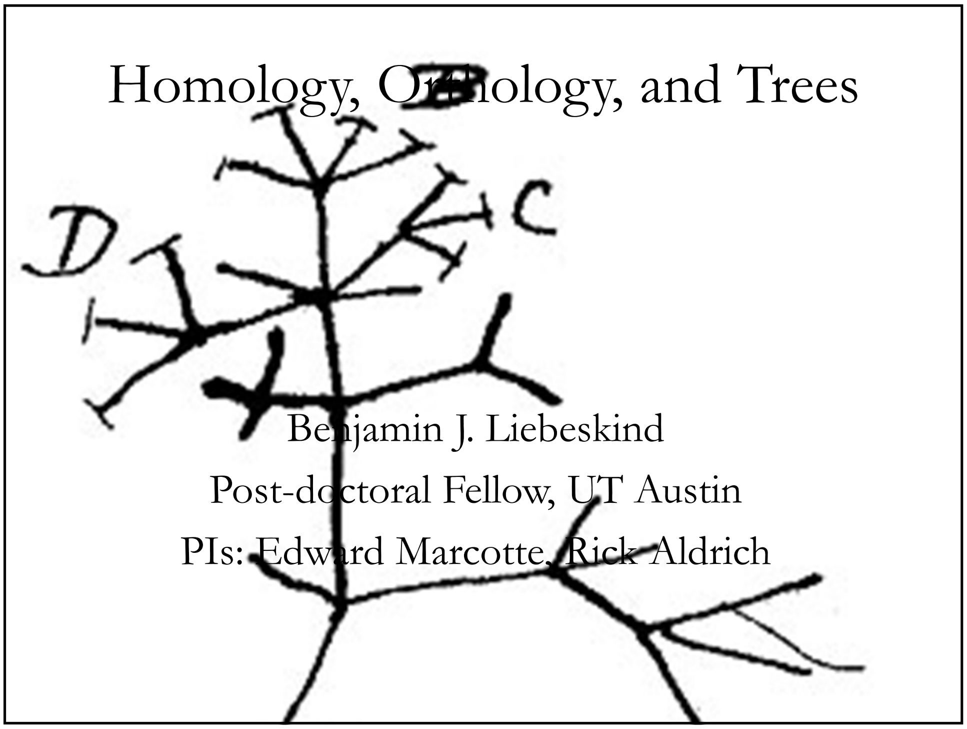
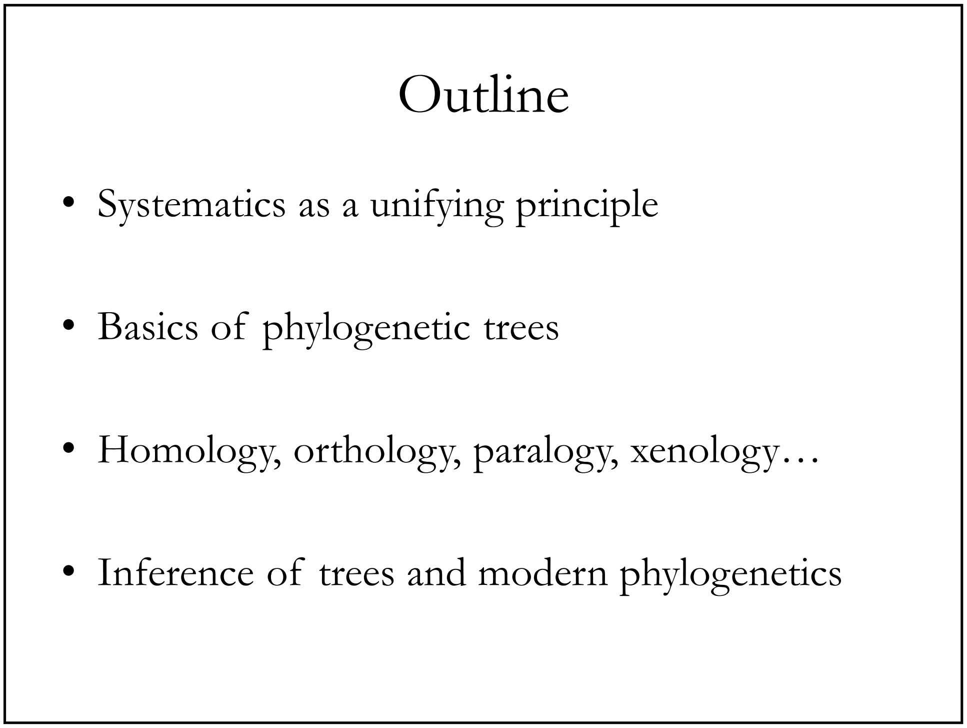


Homology, Orthology, and Trees

Benjamin J. Liebeskind
Post-doctoral Fellow, UT Austin
PIs: Edward Marcotte, Rick Aldrich



Outline

- Systematics as a unifying principle
- Basics of phylogenetic trees
- Homology, orthology, paralogy, xenology…
- Inference of trees and modern phylogenetics

Part I - Systematics

- Biology: ***why are there so many things?***

Systematics

- Biology: ***why are there so many things?***

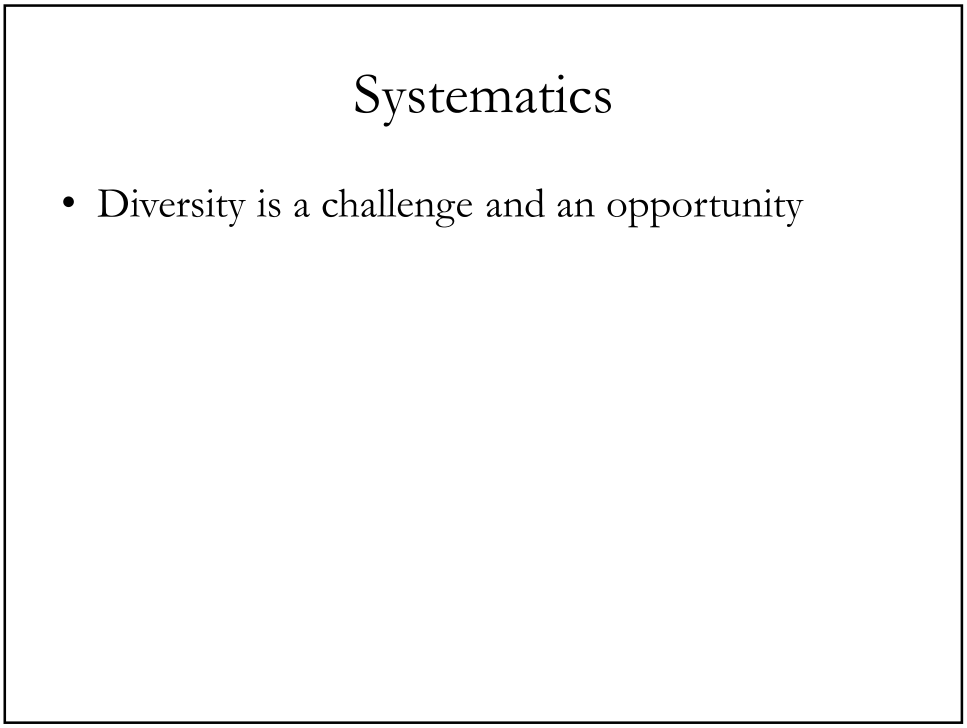
- Diversity is a fundamental fact of biology
  - It is created by a process: Evolution

Systematics

- Biology: ***why are there so many things?***
- Diversity is a fundamental fact of biology
  - It is created by a process: Evolution
- All organisms are the way they are because they evolved to be that way

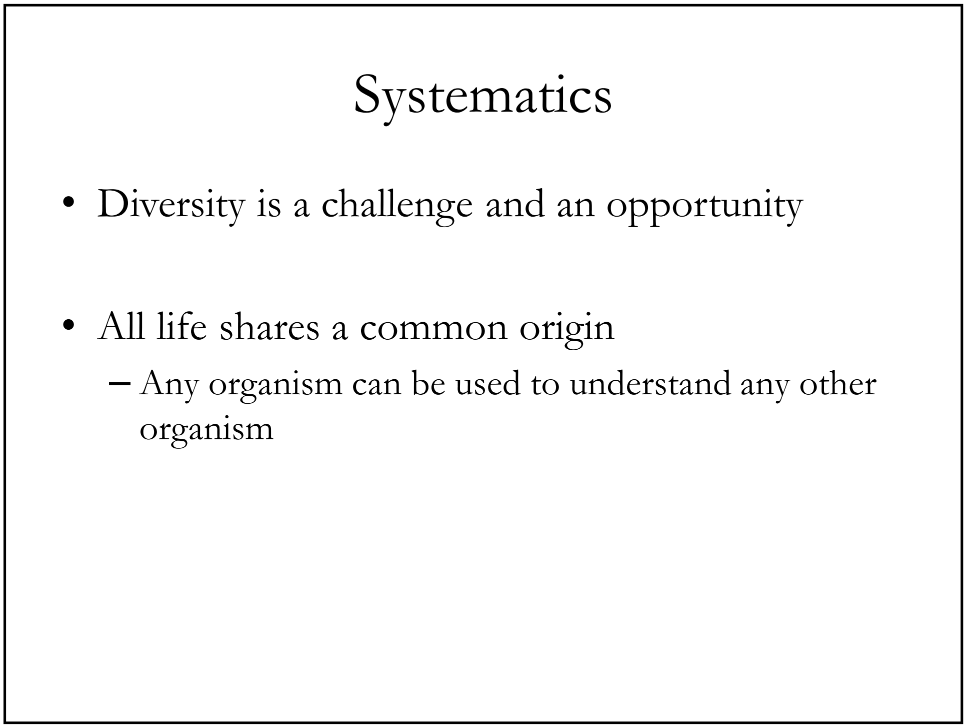
Systematics

- Biology: ***why are there so many things?***
- Diversity is a fundamental fact of biology
  - It is created by a process: Evolution
- "Nothing makes sense except in the light of evolution" – Theodosius Dobzhansky



Systematics

- Diversity is a challenge and an opportunity



Systematics

- Diversity is a challenge and an opportunity
- All life shares a common origin
  - Any organism can be used to understand any other organism

Systematics

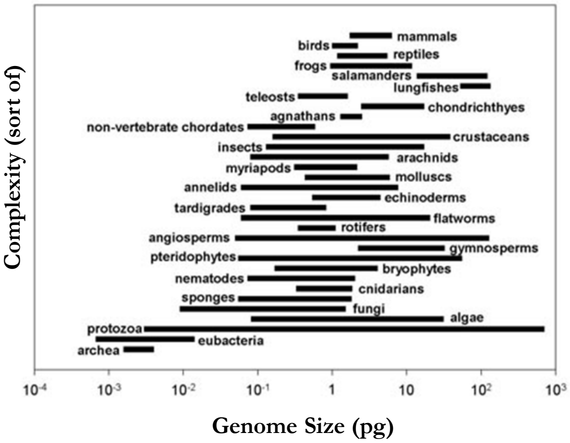
- Diversity is a challenge and an opportunity
- All life shares a common origin
  - Any organism can be used to understand any other organism
- But life forms are radically different
  - Evolution is the key to comparison

Systematics

- Diversity is a challenge and an opportunity

~ 6 orders of magnitude difference in genome size across organisms !!

Gregory, T.R. (2004)

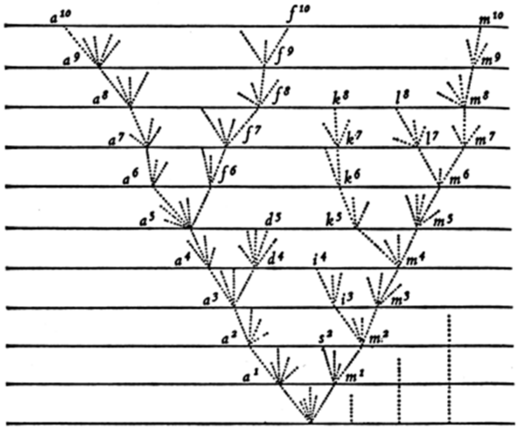


Systematics

- In order to compare organisms, you must *systematize* (group) them.
  – Same goes for parts of organisms.
- Modern systematics uses phylogenetic trees

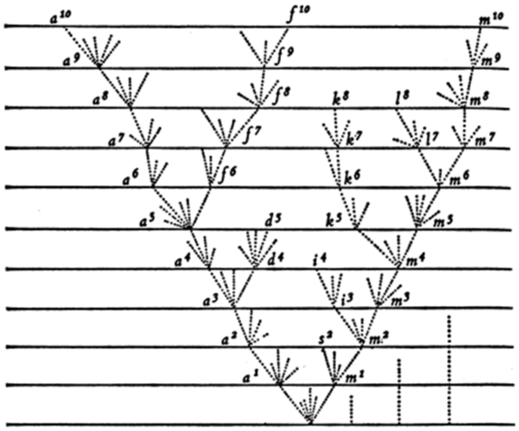
Basics of Phylogenetics

- Only figure in Darwin's "Origin of Species"



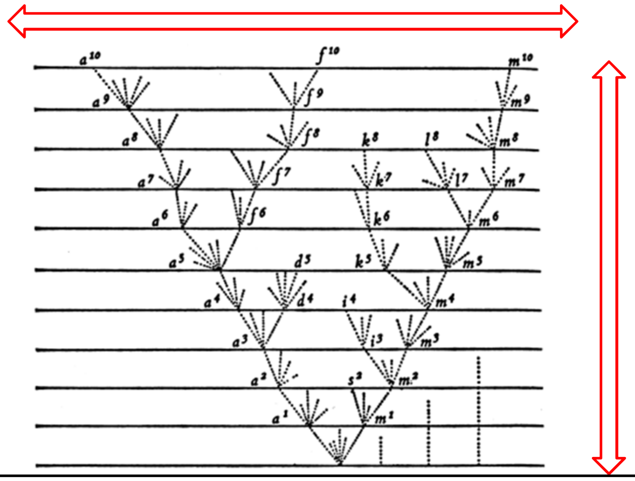
Basics of Phylogenetics

- Trees show the relationship between pattern and process



Basics of Phylogenetics

Pattern



Process

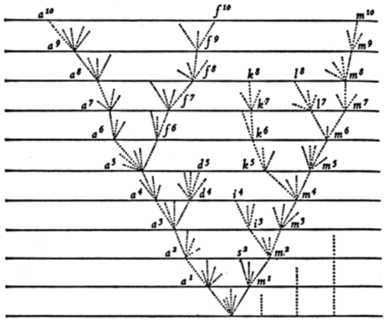
Basics of Phylogenetics

- Phylogenetic systematics (cladistics)
  - Organisms should be grouped by phylogenetic relationships
- Key terms:
  - Clade
  - Monophyly
  - Paraphyly

Willi Hennig (1913 – 1976)

Basics of Phylogenetics

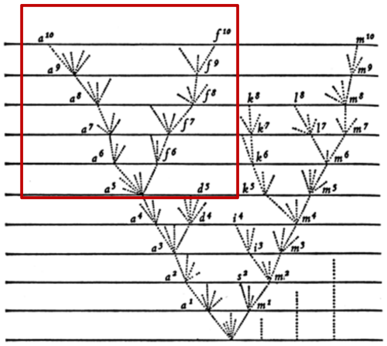
- Clade: an ancestor and all of its descendants



Basics of Phylogenetics

- Clade: an ancestor and all of its descendants

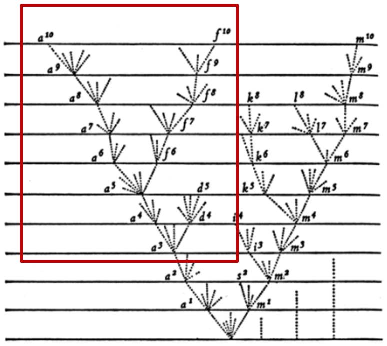
Clade!!



Basics of Phylogenetics

- Clade: an ancestor and all of its descendants

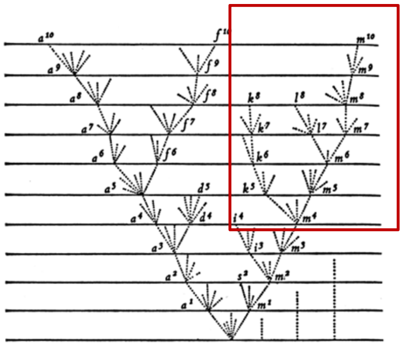
Clade!!



Basics of Phylogenetics

- Clade: an ancestor and all of its descendants

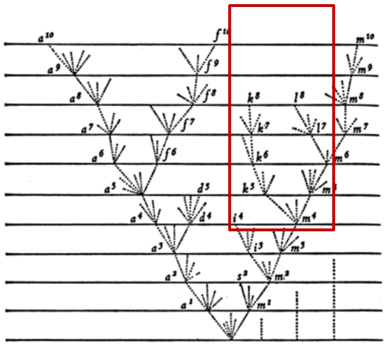
Clade!!



Basics of Phylogenetics

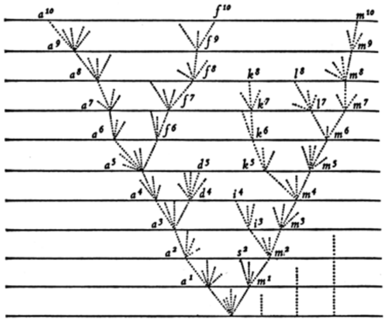
- Clade: an ancestor and all of its descendants

Not A Clade!!



Basics of Phylogenetics

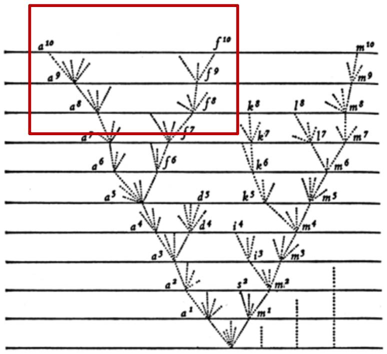
- Monophyletic group: organisms in a clade

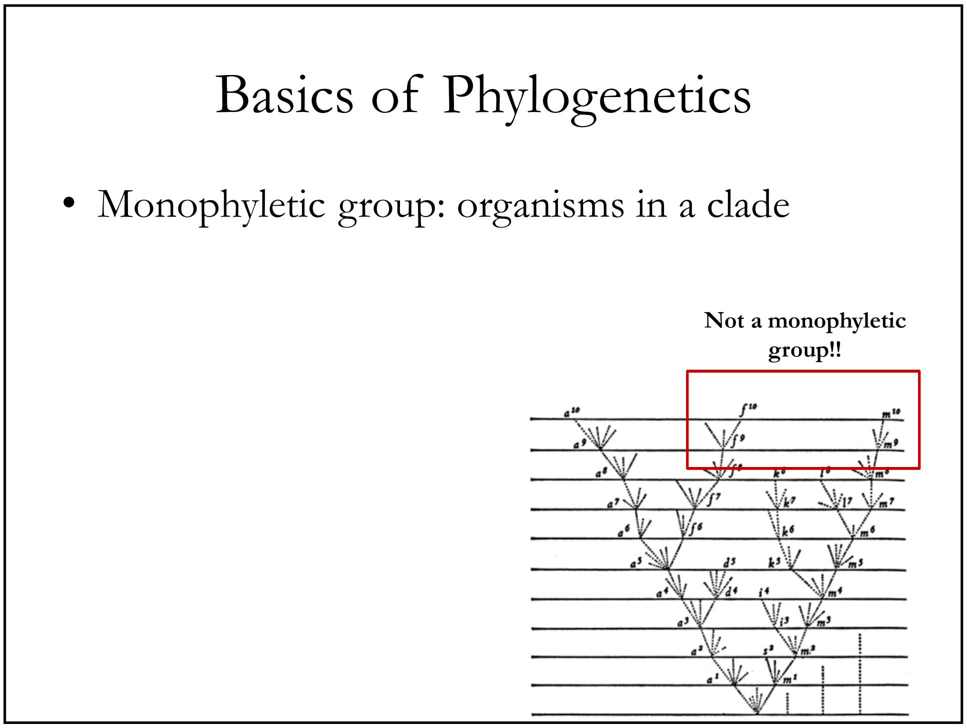


Basics of Phylogenetics

- Monophyletic group: organisms in a clade

Monophyletic group!!

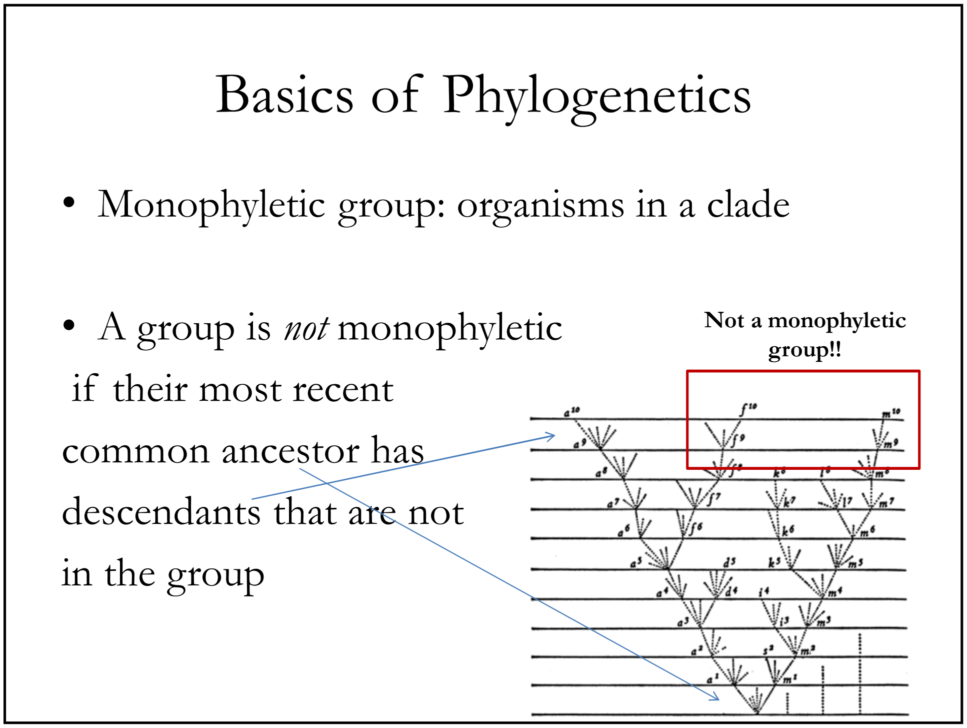




Basics of Phylogenetics

- Monophyletic group: organisms in a clade

Not a monophyletic group!!



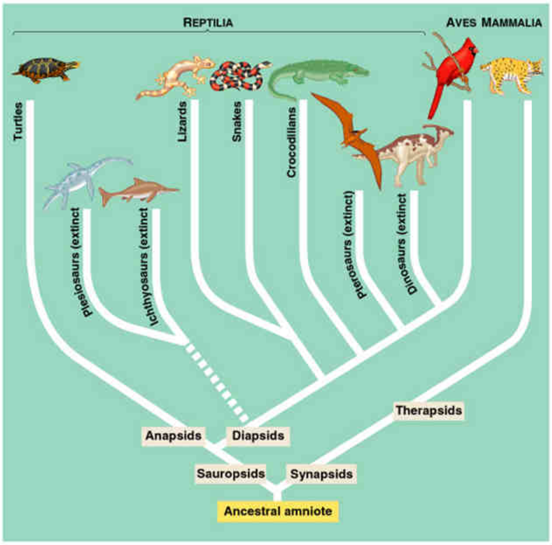
Basics of Phylogenetics

- Monophyletic group: organisms in a clade

- A group is *not* monophyletic if their most recent common ancestor has descendants that are not in the group

Not a monophyletic group!!

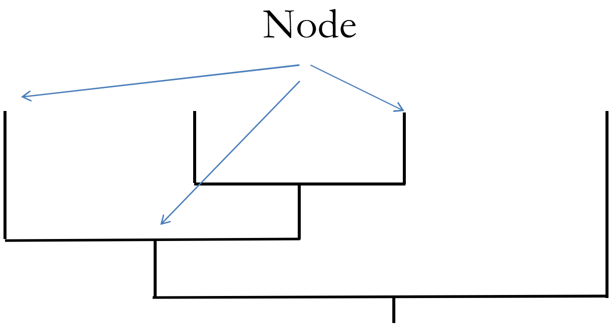
Basics of Phylogenetics



© 1999 Addison Wesley Longman, Inc.

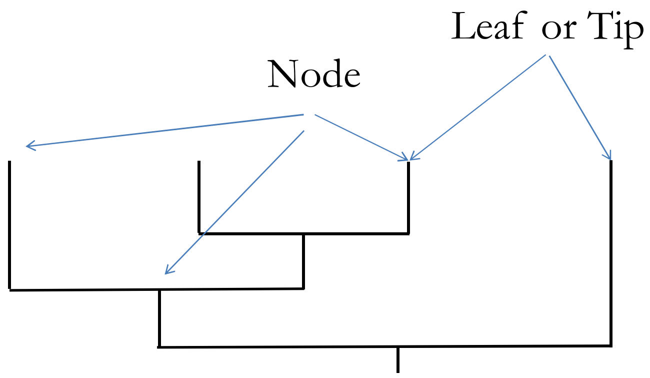
Basics of Phylogenetics

- Tree nomenclature: nodes


Node

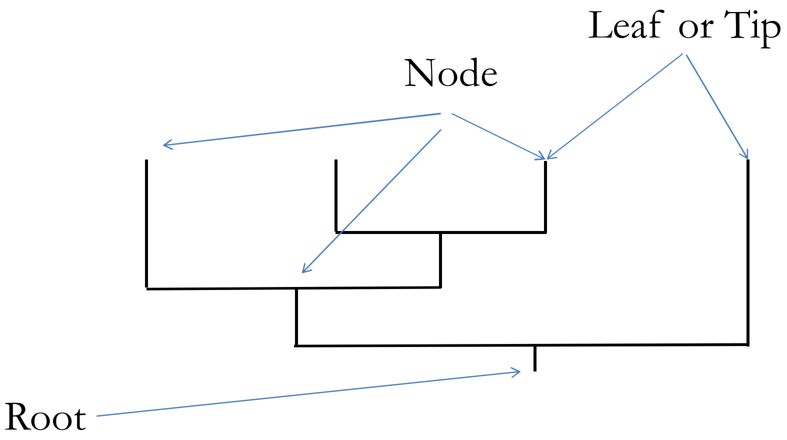
Basics of Phylogenetics

- Tree nomenclature: nodes



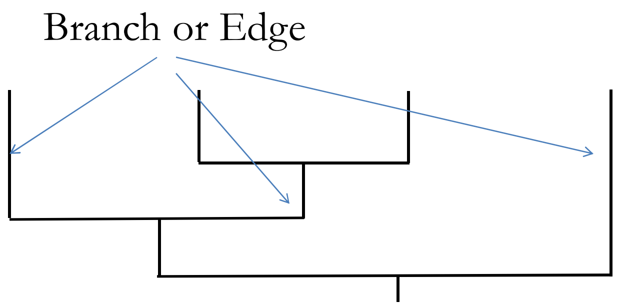
Basics of Phylogenetics

- Tree nomenclature: nodes



Basics of Phylogenetics

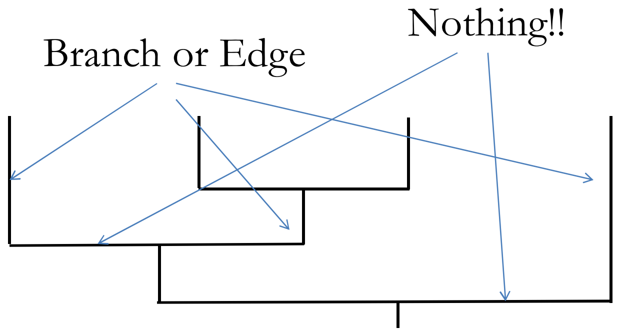
- Tree nomenclature: branches
  - Measures of evolutionary *rate*



Branch or Edge

Basics of Phylogenetics

- Tree nomenclature: branches
  - Measures of evolutionary *rate*

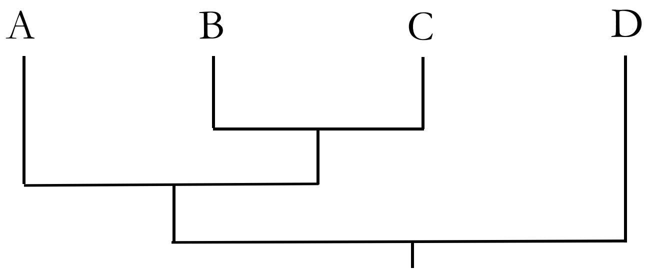


Nothing!!

Branch or Edge

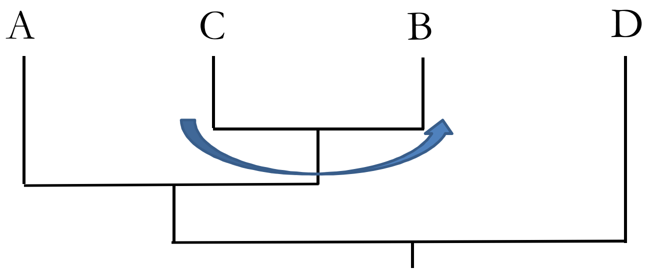
Basics of Phylogenetics

- Topology



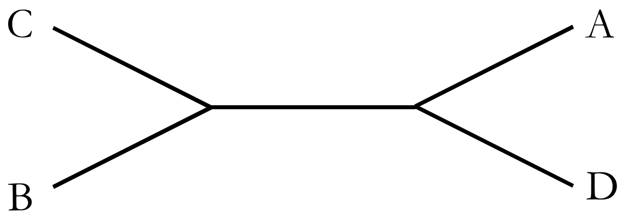
Basics of Phylogenetics

- Any node can be rotated without changing topology



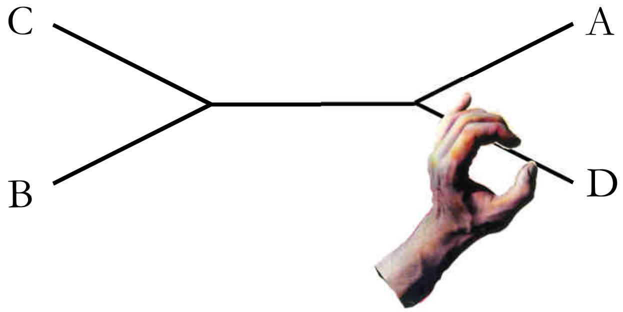
Basics of Phylogenetics

- The tree can be unrooted without changing topology



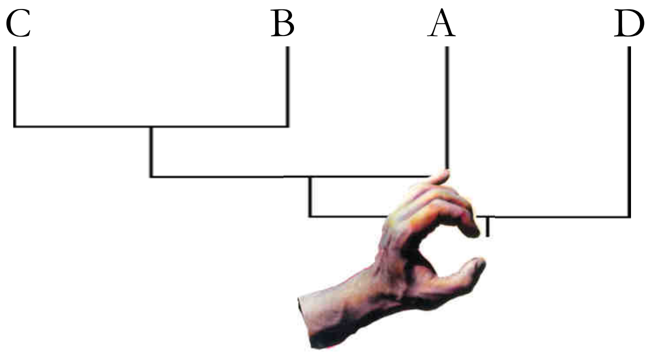
Basics of Phylogenetics

- The tree can be unrooted without changing topology



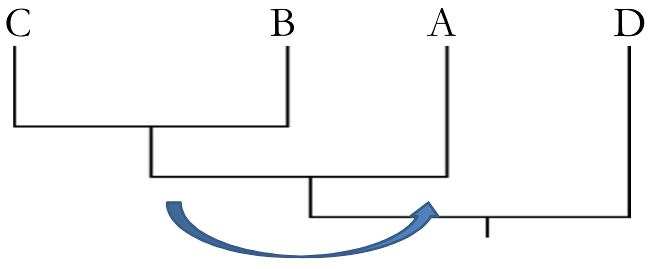
Basics of Phylogenetics

- The tree can be unrooted without changing topology



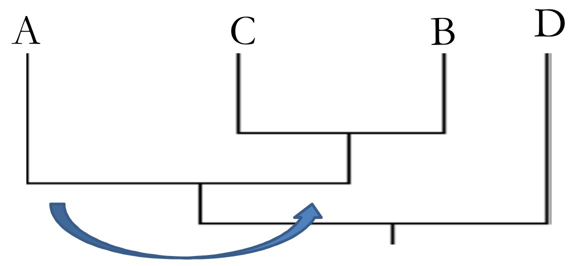
Basics of Phylogenetics

- The tree can be unrooted without changing topology



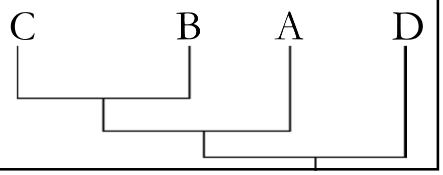
Basics of Phylogenetics

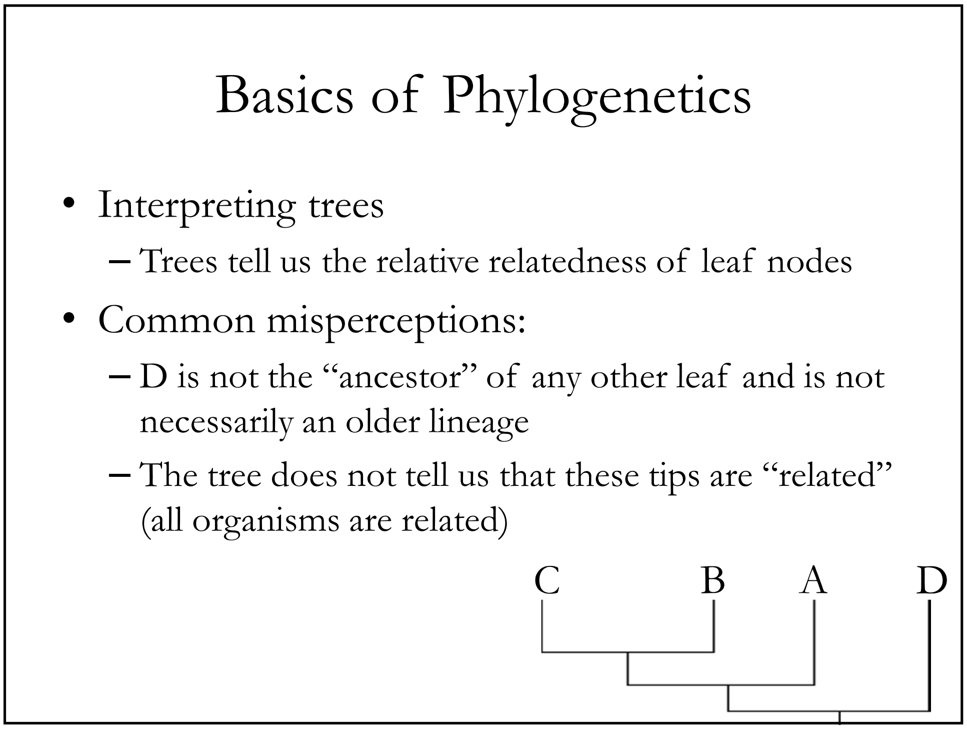
- The tree can be unrooted without changing topology



Basics of Phylogenetics

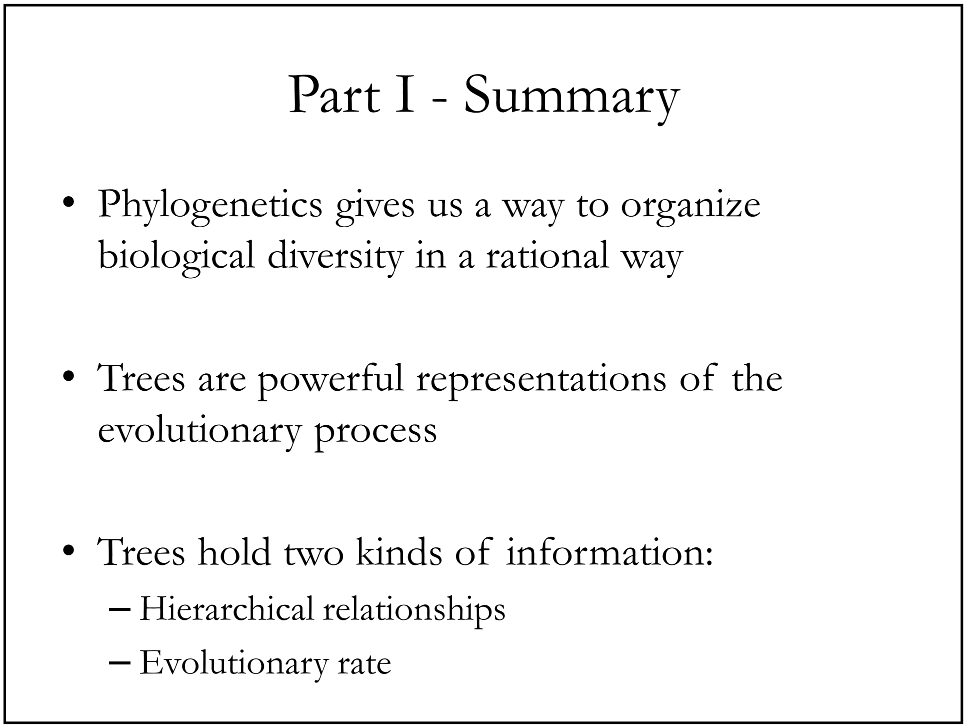
- Interpreting trees
  - Trees tell us the relative relatedness of leaf nodes





Basics of Phylogenetics

- Interpreting trees
  - Trees tell us the relative relatedness of leaf nodes
- Common misperceptions:
  - D is not the "ancestor" of any other leaf and is not necessarily an older lineage
  - The tree does not tell us that these tips are "related" (all organisms are related)



Part I - Summary

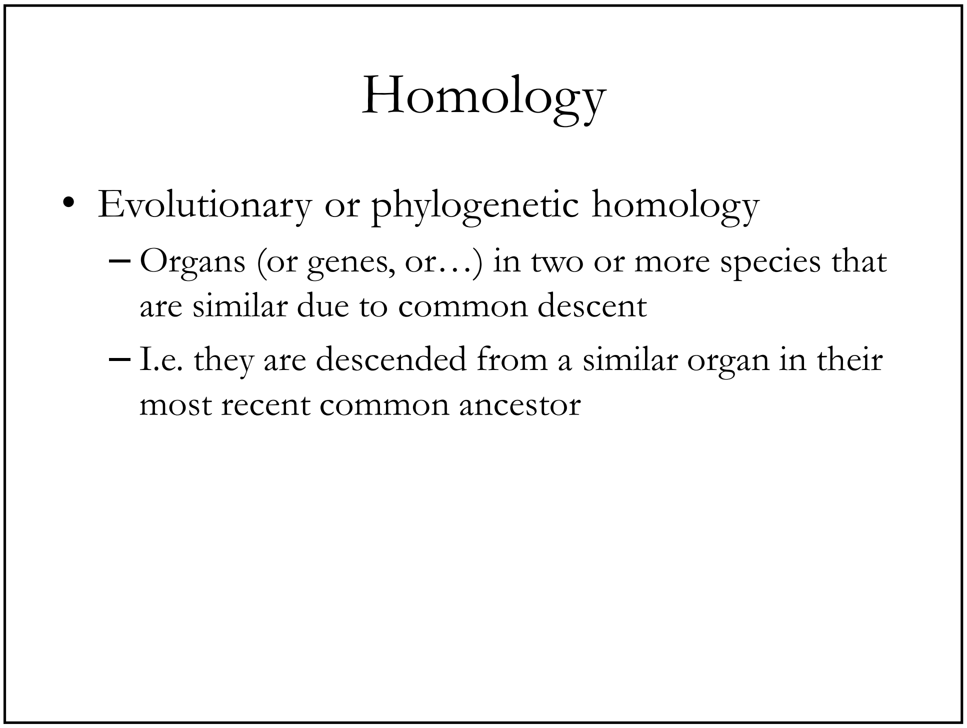
- Phylogenetics gives us a way to organize biological diversity in a rational way
- Trees are powerful representations of the evolutionary process
- Trees hold two kinds of information:
  - Hierarchical relationships
  - Evolutionary rate

Part II – The comparative method

Homology

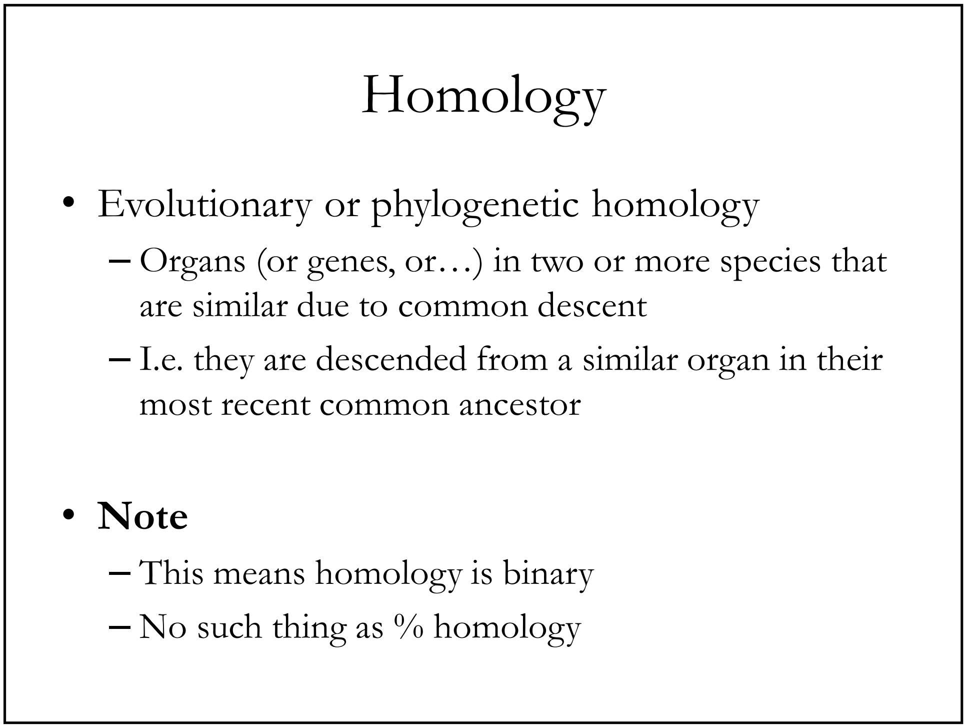
- When comparing parts of organisms, you need a criterion of "sameness"
  - Evolutionary "sameness" is called "homology"

> "Homologue…The same organ in different animals under every variety of form and function…Analogue…A part or organ in one animal which has the same function as another part or organ in a different animal"
>
> - Richard Owen (1843)



Homology

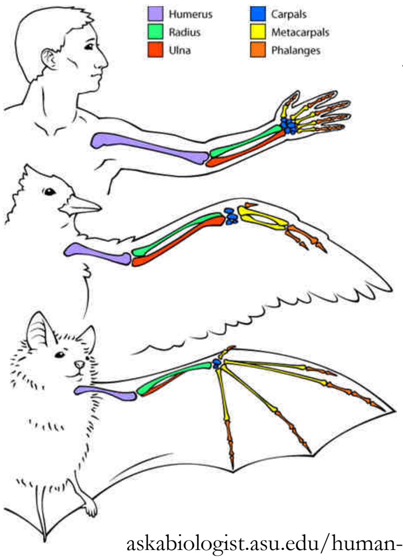
- Evolutionary or phylogenetic homology
  - Organs (or genes, or…) in two or more species that are similar due to common descent
  - I.e. they are descended from a similar organ in their most recent common ancestor



Homology

- Evolutionary or phylogenetic homology
  - Organs (or genes, or…) in two or more species that are similar due to common descent
  - I.e. they are descended from a similar organ in their most recent common ancestor
- **Note**
  - This means homology is binary
  - No such thing as % homology

Homology



This is trickier than it sounds!!

Organs are not monolithic entities!!

Bat wings and bird wings are homologous *as* vertebrate forelimbs

But they are analogous as wings

askabiologist.asu.edu/human-bird-and-bat-bone-comparison

Homology

- Homology also applies to genes
  - How can we tell whether genes are homologous?

Homology

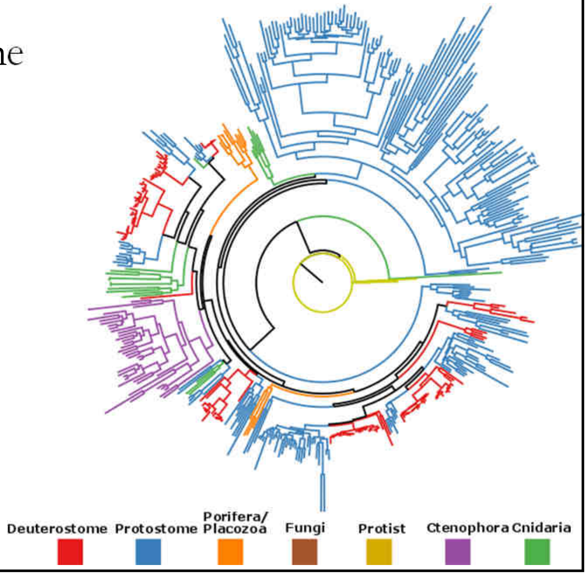
- Homology also applies to genes
  - How can we tell whether genes are homologous?

- Sequence matching scores derived from alignment
  - Null distributions of scores are easily derivable
  - Sequence space is HUGE!
    - Non-homologous gene scores are not distinguishable from random

Homology

- Homologous gene families are very large

- How do we find one-to-one correspondence across species?



Orthology

Walter Fitch

1929 - 2011

What is an Ortholog?

- Need a basis for comparing genes across species
  - *Orthology* is used nearly universally

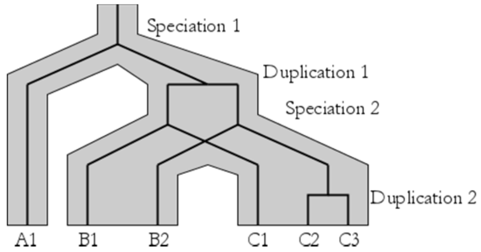
- "Same gene in a different species"

- "Homologous genes are related by speciation, whereas paralogs are related by duplication"

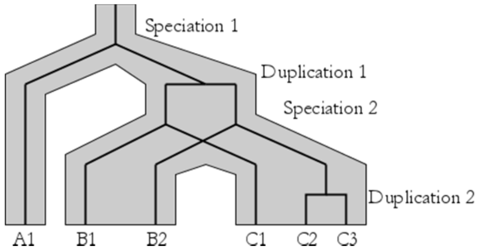
What is an Ortholog?

- Need a basis for comparing genes across species
  - *Orthology* is used nearly universally
- “~~Same gene in a different species~~”
- “~~Homologous genes are related by speciation, whereas paralogs are related by duplication~~”

What is an Ortholog?

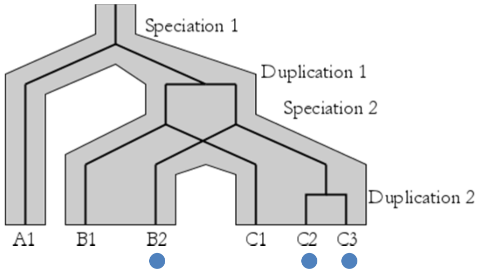


What is an Ortholog?



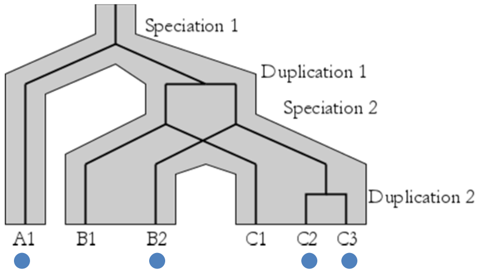
"Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous." – Walter Fitch (2000) *Trends in Genetics*

What is an Ortholog?



"Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous." – Walter Fitch (2000) *Trends in Genetics*

What is an Ortholog?



"Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous." – Walter Fitch (2000) *Trends in Genetics*

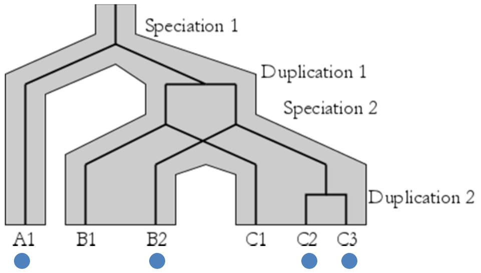
What is an Ortholog?

Deceptive simplicity:

Orthology is a pairwise relationship

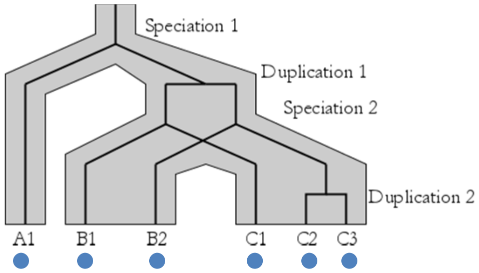
It is not transitive

Co-orthology is often abstracted from, with confusing results

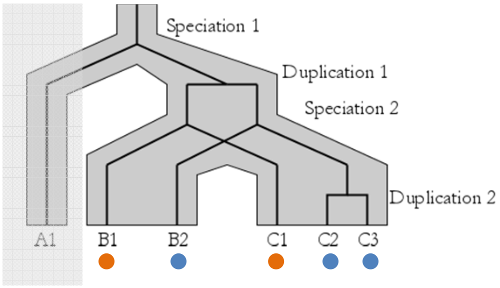


"Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous." – Walter Fitch (2000) *Trends in Genetics*

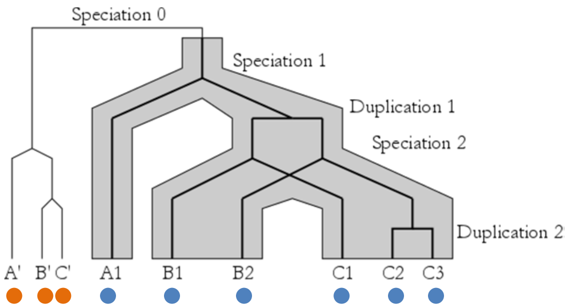
What is an Orthologous Group?



What is an Orthologous Group?



What is an Orthologous Group?



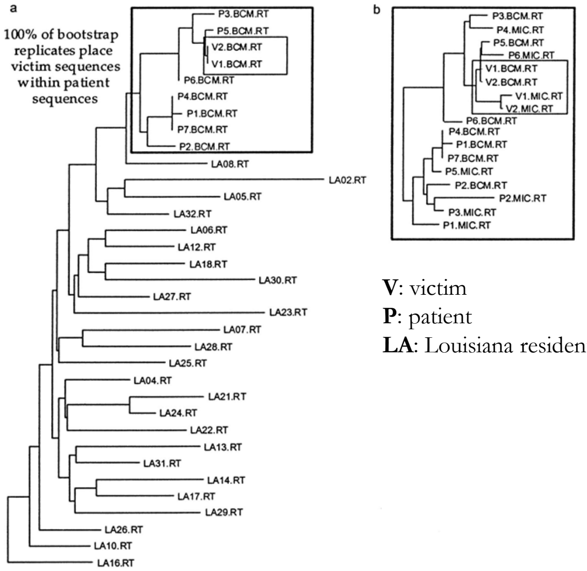
Part II - Summary

- Homology and Orthology give us a way to compare genes in different species
  - Homology contains whole families
  - Orthology is a direct comparison, usually denoting more functional similarity
- Both concepts are potentially slippery

Part III – Inferring Phylogenies

Inferring Phylogenies

- A doctor's girlfriend accuses him of injecting her with HIV. He said it was vitamin B12.
  - Who's right?
- Phylogenetics to the rescue!!



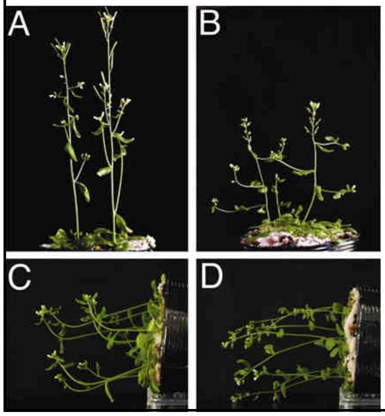
V: victim
P: patient
LA: Louisiana residents with HIV

Michael L. Metzker et al. PNAS 2002;99:14292-14297

David Hillis

Inferring Phylogenies

Gravitropism defects

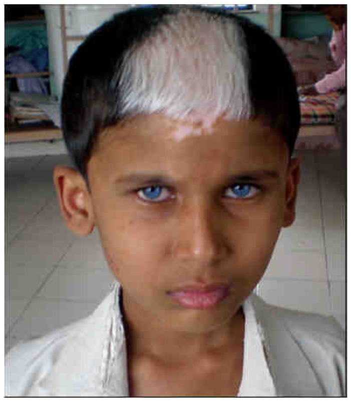


Significantly overlapping sets of orthologs

↔

McGary *et al.* (2013)

Waardenburg syndrome

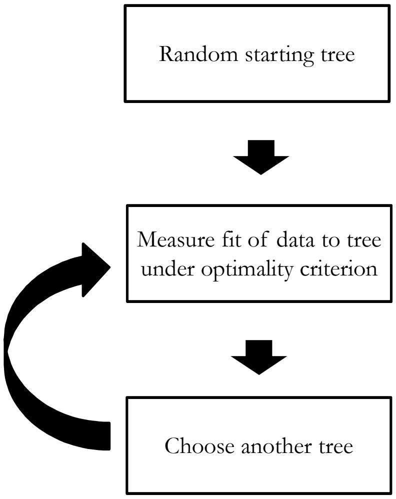


Inferring Phylogenies

How do we infer relatedness between genes?

Inferring Phylogenies

Algorithms:



Random starting tree

Measure fit of data to tree under optimality criterion

Choose another tree

Iterate
(for how long?)

Number of Possible Trees

Number of Taxa	Number of unrooted trees	Number of rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425
20	2.22E+020	8.20E+021
30	8.69E+036	4.95E+038
40	1.31E+055	1.01E+057
50	2.84E+074	2.75E+076
60	5.01E+094	5.86E+096
70	5.00E+115	6.85E+117
80	2.18E+137	3.43E+139

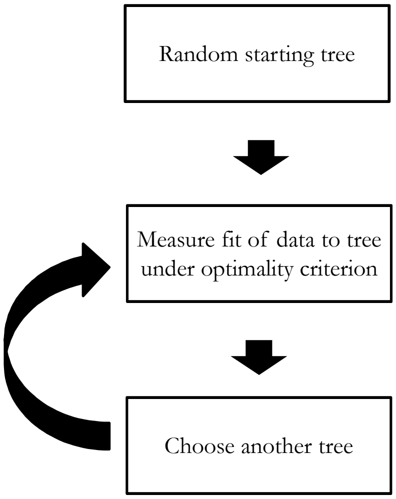
For comparison the universe contains *only* about 10^{89} protons.

(http://www.pagines.ma1.upc.edu/~casanellas/eaca/tree_number.html)

Inferring Phylogenies

Algorithms:

Heuristic Search:
Search a sub-space of trees with a well-defined stopping criterion

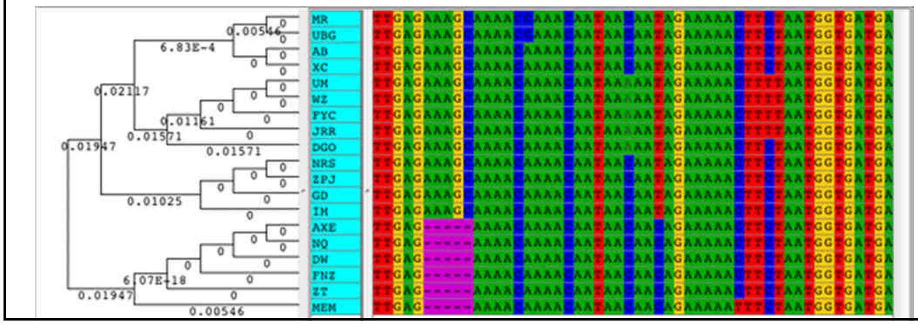


Inferring Phylogenies

- Optimality criteria for inferring trees
  - Pairwise distance methods
  - Maximum parsimony
  - Likelihood/Bayesian methods

Inferring Phylogenies

- Phylogenies are based on alignments
  - Taxa are represented row-wise
  - Columns are sites in the genome
    - Can be nucleotides or amino acids



Parsimony Score

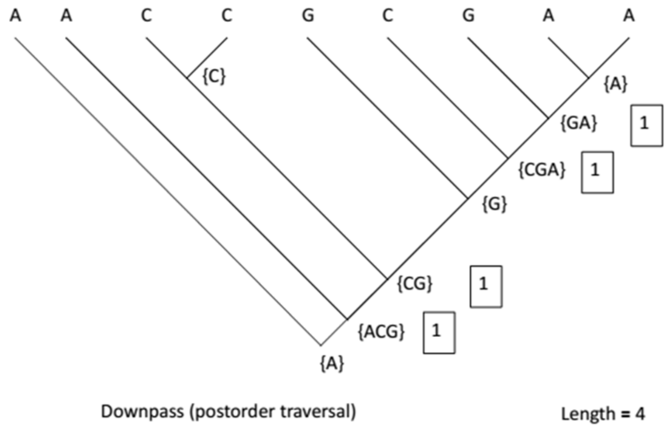


Figure curtesy of David Hillis

Parsimony Score

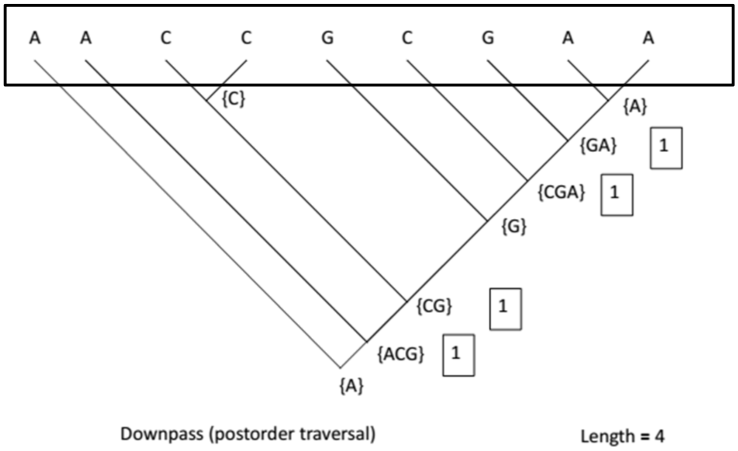


Figure curtesy of David Hillis

Parsimony Score

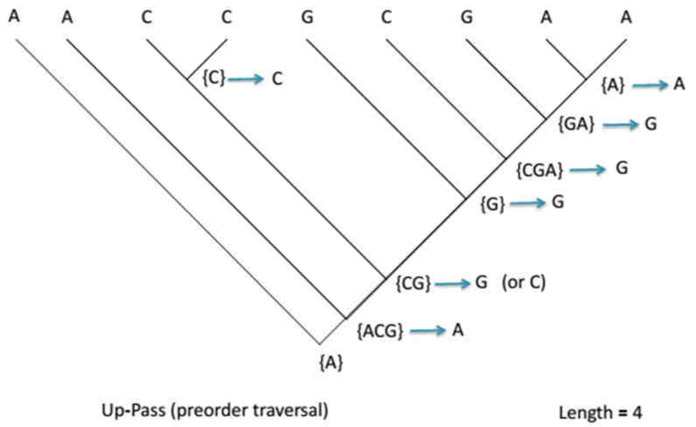


Figure curtesy of David Hillis

Parsimony Score

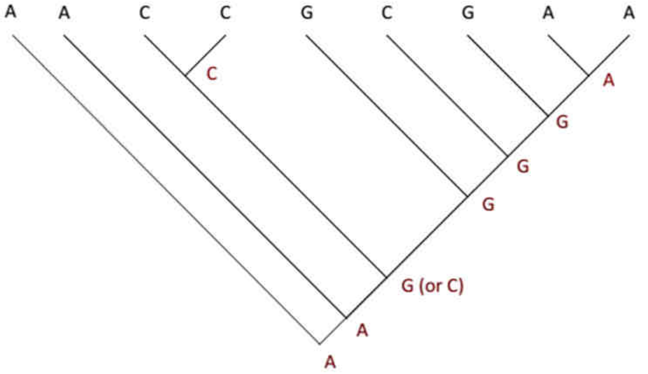


Figure curtesy of David Hillis

Parsimony Score

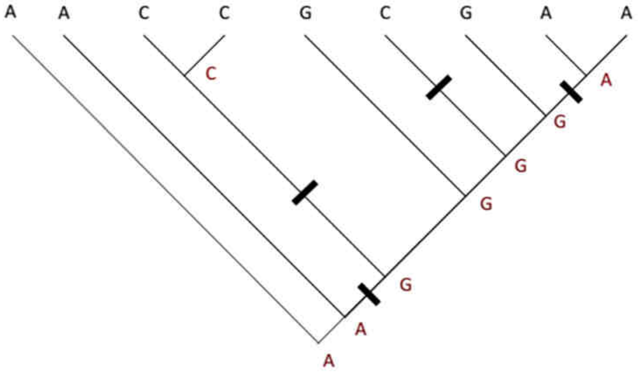


Figure curtesy of David Hillis

Parsimony Score

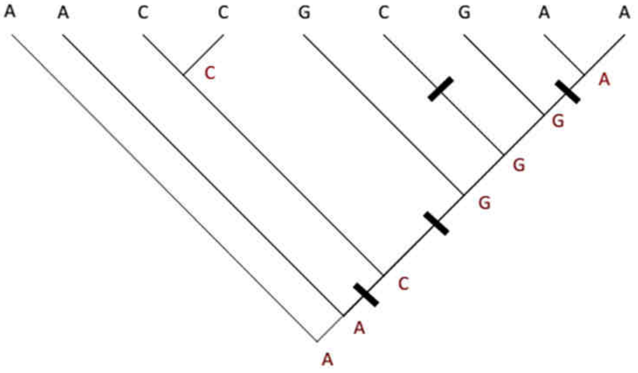


Figure curtesy of David Hillis

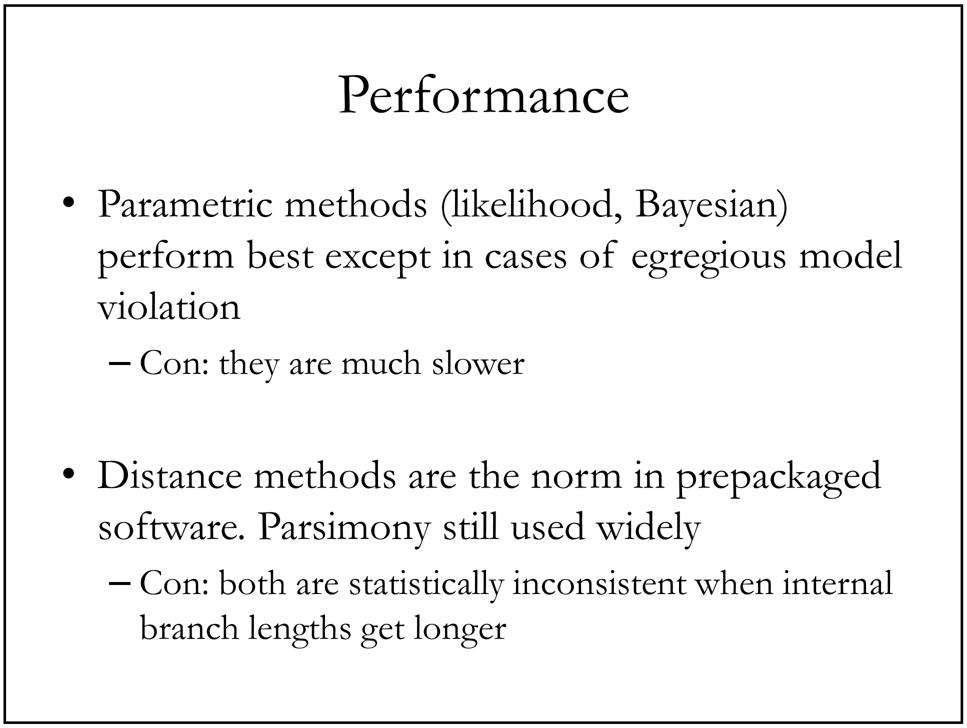
Parametric Methods

- Maximum likelihood methods find the tree that maximizes the probability of a model
  - arg max P(M|T)
- Bayesian methods calculate the probability of a tree given a model
  - P(T|M)
  - Bayes' theorem:

$$P(T|M) = \frac{P(M|T)\ P(T)}{P(M)}$$

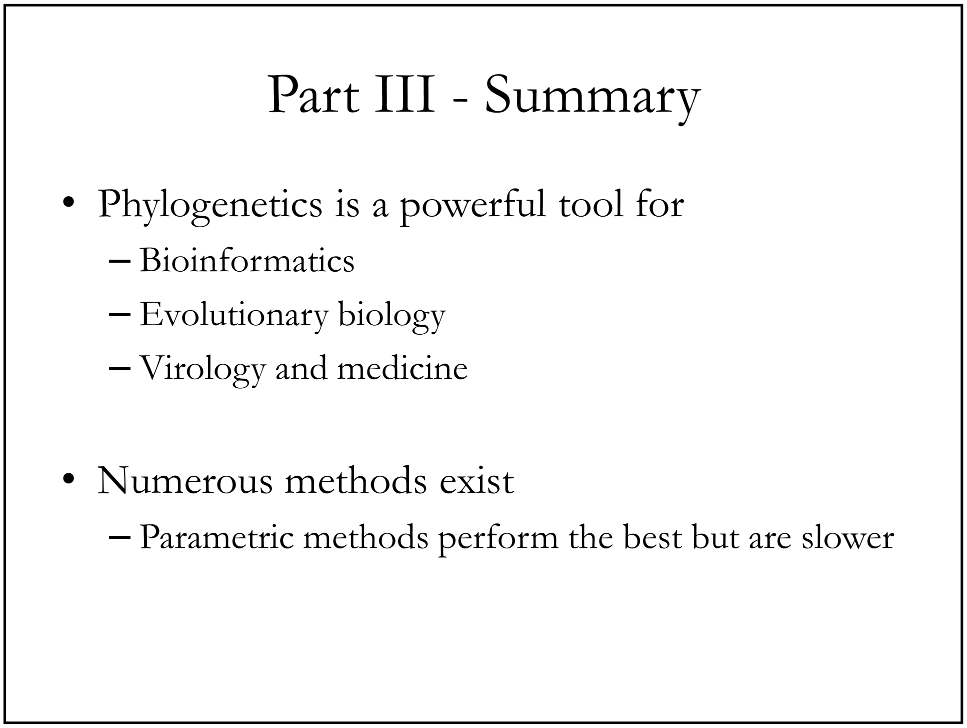
Models of Evolution

- Parametric criteria are evaluated analogously to parsimony
  - One tree is tried at a time!!



Performance

- Parametric methods (likelihood, Bayesian) perform best except in cases of egregious model violation
  - Con: they are much slower
- Distance methods are the norm in prepackaged software. Parsimony still used widely
  - Con: both are statistically inconsistent when internal branch lengths get longer



Part III - Summary

- Phylogenetics is a powerful tool for
  - Bioinformatics
  - Evolutionary biology
  - Virology and medicine
- Numerous methods exist
  - Parametric methods perform the best but are slower

Parametric Methods

- Maximum likelihood methods find the tree that maximizes the probability of a model
  - arg max P(M|T)
- Bayesian methods calculate the probability of a tree given a model
  - P(T|M)
  - Bayes' theorem:

Likelihood

Prior prob. of tree
(usually flat)

$$P(T|M) = \frac{\boxed{P(M|T)}\boxed{P(T)}}{\boxed{P(M)}}$$

Sum over parameter values of model
Evaluated numerically

Models of Evolution

- Continuous time Markov models

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{GA}(t) & p_{CA}(t) & p_{TA}(t) \\ p_{AG}(t) & p_{GG}(t) & p_{CG}(t) & p_{TG}(t) \\ p_{AC}(t) & p_{GC}(t) & p_{CC}(t) & p_{TC}(t) \\ p_{AT}(t) & p_{GT}(t) & p_{CT}(t) & p_{TT}(t) \end{pmatrix}$$

Models of Evolution

- Continuous time Markov models

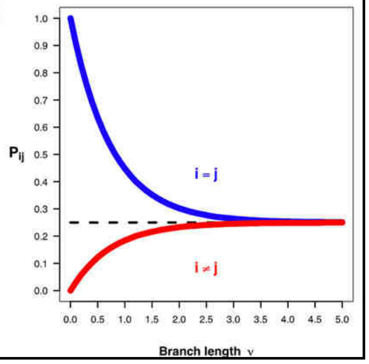
$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{GA}(t) & p_{CA}(t) & p_{TA}(t) \\ p_{AG}(t) & p_{GG}(t) & p_{CG}(t) & p_{TG}(t) \\ p_{AC}(t) & p_{GC}(t) & p_{CC}(t) & p_{TC}(t) \\ p_{AT}(t) & p_{GT}(t) & \boxed{p_{CT}(t)} & p_{TT}(t) \end{pmatrix}$$

Probability of changing from cytosine to thymine
In time *t* along a branch

Models of Evolution

- Continuous time Markov models

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{GA}(t) & p_{CA}(t) & p_{TA}(t) \\ p_{AG}(t) & p_{GG}(t) & p_{CG}(t) & p_{TG}(t) \\ p_{AC}(t) & p_{GC}(t) & p_{CC}(t) & p_{TC}(t) \\ p_{AT}(t) & p_{GT}(t) & p_{CT}(t) & p_{TT}(t) \end{pmatrix}$$



Wait times are exponentially functions
If transition probabilities are equal (=.25),
all probabilities approach .25