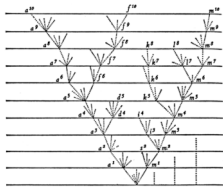
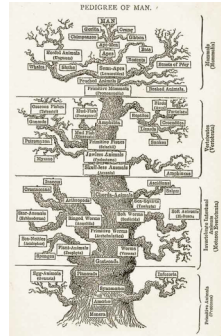


# Phylogenetics



**Todd Vision**  
Biology 522  
March 26, 2007



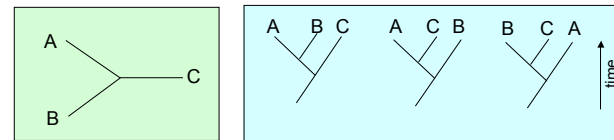
## Applications of phylogenetics

- Studying organismal or biogeographic history
  - Systematics
  - Dating events in the fossil record
  - Conservation biology
- Studying gene and protein families (molecular evolution)
  - Studying functional specificity and divergence
  - Identifying selection at the molecular level
  - Understanding host-parasite coevolution
  - Identifying lateral transfer events

## Outline

- How to read a tree
- How to infer a tree
  - Including how to measure confidence in your answer!
- Applications
  - Testing for lateral gene transfer
  - Studying ancient endosymbiosis
  - Assessing microbial diversity
  - Inferring gene function

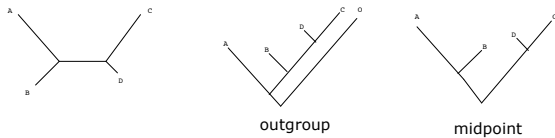
## Unrooted networks vs. rooted trees



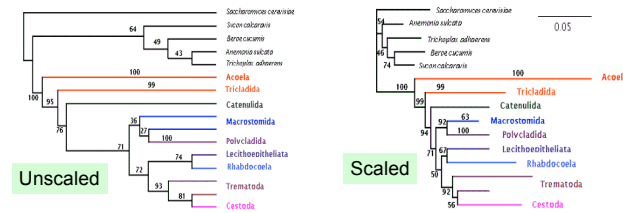
<u>Taxa</u>	<u>Unrooted</u>	<u>Rooted</u>
3	1	3
5	15	105
10	2,027,025	34,459,425

## Root

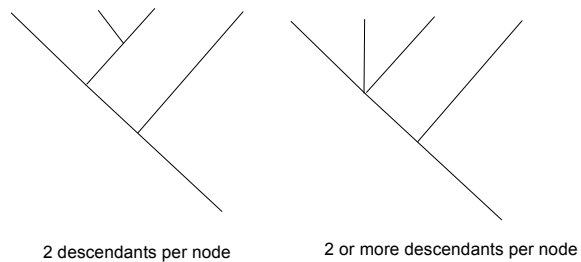
- Without a root, we do not know what direction time is running on a branch!
- Two ways to root a phylogeny
  - Outgroup
  - Midpoint
    - Only valid in the presence of a *clock*



## Unscaled vs. scaled branch lengths

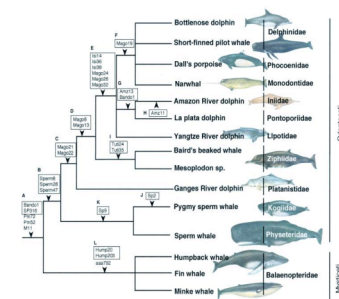


## Bifurcating vs multifurcating nodes



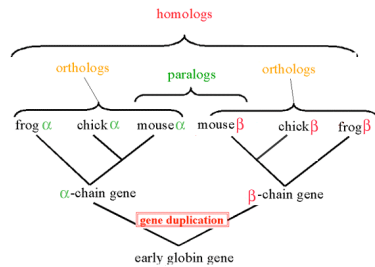
## Clades and monophyly

- Monophyly**
  - Containing all descendants of a common ancestor
  - Defines a *clade*
- Paraphyly**
  - Some descendants not included (e.g. reptiles)
- Polyphyly**
  - Common ancestor not included (e.g. flying animals)

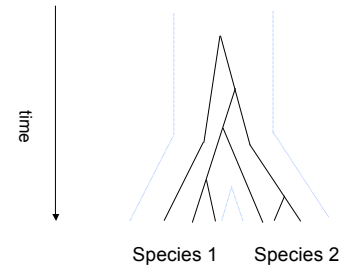


## Evolutionary relationships among genes

- **Homologs:** descended from a common ancestor
- **Orthologs:** homologs that diverged through speciation
- **Paralogs:** homologs that diverged through duplication



## Gene trees and species trees: lineage sorting



## Gene trees and species trees: lateral transfer

- Example: genetic exchange among thermophiles

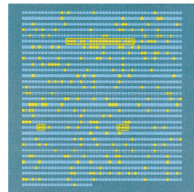


FIGURE 1. Genes of apparent archaean origin in the genome of *Aquifex aeolicus*. Yellow circles represent genes encoding proteins with reliable best hits to archaeal homologs. Gene clusters conserved in *Aquifex* and archaea are boxed. The largest cluster contains genes for a predicted RNA helicase, a nuclease and a zinc-finger-containing nucleic acid-binding protein; the remaining genes encode uncharacterized proteins, most of which are conserved in archaea and *Aquifex* only.

## Phylogenetic methods

- Distance matrix methods
  - Neighbor Joining
- Character-based methods
  - Non-statistical: Maximum Parsimony
  - Statistical: Maximum Likelihood & Bayesian Inference
- They will agree unless there is
  - Convergence, reversion or parallelism (*homoplasy*)
  - Rate heterogeneity
  - Long branches

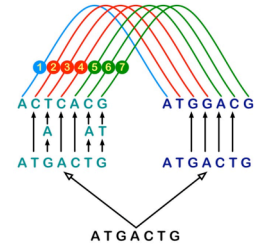
## The starting data: a sequence alignment

```

MYBU  MGLSDQGMGLVLDVWGRVHADIPGHSQSVLILKFRGPPHLDKRFDFRSLKSEDEMRASE
MYCE  -GLSDQGMGLVLDVWGRVHADIPGHSQSVLILKFRGPPHLDKRFDFRSLKSEDEMRASE
MYGO  -GLSDQGMGLVLDVWGRVHADIPGHSQSVLILKFRGPPHLDKRFDFRSLKSEDEMRASE
MYFG  -GLSDQGMGLVLDVWGRVHADIPGHSQSVLILKFRGPPHLDKRFDFRSLKSEDEMRASE
MYWR  -VLSQARMLLILVWGRVHADIPGHSQSVLILKFRGPPHLDKRFDFRSLKSEDEMRASE
MYPH  -GLSDQGMGLVLDVWGRVHADIPGHSQSVLILKFRGPPHLDKRFDFRSLKSEDEMRASE
MYAQ  MGLSDQGMGLVLDVWGRVHADIPGHSQSVLILKFRGPPHLDKRFDFRSLKSEDEMRASE
MYRK  ----TFMZRVRVAVVPPDIPAVGLALILAFKREKRDILPFAEIPVQD-LGWRP
MYTU  ----ADFDKELKQSPFADYTFMGSQVIEFKEPFEKGLPFAEAGAGD-LAGFA
consensus -glsdqwglvldvwgrvhadipghsqsvlilfrgpphldkrfdfRSLKSEDEMRASE

MYBU  DLRKHGAVVLTALGILKRNKGRHARLKPAGSHATKRNIPVYLEPFSRCLIQVLSGRN
MYCE  DLRKHGAVVLTALGILKRNKGRHARLKPAGSHATKRNIPVYLEPFSRCLIQVLSGRN
MYGO  DLRKHGAVVLTALGILKRNKGRHARLKPAGSHATKRNIPVYLEPFSRCLIQVLSGRN
MYFG  DLRKHGAVVLTALGILKRNKGRHARLKPAGSHATKRNIPVYLEPFSRCLIQVLSGRN
MYWR  DLRKHGAVVLTALGILKRNKGRHARLKPAGSHATKRNIPVYLEPFSRCLIQVLSGRN
MYPH  DLRKHGAVVLTALGILKRNKGRHARLKPAGSHATKRNIPVYLEPFSRCLIQVLSGRN
MYAQ  DLRKHGAVVLTALGILKRNKGRHARLKPAGSHATKRNIPVYLEPFSRCLIQVLSGRN
MYRK  DLRKHGAVVLTALGILKRNKGRHARLKPAGSHATKRNIPVYLEPFSRCLIQVLSGRN
MYTU  AIFAGGAVVLTALGILKRNKGRHARLKPAGSHATKRNIPVYLEPFSRCLIQVLSGRN
consensus dLkKHG-VLtaLgYlLkRNkGhHeseLkpLq*HatkRRIPvYleFse-ii-Vl-akh

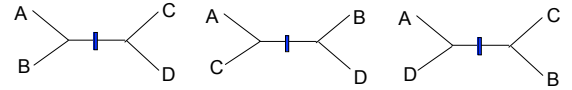
MYBU  PGLPGADAGSAMRKALELFRNIMASHYKELGFGC
MYCE  PGLPGADAGSAMRKALELFRNIMASHYKELGFGC
MYGO  PGLPGADAGSAMRKALELFRNIMASHYKELGFGC
MYFG  PGLPGADAGSAMRKALELFRNIMASHYKELGFGC
MYWR  PGLPGADAGSAMRKALELFRNIMASHYKELGFGC
MYPH  PGLPGADAGSAMRKALELFRNIMASHYKELGFGC
MYAQ  PGLPGADAGSAMRKALELFRNIMASHYKELGFGC
MYRK  PGLPGADAGSAMRKALELFRNIMASHYKELGFGC
MYTU  PGLPGADAGSAMRKALELFRNIMASHYKELGFGC
consensus pglpgadaGyam-kalelfr-Dnns-YKlYfgc
    
```



- 1 no change
- 2 one apparent change:
- 3 single substitution
- 4 multiple substitution
- 5 coincidental substitution
- 6 no apparent change:
- 7 parallel substitution
- 8 convergent substitution
- 9 back substitution

Homoplasy

## Maximum parsimony

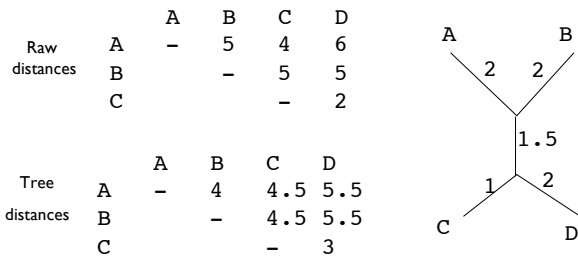


- A acgttgccga
- B acgttactgg
- C cgtaagatcg
- D cgtaaaaccg
- 111112131-

## Neighbor joining

- Clusters taxa based on a matrix of pairwise distance
- The simplest distance is percent divergence
  - DNA or amino acid
  - Only works for very similar sequences
- Better to correct for multiple substitutions
  - Aiming for a distance measure that is linear with time
  - Different methods used for DNA, codons and amino acids
  - Special methods need to be used where GC content is not constant

## Neighbor joining



## Statistical methods

- Two flavors
  - Maximum likelihood
  - Bayesian inference
- Lend themselves to testing hypotheses about trees
  - Getting the correct tree can be less important than testing the hypothesis

## Statistical methods and Bayes' Rule

$$P(M|D) = \frac{P(D|M)P(M)}{\sum_{\text{for all } M} P(D|M)P(M)}$$

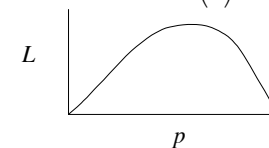
Posterior →  $P(M|D)$  =  $\frac{\text{Likelihood } P(D|M) \times \text{Prior } P(M)}{\sum_{\text{for all } M} P(D|M)P(M)}$

where  
 $M$  = model  
 $D$  = data

## Coin toss example

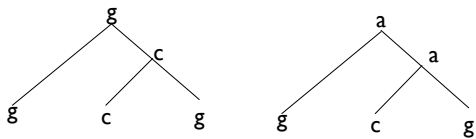
- A coin turns up heads with probability  $p$
- What is the likelihood of the data if we get  $k$  heads out of  $n$  tosses ?

$$L = P(x = k | n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$



## Maximum likelihood

- The phylogeny with the highest likelihood is taken to be correct
- Calculating the likelihood of a phylogeny is computationally intensive
  - For each possible topology, the probability of all possible ancestral states are calculated for each column
  - The topology, branch lengths and substitution model are equivalent to the probability of heads vs. tails

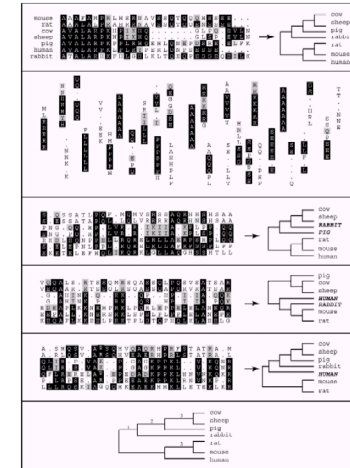


## Bayesian inference

- We may have an *a priori* hypothesis that a coin will be fair
  - $p=0.5$
- Is our belief shaken if
  - we observe 9/10 heads?
  - we observe 900/1000 heads?
- Bayes' Rule gives us the *posterior probability* of the hypothesis
  - the product of the *likelihood* and the *prior probability*
- Pros and cons
  - We get a more intuitive interpretation of probability
  - We get a natural measure of clade support
  - It turns out that larger datasets can be analyzed with Bayesian algorithms
  - But we must specify the (unknown) prior probability of each tree

## Bootstrapping

- How much should you trust any given branch, or clade?
  - With NJ, parsimony and ML, clades do not come with marginal probabilities
- *Bootstrap* by resampling the original alignment
  - Choose an alignment having the same number of columns *with replacement*
  - Compute a new tree
  - Repeat this many times
  - Count the proportion of resampled trees in which each original branch appears
  - Label the branches on the original tree with these proportions



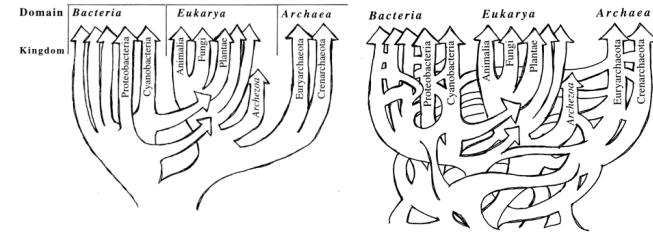
## Some applications of phylogenetics to microbial genetics

- Testing for lateral gene transfer events
- Studying ancient endosymbiosis events
- Characterizing microbial diversity
  - metagenomics
- Functional annotation of microbial genomes

## Lateral gene transfer

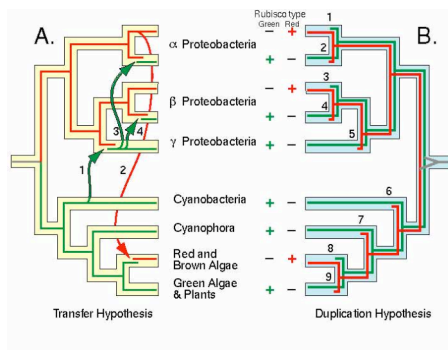
- Non-vertical transmission of genetic material
- Mechanisms
  - Viral transfer
  - Symbiosis/phagocytosis
  - Transformation
- Detecting lateral gene transfer
  - Closest database match
  - Altered base/codon usage
  - *Phylogenetic incongruence*

## Two views of the tree of life

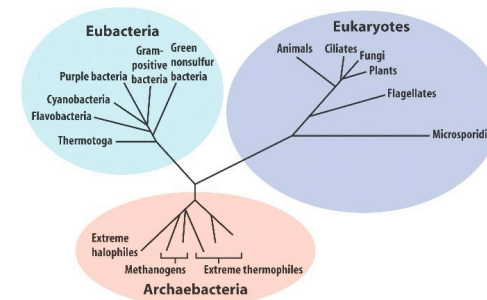


W. Ford Doolittle

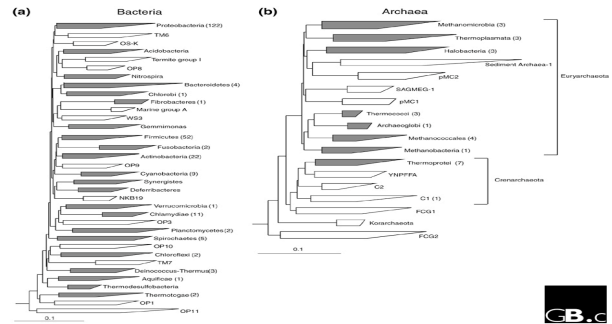
## Stuying ancient endosymbiosis



## One gene's view of the tree of life

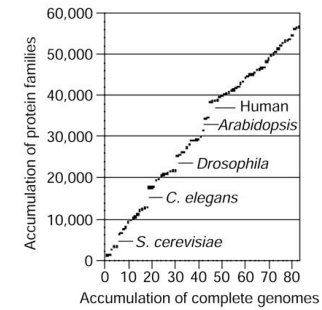


## Uncultured microbial diversity



Hugenholtz P (2000) Genome Biology 3, 1

## Continued discovery of protein families



Kunin et al. (2003) Genome Biol 4, 401

## Points to remember

- Getting a good tree
  - An accurate phylogeny depend on an accurate sequence alignment
  - All methods will give similar results in the absence of homoplasy and rate heterogeneity
  - Use NJ for a quick and dirty tree, ML or Bayesian methods for greater accuracy and to test hypotheses
- To interpret the tree
  - You need to root it (don't be fooled by the graphic!)
  - Obtain some measure of clade support (e.g. bootstrap)
    - Branches with low support should be collapsed to polytomies
  - Think of the phylogeny as a series of nested clades
- Approach the tree with a hypothesis, or at least have a clear goal in mind
  - e.g. what would the tree look like in the absence of horizontal transfer?