Introduction to Phylogenetic trees

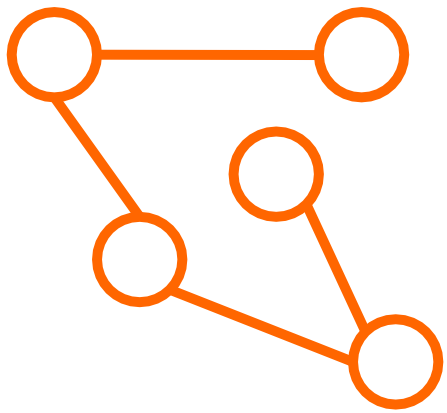
Colin Dewey
BMI/CS 576
www.biostat.wisc.edu/bmi576/
colin.dewey@wisc.edu
Fall 2016

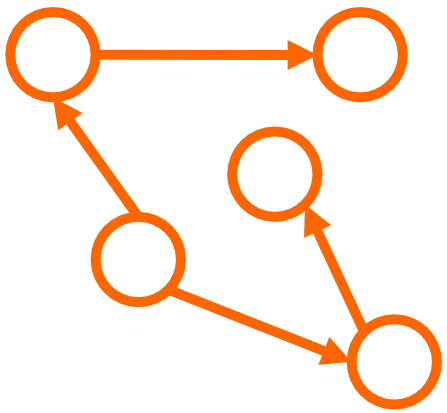
Key concepts in this section

- What are phylogenies or phylogenetic trees?
  - Terminology such as extant, ancestral, branch point, branch length
- Why build phylogenetic trees?
- Algorithms to build phylogenetic trees
  - Distance-based methods
  - Parsimony methods
    - Minimize the number of changes
  - Probabilistic methods
    - Find the tree that best explains the data using probabilistic models

What is a tree?

- Graph theoretically:
  - Undirected case: graph without cycles
  - Directed case: underlying undirected graph is a tree
    - Often it is required that *indegree(v)* ≤ 1 for all *v*



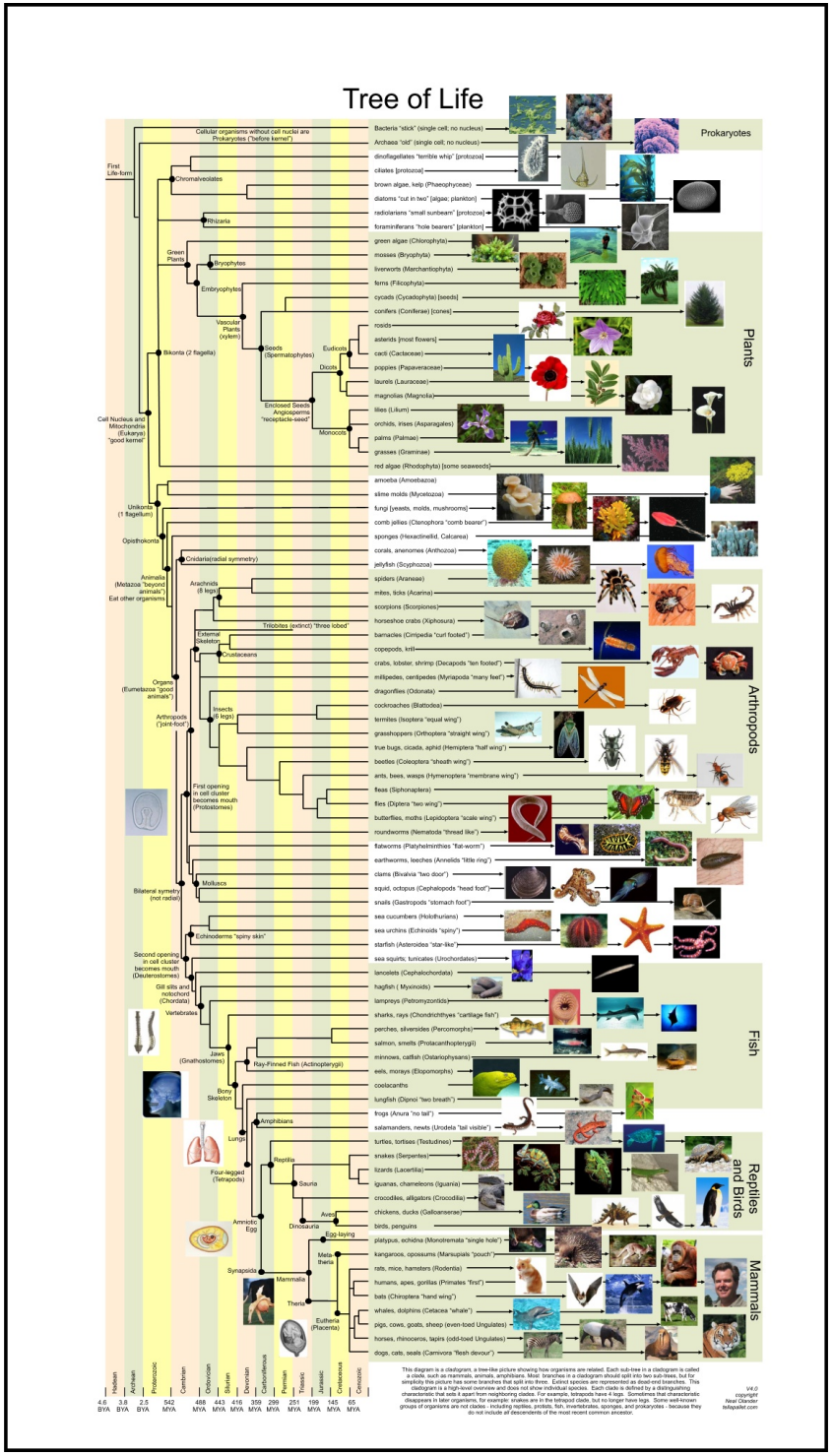


What are phylogenetic trees?

- A tree that describes evolutionary relationships among entities
  - Species, genes, strains
- This relationship is called “phylogeny”
- Leaves represent extant (current day) species
- Internal nodes represent ancestral species
- Phylogenetics:
  - The task for inferring the phylogenetic tree from observations in existing organisms

Why phylogenetic trees?

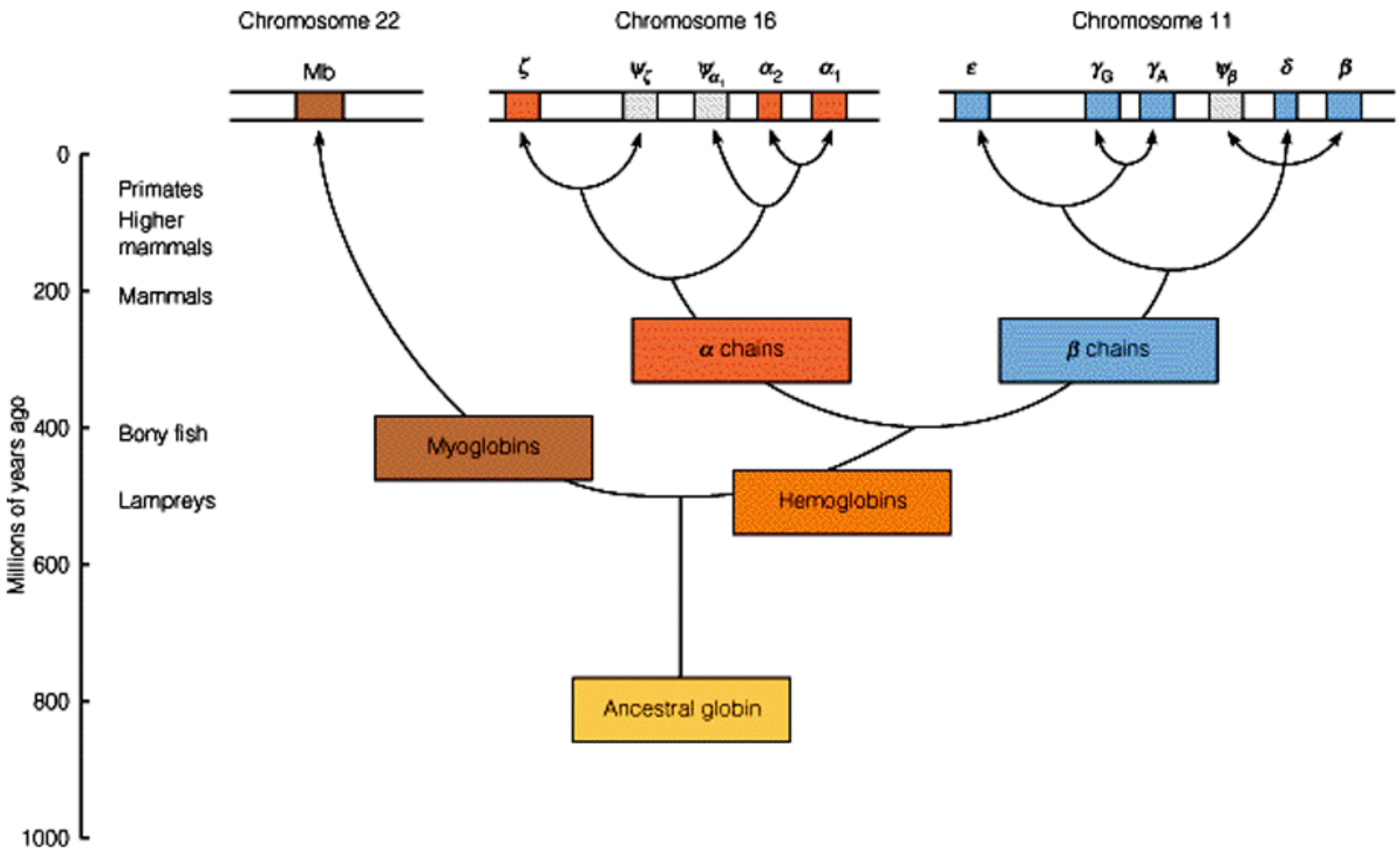
- Inform multiple sequence alignments
- Identify signatures of conservation of sequence
- Understand how organisms are related
  - Do humans and chimpanzees share a more recent common ancestor than do humans and gorillas?
- Ask how closely organisms are related
  - Humans and chimpanzees shard a common ancestor 5mya
- How specific functions/traits have evolved
  - What made us human?



Tree of life aims to represent the phylogeny of all species on earth

http://www.gallery119.com/images/tree_of_life_modern_complete.pdf

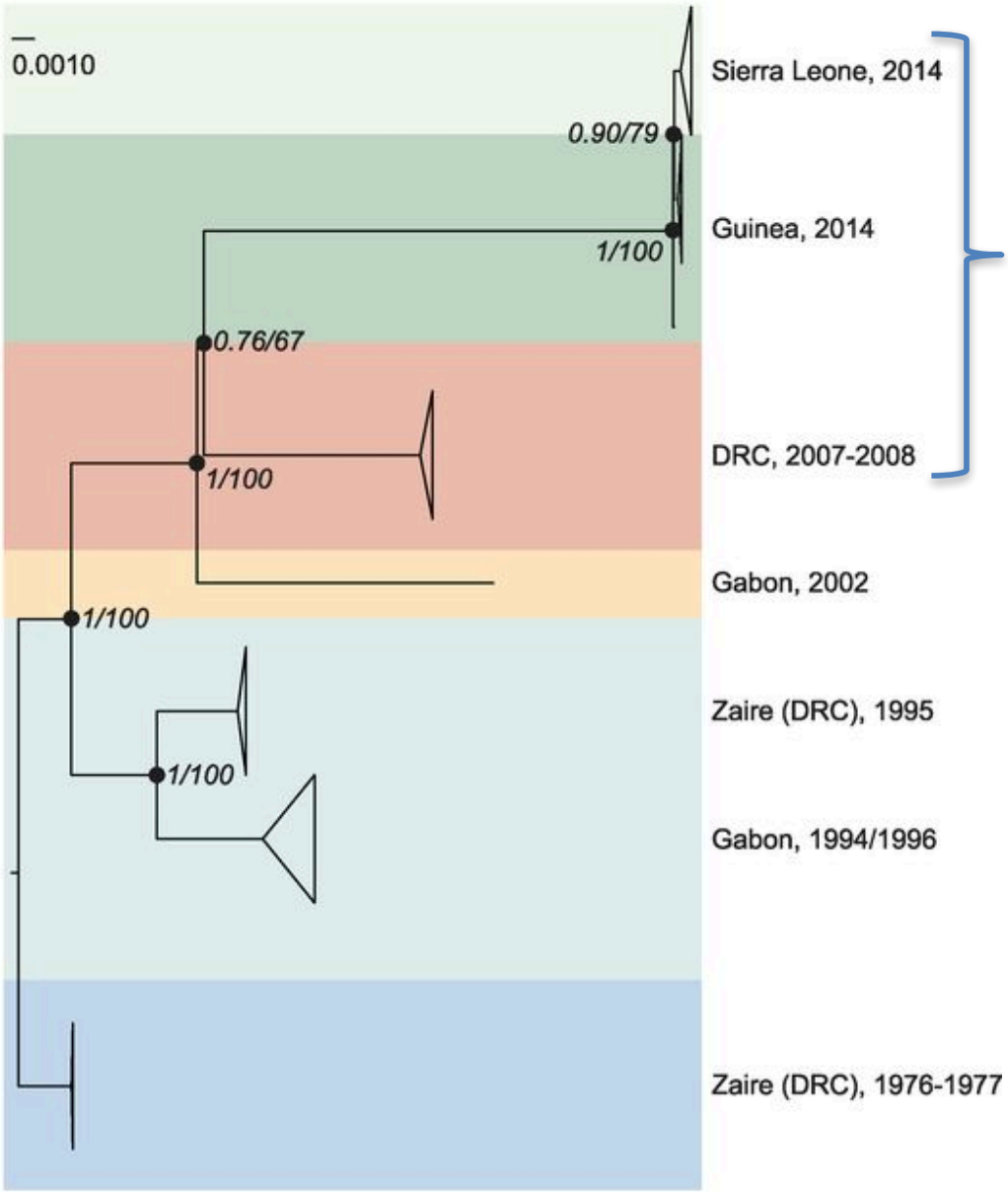
Example Gene Tree: Globins



Tracing the evolution of the Ebola virus

- Ebola virus: a lethal human pathogen, fatality rate 78%
- 2014 Ebola epidemic in Africa
  - Until recently the largest known case happened in 1976 (318 cases)
  - Outbreak reported in Feb 2014
  - 11,310 reported deaths from 2014 outbreak
  - World Health Organization ended declaration of Public Health Emergency in March 2016
- Key questions
  - Where did the pathogen come from?
  - How is it evolving?
- In a 2014 Science paper, researchers reported whole genome sequence alignment of 78 Ebola virus samples

Phylogenetic tree of the Ebola virus



Three recent outbreaks from the same ancestor

Gire et al, Science 2014

Insights gained from sequence comparison

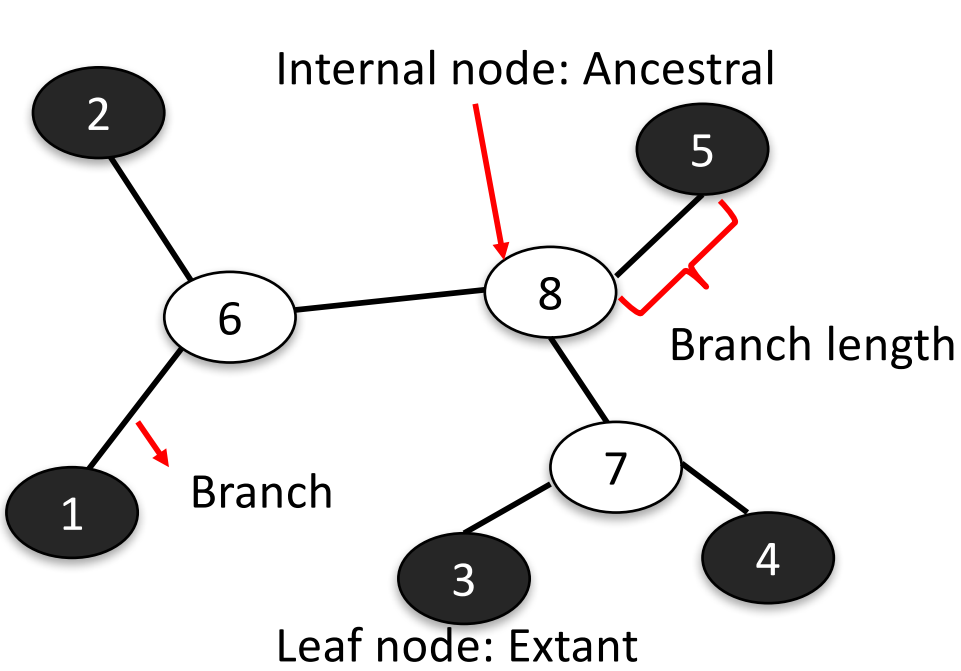
- “Genetic similarity across the sequenced 2014 samples suggests a single transmission from the natural reservoir, followed by human-to-human transmission during the outbreak”
- “..data suggest that the Sierra Leone outbreak stemmed from the introduction of two genetically distinct viruses from Guinea around the same time…”
- “..the catalog of 395 mutations, including 50 fixed nonsynonymous changes with 8 at positions with high levels of conservation across ebola viruses, provides a starting point for such studies”

Gire et al., Science 2014

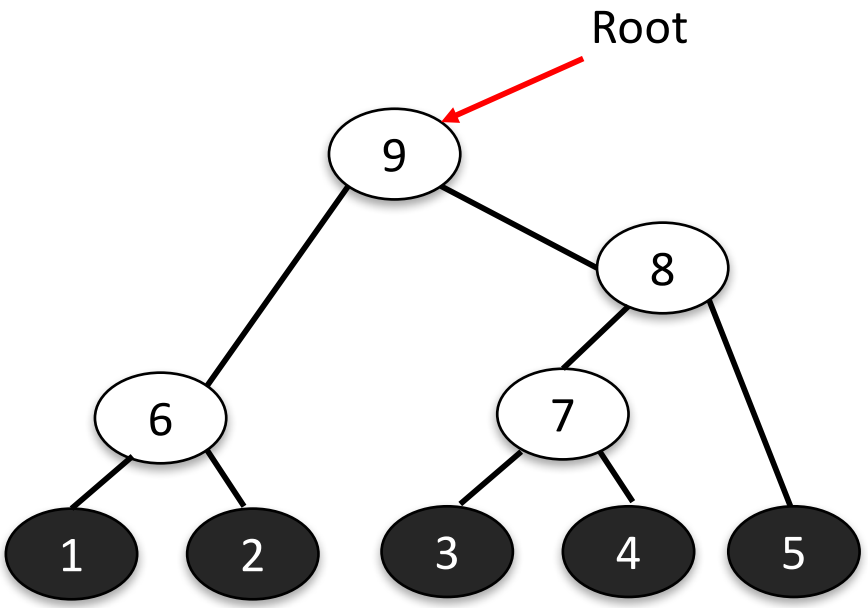
Phylogenetic tree basics

- Leaves represent entities (genes, species, individuals/strains) being compared
  - the term *taxon* (*taxa* plural) is used to refer to these when they represent species and broader classifications of organisms
  - For example if taxa are species, the tree is a species tree
- Internal nodes are ancestral units
- Phylogenetic trees can be rooted or unrooted
  - the root represents the common ancestor
- In a *rooted* tree, path from root to a node represents an evolutionary path
  - Gives directionality to evolutionary time
- An *unrooted* tree specifies relationships among taxa, but lacks directionality information

Tree basics



Unrooted tree



Rooted tree

Each tree topology represents a different evolutionary history

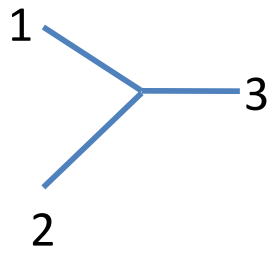
For a species tree, internal nodes represent speciation events

Branch length describes the evolutionary divergence between two nodes

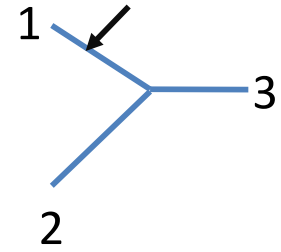
Tree counting

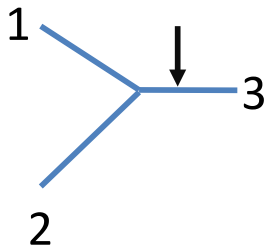
- A rooted binary tree with n leaf nodes has
  - $n-1$ internal nodes
  - $2n-2$ edges/branches
- An unrooted binary tree with n leaf nodes has
  - $n-2$ internal nodes
  - $2n-3$ edges/branches
  - A root can be added to any of these branches to give $2n-3$ rooted trees for any unrooted tree
- E.g. for $n=3$ there is *one* unrooted tree and *three* rooted trees

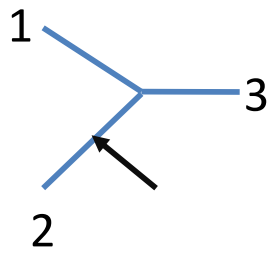
Tree counting



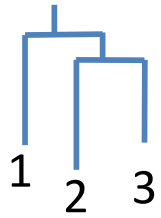
An unrooted tree

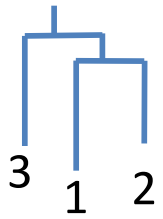


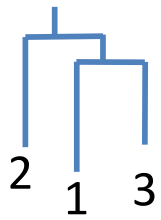




Possible positions for root



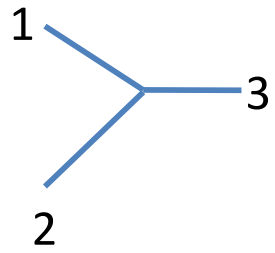


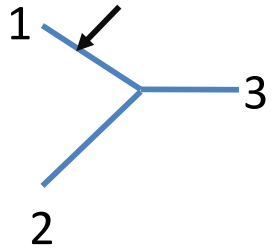


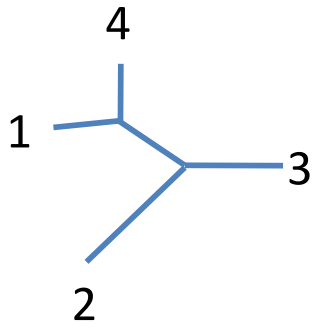
Rooted trees

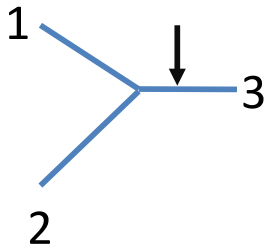
Tree counting

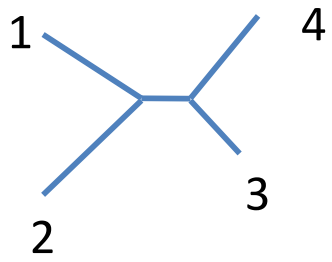
- Instead of adding a root we could add a branch for the n+1th taxon

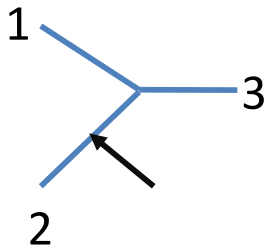


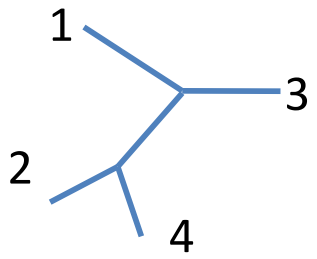












Tree counting

- A tree with 3 leaves can be grown in (2*3)-3=3 ways to make a tree of 4 leaves
  - 3 possible unrooted trees with 4 leaves
- Each tree with 4 leaves can be grown in (2*4)-3=5 ways to make a tree of 5 leaves
  - 3*5 possible unrooted trees with 5 leaves
- Each tree of 5 leaves can be grown in (2*5)-3=7 ways
  - 3*5*7 possible unrooted trees with 6 leaves
- In general for n leaves we can have
  - (1)*(3)*(5)*...(2n-5) unrooted trees

Number of Possible Trees

- given n sequences, there are $\prod_{i=3}^{n}(2i-5)$ possible unrooted trees
- and $(2n-3)\prod_{i=3}^{n}(2i-5)$ possible rooted trees
- This grows very fast
  - For n=10, we have 2 million unrooted trees
  - For n=20, we have $2.2*10^{20}$

Constructing phylogenetic trees

- Phylogenetic tree construction
  - Given observations of n taxonomical units infer the tree that best describes the evolutionary relationships among the units

- Three types of methods
  - Distance based methods
  - Parsimony methods
  - Probabilistic approaches