Why study Phylogenetics?

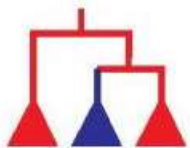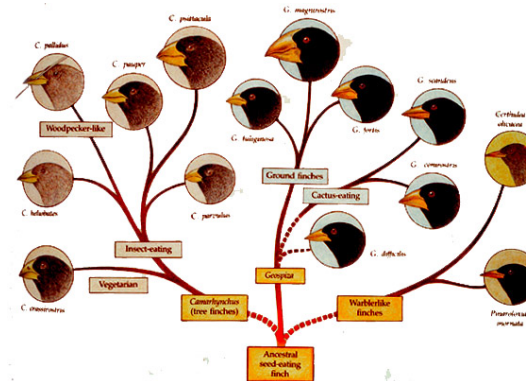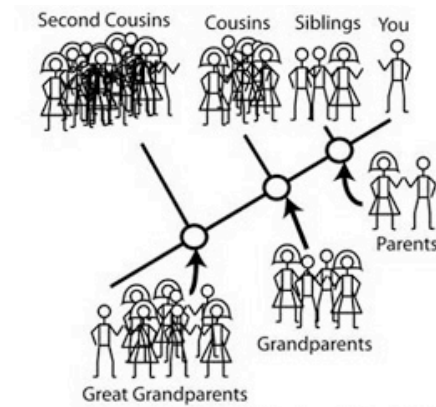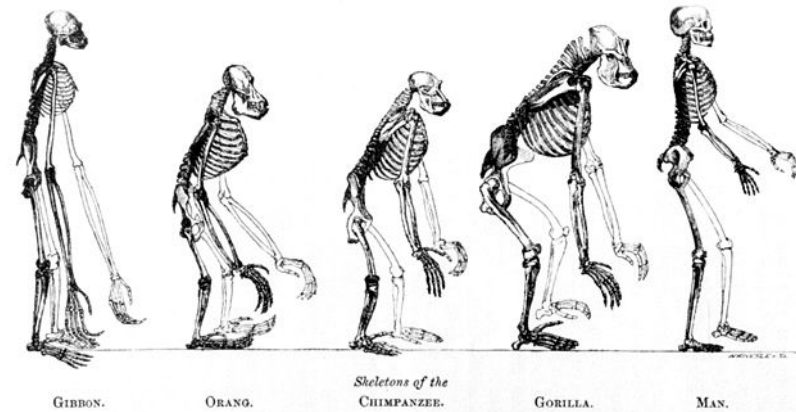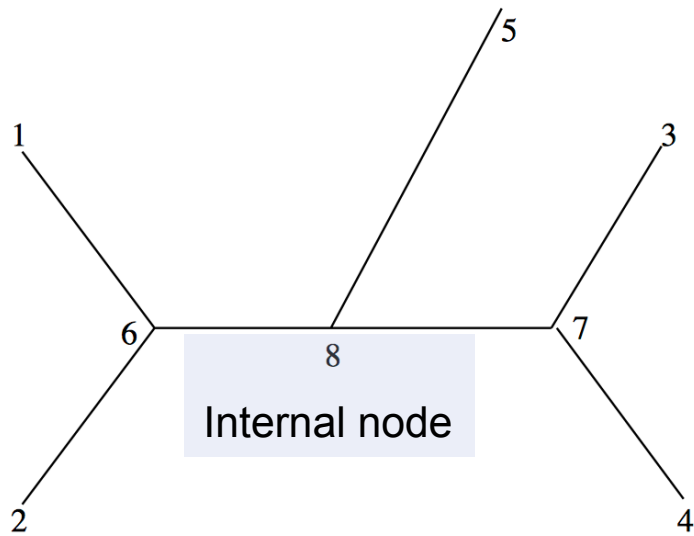Study evolution of species

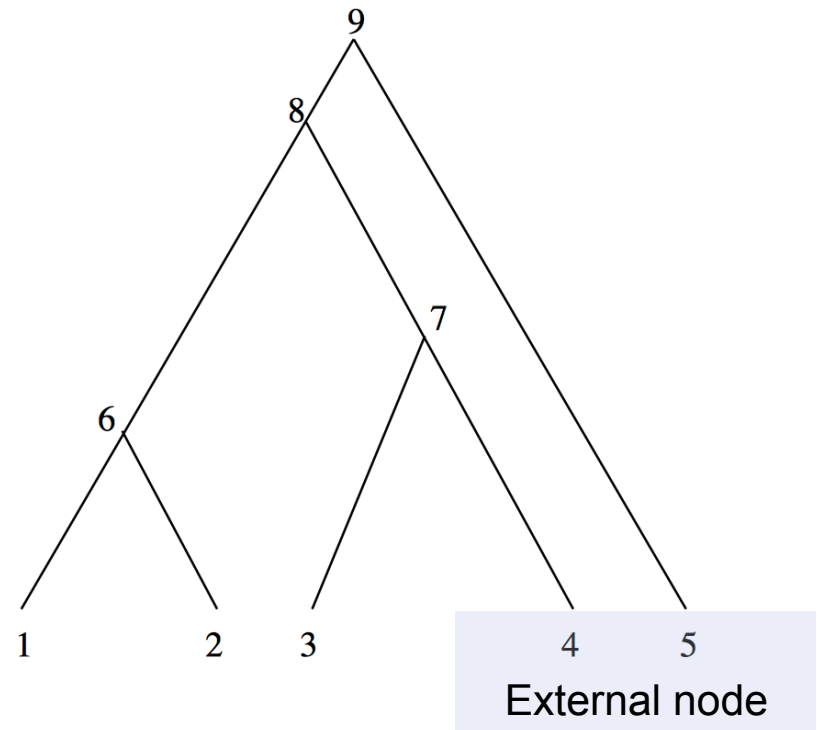Individual relatedness

Geographic variation

Infection transmission
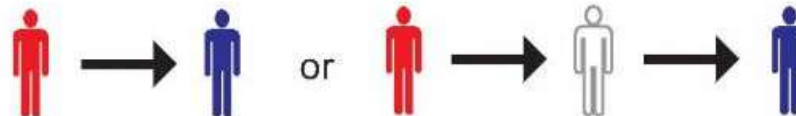
Phylogenetic Tree

Unrooted tree

Rooted tree

Phylogenetic Tree Topologies

$$(2N - 3)!! \sim (\frac{2}{e})^{N-1}(N - 1)^{N-1}\sqrt{2}.$$

N=10, roughly 2 million unrooted tree topology

N=20, 2.2 x 10^{20} unrooted tree topology

Phylogenetic Reconstruction

- Distance based methods – Neighbor Joining (NJ method), UPGMA …

- Maximum parsimony – minimizing total number of evolutionary steps

- ✓ Maximum Likelihood – based on evolutionary model, statistical models

Maximum Likelihood Summary

Data: alignment of sequences

Calculation: likelihood of each site in the data (alignment)
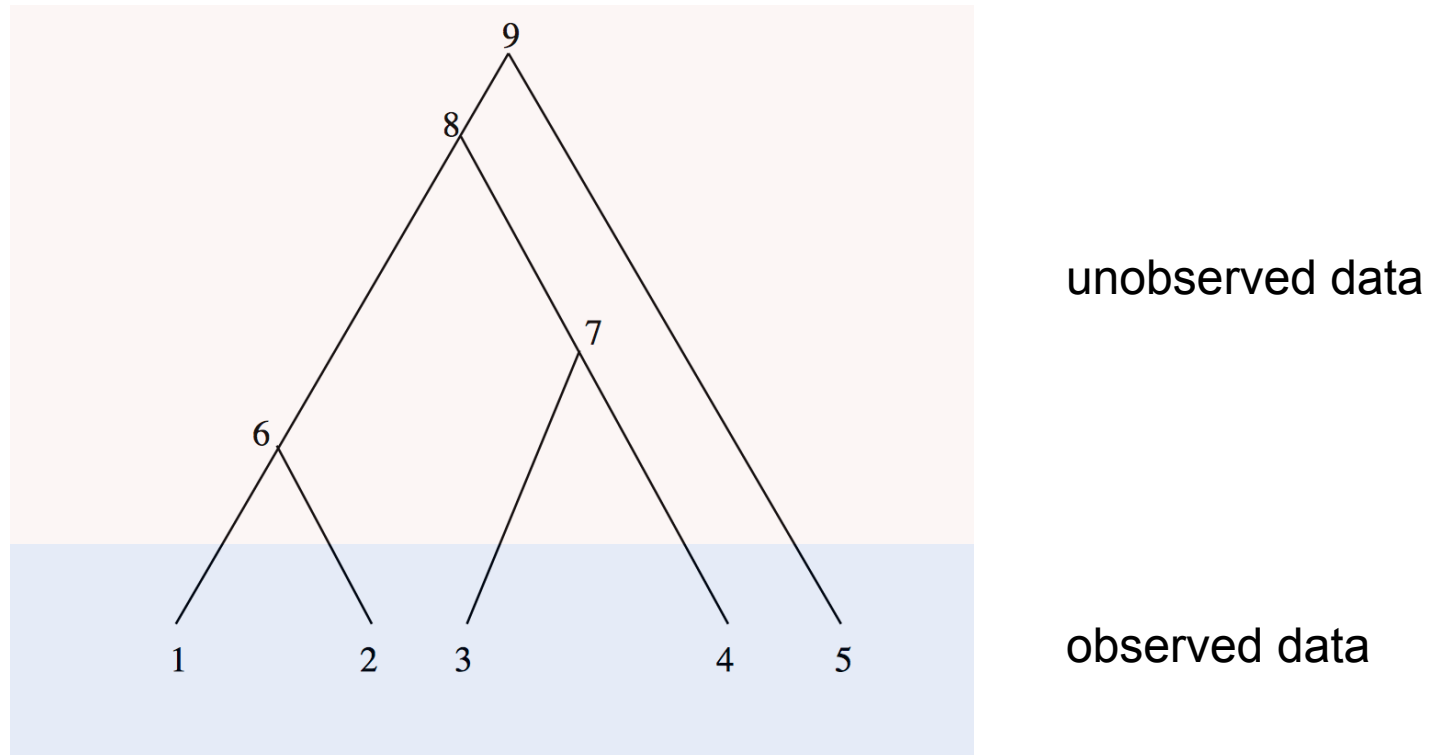 - varies with choice of the model and the initial tree

Optimization: total likelihood (product of the site likelihoods)

Result: ML tree is the tree topology that gives the highest likelihood under the given model

unobserved data

observed data

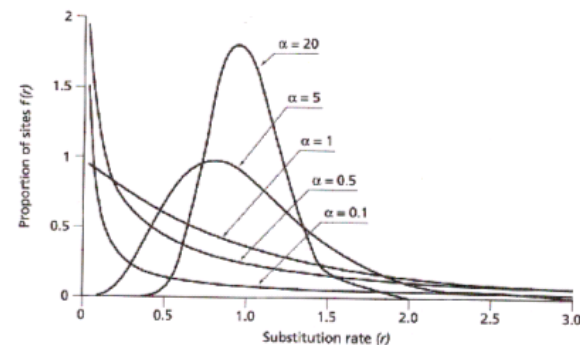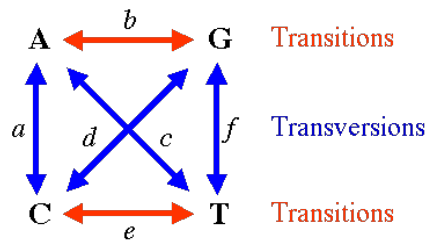Pr (observed data| Tree, evolutionary model)

Parameterization of ML

Select an evolutionary model, describing mutation probabilities

Specify the ratio of variable to invariable sites in the sequence alignment   0.04

Estimate the heterogeneity in mutation rates across sites

Parameterization of ML

Choose a tree topology search algorithm

Specify a starting tree, evidence based or choose an algorithm

Select whether to optimize branch lengths, computationally expensive

Repeat this process, with different starting tress to estimate the reproducibility