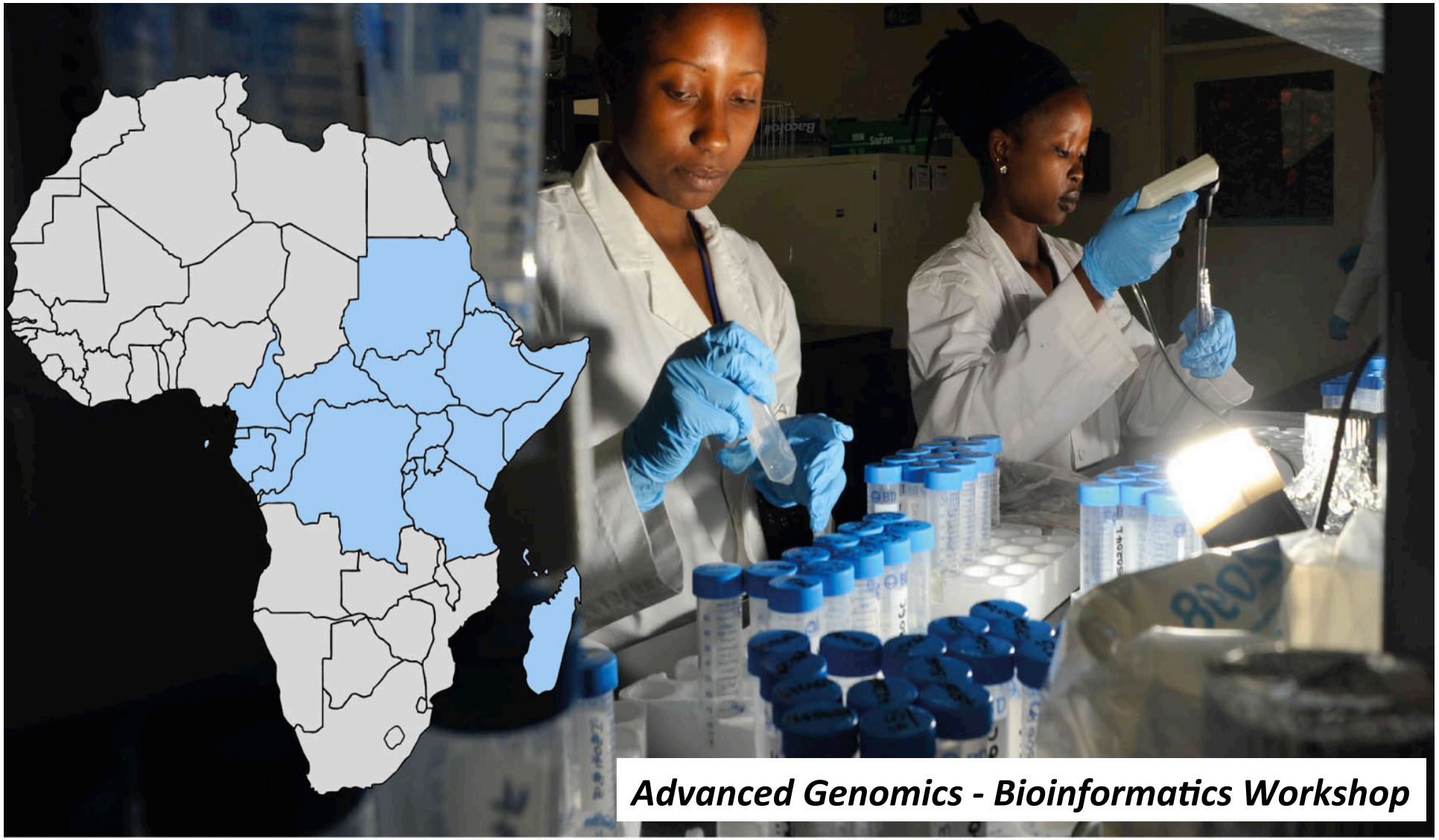


Advanced Genomics - Bioinformatics Workshop

Mark Wamalwa
BecA-ILRI Hub, Nairobi, Kenya
http://hub.africabiosciences.org/
m.wamalwa@cgiar.org

7th – 18th September 2015

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

African Union

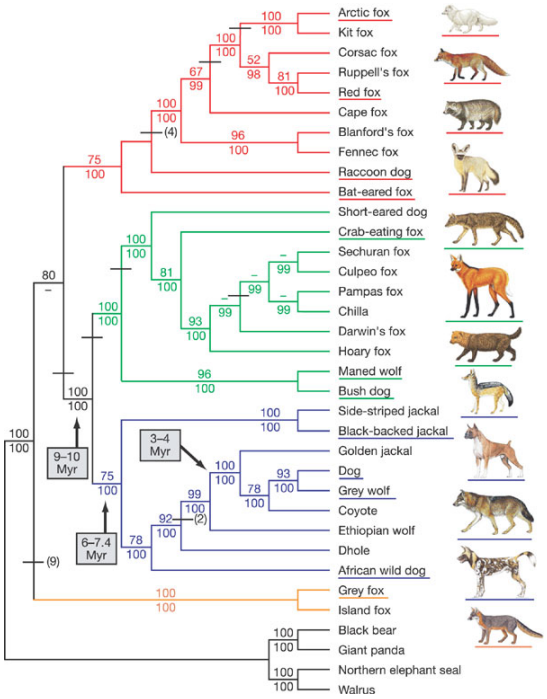
NEPAD A PROGRAMME OF THE AFRICAN UNION

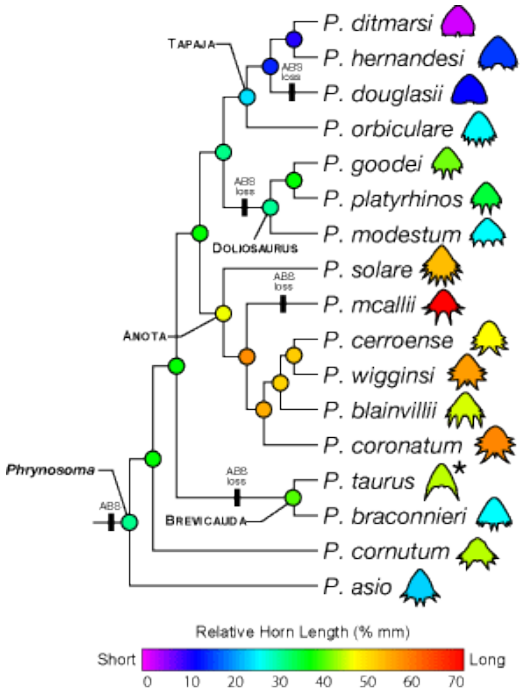
biosciences eastern and central africa

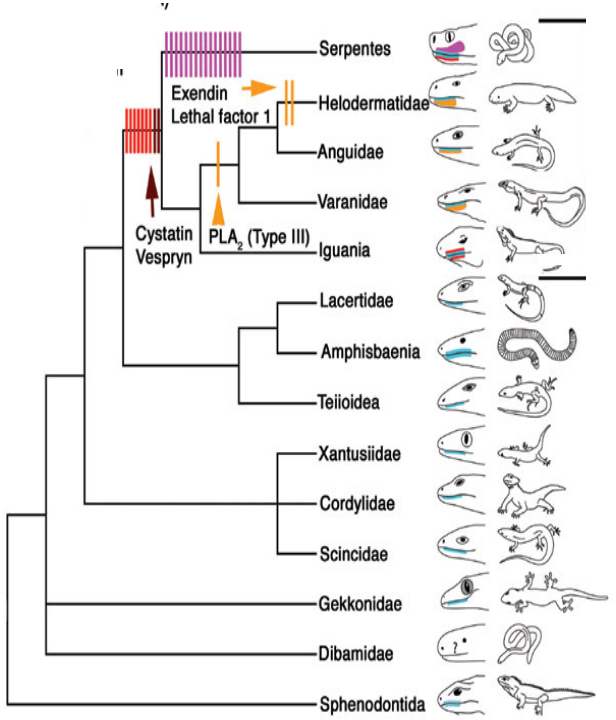
Phylogeny

- **Study of evolutionary relatedness among groups of organisms, achieved by:**
  - **Comparison of molecular data**
  - **Comparison of morphological data**

```
Outgroup     AAGCTTCATAGGAGCAACCATTCTAATAATAAGCCTCATAAAGCC
Species A    AAGCTTCACCGGCGCAGTTATCCTCATAATATGCCTCATAATGCC
Species B    GTGCTTCACCGACGCAGTTGTCCTCATAATGTGCCTCACTATGCC
Species C    GTGCTTCACCGACGCAGTTGCCCTCATGATGAGCCTCACTATGCA
```



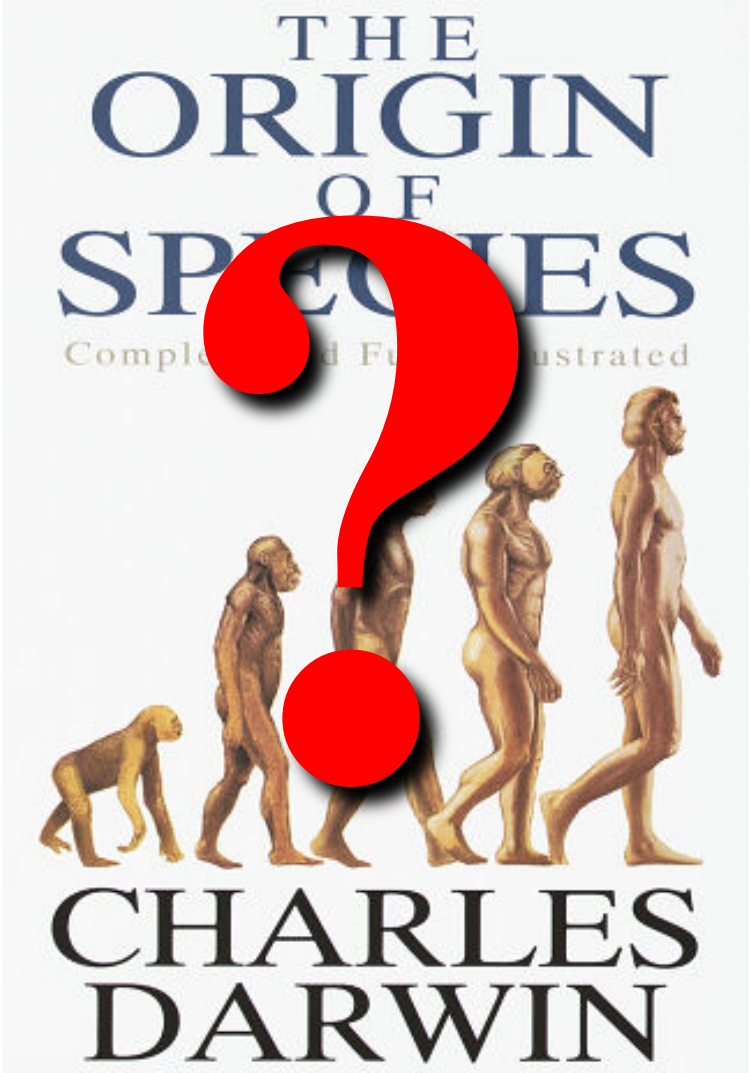




Objectives of phylogenetics

- Reconstruct the correct genealogical ties among biological entities i.e. to reconstruct the evolutionary history of organisms
- Estimate the time of divergence between biological entities
- Chronicle the sequence of events along evolutionary lineages
- To depict the phylogenies, the historical events = evolution of species / genes in tree graphs

Some history…

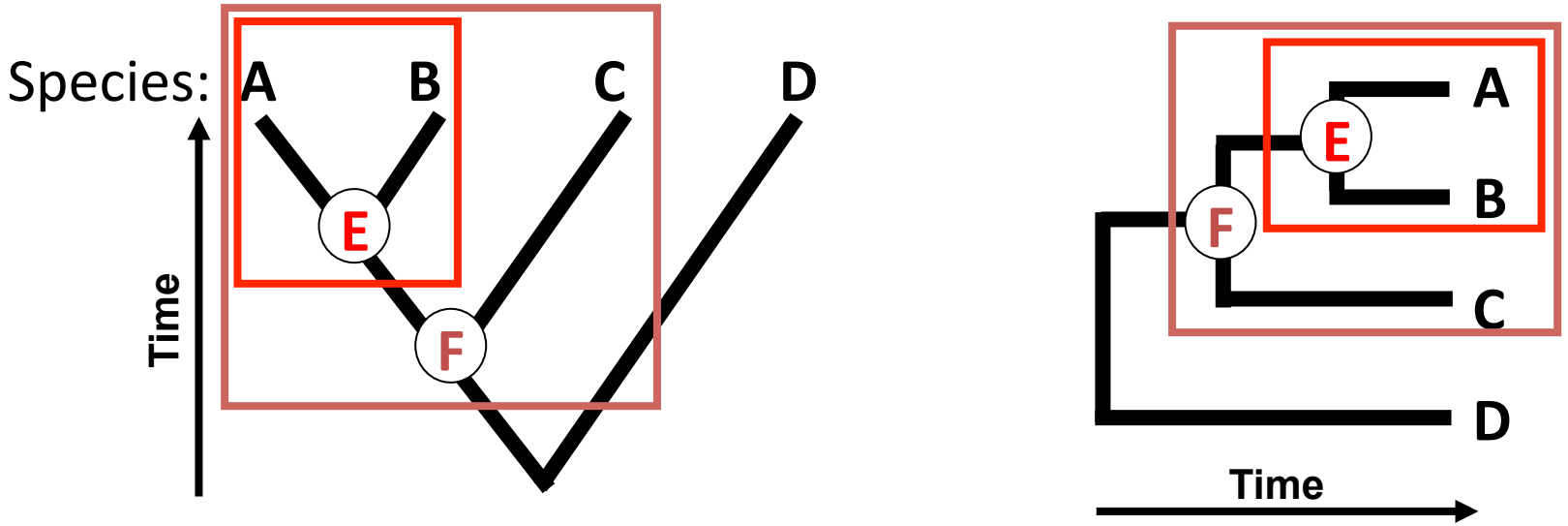


"The time will come I believe, though I shall not live to see it, when we shall have fairly true genealogical **[phylogenetic]** *trees of each great kingdom of nature."*

Phylogeny

What is a phylogenetic tree?

Branching diagram showing relationships between species (or higher taxa) based on their shared common ancestors



A and B are most closely related because they share a common ancestor (call the ancestor “E”) that C and D do not share

A+B+C are more closely related to each other than to D because they share a common ancestor (“F”) that D does not share

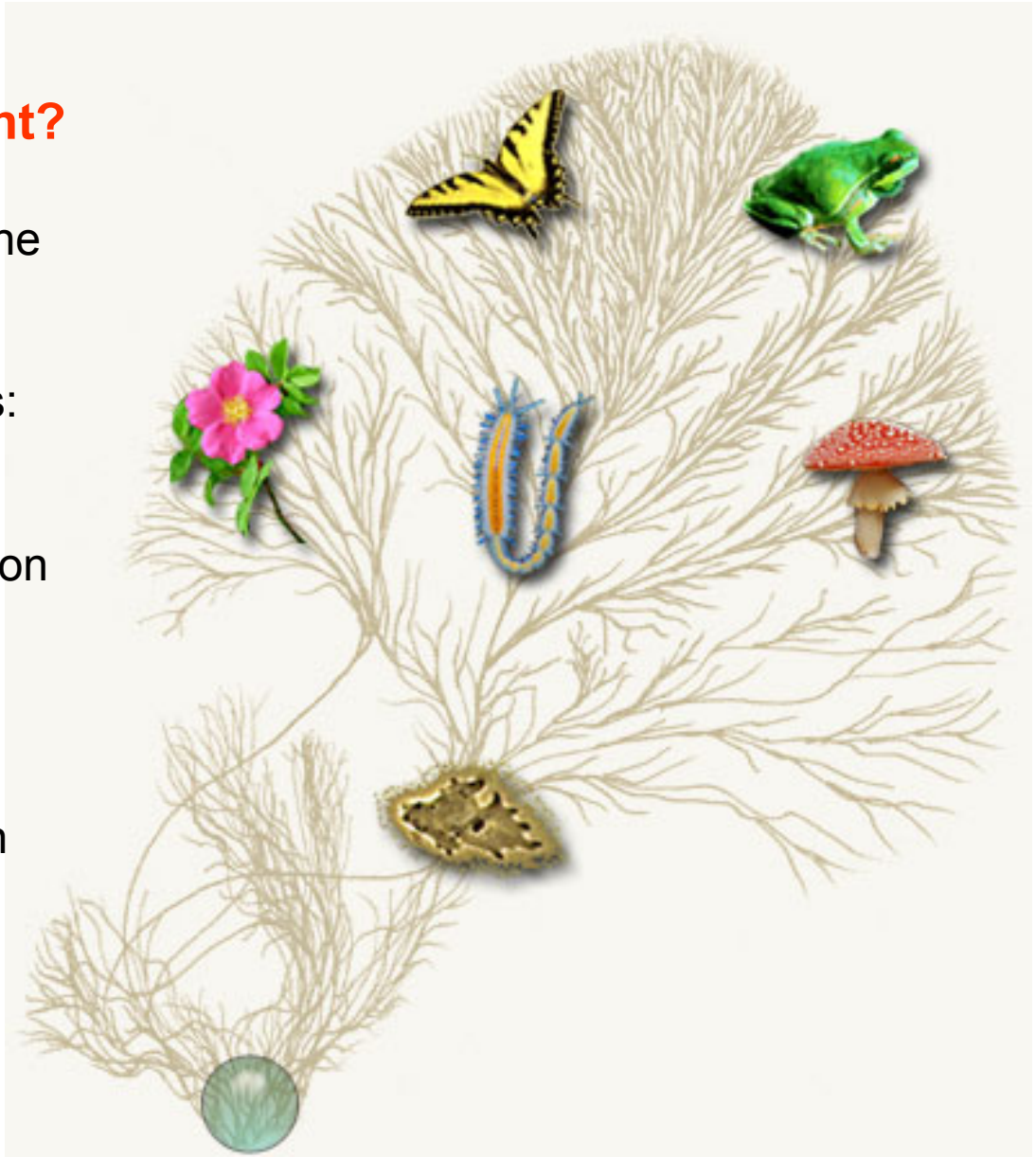
Tree of life

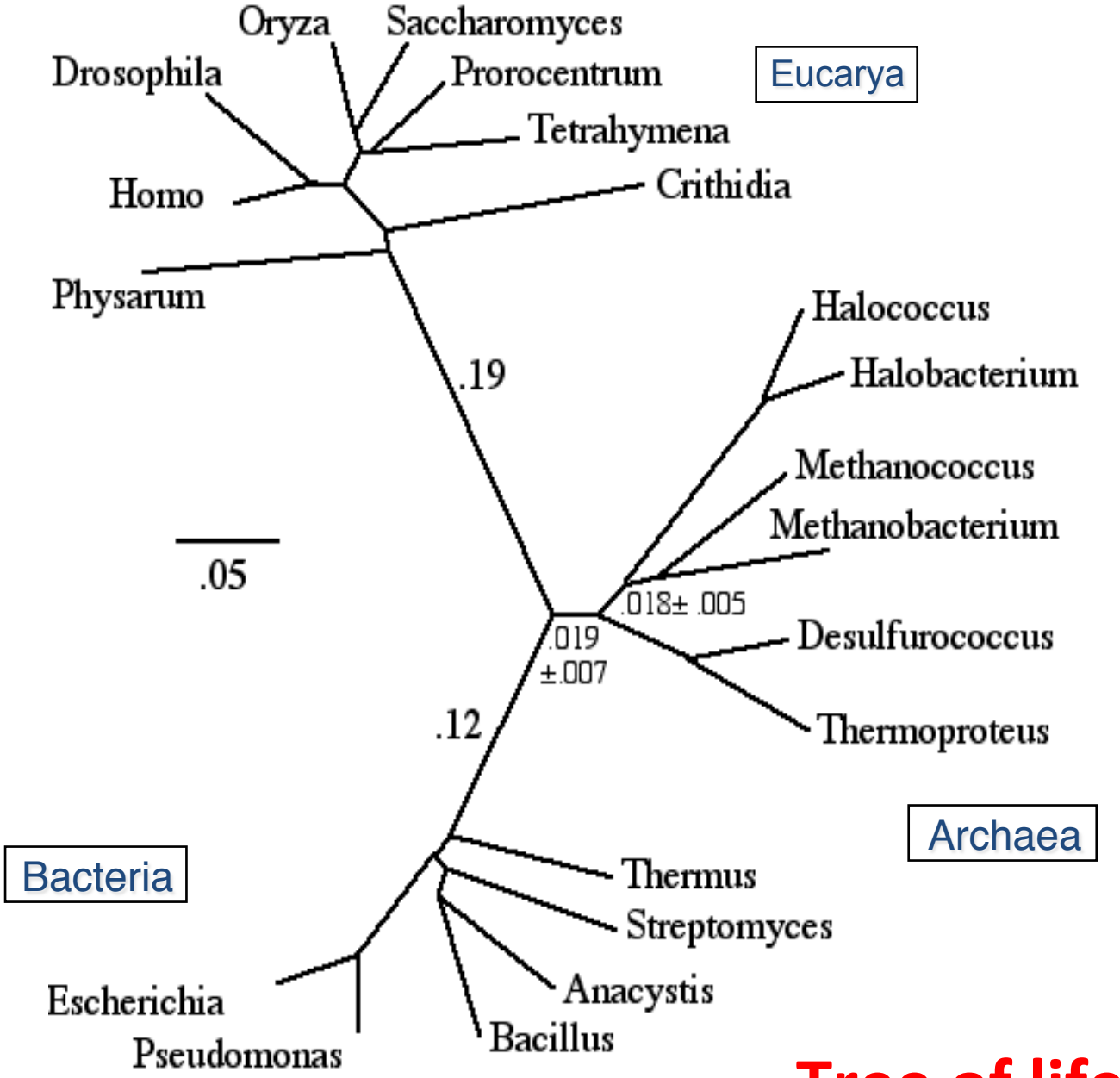
Why is phylogeny important?

Understanding and classifying the diversity of life on Earth

Testing evolutionary hypotheses:
- trait evolution
- coevolution
- mode and pattern of speciation
- correlated trait evolution
- biogeography
- geographic origins
- age of different taxa
- nature of molecular evolution
- disease epidemiology

…and many more applications!





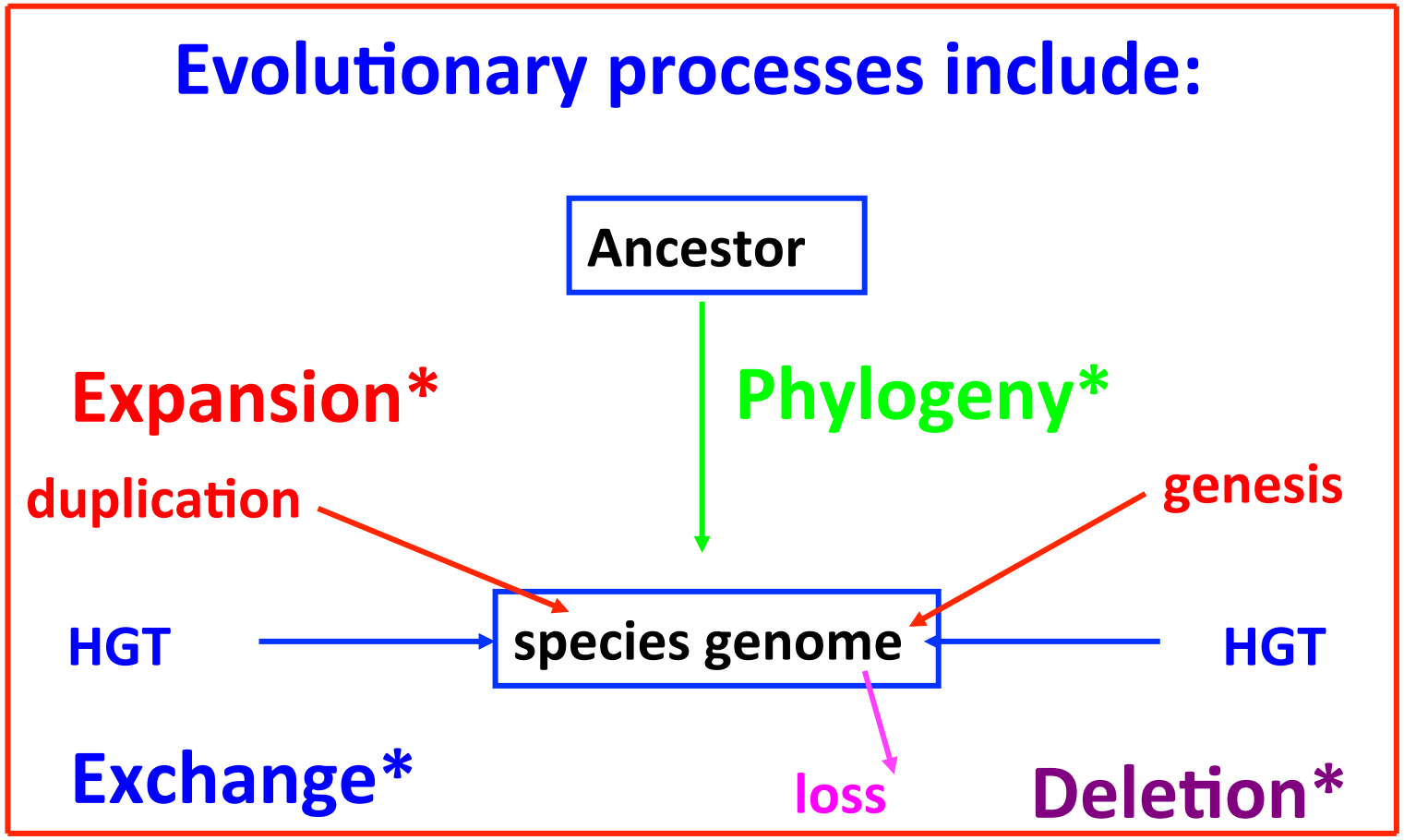
Universal phylogeny

deduced from comparison of SSU and LSU rRNA sequences (2508 homologous sites) using Kimura's 2-parameter distance and the NJ method.

The absence of root in this tree is expressed using a circular design.

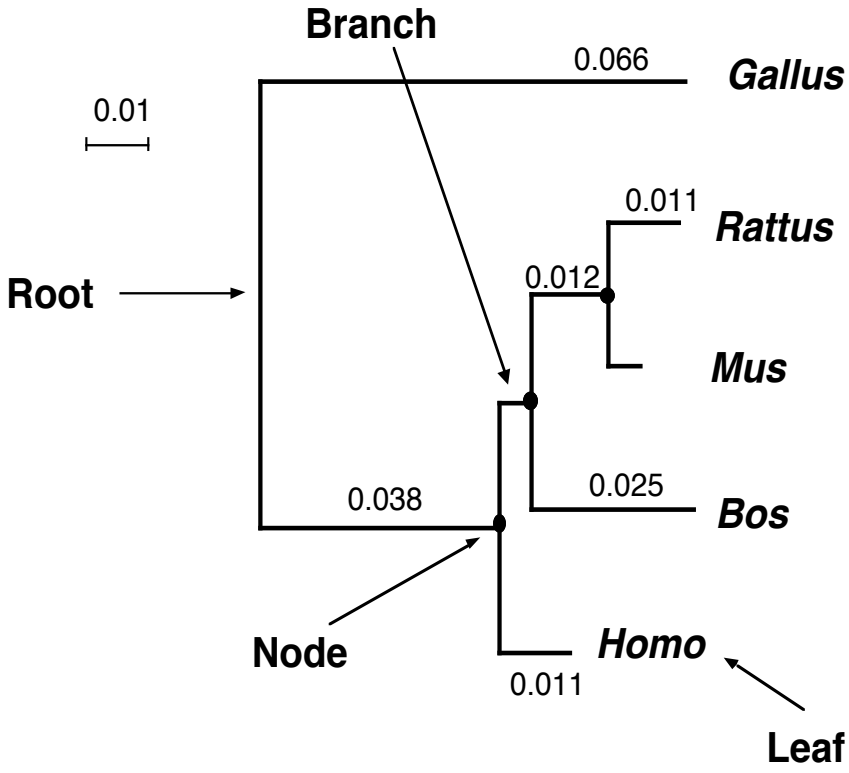
Tree of life

Evolutionary Forces in modifying genetic information



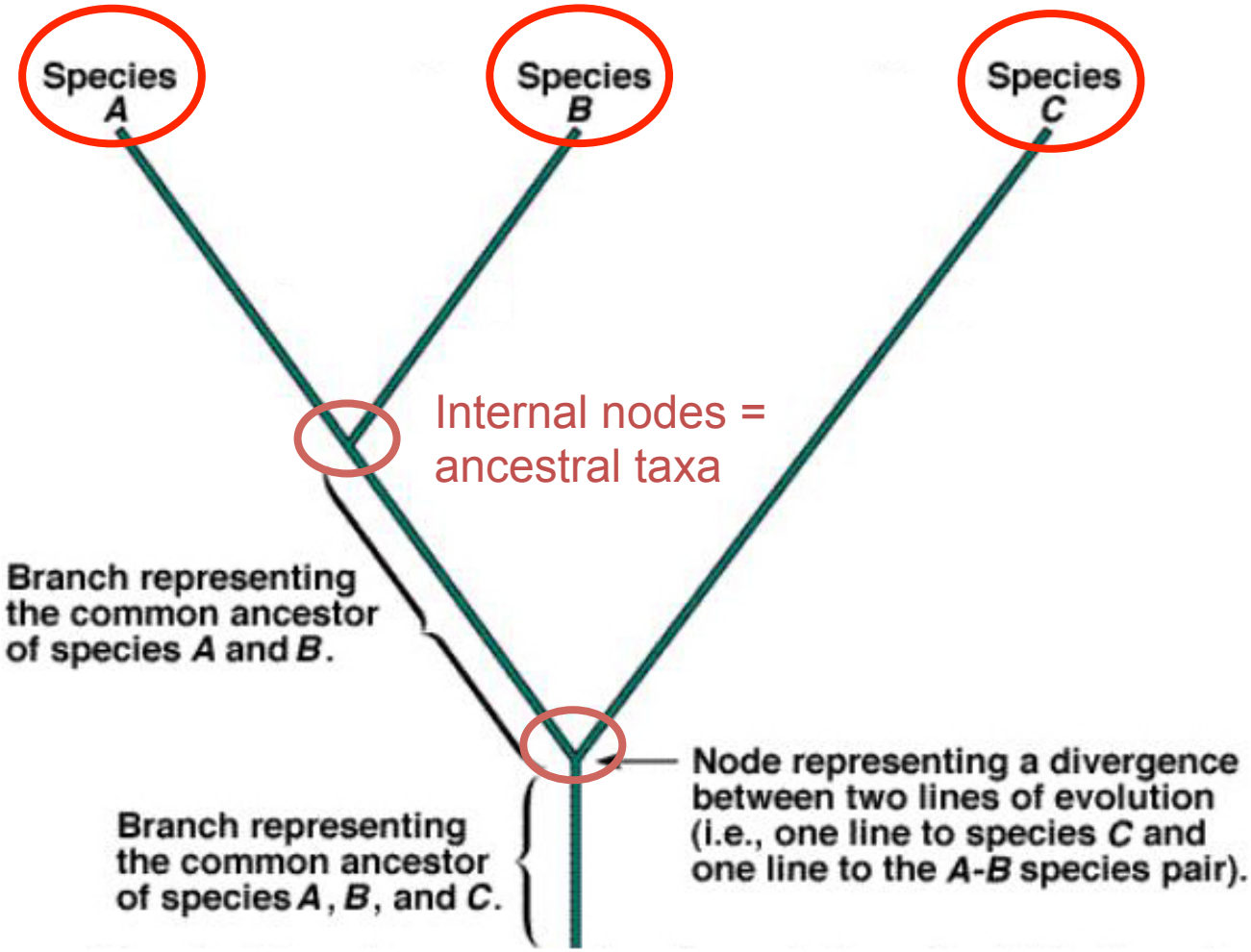
Tree Terminologies

- ✓ Internal branch : between 2 nodes.
- ✓ External branch : between a node and a leaf
- ✓ Horizontal branch length is proportional to evolutionary distances between sequences and their ancestors (unit = substitution / site).
- ✓ Tree Topology = shape of tree = branching order between nodes



Phylogeny

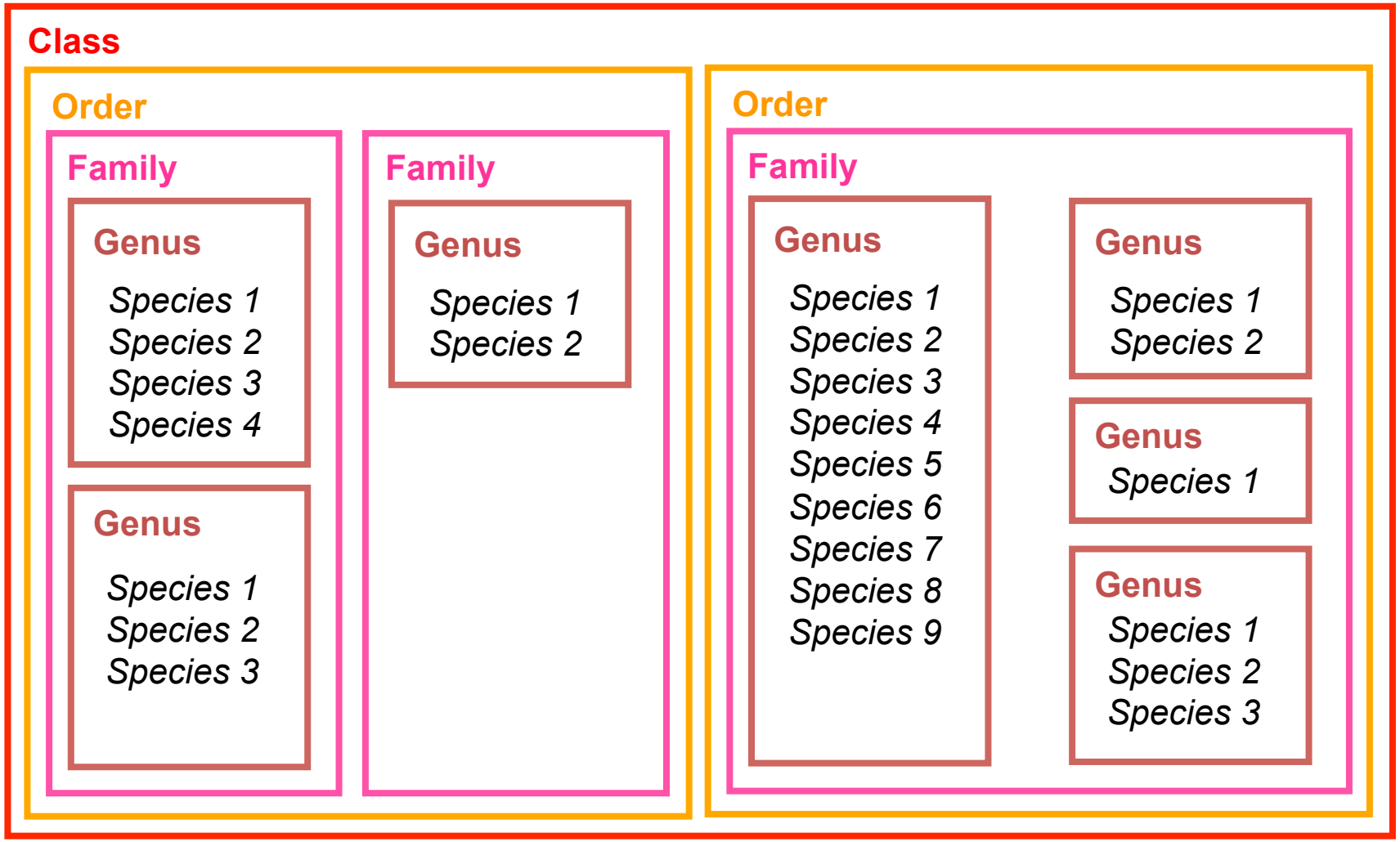
Terminal nodes = contemporary taxa



Phylogeny and classification

Hierarchy

All taxonomic classifications are hierarchical – how does phylogeny differ?
NB: taxa are nested on the basis of shared common ancestors

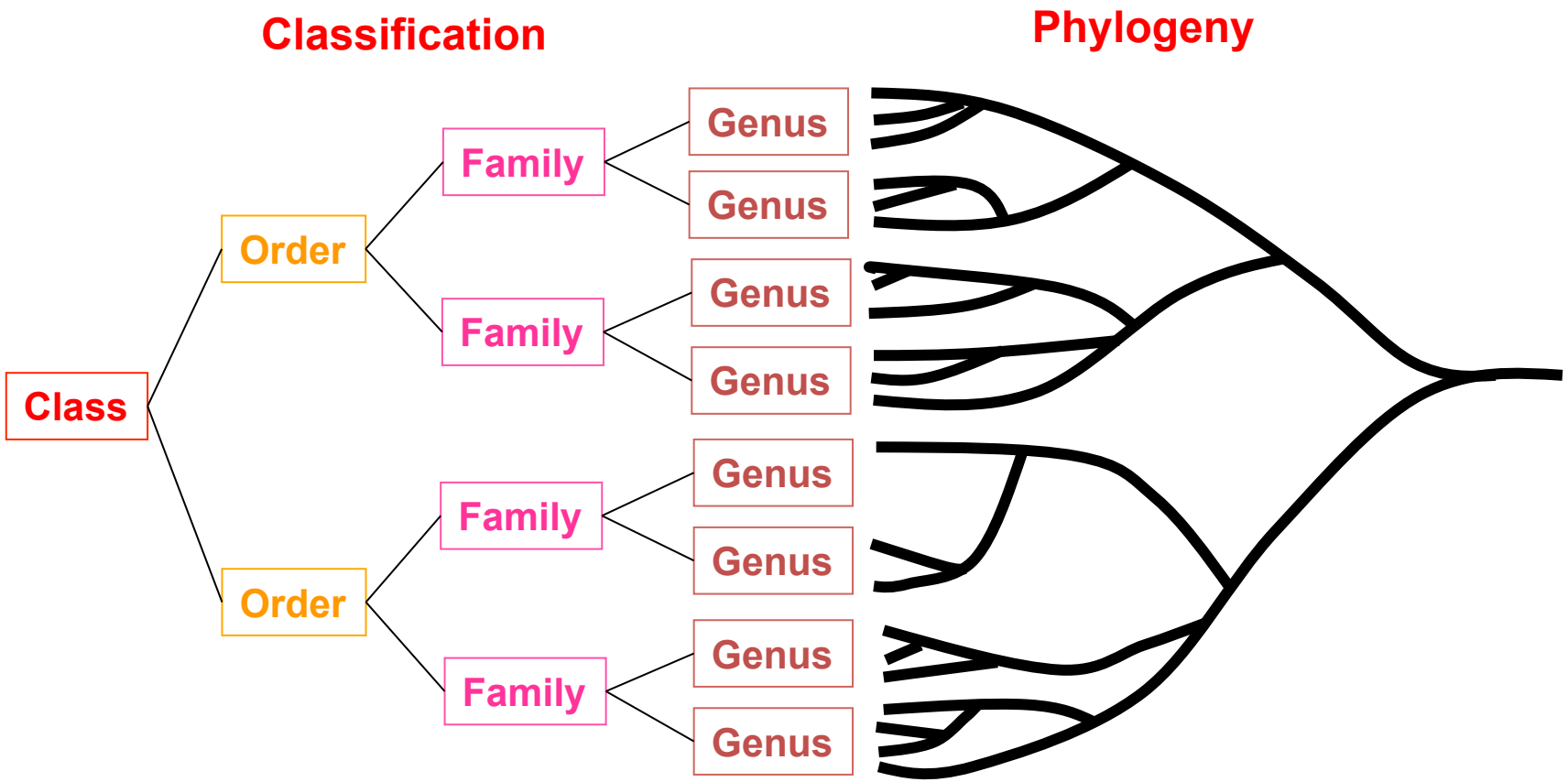


Phylogeny and classification

Hierarchy

Phylogenetic (cladistic) classification reflects evolutionary history

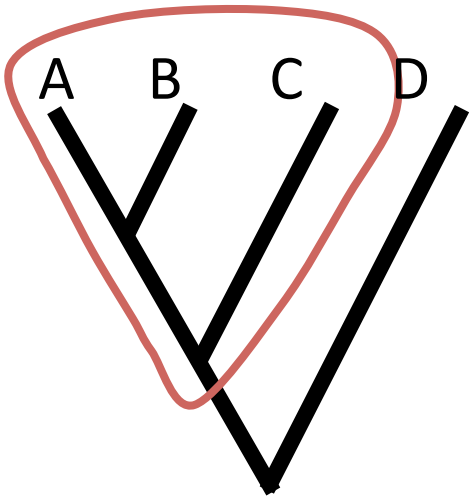
The only objective form of classification – organisms share a true evolutionary history regardless of our arbitrary decisions of how to classify them



Phylogeny and classification

Monophyletic group

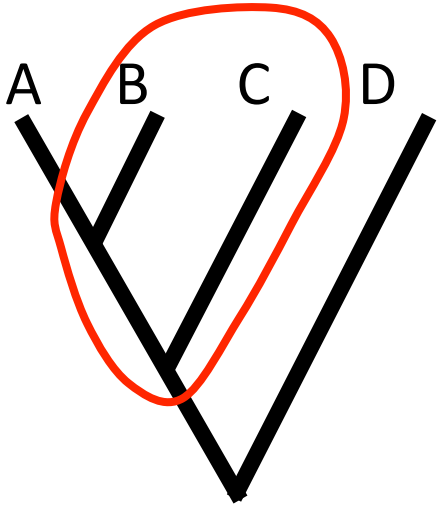
Includes an ancestor all of its descendants



How could this happen?

Paraphyletic group

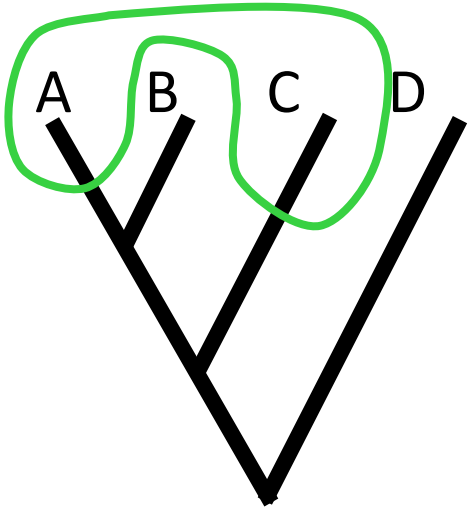
Includes ancestor and some, but not all of its descendants



Taxon A is highly derived and looks very different from B, C, and ancestor

Polyphyletic group

Includes two convergent descendants but not their common ancestor



Taxon A and C share similar traits through convergent evolution

Only monophyletic groups (**clades**) are recognized in cladistic classification

Phylogeny and classification

Monophyly

Each of the colored lineages in this echinoderm phylogeny is a good monophyletic group

Asteroidea

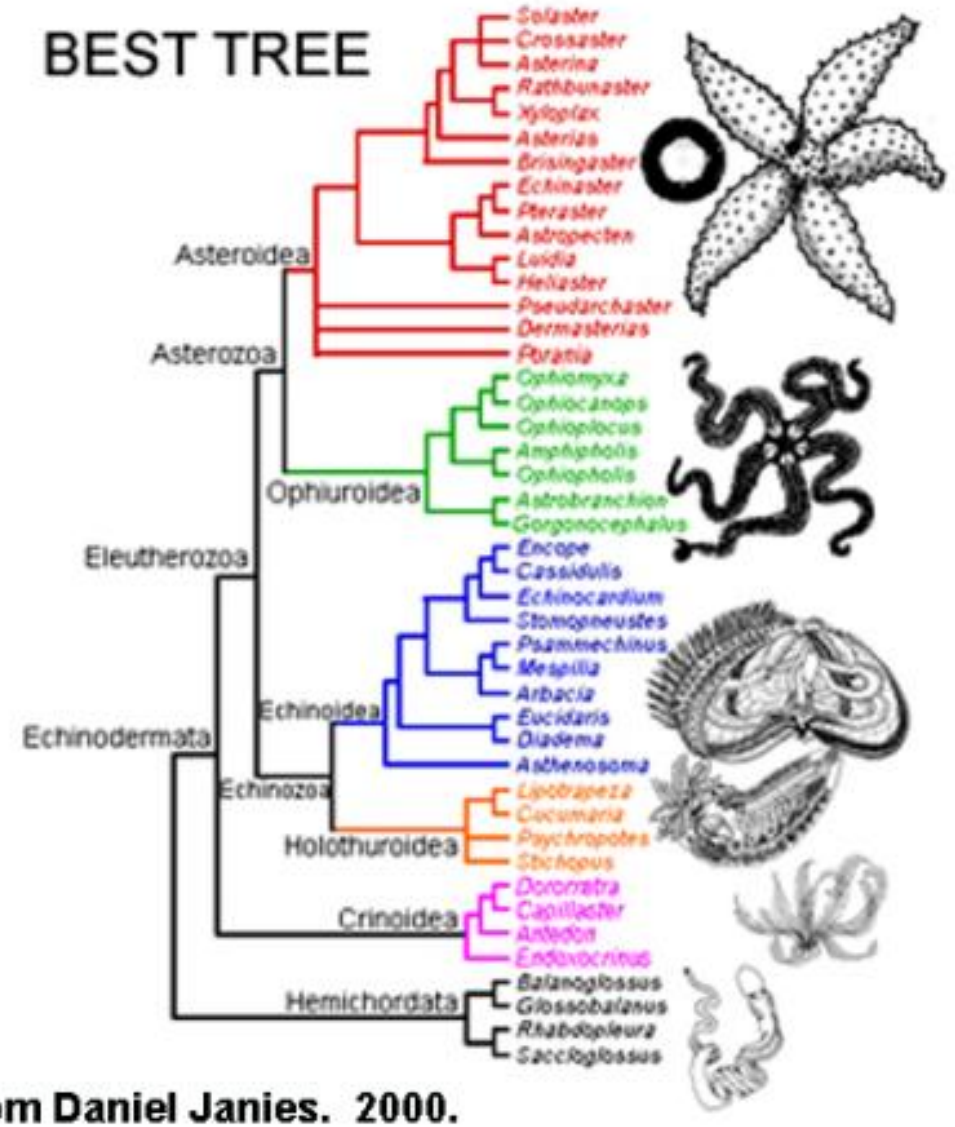
Ophiuroidea

Echinoidea

Holothuroidea

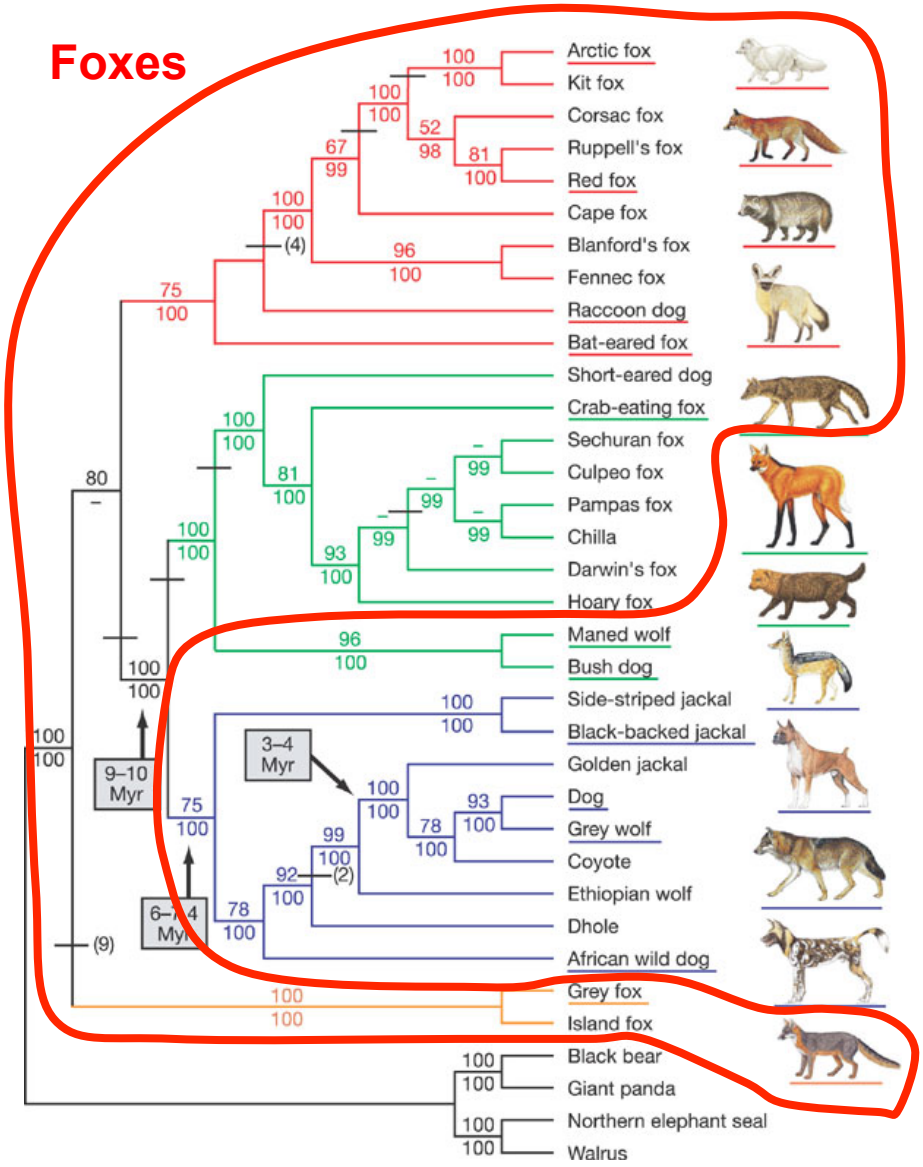
Crinoidea

Each group shares a common ancestor that is not shared by any members of another group



From Daniel Janies. 2000.

Paraphyletic groups



Lindblad-Toh et al. (2005) *Nature* 438: 803-819

Paraphyly

"Foxes" are paraphyletic with respect to dogs, wolves, jackals, coyotes, etc.

This is a trivial example because "fox" and "dog" are not formal taxonomic units, but it does show that a dog or a wolf is just a derived fox in the phylogenetic sense

Testing evolutionary hypotheses

Mapping evolutionary transitions

Some horned lizards squirt blood from their eyes when attacked by canids

How many times has blood-squirting evolved?

Blood squirting? **No** **Yes**

P. modestum

P. platyrhinos

P. mcallii

P. solare

P. cornutum

P. coranatum

P. hernandesi

P. douglasi

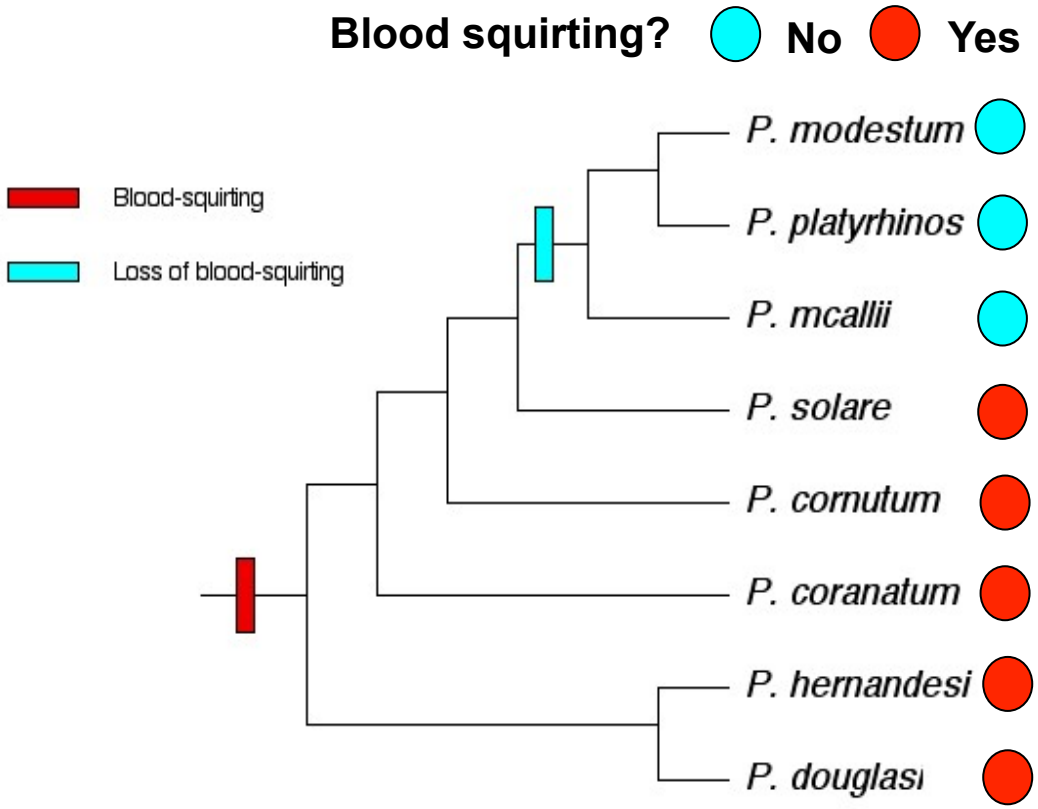
Testing evolutionary hypotheses

Mapping evolutionary transitions

Some horned lizards squirt blood from their eyes when attacked by canids

How many times has blood-squirting evolved?

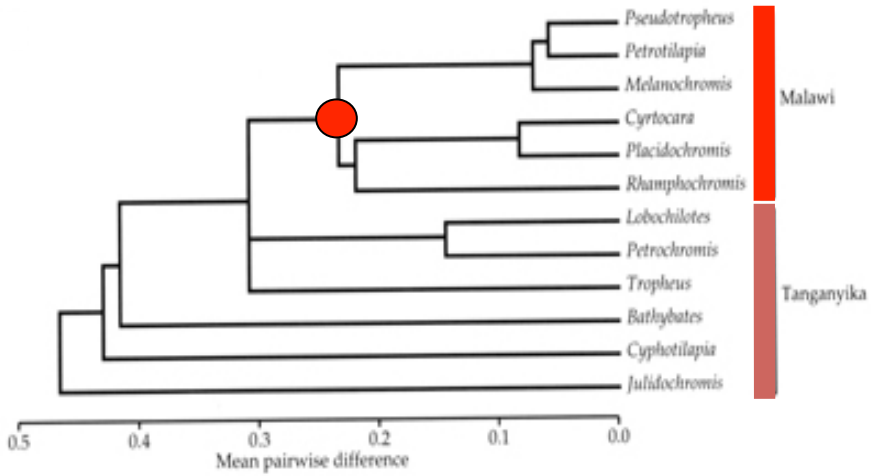
This phylogeny suggests a single evolutioary gain and a single loss of blood squirting

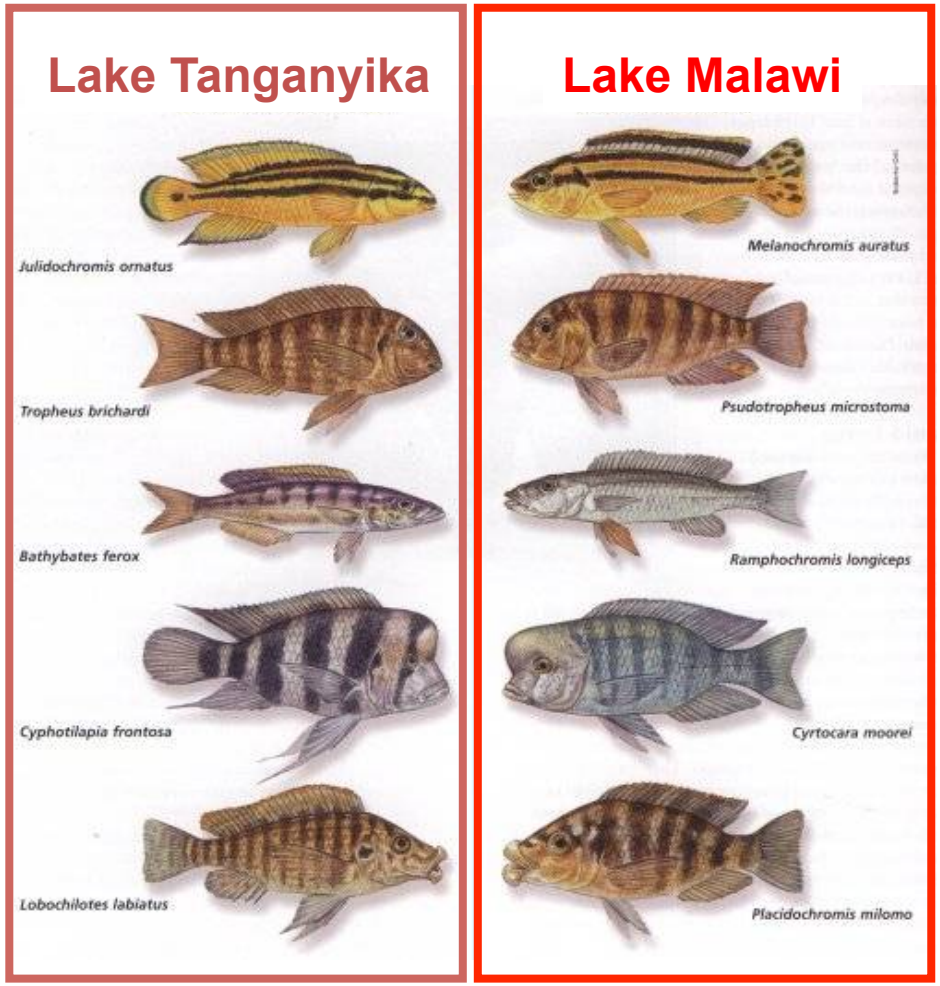


Testing evolutionary hypotheses

Convergence and modes of speciation

What can this phylogeny tell us about homology/analogy and speciation?

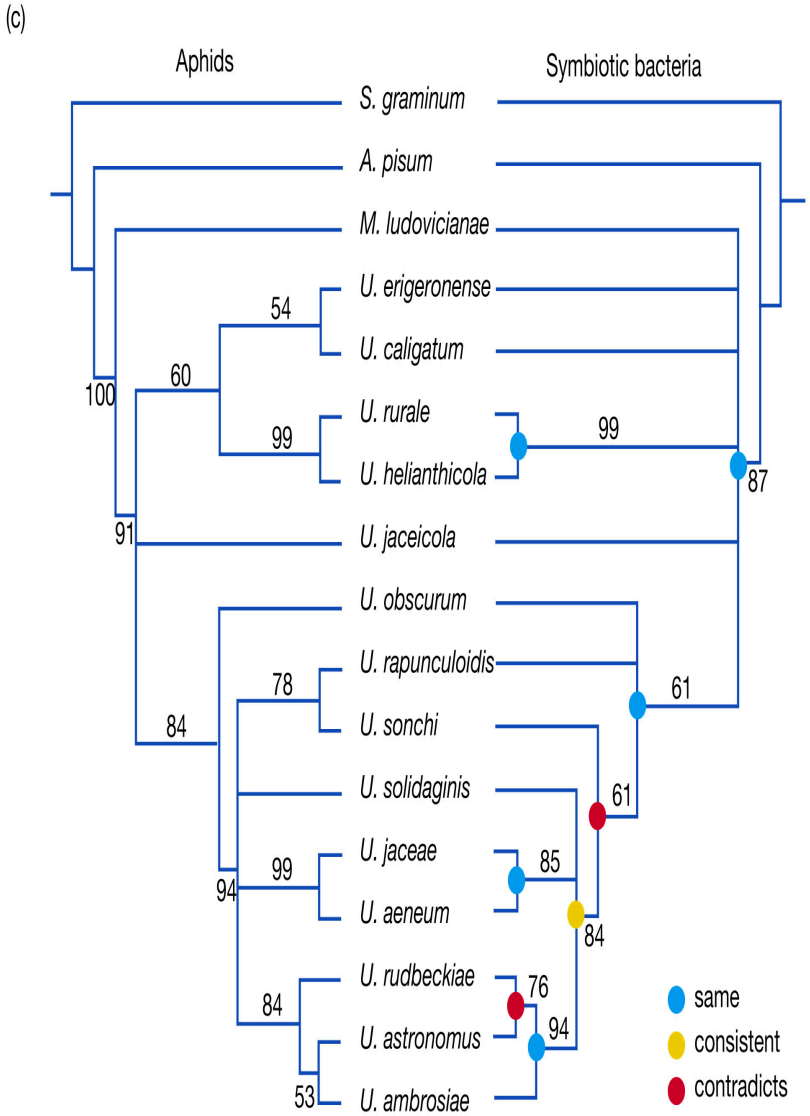




1. Similarities between each pair are the result of **convergence**

2. **Sympatric speciation** more likely than allopatric speciation

Testing evolutionary hypotheses



Clark et al. (2000)

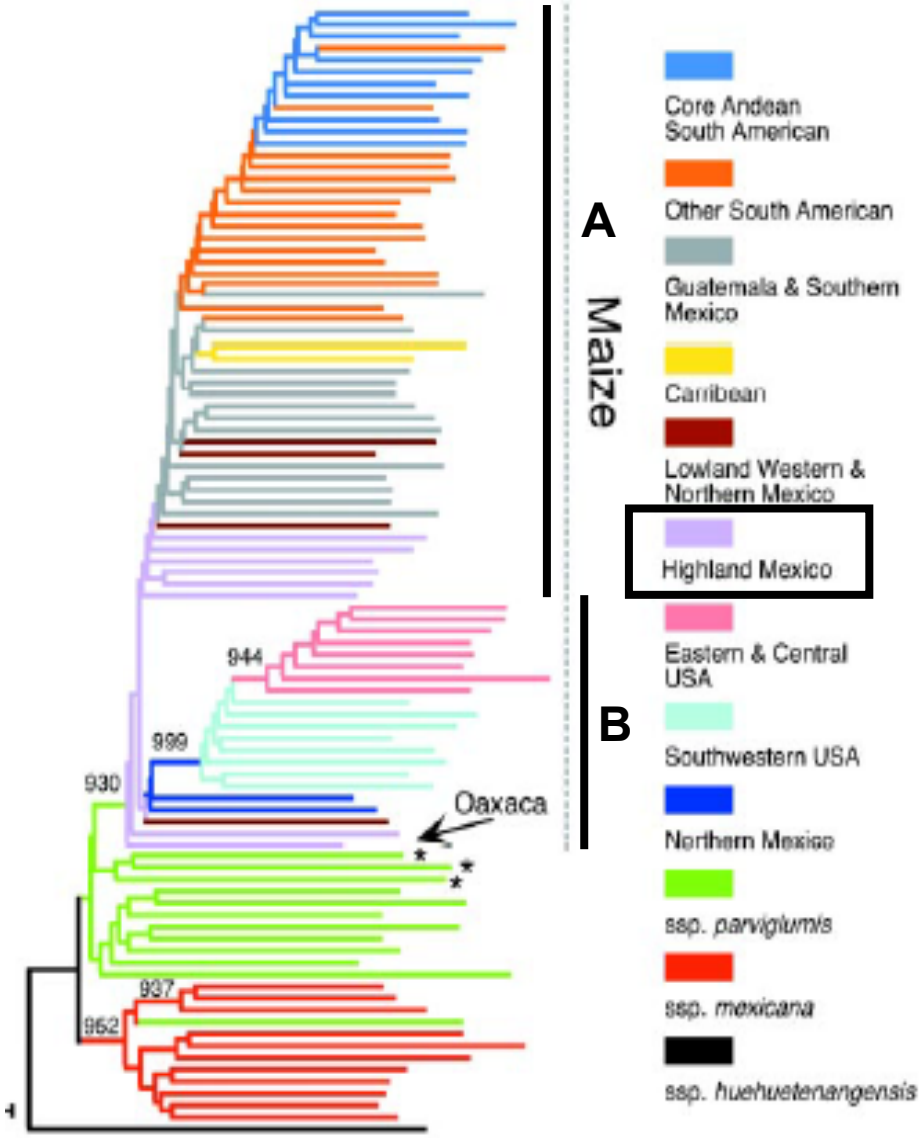
Coevolution

Aphids and bacteria are symbiotic

Given this close relationship, we might expect that speciation in an aphid would cause parallel speciation in the bacteria

When comparing phylogenies for each group we see evidence for **reciprocal cladogenesis** (but also contradictions)

Testing evolutionary hypotheses



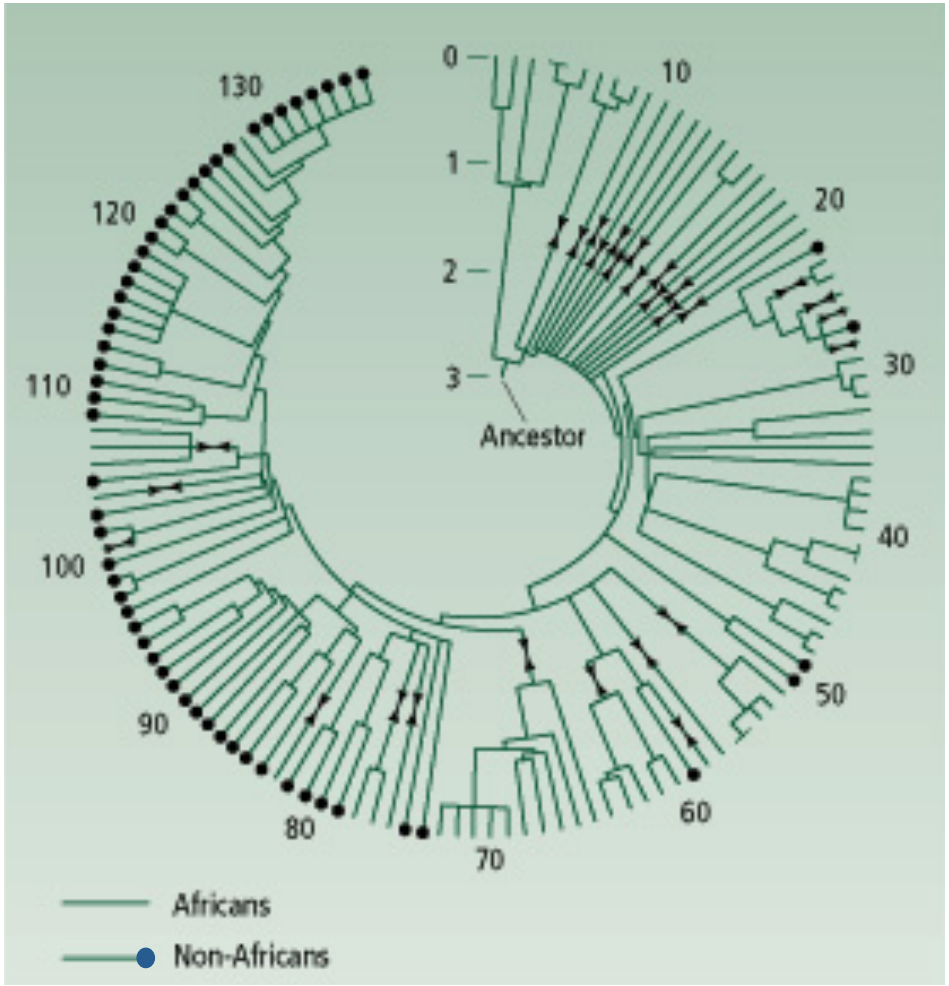
Matsuoka et al. (2002)

Geographic origins

Where did domestic corn (*Zea mays maize*) originate?

Populations from **Highland Mexico** are at the base of each maize clade

Testing evolutionary hypotheses



Vigilant et al. (1991) *Science*

Geographic origins

Where did humans originate?

Each tip is one of 135 different mitochondrial DNA types found among 189 individual humans

African mtDNA types are clearly basal on the tree, with the non-African types derived

Suggests that humans originated in Africa

Molecular phylogenetics

Study of evolutionary relationships between genes and species

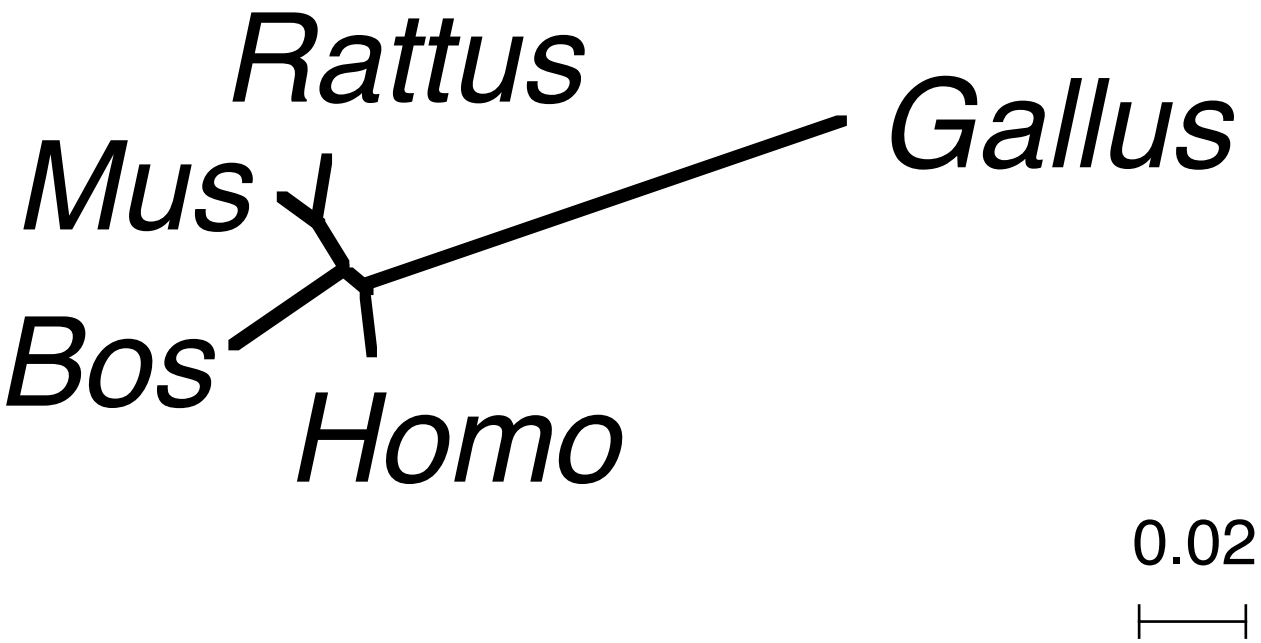
- The actual pattern of evolutionary history is the phylogeny or evolutionary/phylogenetic tree which we try to estimate.

- **A tree is a mathematical structure which is used to model the actual evolutionary history of a group of sequences or organisms.**

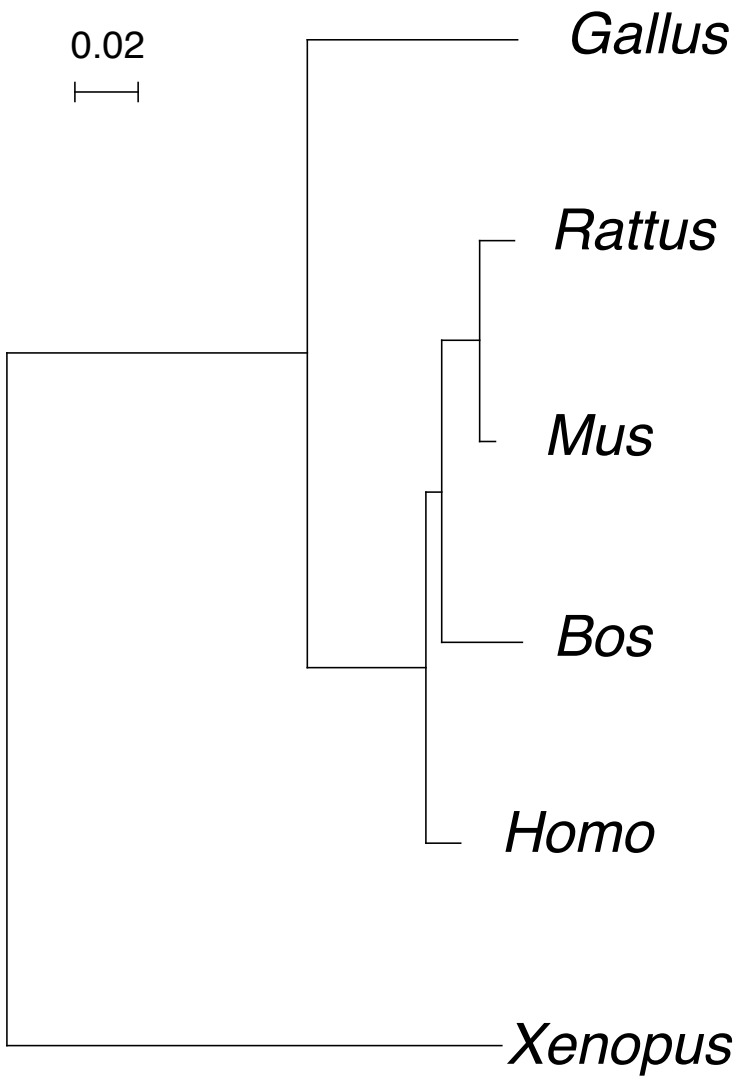
Rooted and Unrooted Trees

- Most phylogenetic methods produce unrooted trees. This is because they detect differences between sequences, but have no means to orient residue changes relatively to time.
- Two means to root an unrooted tree :
  - The outgroup method : include in the analysis a group of sequences known *a priori* to be external to the group under study; the root is by necessity on the branch joining the outgroup to other sequences.
  - Make the molecular clock hypothesis : all lineages are supposed to have evolved with the same speed since divergence from their common ancestor. The root is at the equidistant point from all tree leaves.

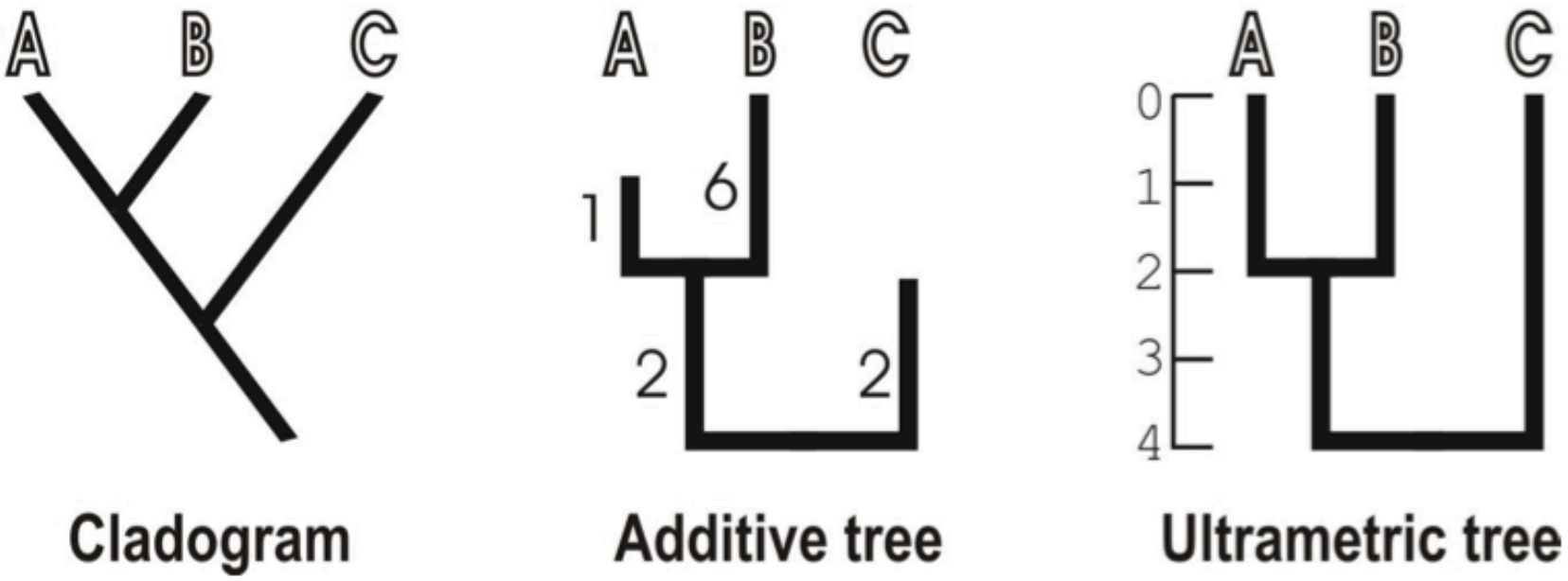
Unrooted Tree



Rooted Tree



Cladograms, ultrametric and additive trees



Cladogram Additive tree Ultrametric tree

a) **CLADOGRAM**: tree shows the relative recency of common ancestry;
Do not tell us anything about the evolutionary distances between taxa.
b) **ADDITIVE TREES**: Branch lengths in these trees are related to attributes such as number of nucleotide change per site.
c) **ULTRAMETRIC TREES:** have terminal nodes that are the same distance from the root

Types of phylogenetic analysis methods

- Phenetic: trees are constructed based on observed characteristics, not on evolutionary history
- Cladistic: trees are constructed based on fitting observed characteristics to some model of evolutionary history

Distance methods

Parsimony and Maximum Likelihood methods

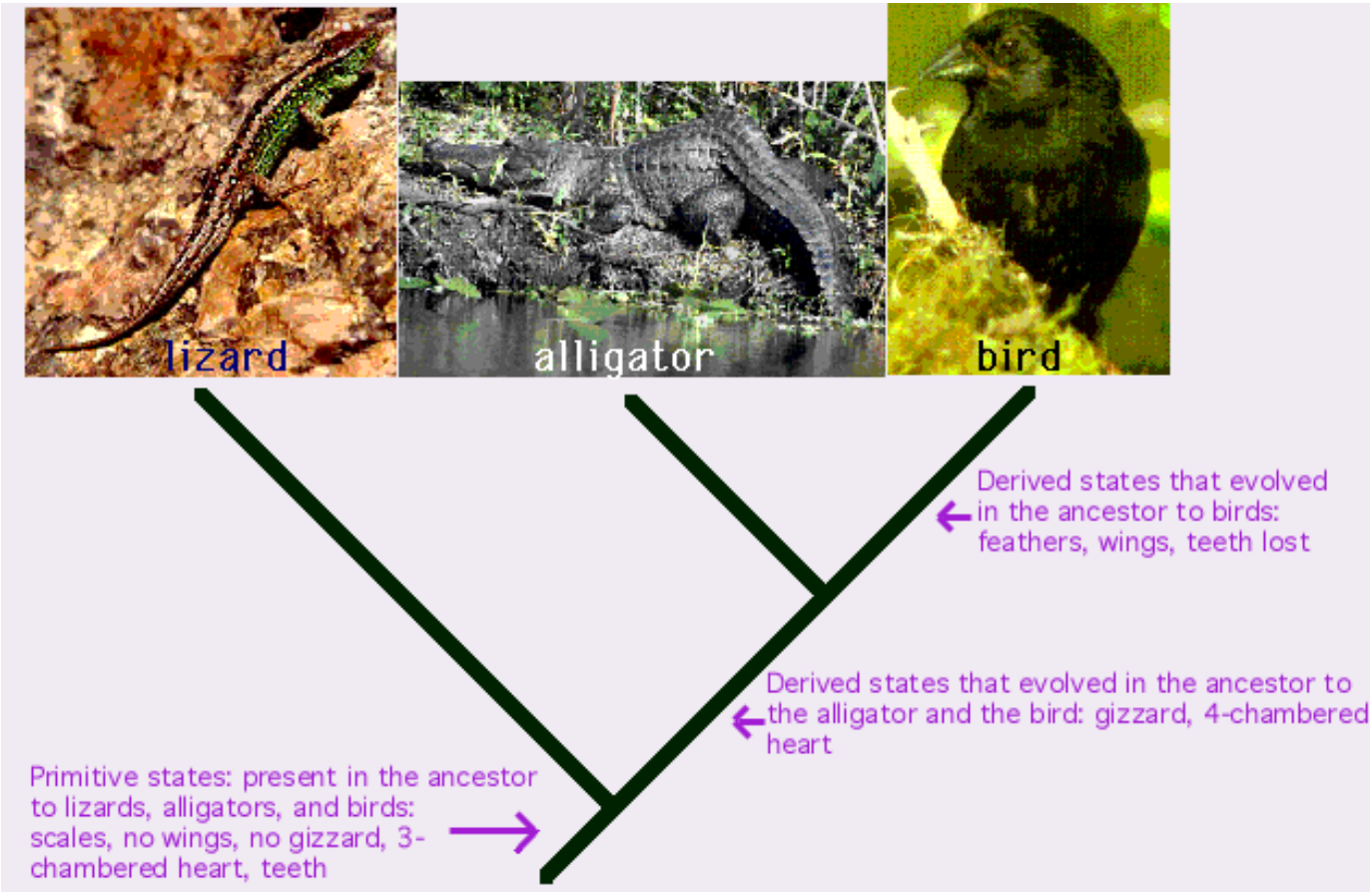
Phenetics

- Study of phylogenetics based on shared traits

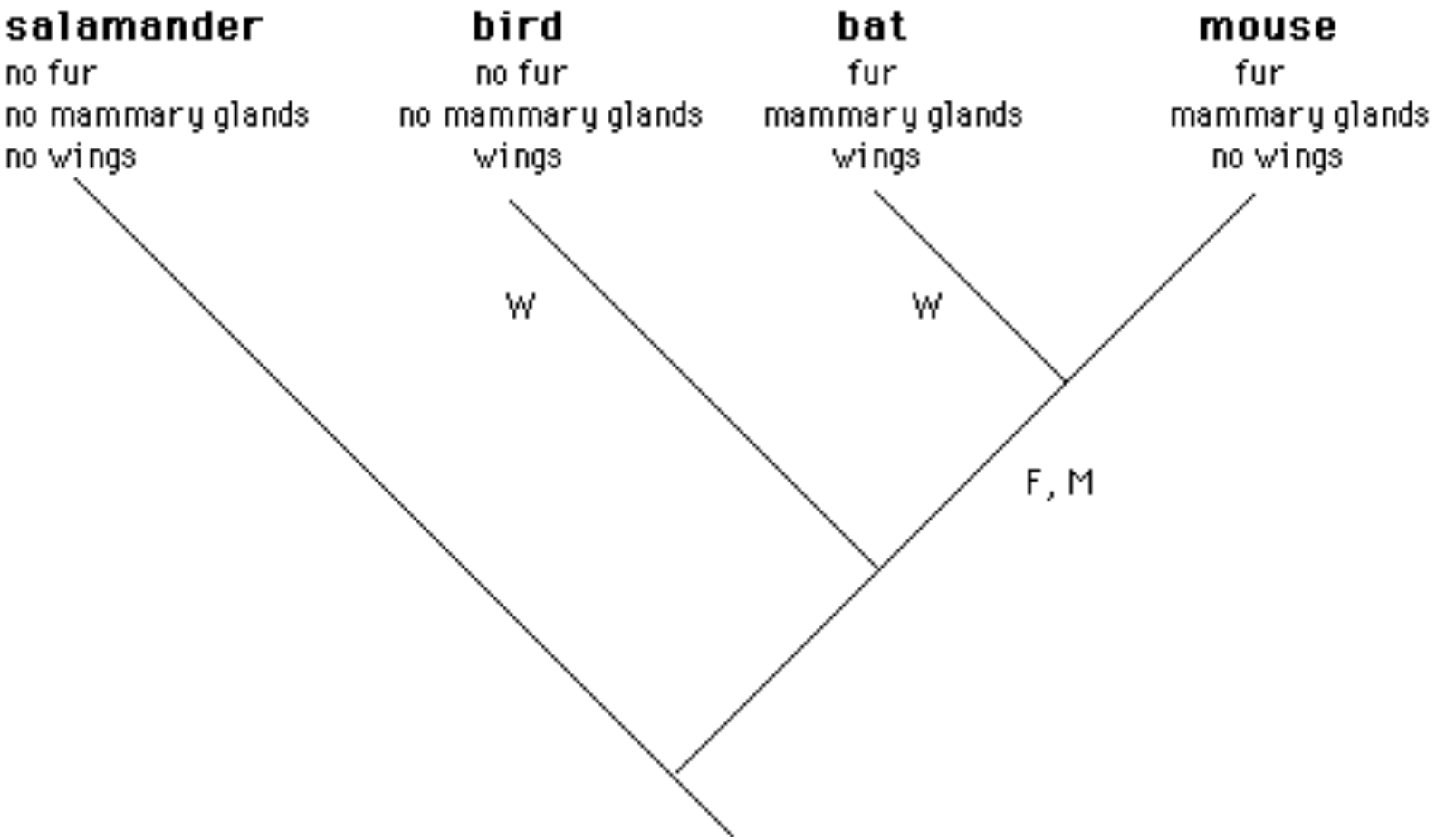
- Problem: not all shared traits are informative about evolutionary relationships

Cladistics

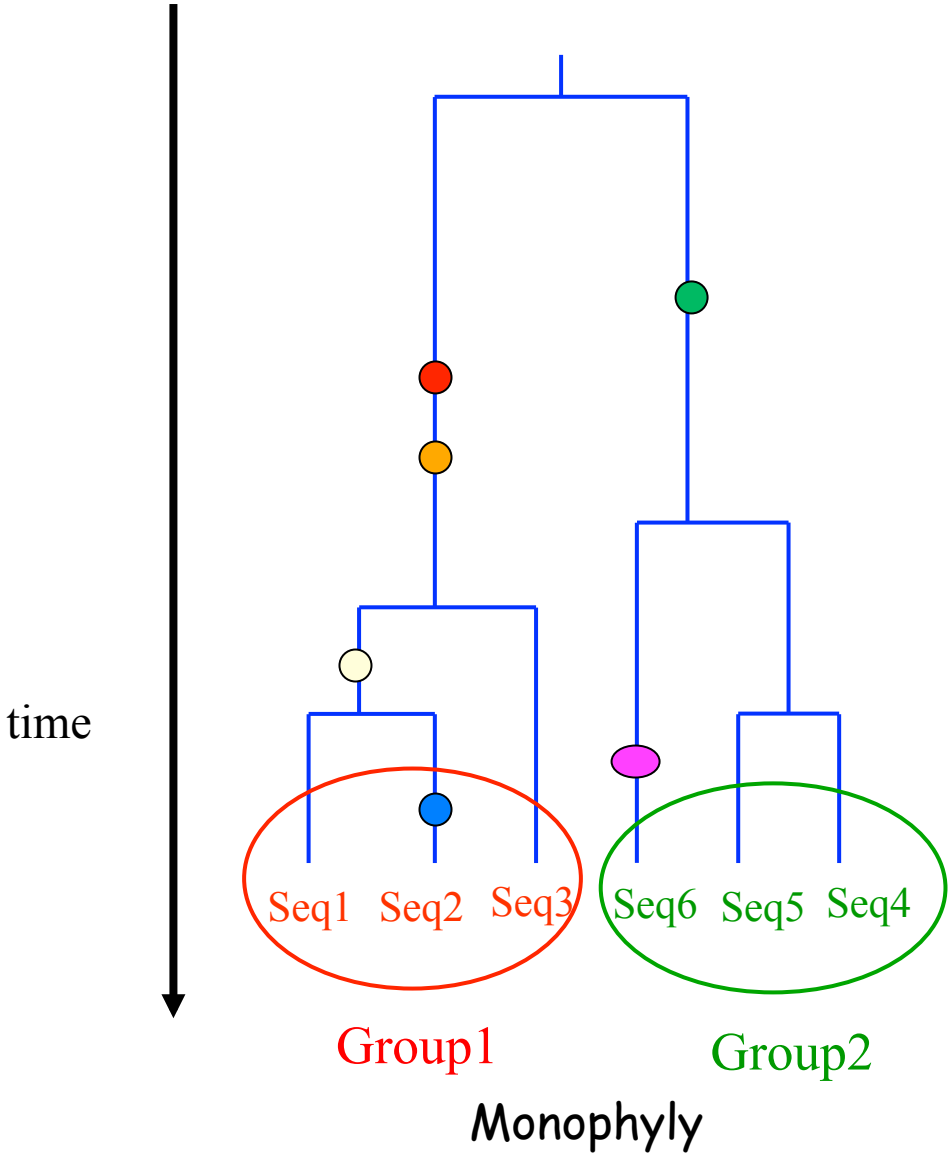
- Study of phylogenetic relationships based on shared **derived** traits

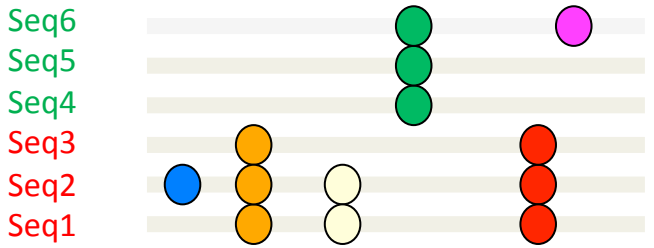


Caveat: homoplasy - independent evolution of the same character



Sequences Reflect Relationships





… =derived mutations

What Sequences to Study?

- Study more than just one region!
- Different sequences evolve at different rates
  - Proteins
  - Highly variable sequences (ex: immunoglobulin genes)
  - Highly conserved (actin, rRNA coding regions)
  - Different regions within a single gene can evolve at different rates (conserved vs. variable domains)

Molecular Phylogenies

- The gene compared must evolve at a rate comparable to the divergence time of the organism; for example:
  - 18S rRNA gene for phylum-level divergences since it evolves very slowly.
  - Hemoglobin genes for mammalian orders.
  - Mitochondrial DNA for species divergences within a genus.
  - Repetitive DNA sequences (e.g. microsatellites) for individuals within species.

DNA is a good tool for taxonomy

DNA sequences have advantages over taxonomic characters:

- Unambiguous
- Large numbers of characters can be scored for each individual

DNA is a good tool for taxonomy

```
A    aat tcg ctt cta gga atc tgc cta atc ctg
B    ... ..a ..g ..a .t. ... ... t.. ... ..a
C    ... ..a ..c ..c ... ..t ... ... ... t.a
D    ... ..a ..a ..g ..g ..t ... t.t ..t t..
```

Each nucleotide difference is a character

Molecular Phylogeny

- Starting point: a set of homologous, aligned DNA or protein sequences
- Result of the process: a tree describing evolutionary relationships between studied sequences
  = a genealogy of sequences
  = a phylogenetic tree

```
CLUSTAL W (1.74) multiple sequence alignment

Xenopus         ATGCATGGGCCAACATGACCAGGAGTTGGTGTCGGTCCAAACAGCGTT---GGCTCTCTA
Gallus          ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC---CAAAATAACACCAACATGCAAATG
Bos             ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACCCAAAACAGCACCAACGTGCAAATG
Homo            ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACTCAAAACAGCACCAACGTGCAAATG
Mus             ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACTCAAAACAGCACCAACGTGCAAATG
Rattus          ATGCATCCGCCACCATGACCAGCGGGAGGTAGCTCTCAAAACAGCACCAACGTGCAAATG
                ******  **** *********  *  ***  *   * *** * *            *
```

Alignment and Gaps

- The quality of the alignment is essential : each column of the alignment (site) is supposed to contain homologous residues (nucleotides, amino acids) that derive from a common ancestor.

  ==> Unreliable parts of the alignment must be omitted from further phylogenetic analysis.

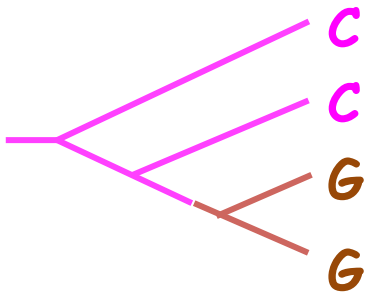
- Most methods take into account only substitutions ; gaps (insertion/deletion events) are not used.

  ==> gaps-containing sites are ignored.

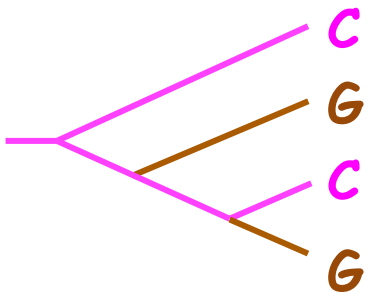
```
Xenopus     ATGCATGGGCCAACATGACCAGGAGTTGGTGTCggtCCAAACAGCGTT---GGCTCTCTA
Gallus      ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC---CAAAATAACACCaacATGCAAATG
Bos         ATGCATCCGCCACCATGACCAGCAGGAGGTAGCagtCAAAACAGCACCaacGTGCAAATG
Homo        ATGCATCCGCCACCATGACCAGCAGGAGGTAGCagtCAAAACAGCACCaacGTGCAAATG
Mus         ATGCATCCGCCACCATGACCAGCAGGAGGTAGCactCAAAACAGCACCaacGTGCAAATG
Rattus      ATGCATCCGCCACCATGACCAGCGGGAGGTAGCtctCAAAACAGCACCaacGTGCAAATG
```

Caveat: homoplasy: independent evolution of the same character

Evolutionary relationship:
Shared ancestral characters
Shared derived characters

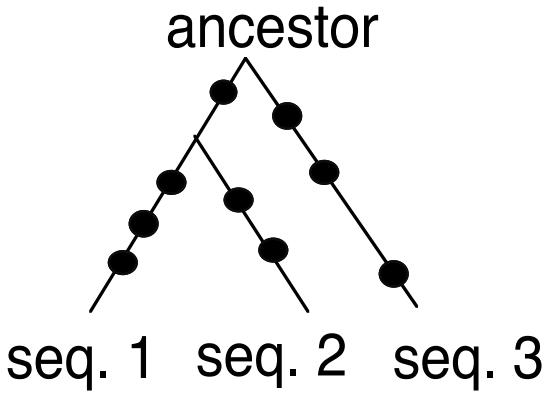


Homoplasy (independent evolution of the same character):



Caveat: Saturation i.e. loss of phylogenetic signal

- When compared homologous sequences have experienced too many residue substitutions since divergence, it is impossible to determine the phylogenetic tree, whatever the tree-building method used.



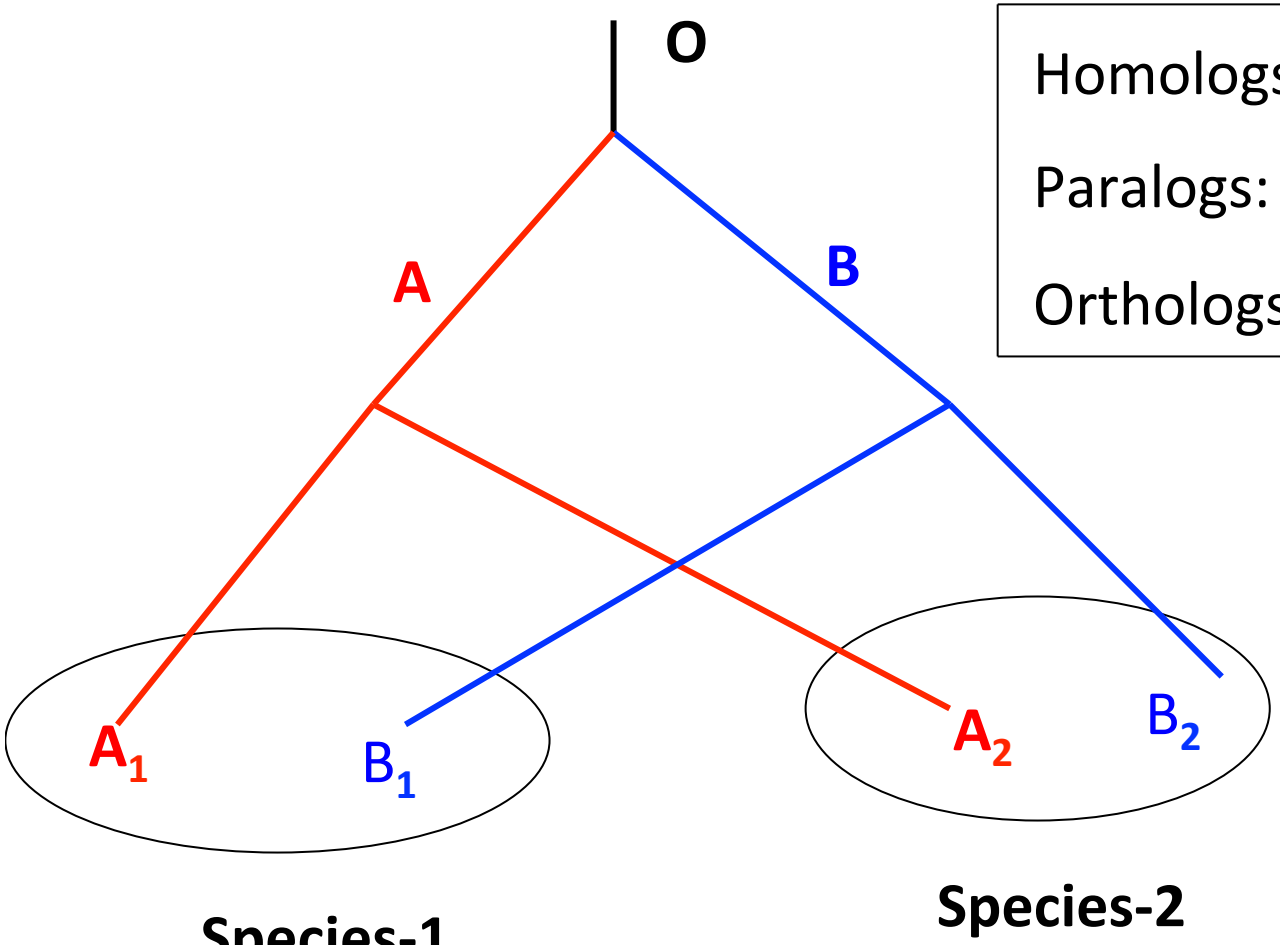
- NB: often saturation may not be detectable

Caveats: horizontal gene transfer

- Common in prokaryotes (bacteria) and viruses: an organism incorporates genetic material from another organism

- Exists in Eukaryotes: mitochondrial DNA or chloroplast DNA transferred to the nucleus

Homolog - Paralog - Ortholog

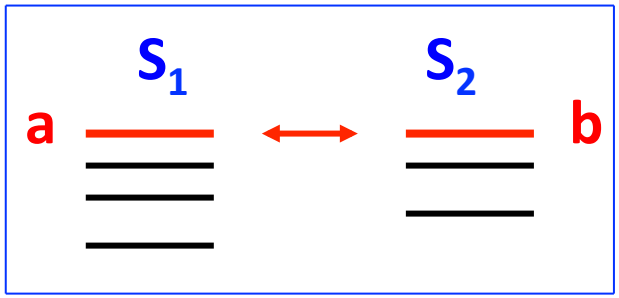


Homologs: A_1, B_1, A_2, B_2

Paralogs: A_1 vs B_1 and A_2 vs B_2

Orthologs: A_1 vs A_2 and B_1 vs B_2

Sequence analysis



Caveats: Orthologs vs. Paralogs

Tree Building Goals

- Maximize shared derived character states in a lineage
- Minimize homoplasies
  - Parallel changes, convergences, and reversals of character states between and within lineages

Tree building methods

- Distance methods
- Maximum Parsimony
- Maximum likelihood
- Bayesian inference

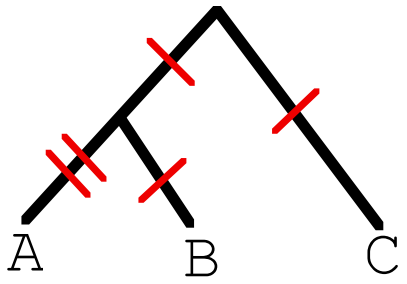
All methods can be used with different substitution models

Distance methods

- Starts from a multiple sequence alignment
- Makes a matrices of pairwise sequence distances (number of differences)
- Builds a phylogenetic tree

Correspondence between trees and distance matrices

- Any phylogenetic tree induces a matrix of distances between sequence pairs
- "Perfect" distance matrices correspond to a single phylogenetic tree



tree

	A	B	C
A	**0**		
B	**3**	**0**	
C	**4**	**3**	**0**

distance matrix

Building phylogenetic trees by distance methods

General principle :

Sequence alignment

↓ (1)

Matrix of evolutionary distances between sequence pairs

↓ (2)

(unrooted) tree

- (1) Measuring evolutionary distances.
- (2) Tree computation from a matrix of distance values.

Multiple sequence alignment

Species A ATGGCTATTCTTATAGTACG
Species B ATCTAGTCTTATATTACA

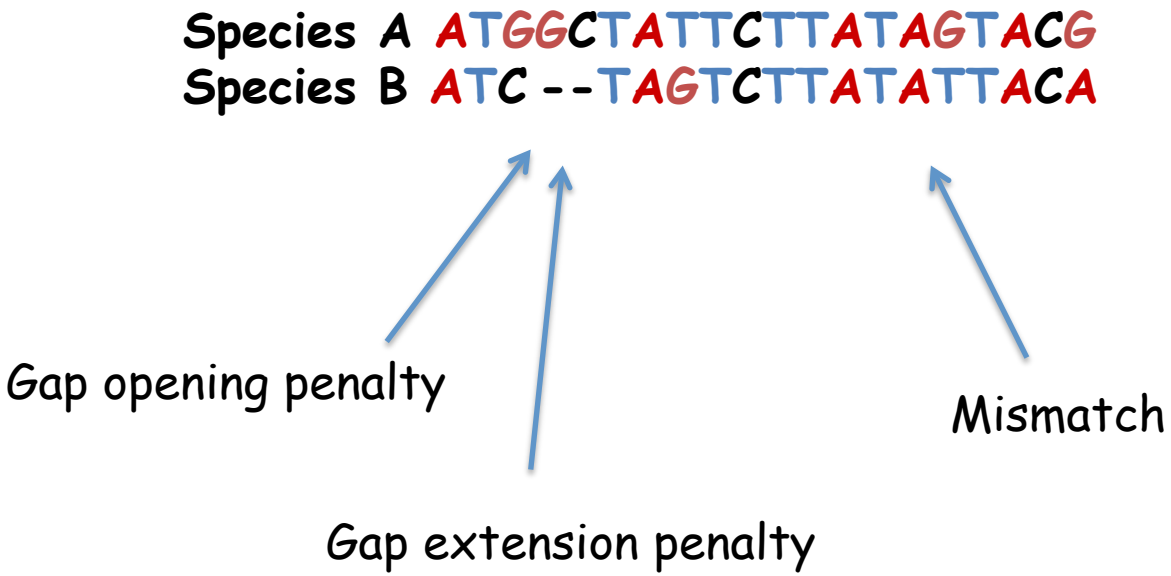
Aligned sequences

Species A ATGGCTATTCTTATAGTACG
Species B ATC--TAGTCTTATATTACA

Multiple sequence alignment

- Different softwares: ClustalW, ClustalX, Muscle

Minimize total score



Principle of distance methods

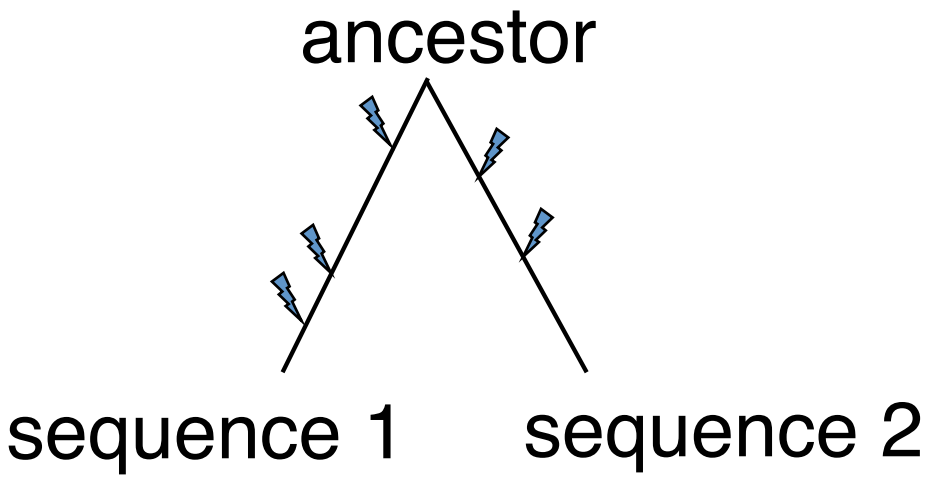
Taxa	Characters
Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTCG

Transform the sequence data into pairwise distances

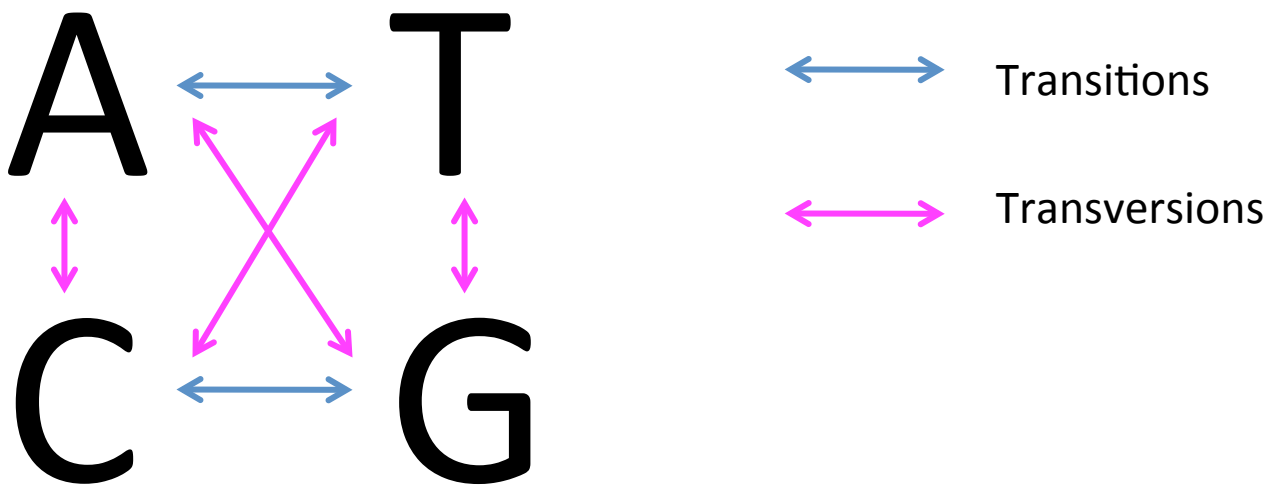
	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	----	----	0.40	0.55	0.50
Species C	----	----	----	0.15	0.40
Species D	----	----	----	----	0.25
Species E	----	----	----	----	----

Evolutionary Distances

- They measure the total number of substitutions that occurred on <u>both</u> lineages since divergence from last common ancestor.
- Divided by sequence length.
- Expressed in substitutions / site



Substitution models



Variation in rates

Quantification of evolutionary distances (2): Kimura's two parameter distance (DNA)

- Hypotheses of the model :

  (a) All sites evolve independently and following the same process.

  (b) Substitutions occur according to two probabilities :

  One for transitions, one for transversions.

  Transitions : G <–>A or C <–>T Transversions : other changes

  (c) The base substitution process is constant in time.

- Quantification of evolutionary distance (*d*) as a function of the fraction of observed differences (*p*: transitions, *q*: transversions):

$$d = -\frac{1}{2}\ln[(1-2p-q)\sqrt{1-2q}\,]$$

Kimura (1980) J. Mol. Evol. 16:111

Quantification of evolutionary distances (3): PAM and Kimura's distances (proteins)

- Hypotheses of the model (Dayhoff, 1979) :

  (a) All sites evolve independently and following the same process.

  (b) Each type of amino acid replacement has a given, empirical probability :
  Large numbers of highly similar protein sequences have been collected; probabilities of replacement of any a.a. by any other have been tabulated.

  (c) The amino acid substitution process is constant in time.

- Quantification of evolutionary distance (d) :
  the number of replacements most compatible with the observed pattern of amino acid changes and individual replacement probabilities.

- Kimura's empirical approximation : $d = - \ln(1 - p - 0.2\, p^2)$

  (Kimura, 1983) where p = fraction of observed differences

Quantification of evolutionary distances (4): Other distance measures

- Several other, more realistic models of the evolutionary process at the molecular level have been used :
  - Accounting for biased base compositions (Tajima & Nei).
  - Accounting for variation of the evolutionary rate across sequence sites.
  - etc ...

Quantification of evolutionary distances (5): Synonymous and non-synonymous distances (coding DNA): Ka, Ks

- Hypothesis of previous models :
  (a) All sites evolve independently and following the same process.
- Problem: in protein-coding genes, there are two classes of sites with very different evolutionary rates.
  - non-synonymous substitutions (change the a.a.): slow
  - synonymous substitutions (do not change the a.a.): fast
- Solution: compute two evolutionary distances
  - Ka = non-synonymous distance
    = nbr. non-synonymous substitutions / nbr. non-synonymous sites
  - Ks = synonymous distance
    = nbr. synonymous substitutions / nbr. synonymous sites

Distance methods

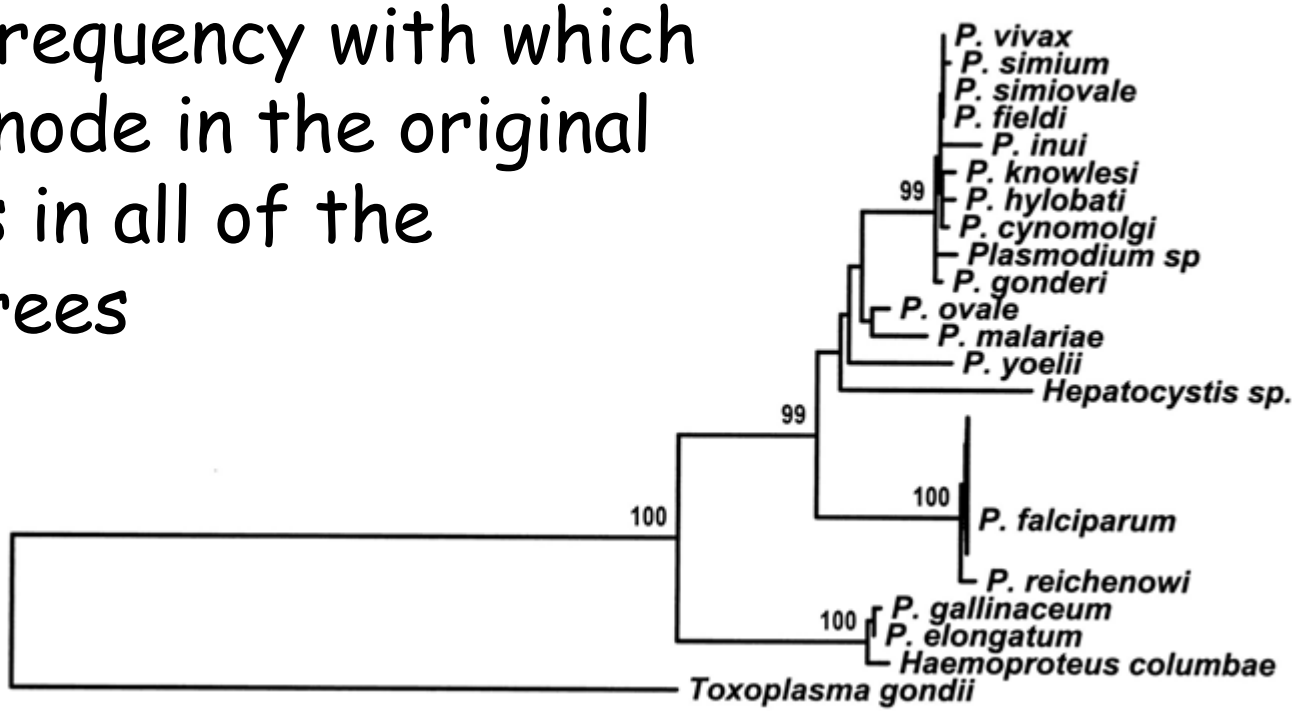
- UPGMA (Unweighted Pair Group Method with Arithmetic mean): same rate of evolution on each branch

- The Neighbor Joining method = most popular method

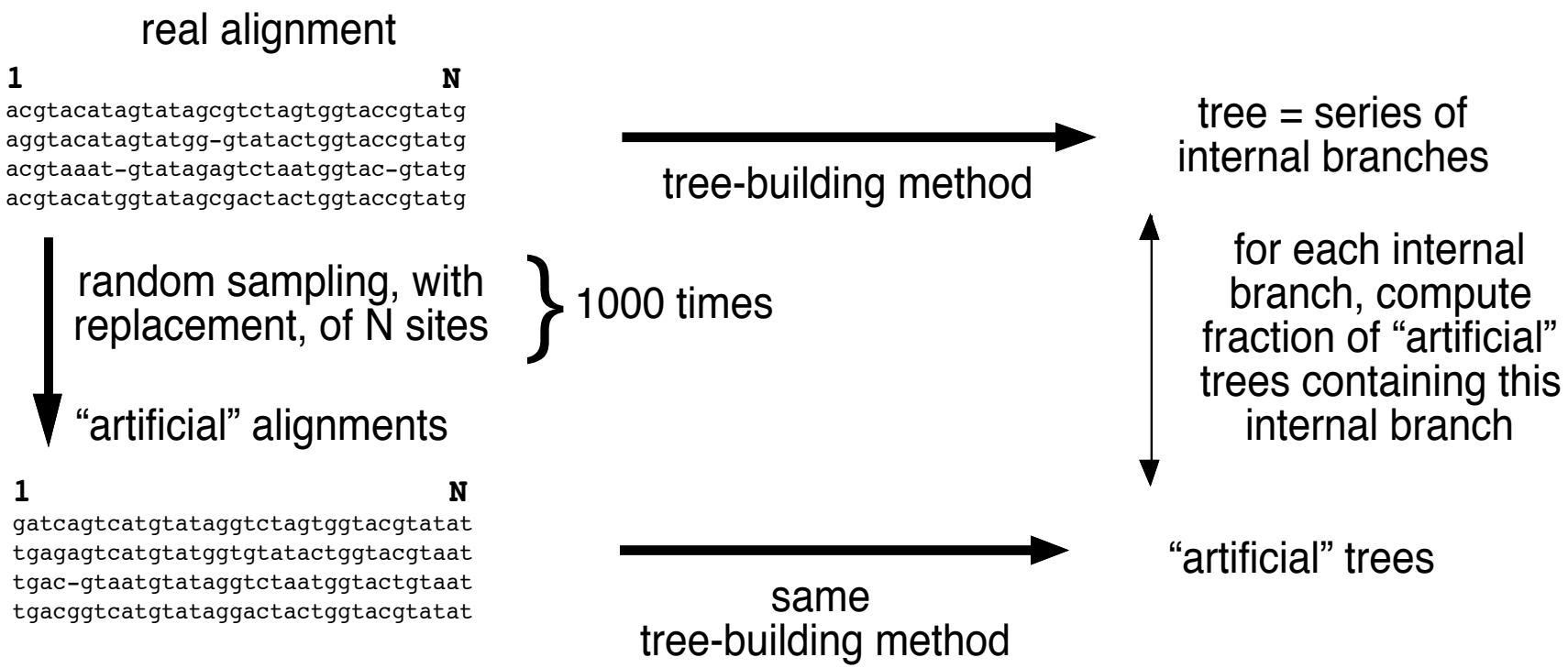
  does not assume the same rate of evolution on each branch of a tree

Resampling procedures: The Bootstrap

- Randomly resample the data X-times and infer a phylogeny
- Derive the frequency with which a particular node in the original tree appears in all of the resampled trees

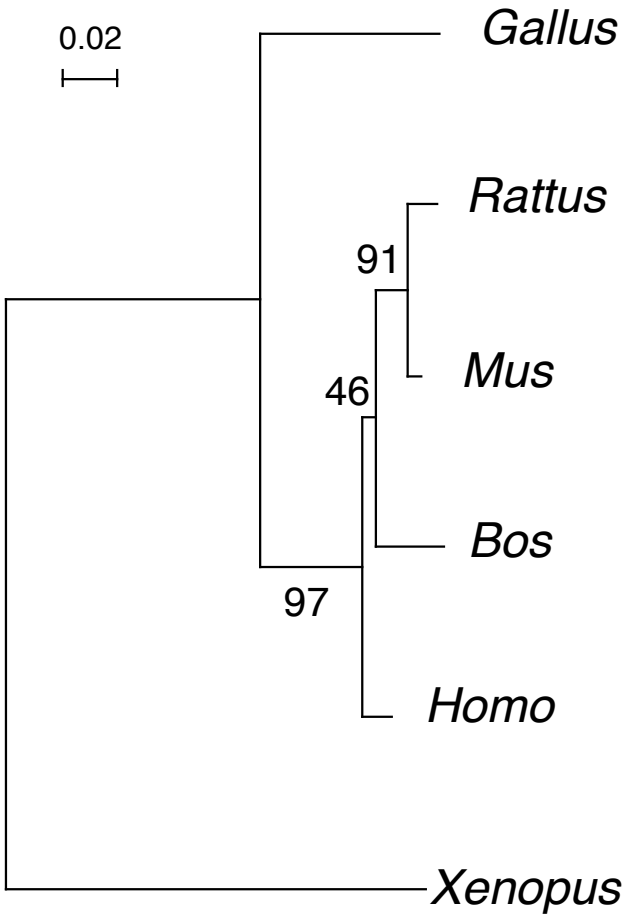


Bootstrap procedure



The support of each internal branch is expressed as percent of replicates.

"bootstrapped" tree



Bootstrap procedure : properties

- Internal branches supported by ≥ 90% of replicates are considered as statistically significant.
- The bootstrap procedure only detects if sequence length is enough to support a particular node.
- The bootstrap procedure does not help determining if the tree-building method is good. A wrong tree can have 100 % bootstrap support for all its branches!

Building Trees with Parsimony

- **Parsimony** involves evaluating all possible trees and giving each a score based on the number of evolutionary changes that are needed to explain the observed data.

- The best tree is the one that requires the fewest base changes for all sequences to derive from a common ancestor.

Maximum likelihood and bayesian methods

- Allows for substitution rates to differ on lineages and sites: appropriate for distantly related species
- Estimates the likelihood of a tree = probability of the data given an evolutionary model
- Complex and computationally intensive!

Gene tree vs. Species tree

- The evolutionary history of genes reflects that of species that carry them, except if :
  - horizontal transfer = gene transfer between species (*e.g.* bacteria, mitochondria)
  - Gene duplication : orthology/ paralogy

WWW resources for molecular phylogeny (1)

- Compilations

⇨ A list of sites and resources:
`http://www.ucmp.berkeley.edu/subway/phylogen.html`

⇨ An extensive list of phylogeny programs
`http://evolution.genetics.washington.edu/phylip/software.html`

- Databases of rRNA sequences and associated software

⇨ The rRNA WWW Server - Antwerp, Belgium.
`http://rrna.uia.ac.be`

⇨ The Ribosomal Database Project - Michigan State University
`http://rdp.cme.msu.edu/html/`

WWW resources for molecular phylogeny (2)

- Database similarity searches (Blast) :

`http://www.ncbi.nlm.nih.gov/BLAST/`
`http://www.infobiogen.fr/services/menuserv.html`
`http://bioweb.pasteur.fr/seqanal/blast/intro-fr.html`
`http://pbil.univ-lyon1.fr/BLAST/blast.html`

- Multiple sequence alignment

⇨ClustalX : multiple sequence alignment with a graphical interface (for all types of computers).
`http://www.ebi.ac.uk/FTP/index.html` and go to 'software'

⇨Web interface to ClustalW algorithm for proteins:
`http://pbil.univ-lyon1.fr/` and press "**clustal**"

WWW resources for molecular phylogeny (3)

- Sequence alignment editor

⇨ SEAVIEW : for windows and unix
`http://pbil.univ-lyon1.fr/software/seaview.html`

- Programs for molecular phylogeny

⇨ PHYLIP : an extensive package of programs for all platforms
`http://evolution.genetics.washington.edu/phylip.html`

⇨ CLUSTALX : beyond alignment, it also performs NJ

⇨ PAUP* : a very performing commercial package
`http://paup.csit.fsu.edu/index.html`

⇨ PHYLO_WIN : a graphical interface, for unix only
`http://pbil.univ-lyon1.fr/software/phylowin.html`

⇨ MrBayes : Bayesian phylogenetic analysis `http://morphbank.ebc.uu.se/mrbayes/`

⇨ PHYML : fast maximum likelihood tree building `http://www.lirmm.fr/~guindon/phyml.html`

⇨ WWW-interface at Institut Pasteur, Paris
`http://bioweb.pasteur.fr/seqanal/phylogeny`

WWW resources for molecular phylogeny (4)

- Tree drawing

  NJPLOT (for all platforms)

  `http://pbil.univ-lyon1.fr/software/njplot.html`

- Lecture notes of molecular systematics

  `http://www.bioinf.org/molsys/lectures.html`

END

Thank you!