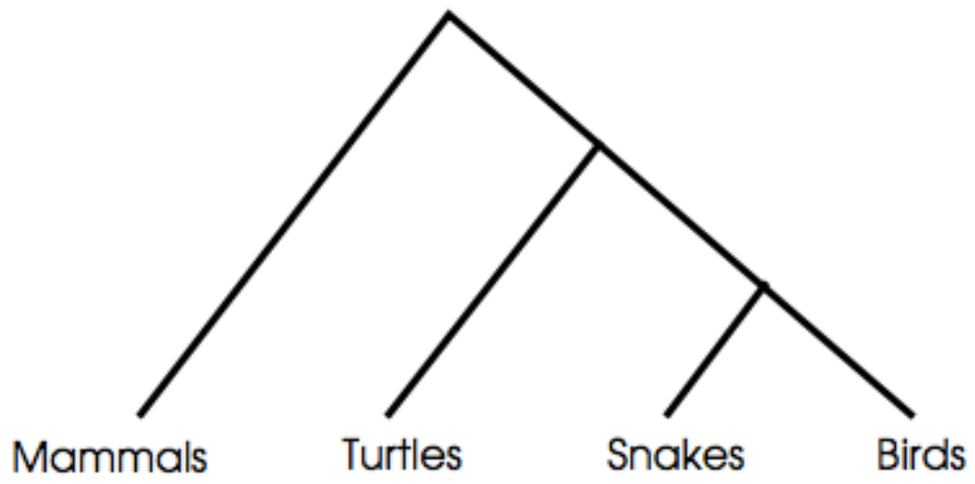
Topic 8: Phylogenomics with SNPs



Overview

- What is phylogenetic
- Terms and outline
- Methods of trees
    - UPGMA
    - Neighbour joining
    - Maximum parsimony
    - Maximum likelihood
- Distance calculations
- Considerations for SNPs
- Reticulate networks

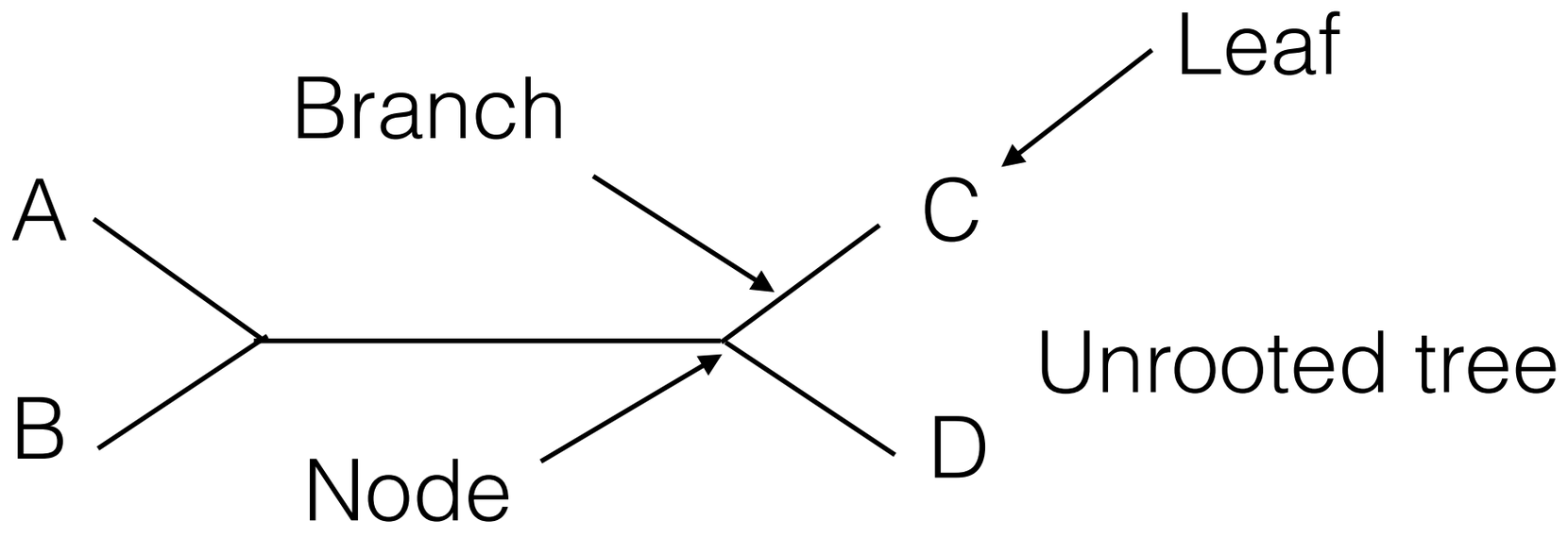
Learning Goals

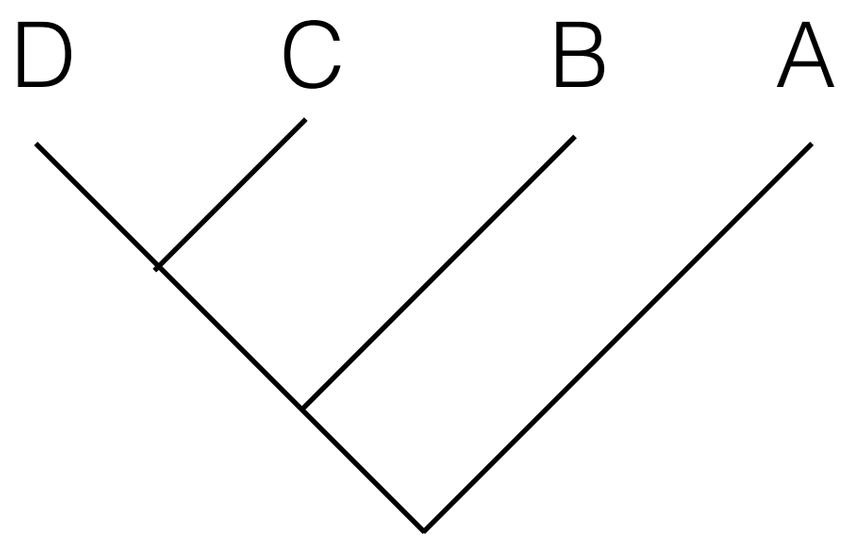
- What are the different methods of building a phylogenetic tree?
- What are the different methods of calculating phylogenetic distance?
- How can confidence be measured in phylogenetic trees?
- How can you use SNPs for phylogenetic analysis?

Phylogenetics

- Reconstructs evolutionary ties between organisms.
- Estimates divergence times between organisms.
- Can use morphological or genetic data.

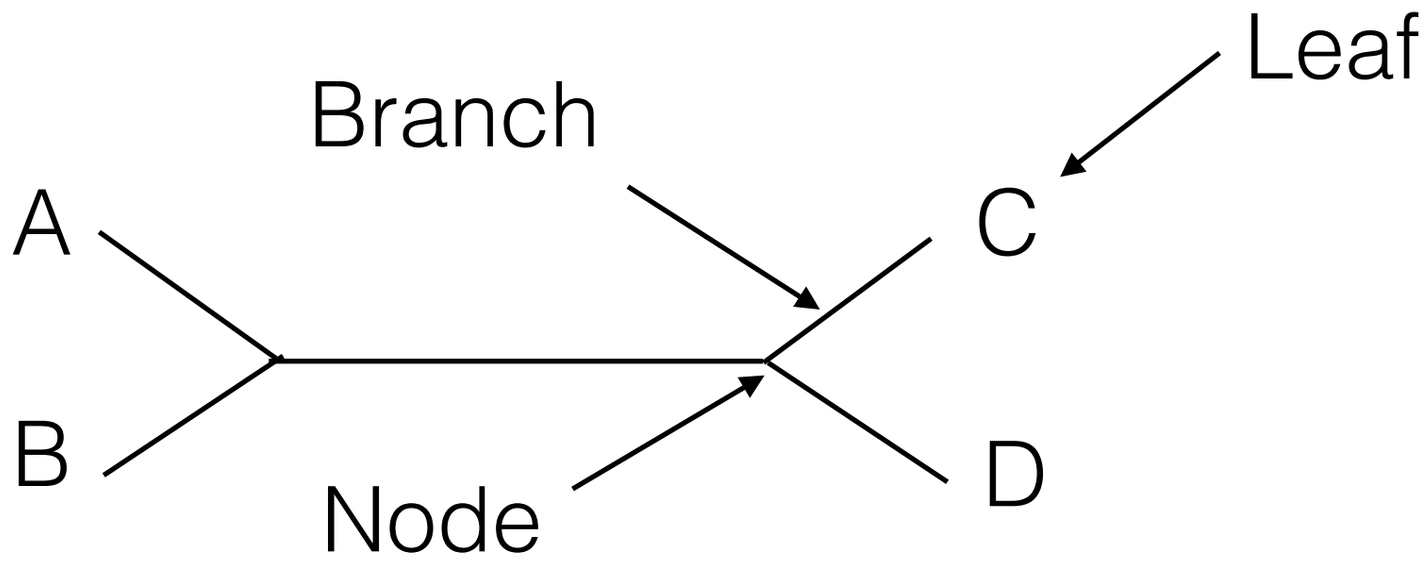
Terms





Rooted tree

Terms



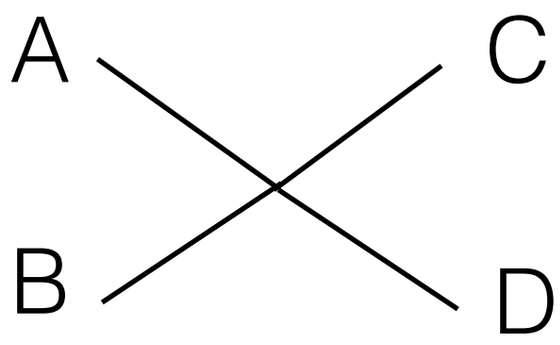
Operational Taxonomic Unit (OTU): An external node representing a monophyletic group

Distance methods

- Tries to build a tree where the distances measured between leaves on the tree correspond to the actual distance between objects

Distance methods

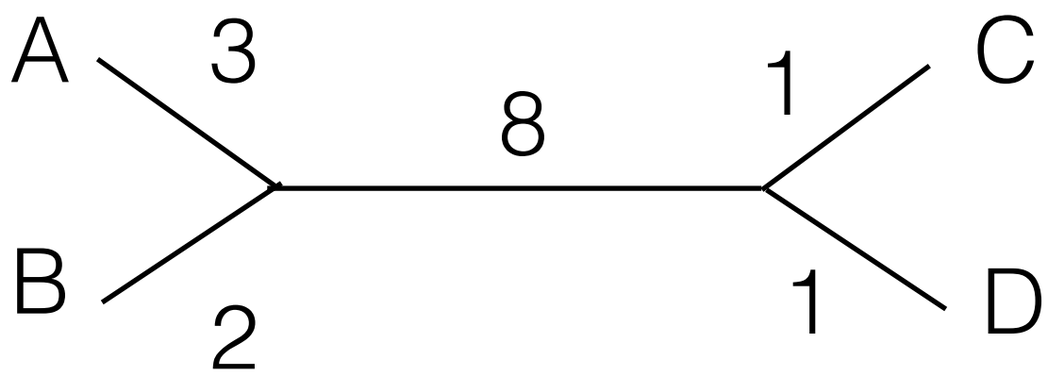
- Tries to build a tree where the distances measured between leaves on the tree correspond to the actual distance between objects



	A	B	C	D
A	0			
B	5	0		
C	12	11	0	
D	12	11	2	0

Distance methods

- Tries to build a tree where the distances measured between leaves on the tree correspond to the actual distance between objects



	A	B	C	D
A	0			
B	5	0		
C	12	11	0	
D	12	11	2	0

Distance methods

- Tries to build a tree where the distances measured between leaves on the tree correspond to the actual distance between objects

- Easy to calculate when distance matrix is additive, but often is not.

- Need to use heuristics to pick best fitting tree because there are too many to try all.

UPGMA

- Unweighted Pair Group Method with Arthmetic Mean
- A sequential clustering algorithm
- Very fast and often inaccurate. Used as a starting point for other methods.

UPGMA

- Pick pair with lowest distance and build a composite OTU.

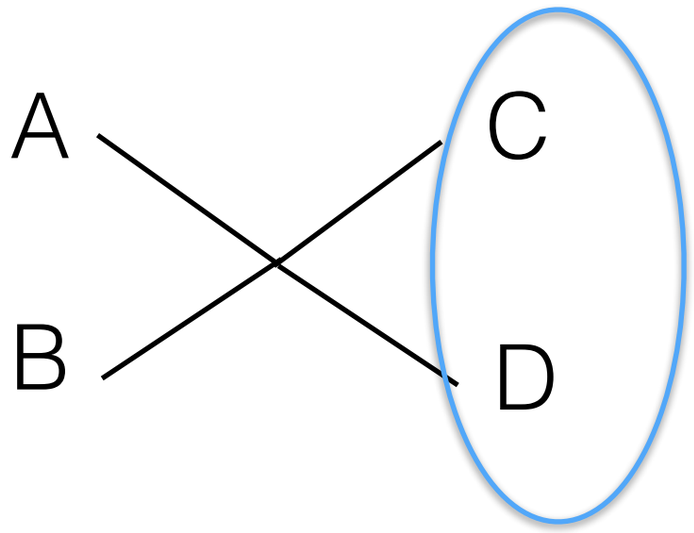
- With new group OTU, find next pair and make and OTU

UPGMA

- Pick pair with lowest distance and build a composite OTU.

- With new group OTU, find next pair and make and

OTU



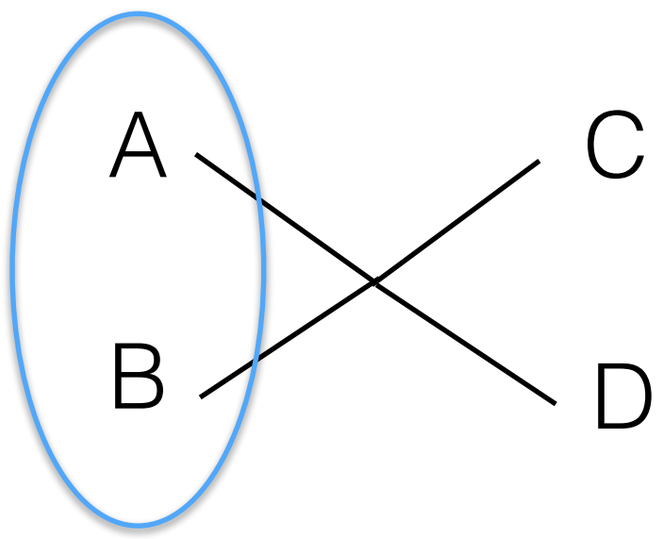
	A	B	C	D
A	0			
B	5	0		
C	10	7	0	
D	9	8	2	0

UPGMA

- Pick pair with lowest distance and build a composite OTU.

- With new group OTU, find next pair and make and

OTU

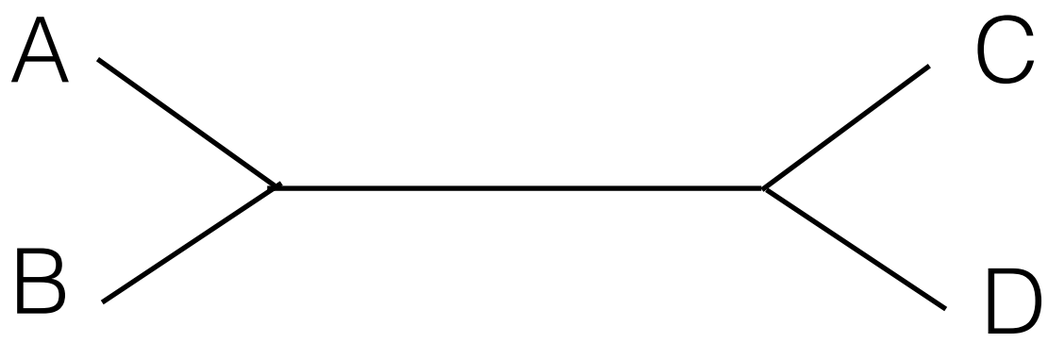


	(AB)	(CD)
(AB)	0	
(CD)	8.5	0

UPGMA

- Pick pair with lowest distance and build a composite OTU.

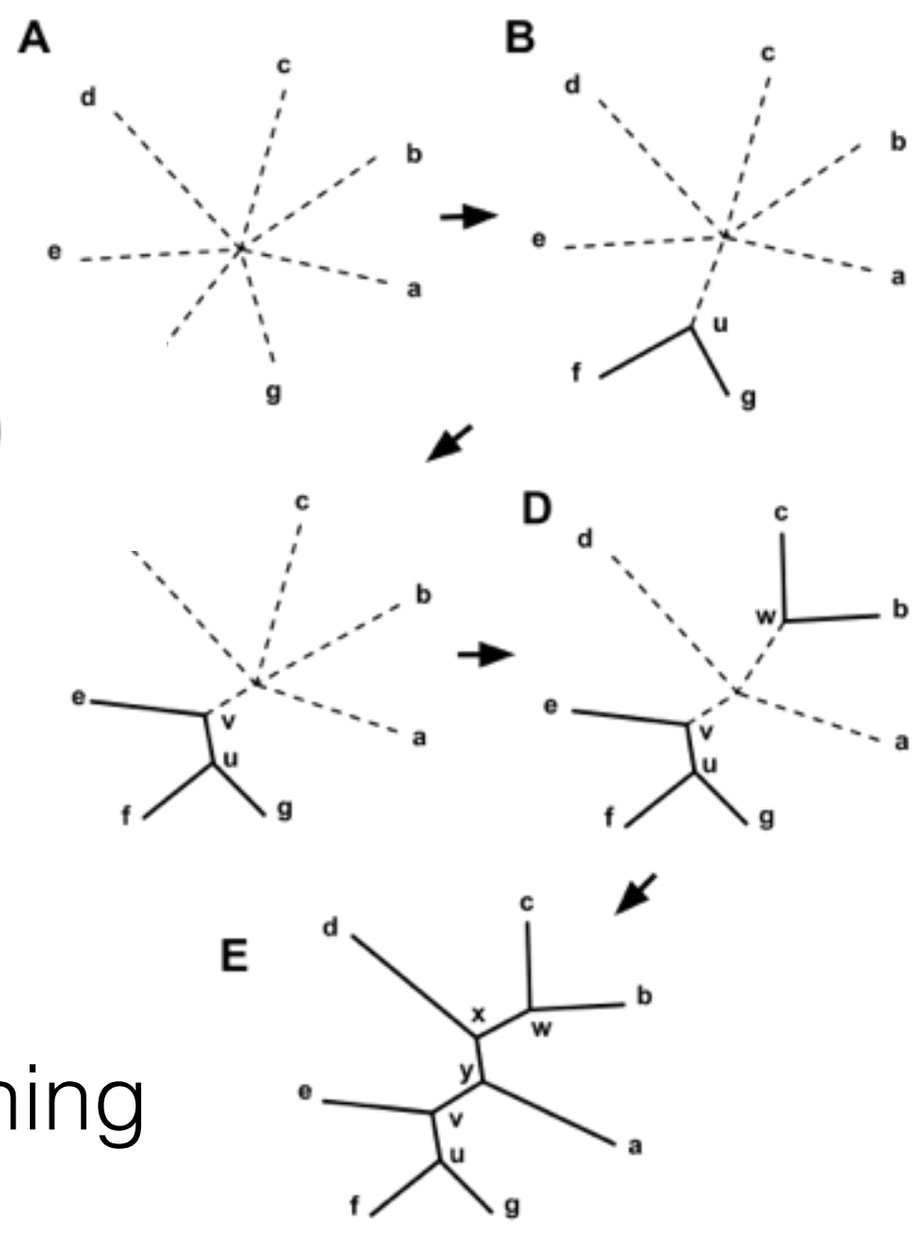
- With new group OTU, find next pair and make and OTU



	AB	B	(CD)
A	0		
B	5	0	
(CD)	9.5	7.5	0

Neighbour-joining

- Similar to UPGMA, but uses a Q-matrix.



$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

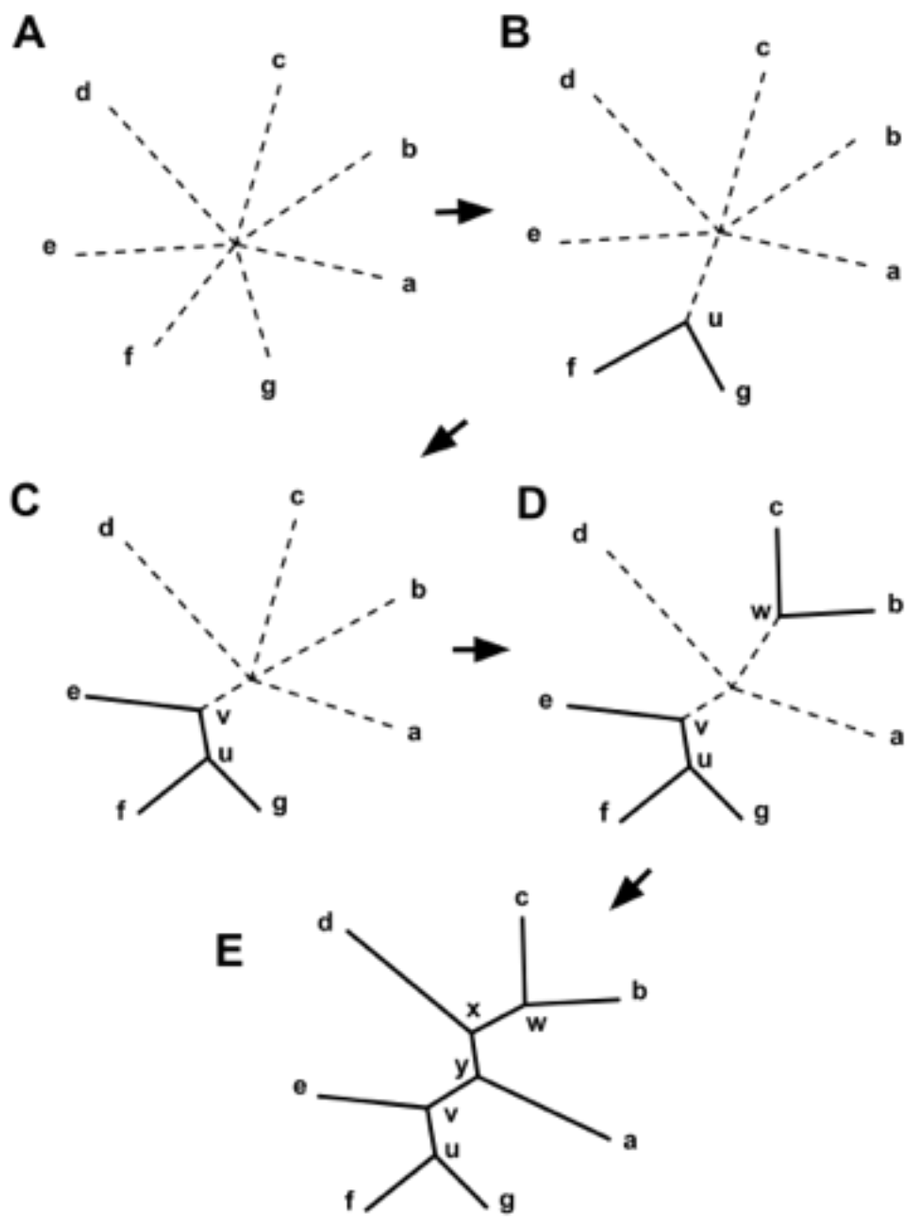
Distance i to j

Distance i to everything

Distance j to everything

Neighbour-joining

- Similar to UPGMA, but uses a Q-matrix.
- Allows uneven branches when grouping.
- Fast



Maximum parsimony

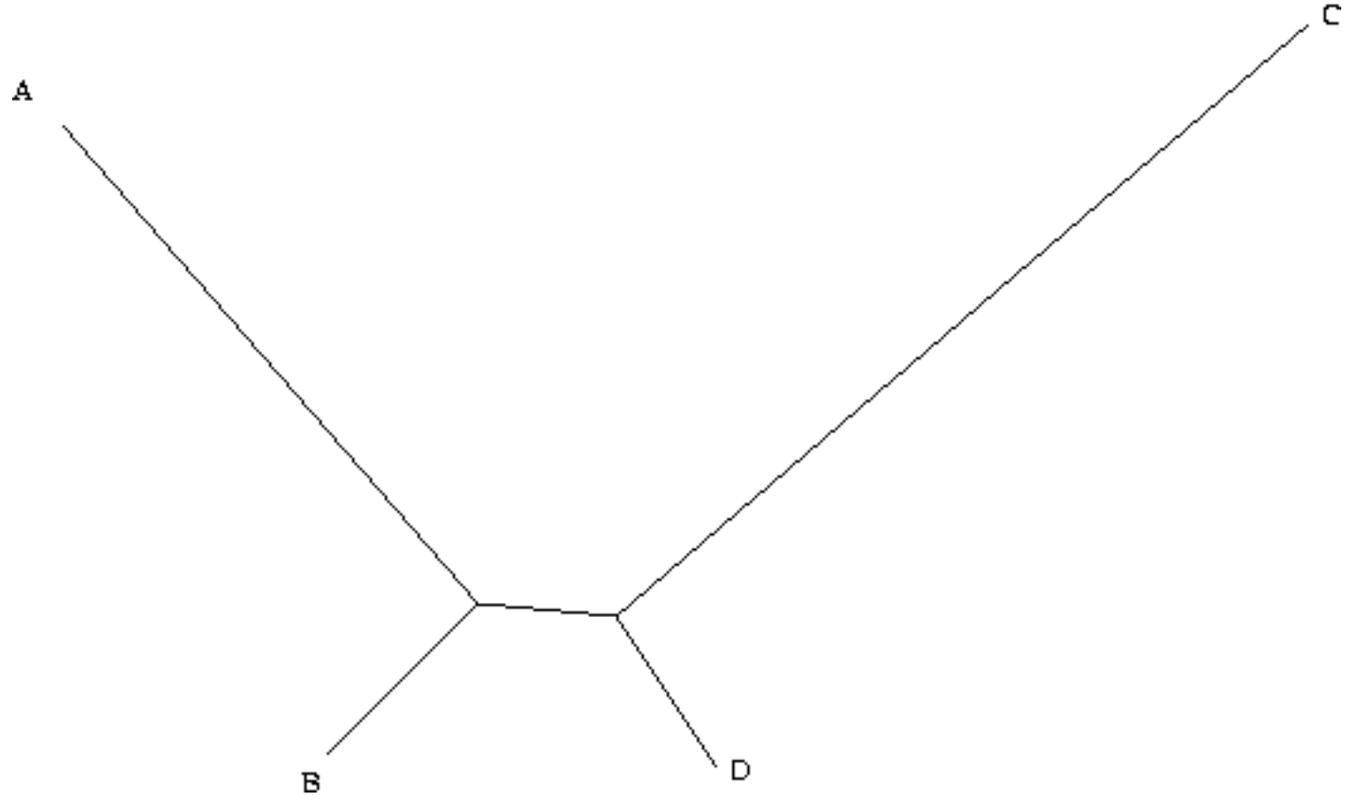
- Looking for a phylogenetic tree with the minimal number of character state changes.

- No easy way to find the most parsimonious tree, must search through many trees.

- Can have problems with long branch attraction

Maximum parsimony

- Can have problems with long branch attraction



Maximum Likelihood

- Calculates the likelihood of trees based on substitution models and picks the model with the highest likelihood.

- Uses heuristics to search through the possible tree space.

Bayesian Trees

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

- Finds the tree with the highest posterior probability based on a model of evolution and prior probabilities.

- Can use MCMC to search the tree space.

Substitution models

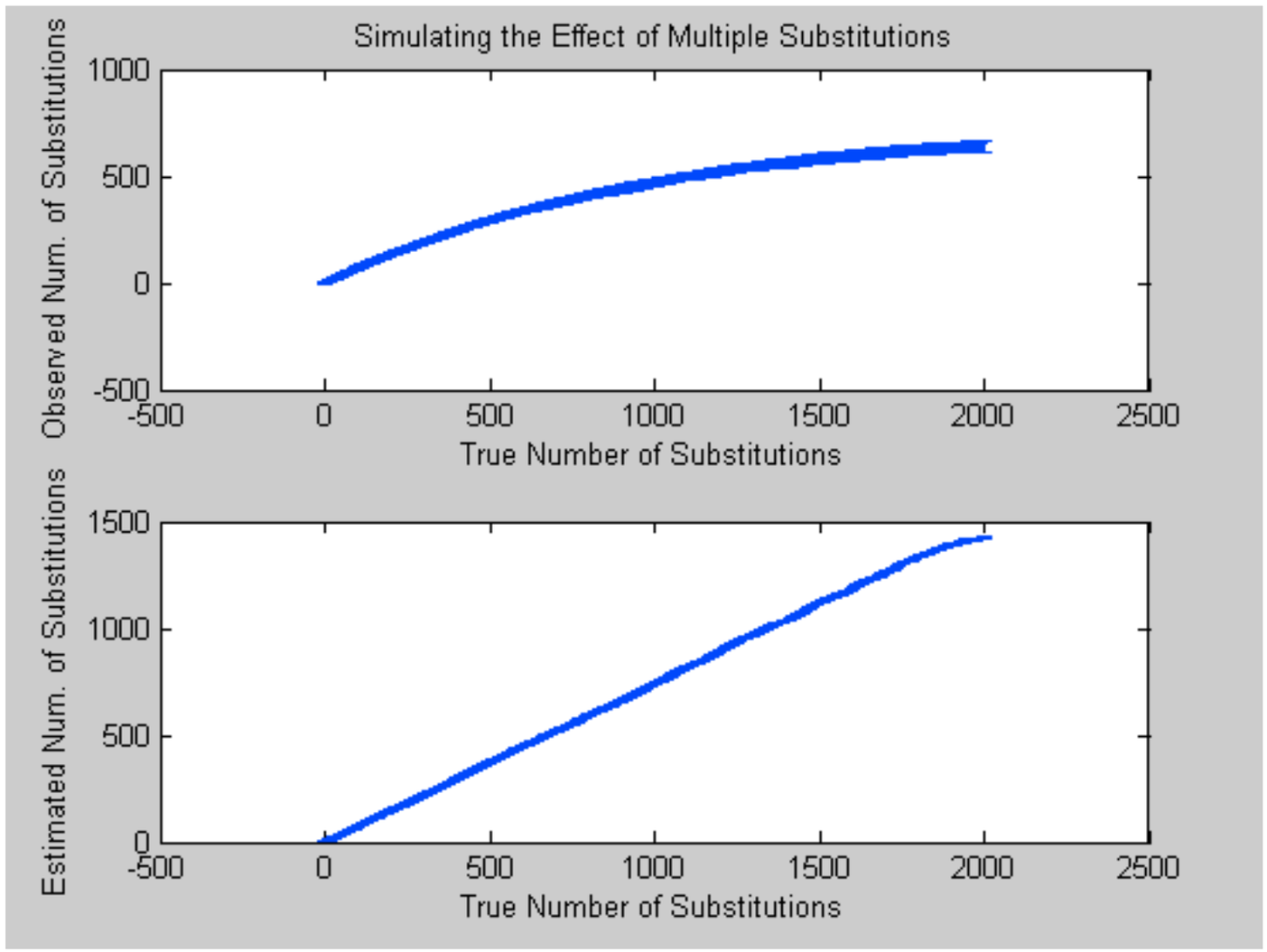
- Many different models of varying complexity.
  - Equal or unequal mutation rates
  - Equal or unequal base frequencies
  - GC bias or not
- More parameters not always better, can overfit to your data, so you should use a program to pick the best model.

Substitution models

Hidden Changes

Seq1 A -> T -> C
Seq2 A -> C

- Also helps control for saturation of mutations and hidden changes



Bootstrapping

- Repeat your analysis with a bootstrapped version of your dataset X1000 times.
    - Bootstrapping involves building a dataset of the same number of characters by sampling with replacement from your original dataset.

- The percent of bootstrap datasets that produce the same tree is your confidence value.

Considerations for SNPs

- Generally you only keep variable sites, while many phylogenetic algorithms assume invariant sites are included.

- Need to use models that explicitly control for the ascertainment bias, or include invariant sites.

Reticulate networks

- Phylogenetic history is not always perfectly bifurcating. Gene flow can occur between species and different loci can have different phylogenetic histories.

Reticulate networks

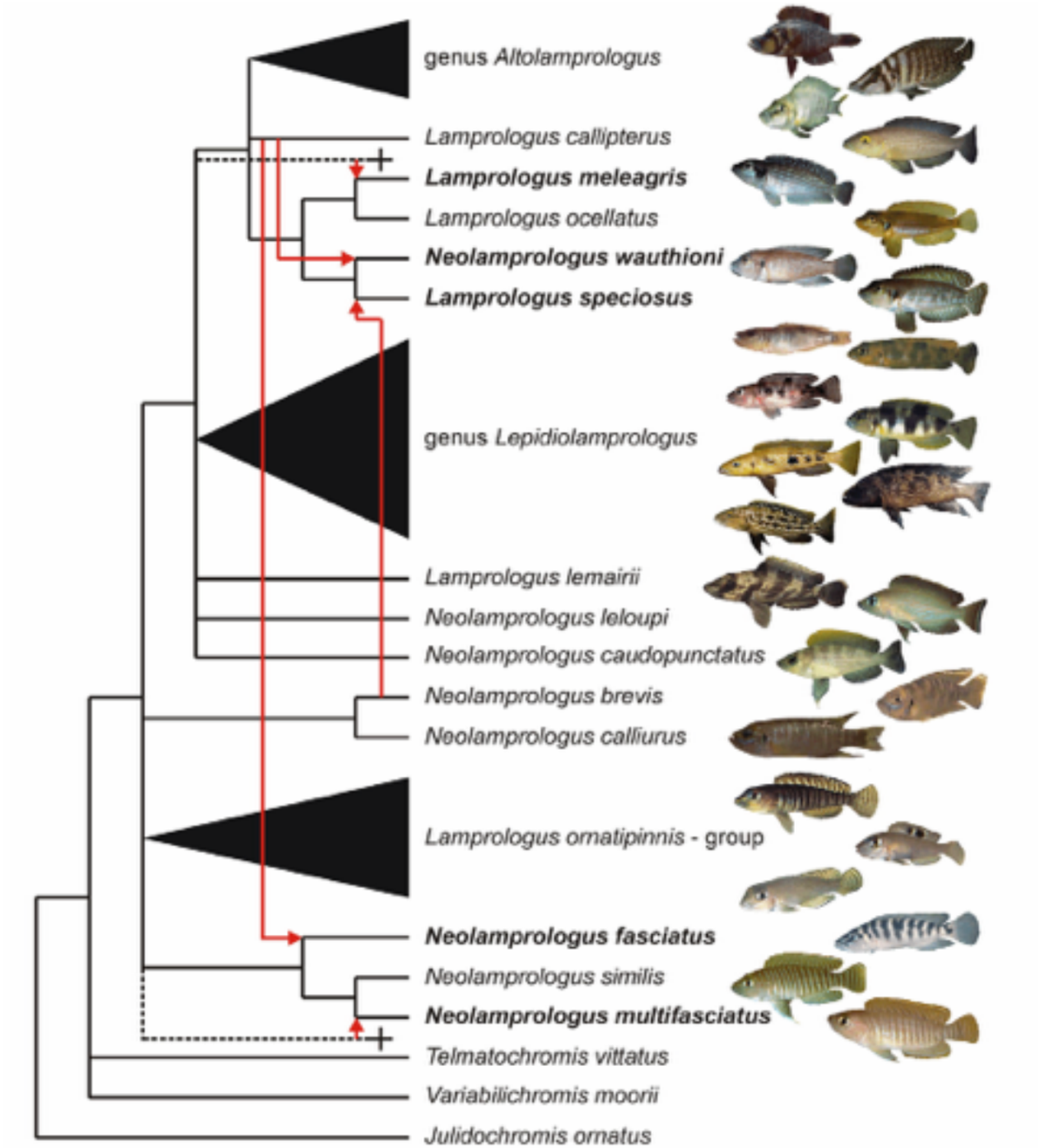
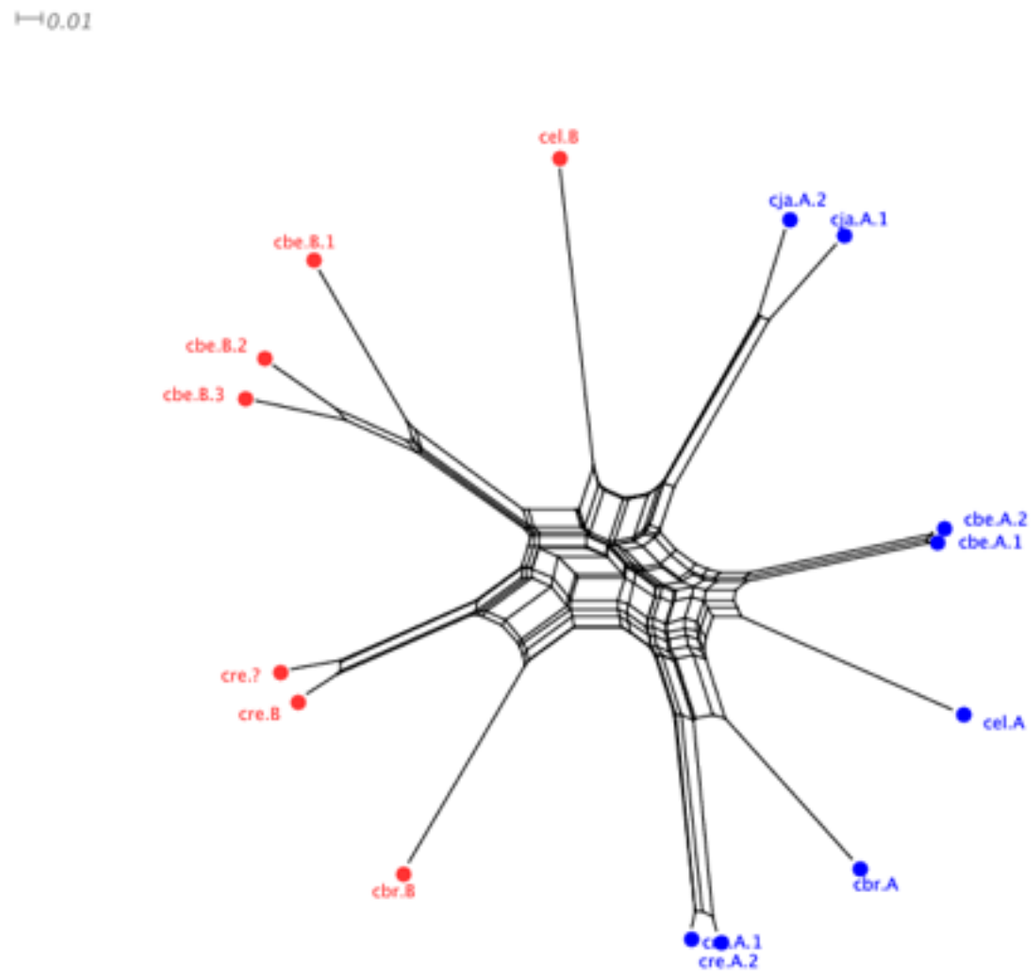


Figure 4
Reticulation of the species phylogeny by hybrid speciation. A strict consensus of mitochondrial and nuclear phylogenies was constructed from the subset of species that was assumed to have undergone bifurcating speciation. The inferred hybrid species (indicated in bold) were added according to their positions in the nuclear phylogeny. Stippled branches indicate hypothesized, now extinct lineages; red arrows indicate the direction of introgression of the mitochondrial genome into the hybrid species. Photographs show the large degree of morphological diversity in the ossified-group lamprologines.

Koblmuller et al 2007

Splitstree


0.01
cel.8
cja.A.2
cja.A.1
cbe.B.1
cbe.B.2
cbe.B.3
cbe.A.2
cbe.A.1
cre.?
cre.B
cel.A
cbr.B
cbr.A
cre.A.2

SNP phylogenetic programs

- IQ tree
- starBEAST
- SplitsTree