

Tree-Thinking Skills

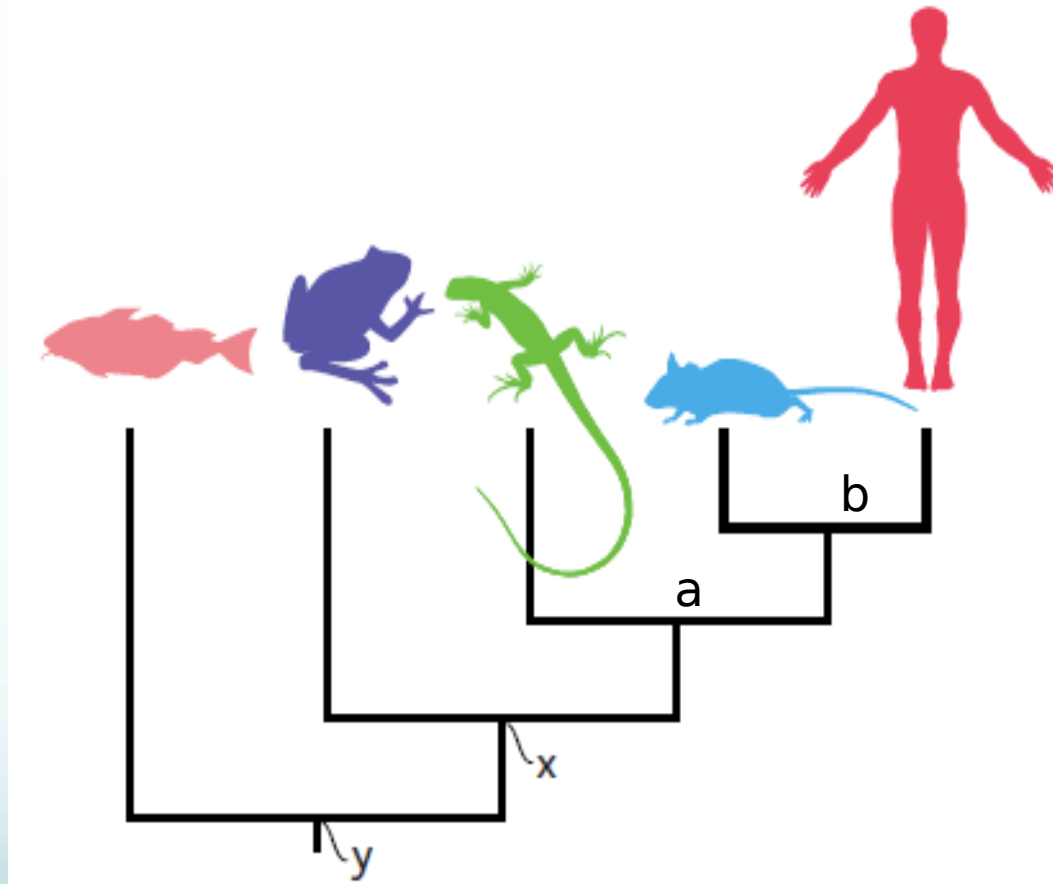
KIZ, Summer 2026

Prof Weigang Qiu

Department of Biological Sciences
Hunter College of City University of New York

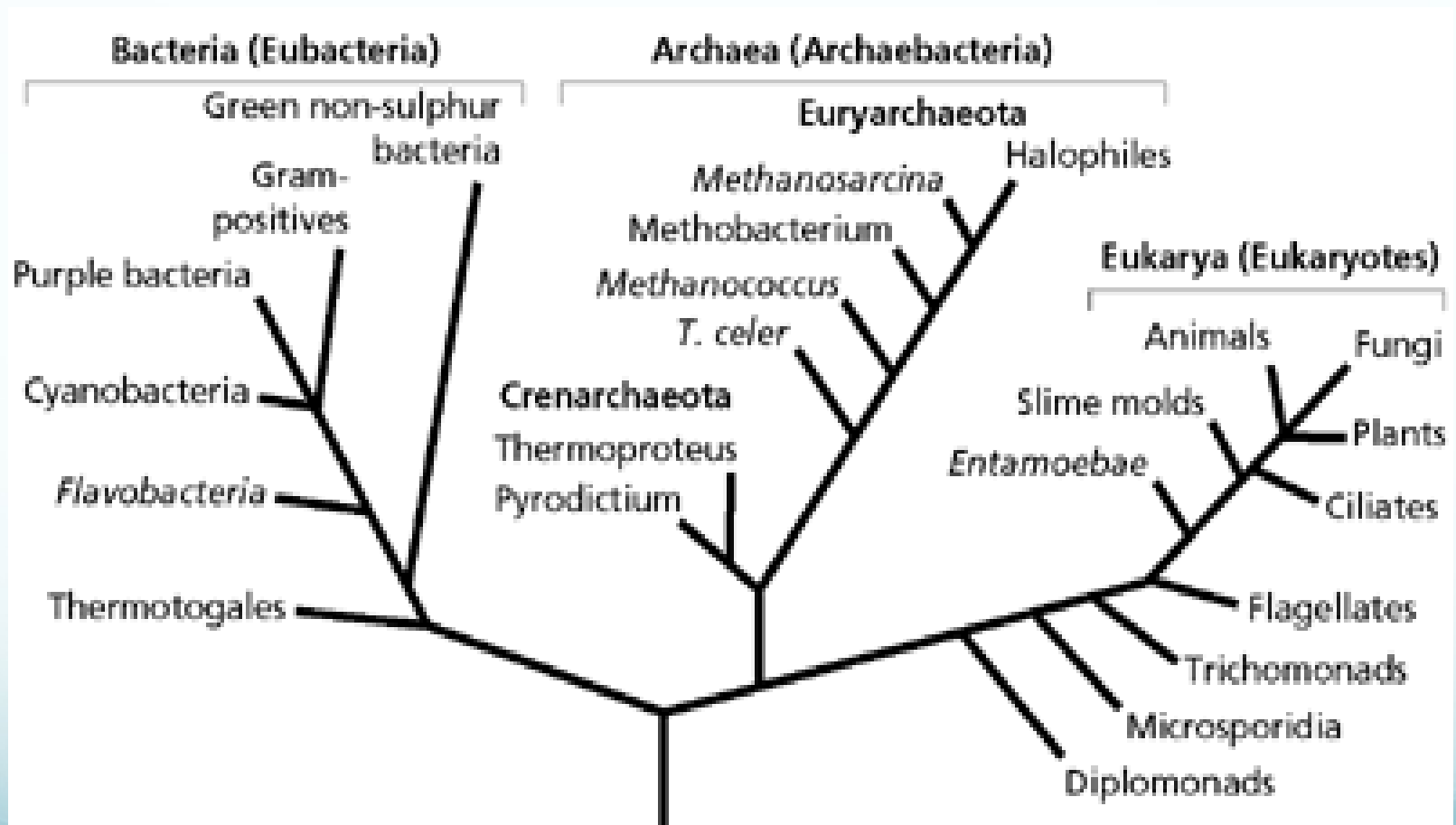
Session I. How to Read a Tree?

Determine Relatedness (**solely**) by Level of Last Common Ancestors



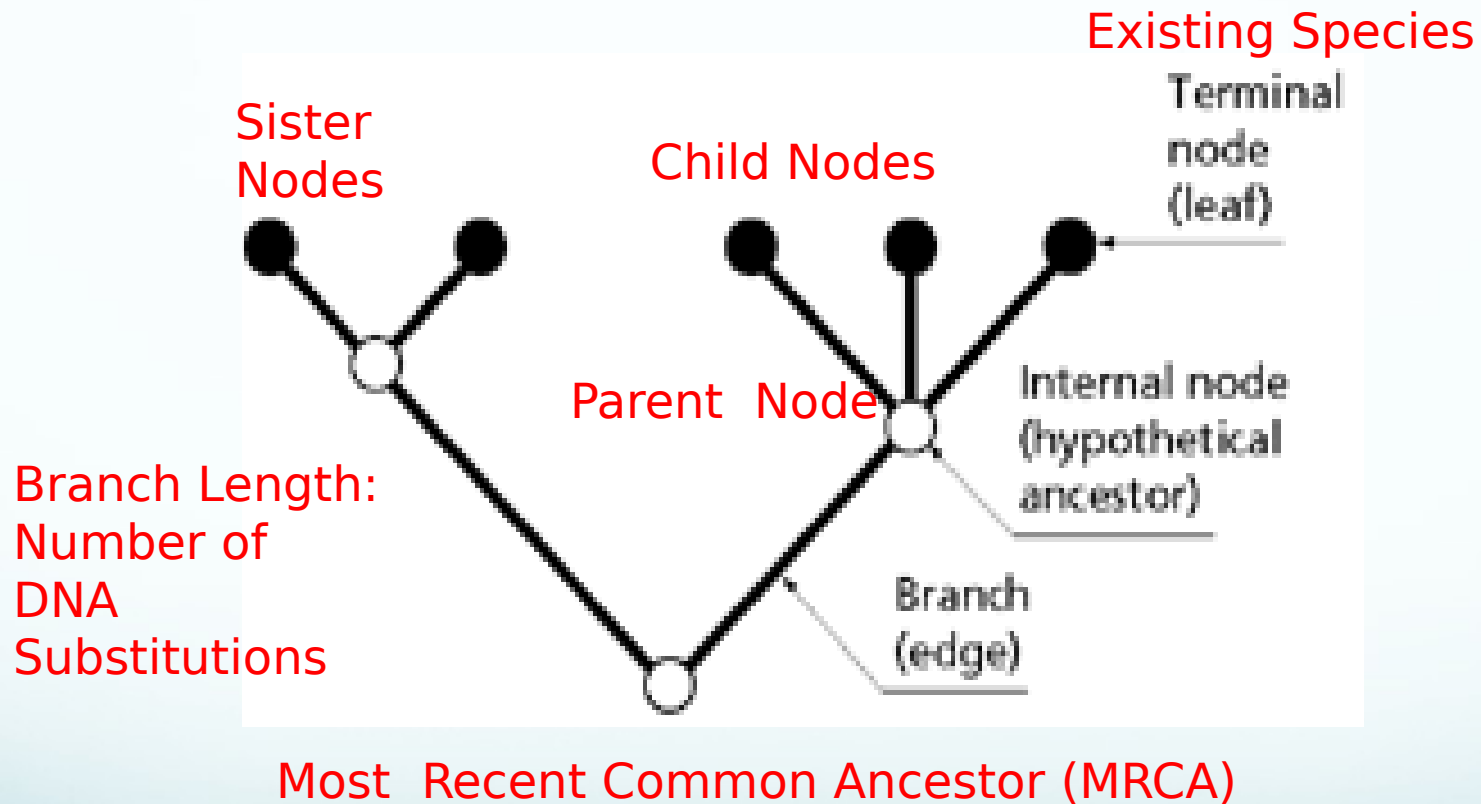
and **label** LCA (last common ancestor) or MRCA (most recent common ancestor)

Tree of Life



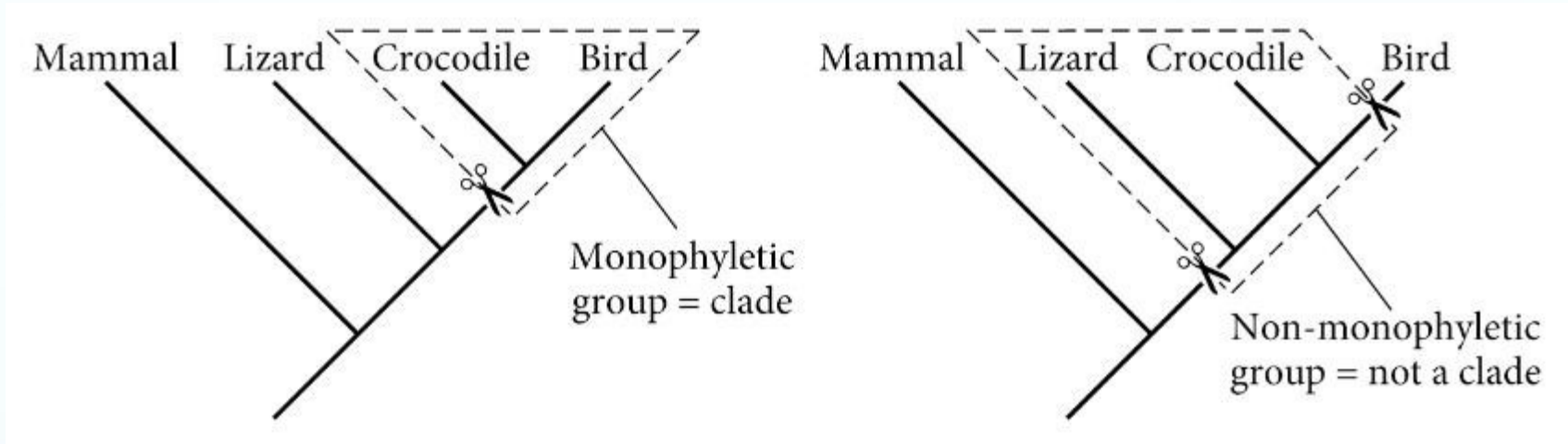
Homology (common ancestry) $\neq$ Similarity (e.g., fish & whale)

Tree Terms



Clades

ALL descendants of a node

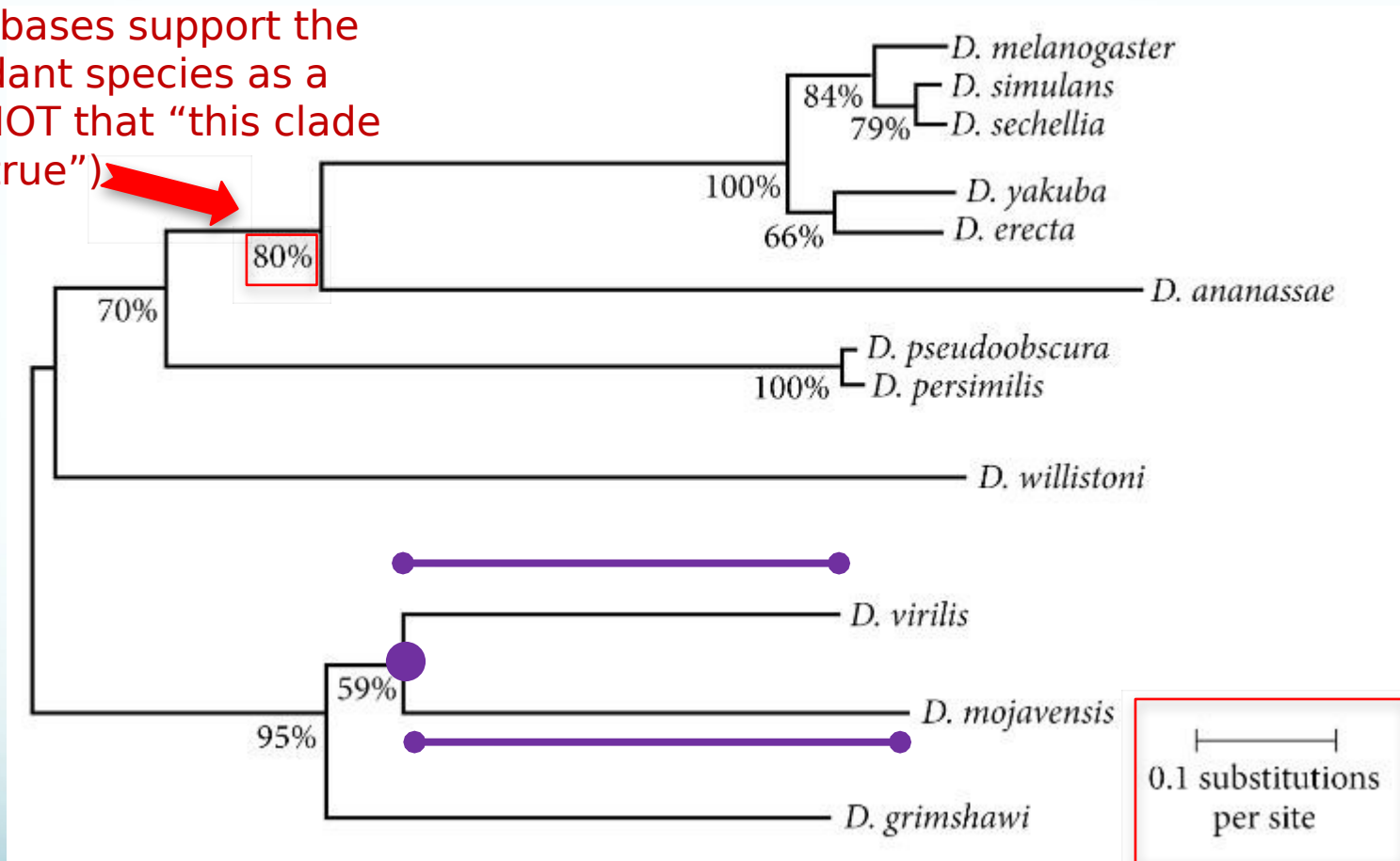


Groups that are not a clade:

- Apes excluding humans
- Prokaryotes
- Dinosaur
- Reptile
- Gram-Negatives
- Animals excluding humans

Tree Distances & Branch Support

Branch support: 80% of aligned bases support the descendant species as a clade (NOT that “this clade is 80% true”)

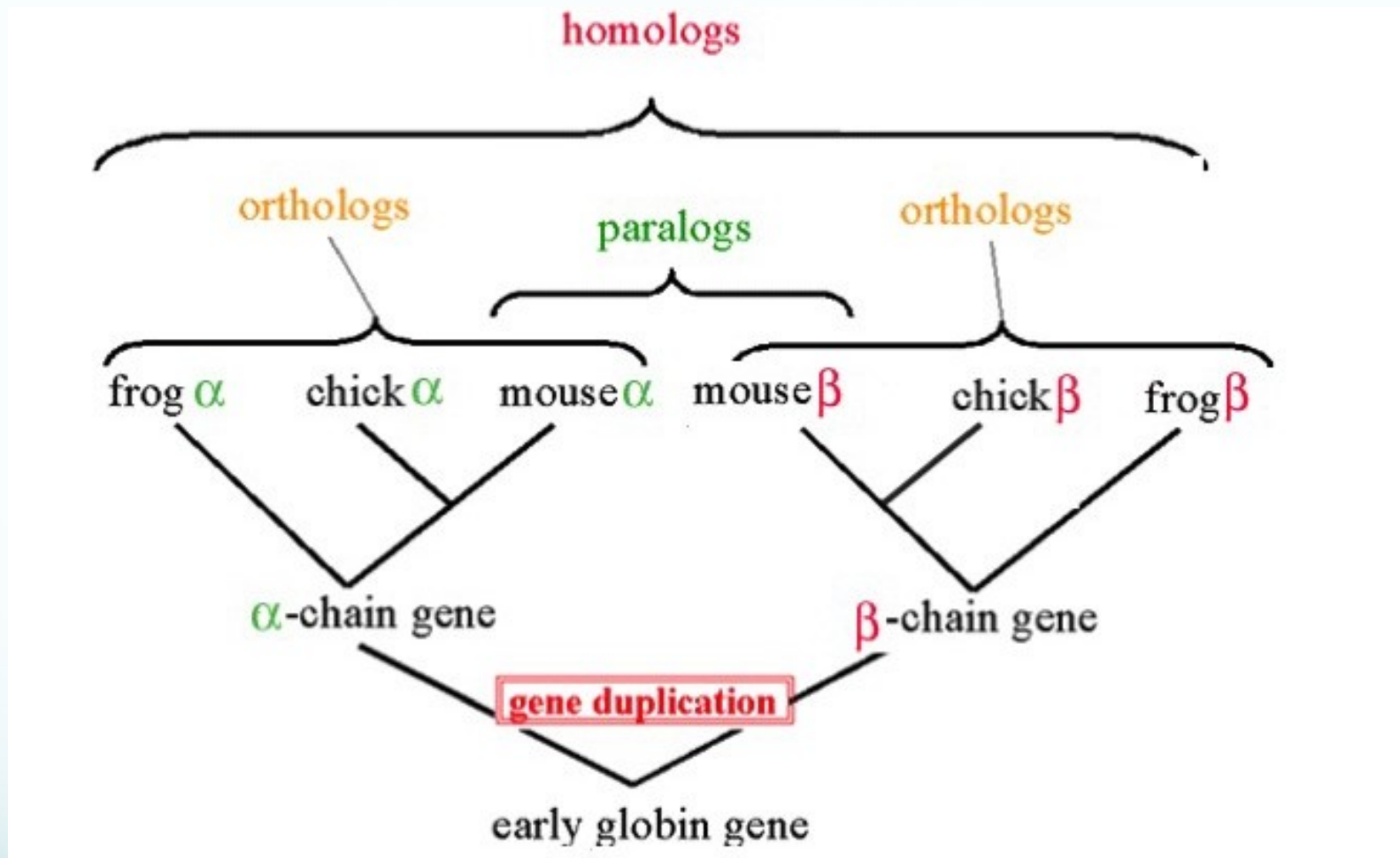


Evolutionary distance between *D. virilis* & *D. mojavensis* is ~ 0.7 sub/site, estimated by adding two distances since their last common ancestor (LCA, dot) & then multiplying it by the scale bar (0.1 per site = 10% sites)

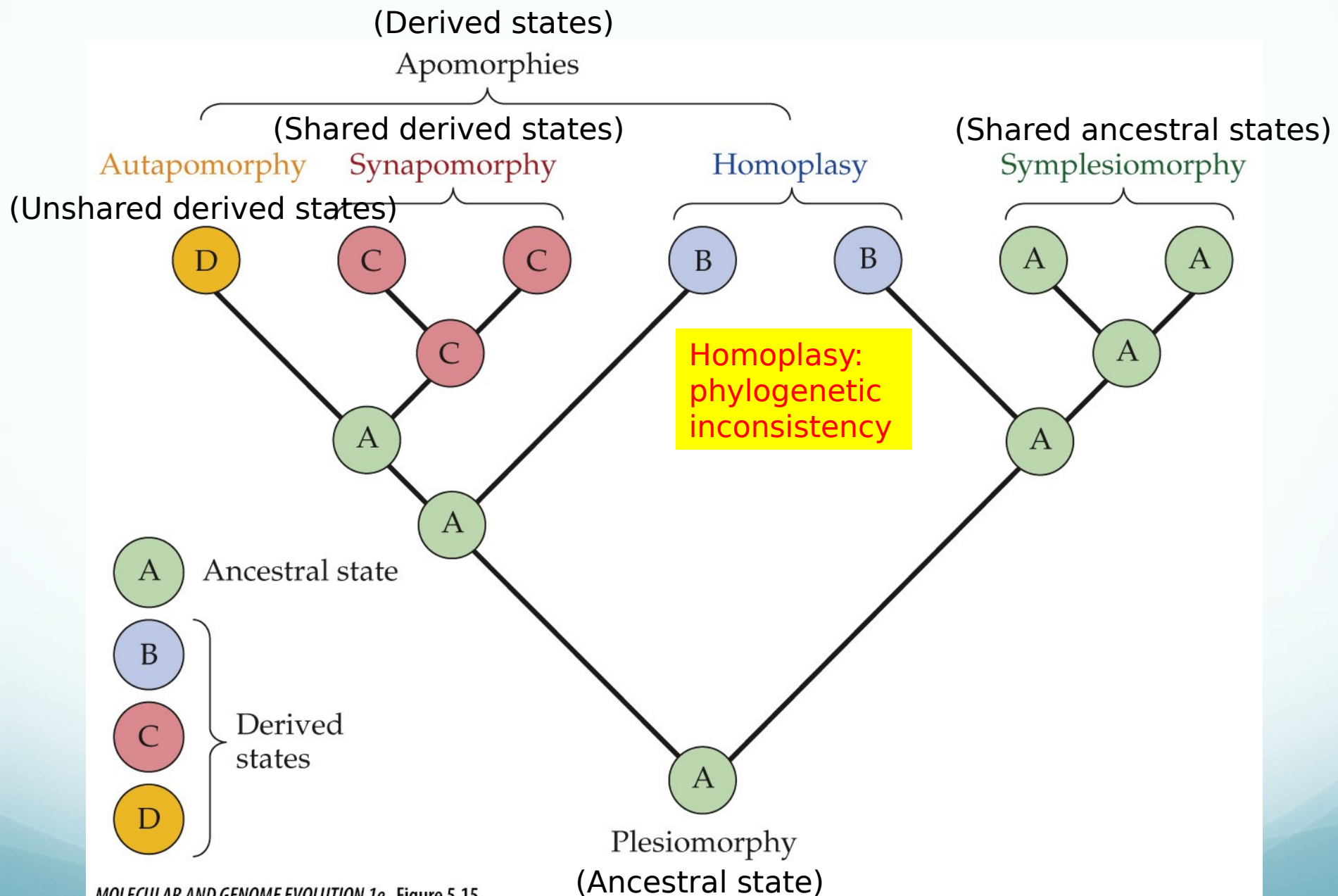
Session II. How to Build a Tree?

- Choice of molecular markers:
 - **Slow-evolving** sequences for **distantly related** species/individuals (e.g., protein sequence for vertebrates, 16S rRNA for bacteria)
 - **Fast-evolving** sequences for **closely related** species/individuals (e.g., mtDNA or non-coding sequences for humans, DNA sequences for primates)
- Identify homologs (BLAST, HMMER)
- Obtain multiple alignment (“muscle”, “clustal”)
- Infer tree (Neighbor-Joining; Parsimony; Likelihood; Bayesian)
- Obtain branch support (Bootstrap; Posterior support)

Gene Tree vs Species Tree

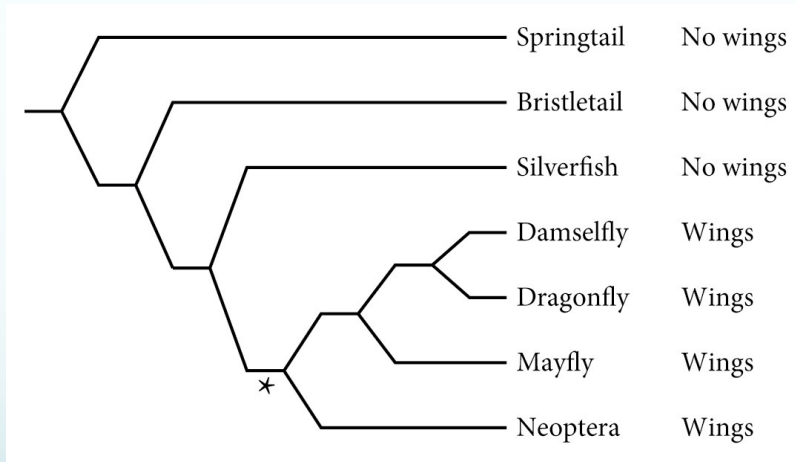


- Orthologs: Descendants of a speciation event; similar mol. function
- Paralogs: Descendants of a gene duplication event; often differ in function
- Paralogs from the same species (genome) are often more divergent than orthologs from different species (genomes) because duplication predates speciation



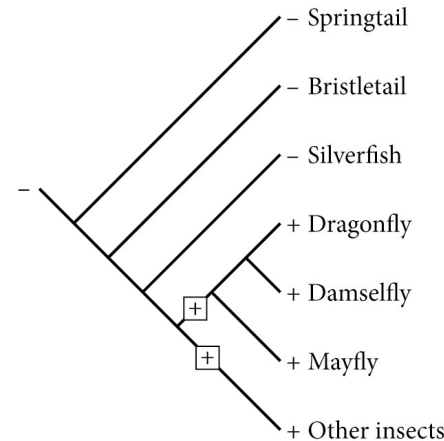
Phylogenetic Consistency & Parsimony

$$C.I. = \frac{N_{\text{smallest_number_of_changes}}}{N_{\text{observed_number_of_changes}}}$$

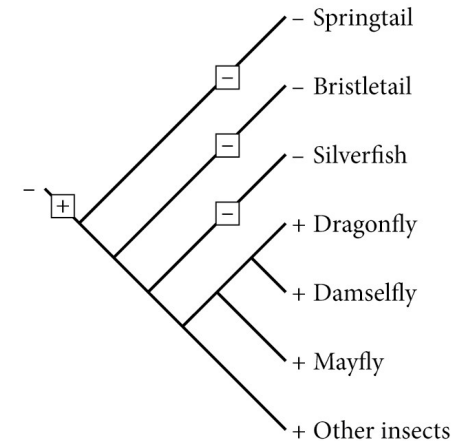


Consistency Index (C.I.)=1

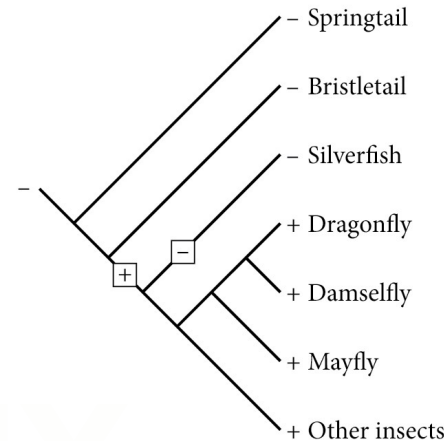
PARSIMONY



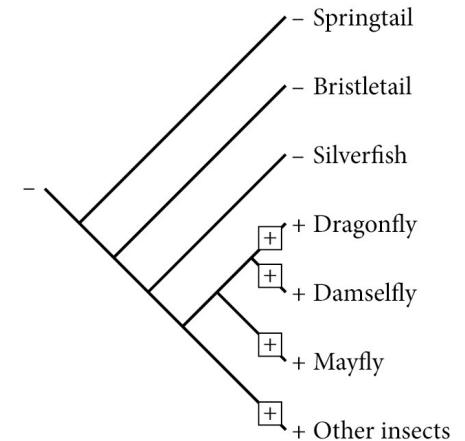
Number of changes = 2
CI = 0.5



Number of changes = 4
CI = 0.25

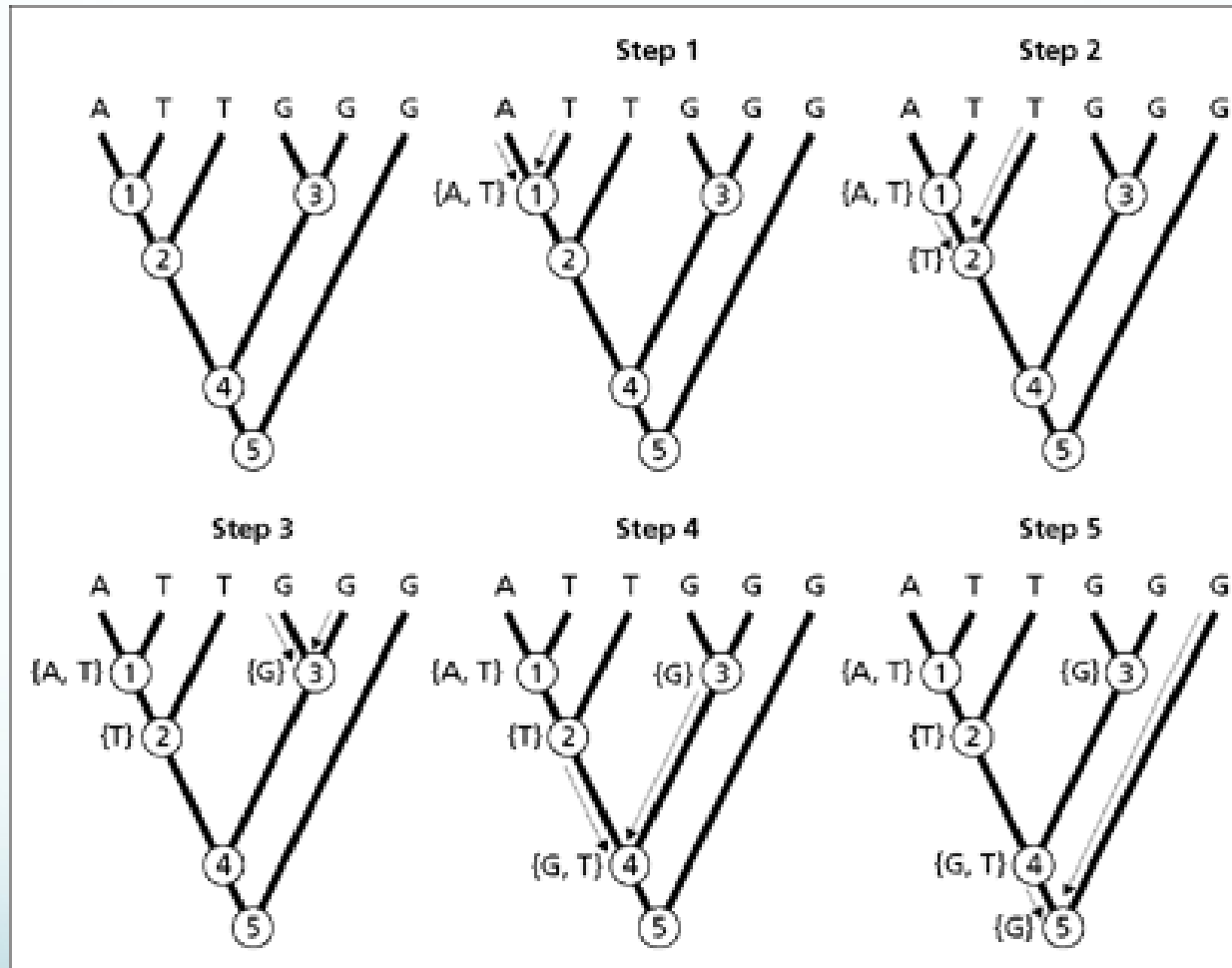


Number of changes = 2
CI = 0.5



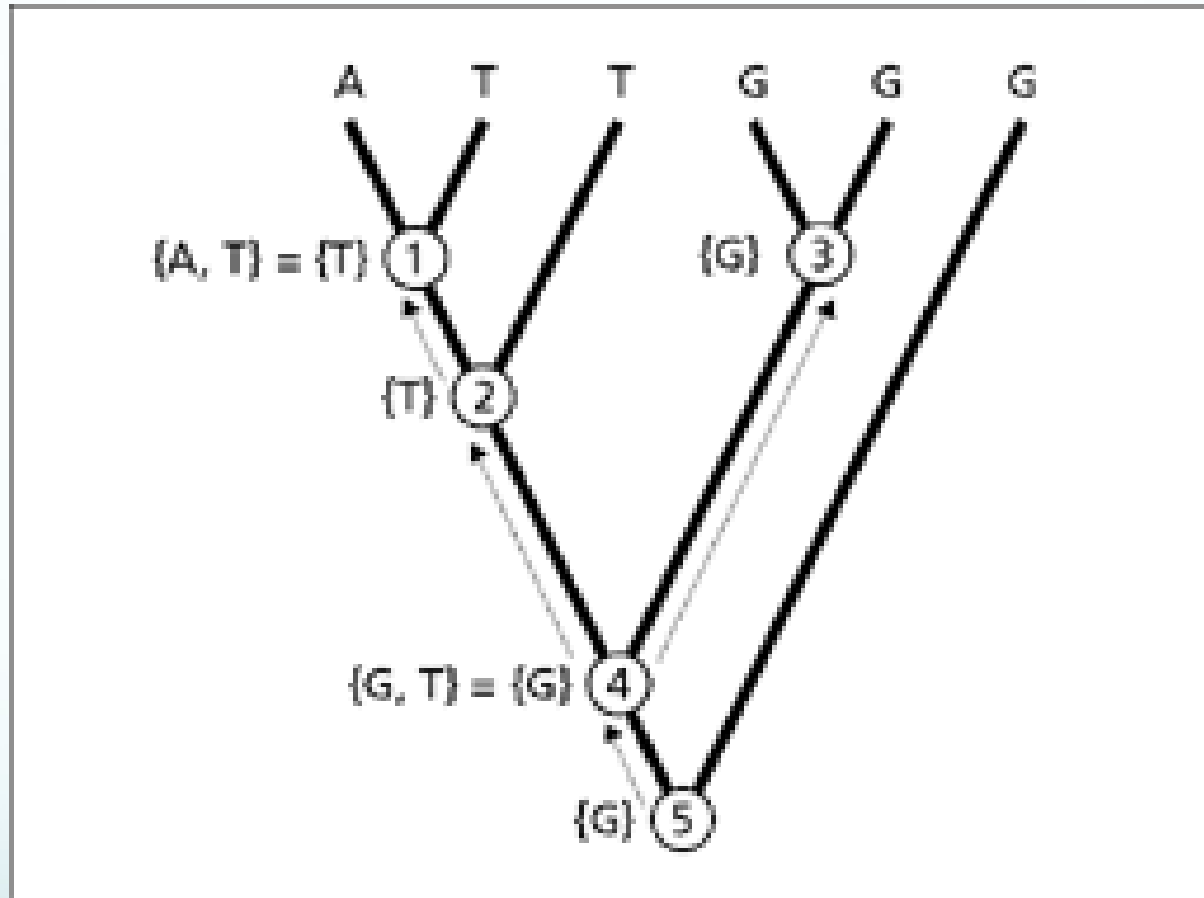
Number of changes = 4
CI = 0.25

Parsimony Reconstruction (1)



Rule 1: From two sister nodes, infer parent state by choosing intersection, if any. Otherwise, take the union.

Parsimony Reconstruction (2)



Rule 2: From a parent node, infer child state by choosing intersection, if any. Otherwise, select an arbitrary state

Algorithm Challenge: Recursive Tree Walk

Algorithm 20: Recursive tree walk

input : *tree*, a phylogenetic tree; *node*, starting node for tree walk

output: None

```

1 Function walk.tree(tree, node) :
2   if node is a tip then
3     return
4   else
5     children ← descendants(tree, node);
6     for child in children do
7       walk.tree(tree, child);
8     end
9   end
10 root ← get.root(tree);
11 walk.tree(tree, root);
  
```

The algorithm constructs a function `walk.tree`. This function finds the direct descendants of a node and then calls itself with each child as the input. The function returns when it reaches a tip. To walk the entire tree, the function is first called with the root node.

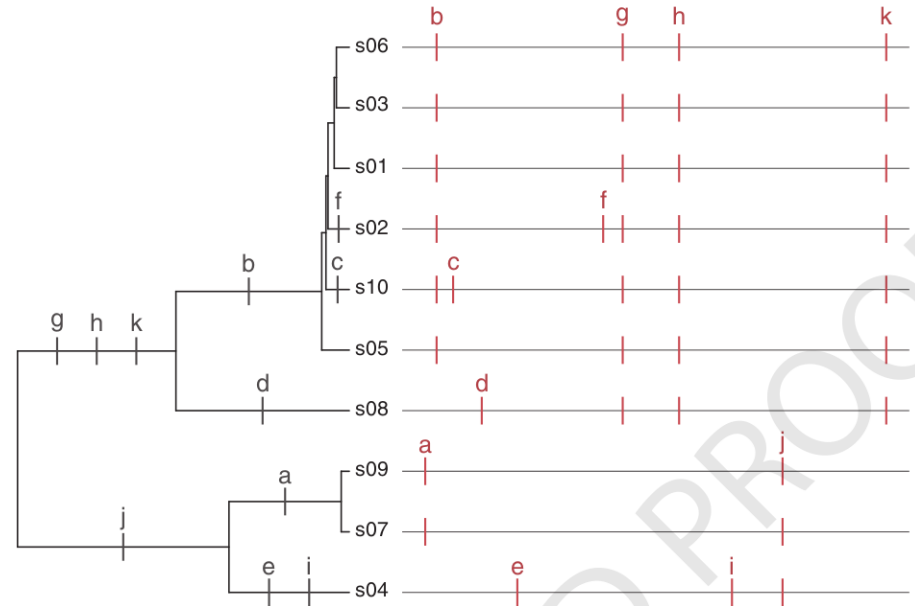


Figure 3.3 Sequences in a population are related by common descent (Algorithm 10). A sample of $n = 10$ haplotypes (s01–s10) were simulated with $\theta = 2N\mu = 10$ using the `ms` program (Hudson, 2002). The program outputs a coalescent tree (left panel), SNP positions (uniformly distributed along the sequence), and the haplotype sequences (right panel, with the red tick marks representing the derived alleles a–k). Mutations occurring on the tree branches (in gray) were inferred according to the phylogenetic distribution of the derived alleles.