Introduction to Phylogenetic Trees

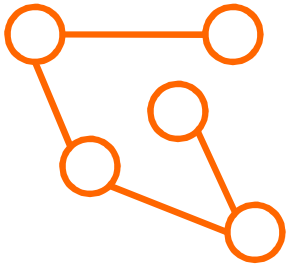
BMI/CS 576
www.biostat.wisc.edu/bmi576.html
Mark Craven
craven@biostat.wisc.edu

Phylogenetic inference: task definition

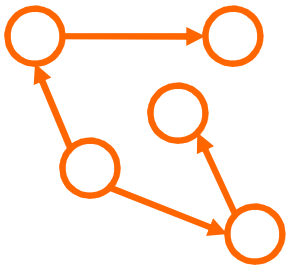
- Given
  - data characterizing a set of species/genes
- Do
  - infer a *phylogenetic tree* that accurately characterizes the evolutionary lineages among the species/genes

What is a tree?

- undirected case: a graph without cycles



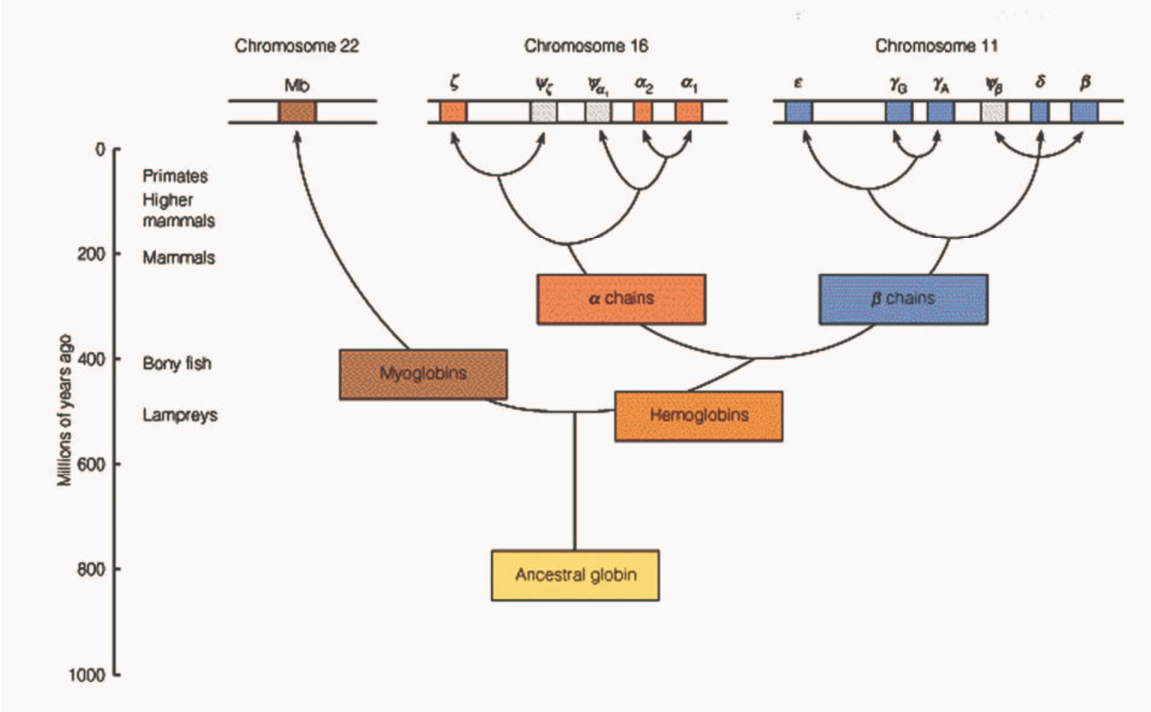
- directed case: underlying undirected graph is a tree (sometimes it is required that $\text{indegree}(v) \leq 1$ for all v)



Phylogenetic tree basics

- leaves represent things (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
- internal nodes are hypothetical ancestral units
- in a *rooted* tree, path from root to a node represents an evolutionary path
  - the root represents the common ancestor
- an *unrooted* tree specifies relationships among things, but not evolutionary paths

Example gene tree: globins



Example species tree: baboons

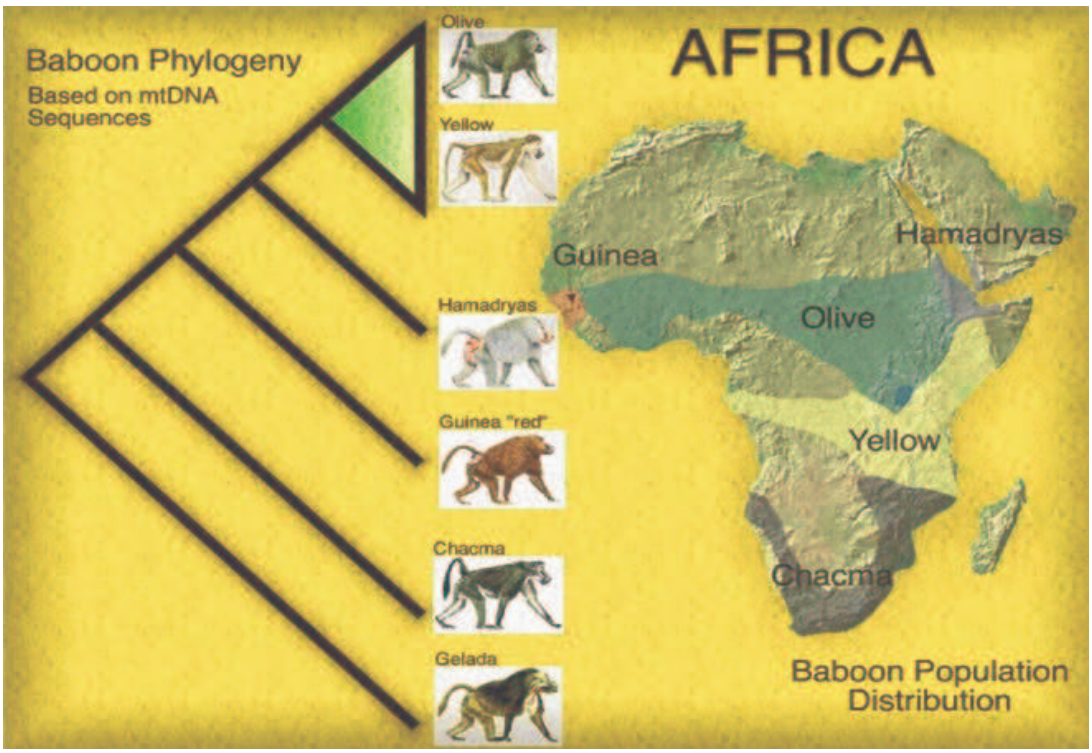


Image from Southwest National Primate Research Center

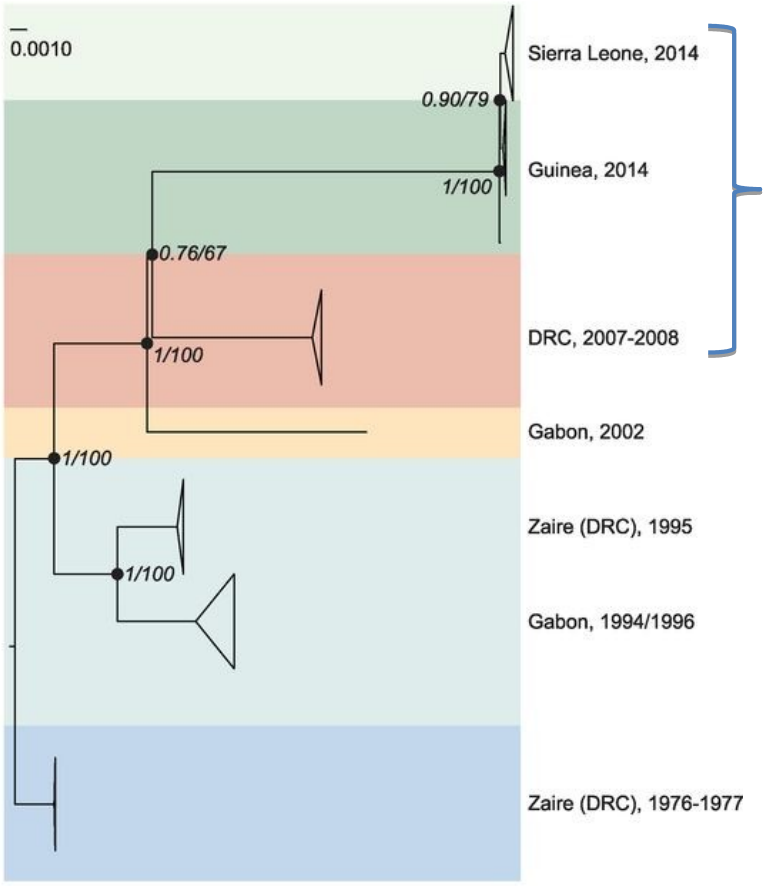
Motivation

- why construct trees?
  - to understand lineage of various species
  - to understand how various functions evolved
  - to inform multiple alignments
  - to identify what is most conserved/important in some class of sequences

Tracing the evolution of the Ebola virus

- Ebola virus: a lethal human pathogen
- 2014 Ebola epidemic in Africa
  - until recently the largest case in 1976 (318 cases)
  - outbreak reported in Feb 2014
  - 11,315 deaths, fatality rate 78%
- Key questions
  - where did the pathogen come from?
  - how is it evolving?
- In a 2014 Science paper
  - whole genome sequence alignment of 78 Ebola virus samples

Phylogenetic tree of the Ebola virus



Gire et al, Science 2014

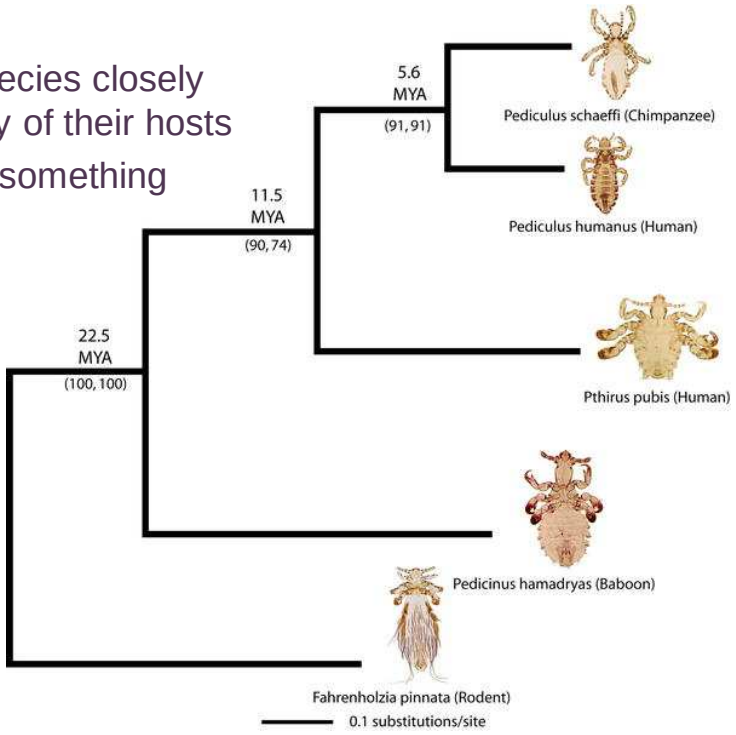
Insights gained from sequence comparison

- "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 non-synonymous 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

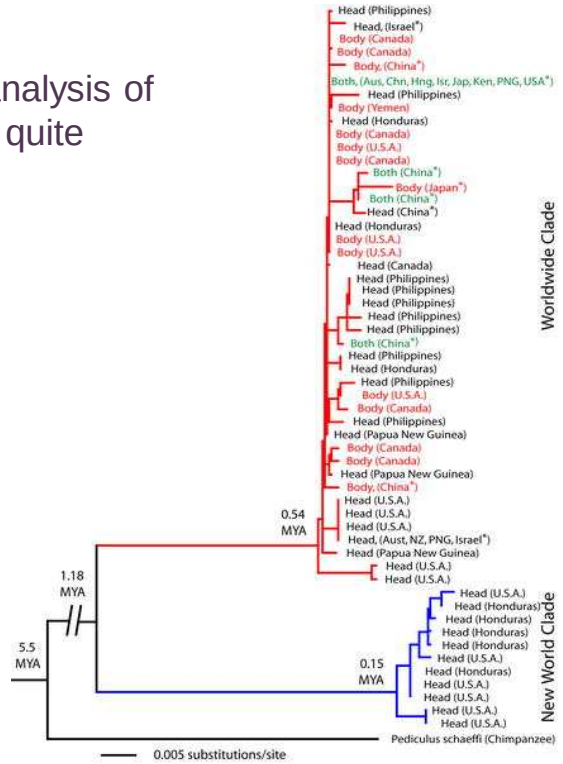
Genetic Analysis of Lice Supports Direct Contact between Modern and Archaic Humans
D. Reed et al., *PLoS Biology* 2(11), November 2004.

- inferred phylogeny of lice species closely parallels accepted phylogeny of their hosts
- can phylogeny of lice tell us something about evolution of hosts?



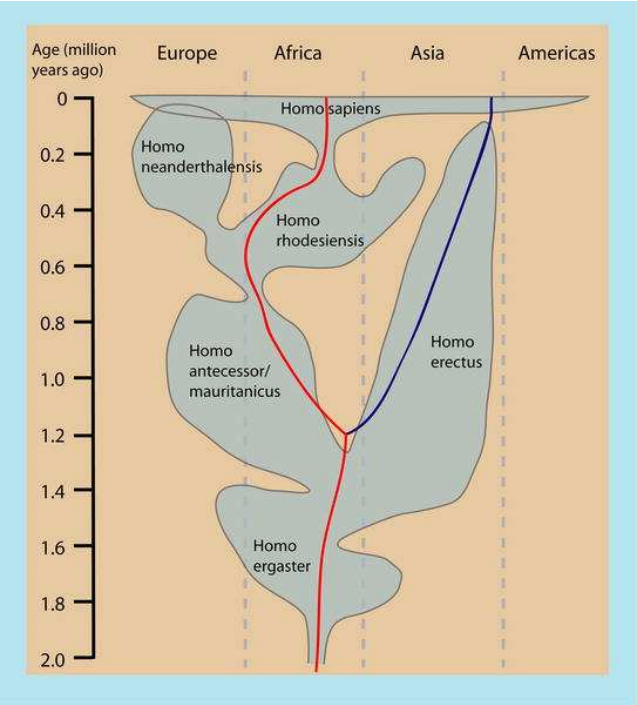
Genetic Analysis of Lice Supports Direct Contact between Modern and Archaic Humans
D. Reed et al., *PLoS Biology* 2(11), November 2004.

- a more detailed phylogenetic analysis of human lice species shows two quite separate *clades* (subtrees)



Genetic Analysis of Lice Supports Direct Contact between Modern and Archaic Humans
D. Reed et al., *PLoS Biology* 2(11), November 2004.

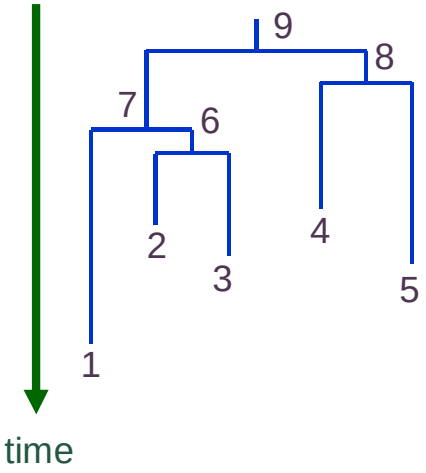
- this phylogeny supports a theory of human evolution in which
  - *H. erectus* and the ancestors of *H. sapiens* had little or no contact for a long period of time
  - there was contact between *H. erectus* and *H. sapiens* as late as 30,000 years ago

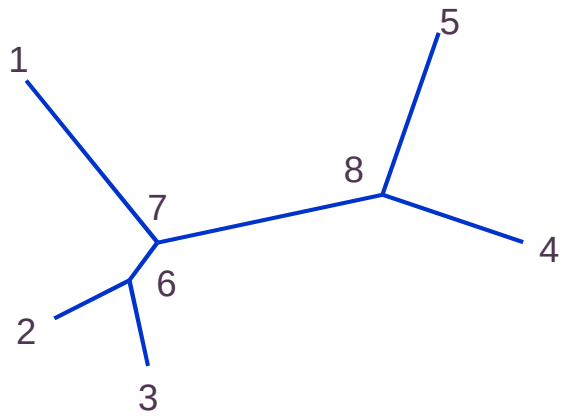


Data for building trees

- trees can be constructed from various types of data
  - *distance-based*: measures of distance between species/genes
  - *character-based*: morphological features (e.g. # legs), DNA/protein sequences
  - *gene-order*: linear order of orthologous genes in given genomes

Rooted vs. 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

Number of possible trees

# taxa (n)	# unrooted trees	# rooted trees
4	3	15
5	15	105
6	105	945
8	10,395	135,135
10	2,027,025	34,459,425

Phylogenetic tree approaches

- three general types of methods
  - *distance*: find tree that accounts for estimated evolutionary distances
  - *parsimony*: find the tree that requires minimum number of changes to explain the data
  - *maximum likelihood*: find the tree that maximizes the likelihood of the data