

A Maximum Parsimony Formulation

* Input: A collection of gene trees G.
* Output: A species tree that results in the smallest number of extra lineages once all trees in G are reconciled with it under the lca mapping.

A Maximum Parsimony Formulation

* The problem is very hard.
* Heuristics and (worst-case exponential) algorithms exist.