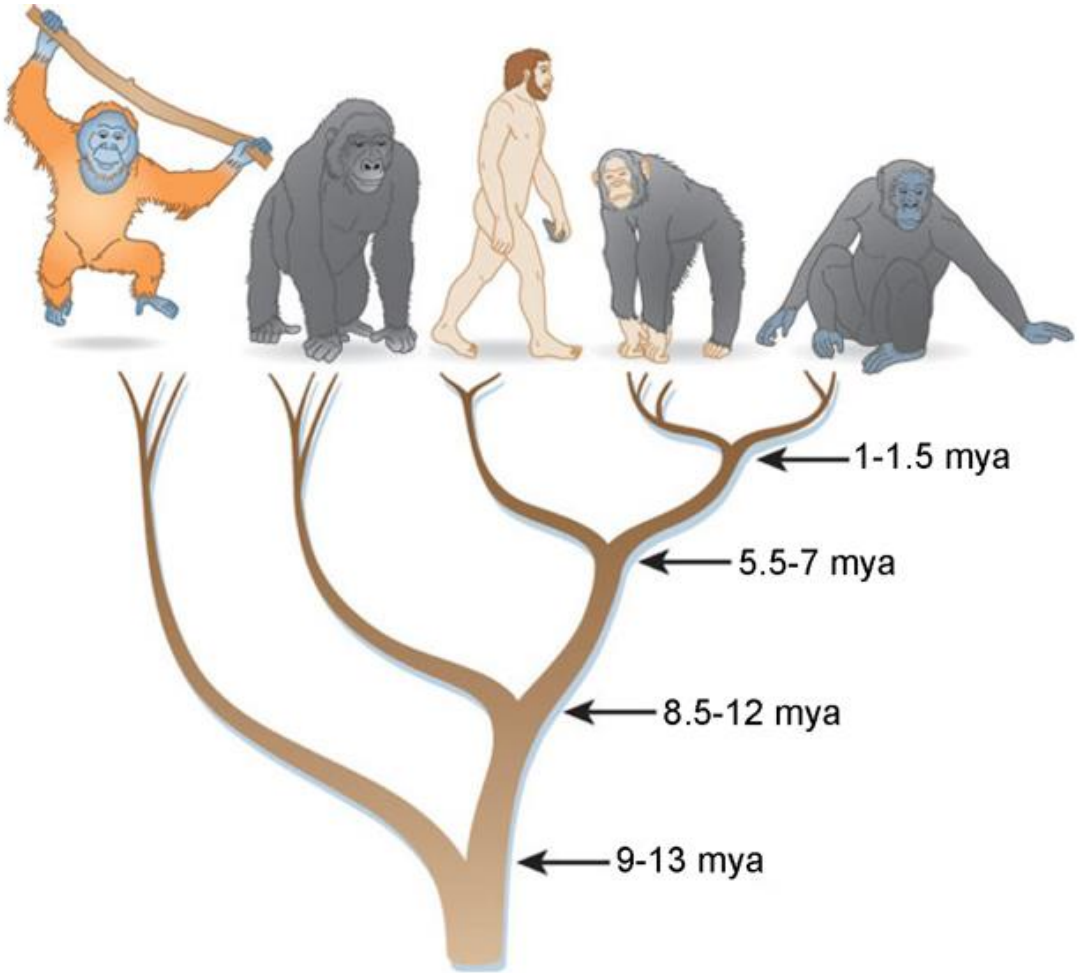
Phylogenetics

Sebastian Zoellner (szoellne@umich.edu)

Example Phylogeny

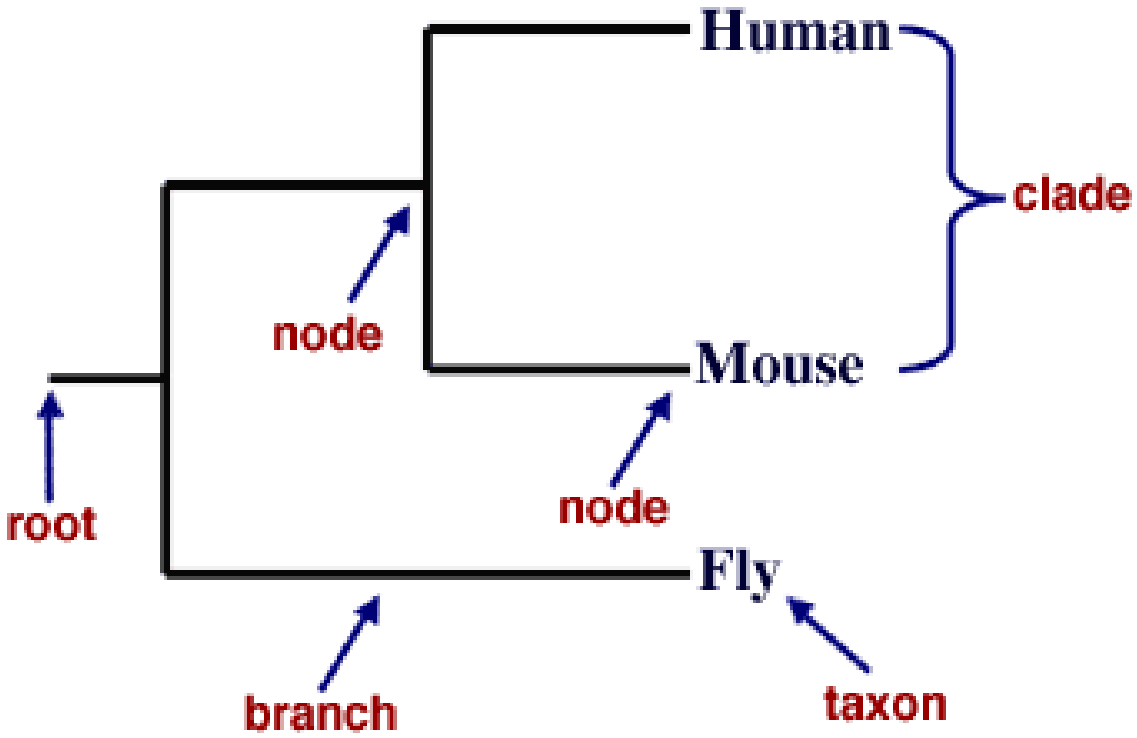


Pääbo, Svante. The mosaic that is our genome. Nature 421, 409-412 (2003)

Phylogenetics

- Phylogenetics: Use genetic variation between taxa (species, populations) to infer evolutionary relationships.
  - Each taxon is represented by a single sequence.
- Goal is to infer a specific tree.
  - Each locus in the genome evolved under the same tree.
  - Long timescales make inferences challenging.

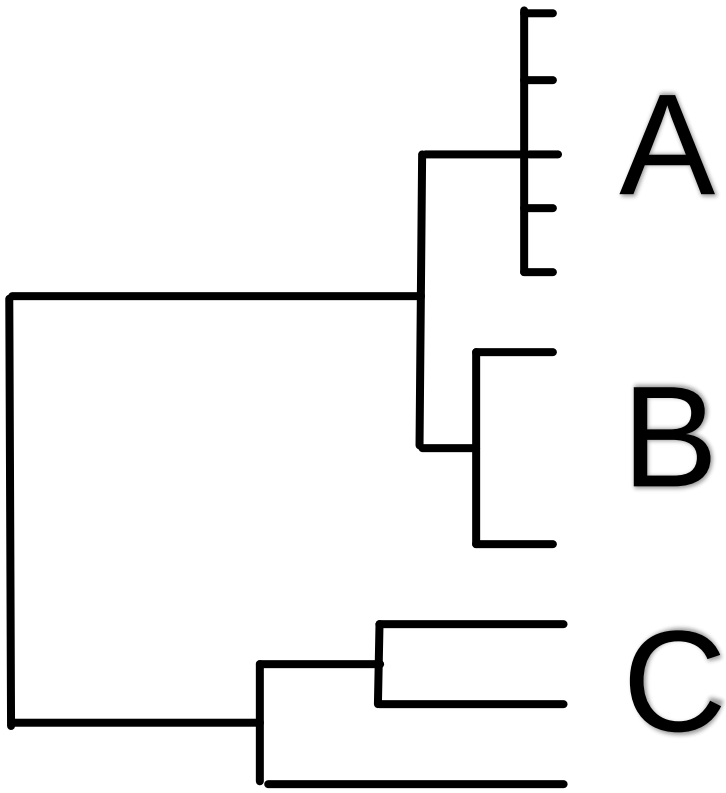
Definition of Terms



Phylogenetics in Epidemiology

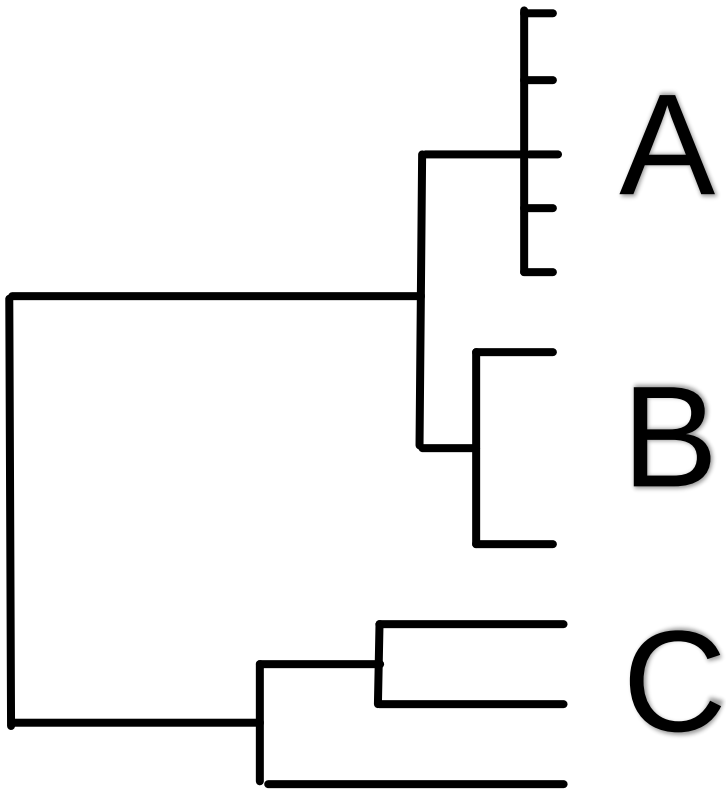
- Relationship between pathogens tells us about
  - Transmission (Closely related pathogens were transmitted from a common host in the recent past).
  - Horizontal Transfer
  - Selection

Learning from Pathogen-Trees



- Assume all samples from clade A were collected in the UM hospital, clades B and C are from Washtenaw county.
- Did infections occur at the hospital ?

Learning from Pathogen-Trees



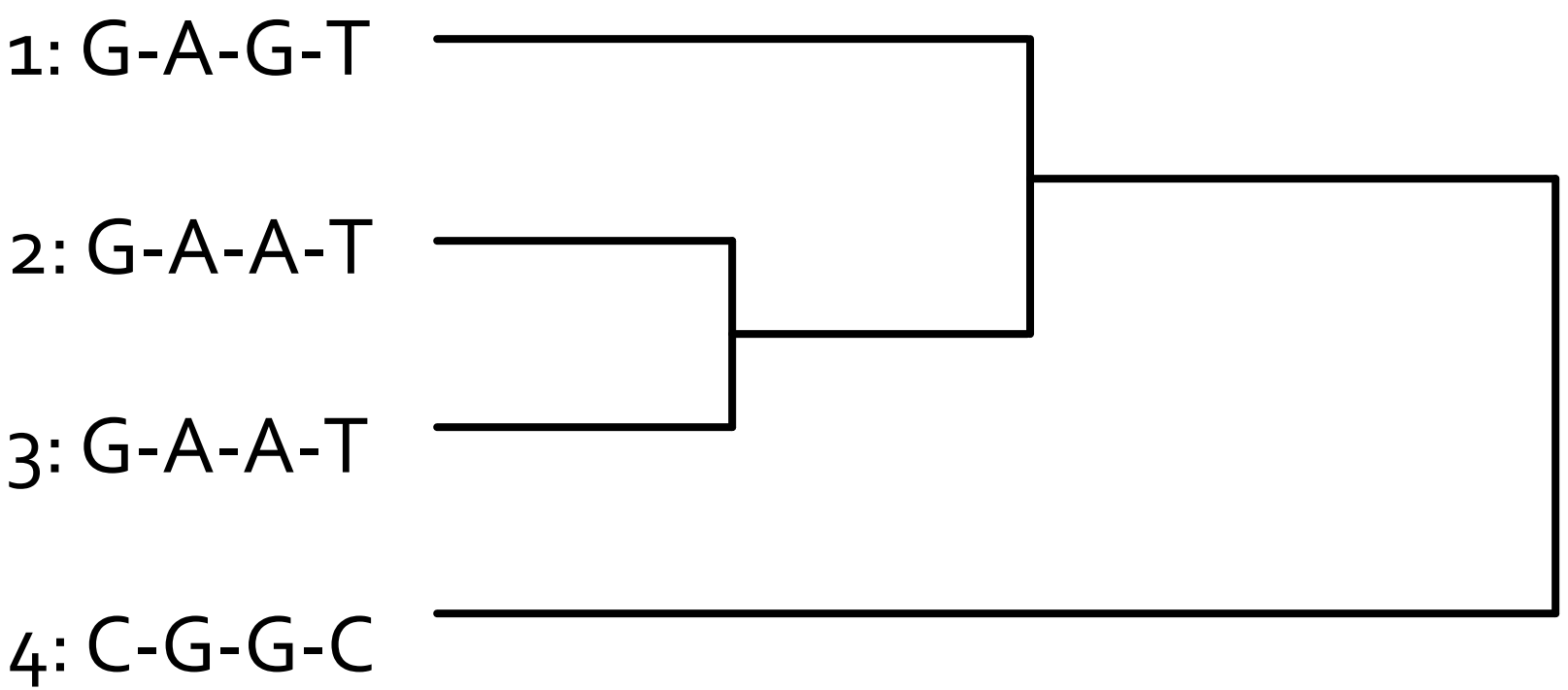
- Assume all samples from clade A were collected in Washtenaw, clade B in Bethesda, Md and clade C in Detroit.
- How did the pathogen come to AA?

Where did the tree come from?

- We select genetic sequences that are descendant from the same ancestral sequence.
- We consider the positions that differ between the pathogens.
- On average, more similar sequences have a more recent common ancestor.

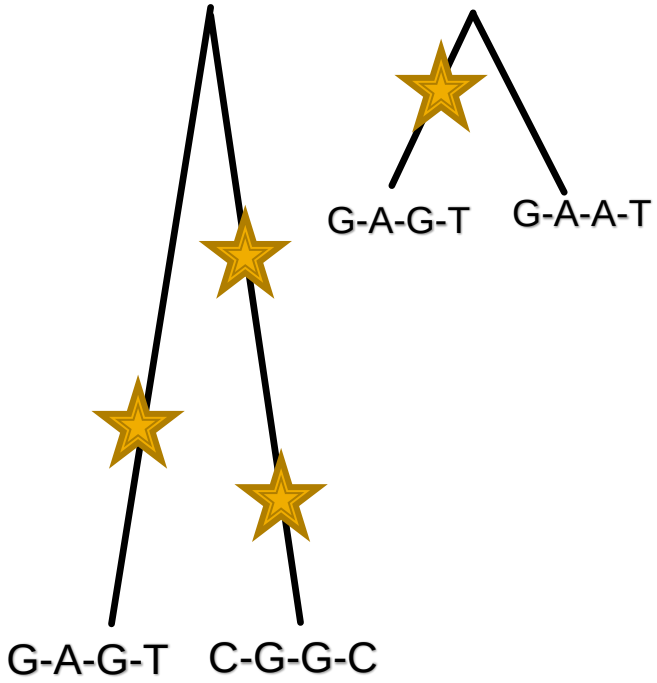
A simple example

- Consider 4 species and their DNA sequence:

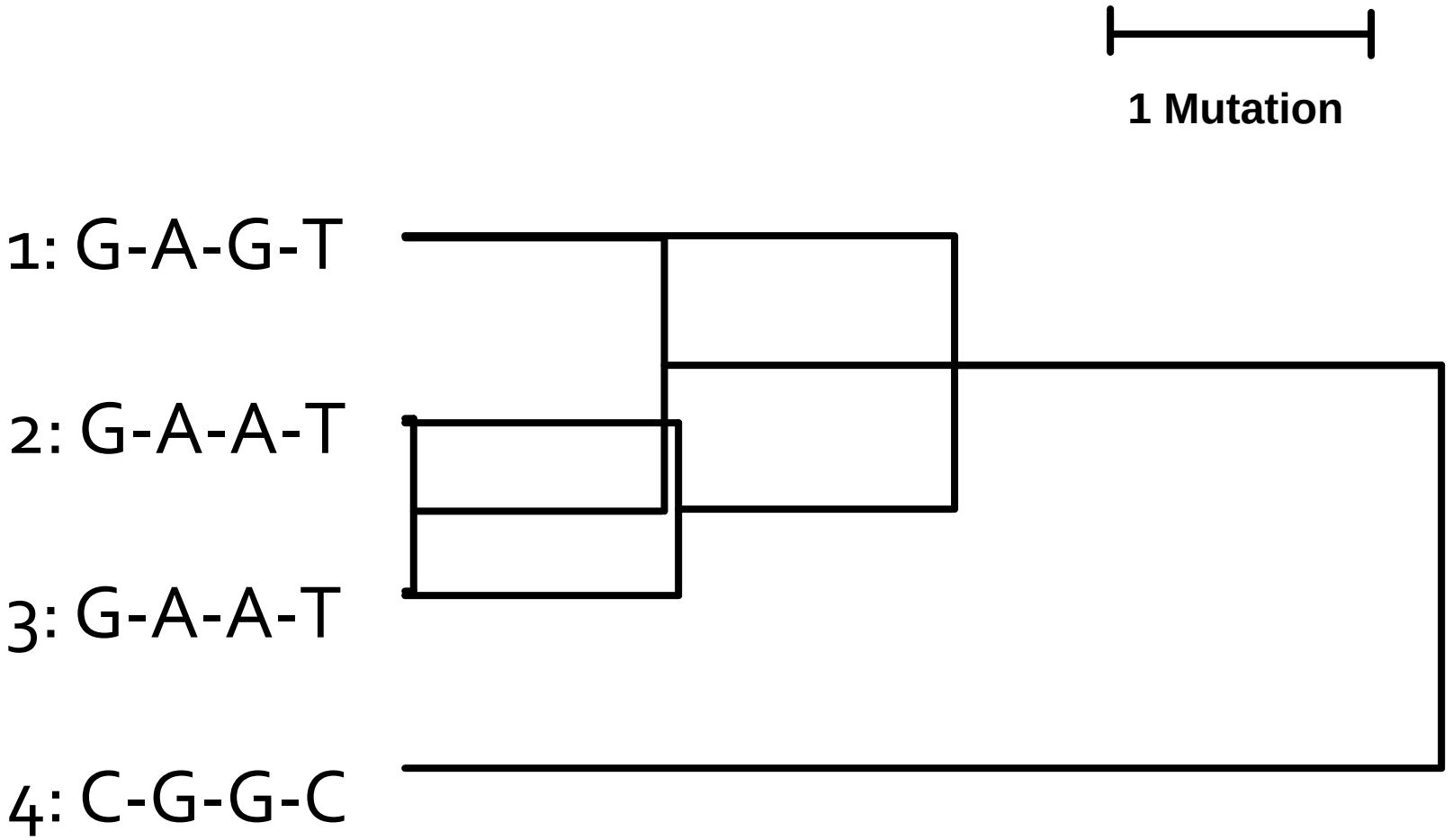


Estimating times

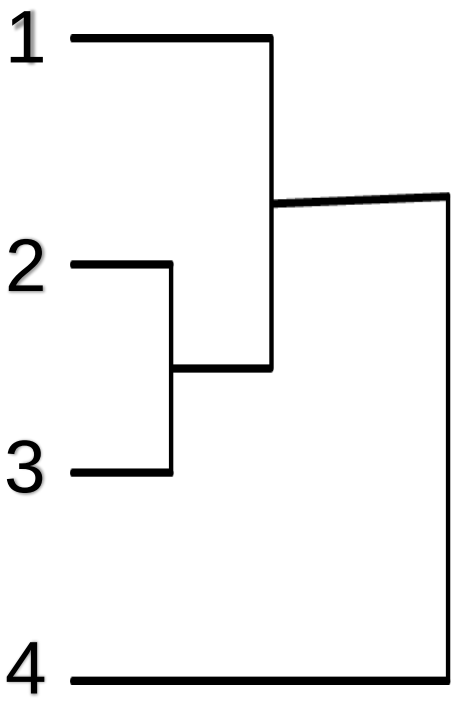
- We want to know when two taxa had a common ancestor.
- Taxa that share a common ancestor further in the past typically have more differences.
- Mutation models account for the possibility of back mutations (Jukes-Cantor).
- Advanced models account for mutation rate differences (e.g. Felsenstein).

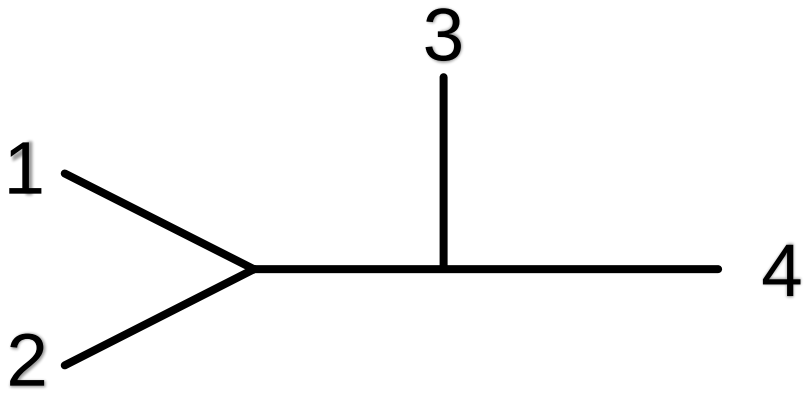


Example for estimating times



Rooted and unrooted trees



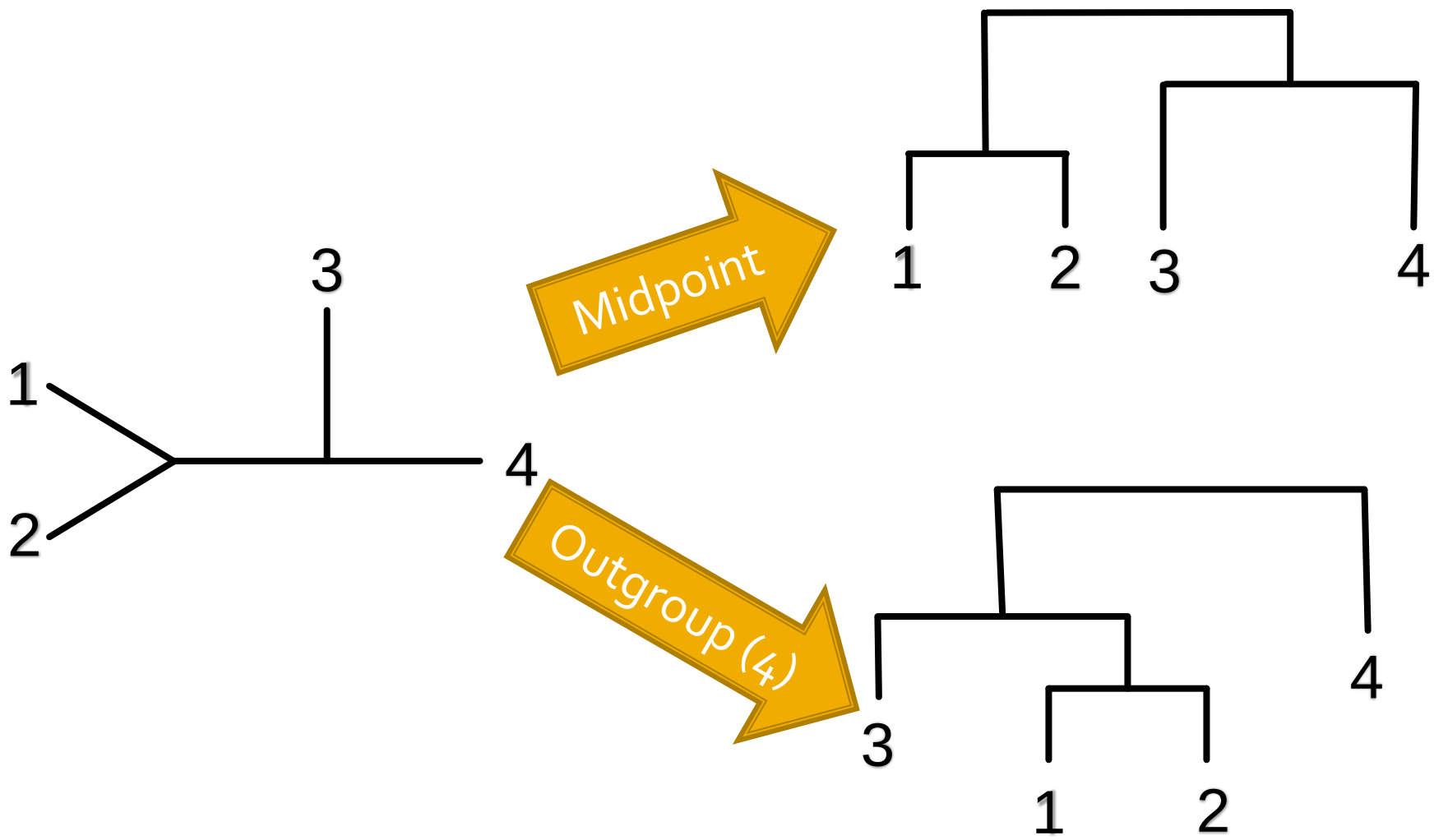


How to root a tree

- Midpoint rooting: If we assume evolution happens at a constant speed, all taxa should be at the same distance from the root. Assumes a molecular clock.

- Outgroup rooting: Include one taxon where you know is only distantly related to all taxa in your sample. That taxon defines the outgroup.

Rooting approach my determine result



Algorithms for Phylogenetics

- As long as every site mutates only once, generating trees is easy.
- Sites mutate more than once.
- So we need to find the tree that fits the data best.
- There are (2n-3)!! rooted trees and (2n-5)!! unrooted trees.
- Typically, we aim to identify an unrooted tree.

Maximum Parsimony

- Inspired by Occam's Razor.
- Analogous to morphologic phylogeny.
- Minimize the amount of homoplasy.
- No efficient search algorithm, requires heuristic for >20 taxa.
- Biased towards too short trees.
- Long branch attraction (Independent outgroups may look related due to convergent evolution, not consistent).

Neighbor Joining

- One of several distance-based methods.
- Bottom-up clustering method.
- Computationally efficient (polynomial).
- Correct if distance matrix is correct.
- Consistent under many models of evolution.
- Not dependent on molecular clock.
- But needs a distance matrix.

Maximum likelihood

- Goal: Maximize the probability of the sequence data conditional on the tree and the mutational model(s).
- Solution is typically attained with Felsenstein's pruning algorithm.
- Long branch attraction (Independent outgroups may look related due to convergent evolution, not consistent).
- Many computational tricks, e.g. quartet puzzling.
- Provides some measure for "goodness" of tree.
- Software: PHYLIP, MEGA

Bayesian methods

- Natural way to deal with missing data.
- Monte-Carlo Integration may allow dealing with large tree space.
- Choosing a prior is always an art.
- Bootstrap support values tend to be smaller than posterior probabilities.
- Software: Mr Bayes

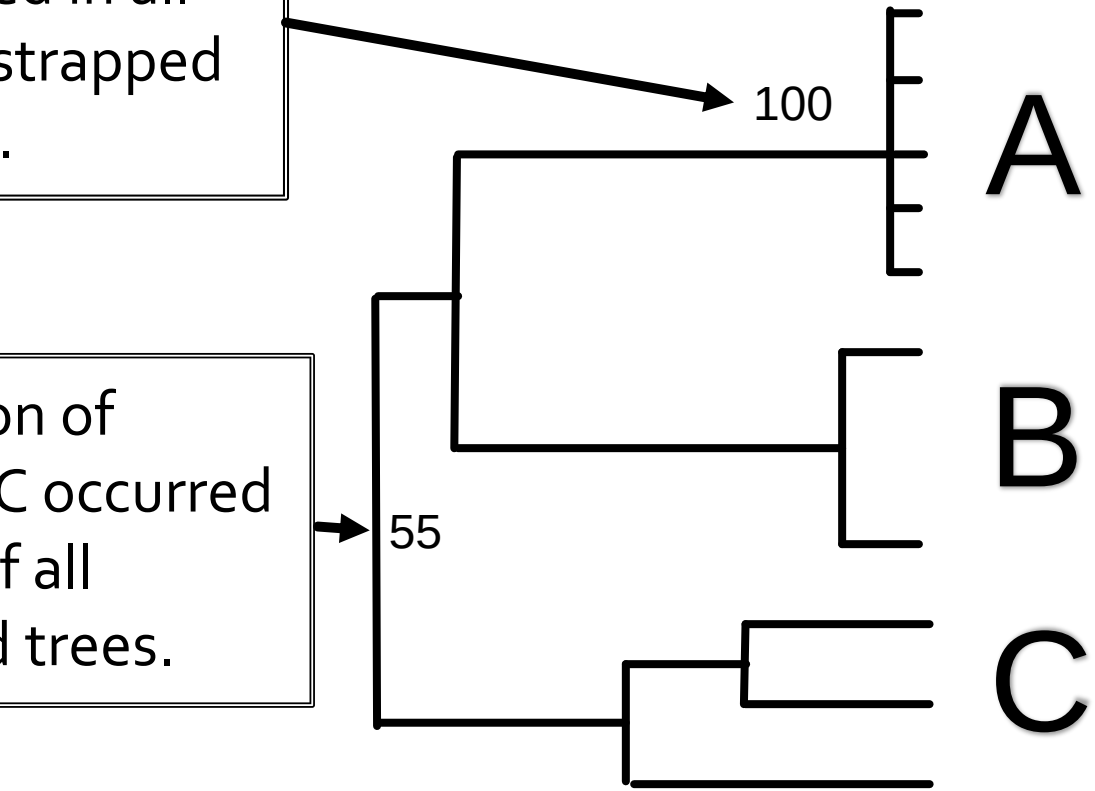
Assigning Confidence: Bootstrapping

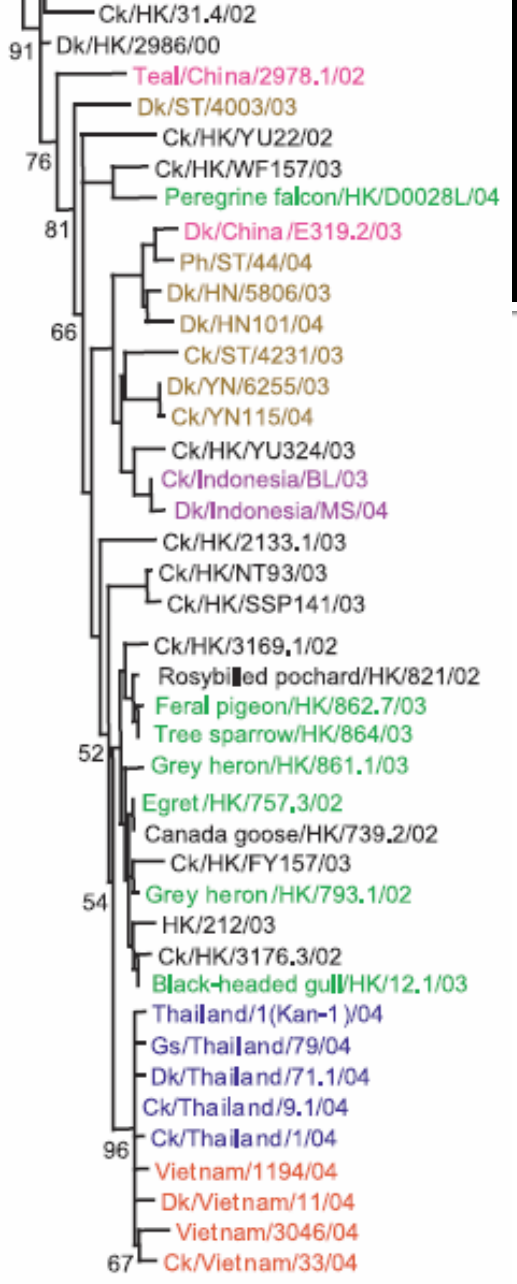
- The tree represents the best guess.
- We need to know how good this guess is: How much could random events have formed the tree?
- Assess via bootstrap: Resample sites with replacement and re-generate the tree.
- For each internal branch count how often you see the same branch.
- Careful: Garbage in-garbage out applies!

Bootstrap example

The clade A existed in all bootstrapped trees.

The separation of clades B and C occurred only in 55% of all bootstrapped trees.





H5N1 in East Asia

- Li et al (04) studied the relationship of Samples of H1N1 in birds.
- Rooted w. outgroup.
- Cases in Vietnam/Thailand form a clade with high confidence.
- Gene-specific analysis show reservoir in wild ducks for pathogenicity gene,

Caveats for pathogen phylogenies.

- Horizontal gene transfer-> Tree is a less appropriate model.
- Phylogenetics assumes internal nodes are not among the taxa.
- UPGMA is used for trees but makes assumptions that do not hold.
- Choice of genetic data matters
  - Multilocus sequence typing (MLST) is standard
  - Uses little DNA, less resolution in time estimates.
  - Full sequencing is cheap but the analysis can be a little challenging.

Summary

- Phylogenetic analysis can elucidate transmission of pathogens.
- Multiple methods exist for generating the tree and for rooting it; the choice of method can affect the conclusions. Model-based methods typically perform best.
- Bootstrap can assess the support for aspects of the tree if a model is correct.
- More reading: Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. Hall BG, Barlow M. Ann Epidemiol. 2006 Mar;16(3):157-69. Epub 2005 Aug 15. PMID: 16099674

Thank you for your attention!

A simple example