# Basic phylogenetics

**Praveen Karanth**
**CES, IISc**

**ICTS School and discussion meeting:**
**Population Genetics and Evolution Draft**
**24 Jan 2014**

## *History of phylogeny (Why build evolutionary trees ?)*

Biological diversity needs to be organized and catalogued i.e., classified.

For better communication between biologists.

This exercise of classifying biodiversity is of immense importance because we need to know what's out there and how they are related to each other to better understand the underlying biological processes that have generated them.

This information in turn can be used to our benefit, such as, to develop cure to various diseases or to manage and control of various pests, and for efficient use of economically important species (domesticated plants and animals).

Moreover this information is central for the management and conservation of our biological heritage. For example, what do we conserve if we don't know what's out there.

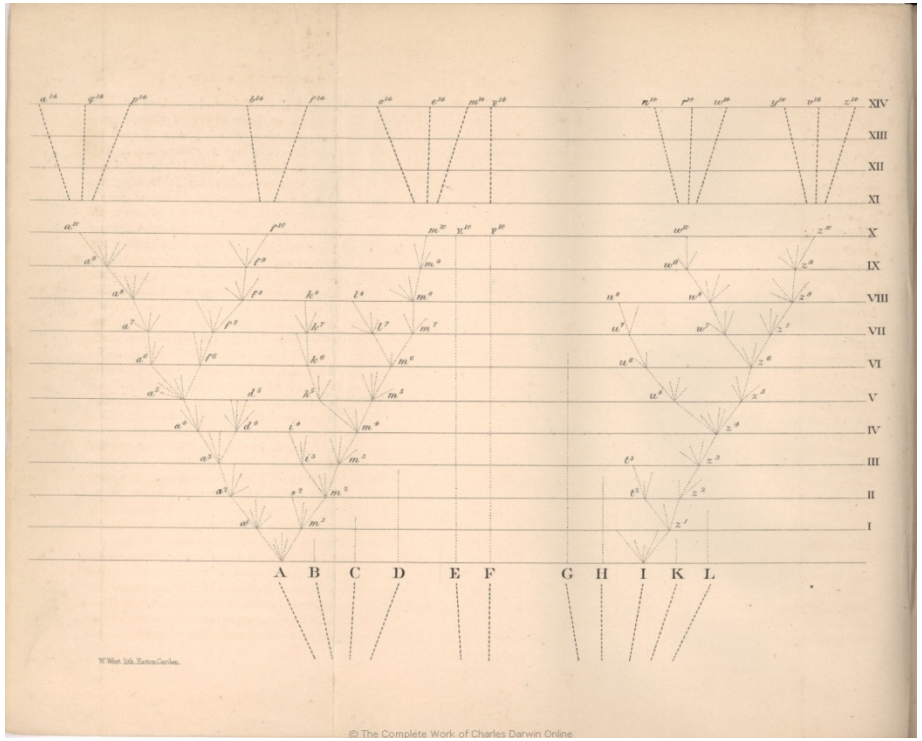
Biological diversity needs to be organized and catalogued: classified

Classification
Taxonomy: Carolus Linnaeus (1759) : naming, ranking, and classifying organisms, hierarchical taxonomy

After Darwin's book "On the origin of species" that was published in 1859, it became apparent that classification of biological entities should be based on evolutionary relationships.

Thus was born the area of Systematics wherein organisms are classified based on evolutionary relationships.

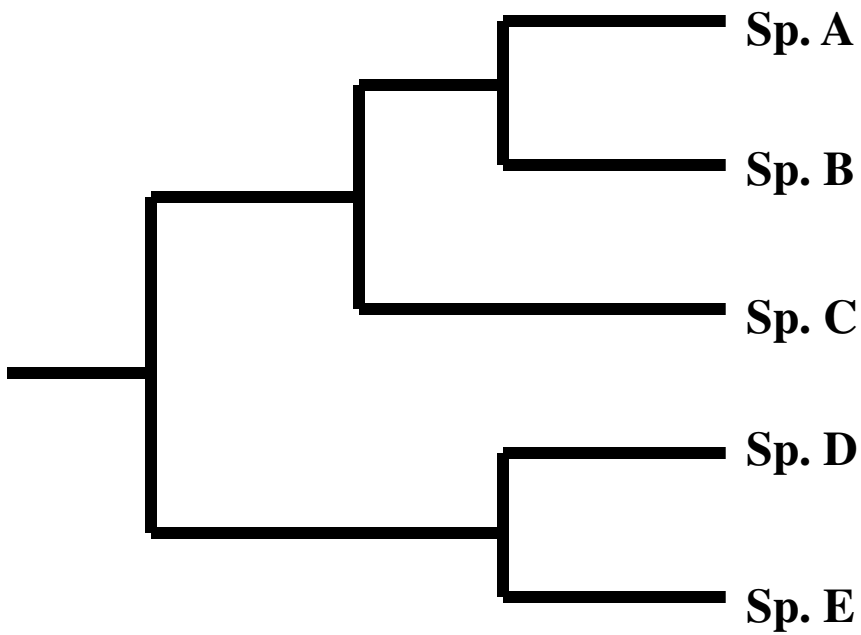
PHYLOGENY




http://darwin-online.org.uk/

Systematic: classifying organisms based on evolutionary relationships

**Phylogeny: It's a evolutionary tree that shows how species are related to each other**




"genesis and evolution of a phylum," 1872 (in Darwin),
from Ger. *Phylogenie,* coined 1866 by Ger. biologist Ernst Heinrich Haeckel (1834-1919)
from Gk. *phylon* "race" + *-geneia* "origin," from *-genes* "born."
http://www.etymonline.com/

Systematics/taxonomy often lack objective criteria and evolutionary relationships between species are based on assertions by experts on particular taxonomic groups. There was a lack of unifying or standardized classification method, no formalized procedure, informed opinion!

Two schools, systematic classification of organisms

**Phenetics**: Classification based on overall similarity in as many characters as possible (Sneath and Sokal. 1973). Numerical taxonomy

***Phenogram***: A branching diagram that links species by estimates of overall similarity.

**Cladistics**: Method of classifying organisms into groups (taxa) based on subset of similarity attributable to shared derived (synapomorphies) characters. Does not use all characters (Willi Hennig 1950). Phylogenetic systematics

Principle of parsimony: optimal cladogram, tree requiring the minimum number of steps to explain the character distribution.
Parsimony: the principle of endorsing the simplest explanation that covers a case

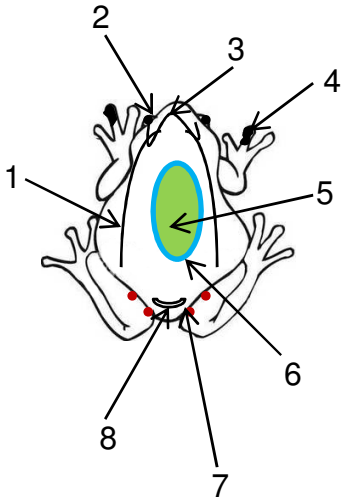
***Cladogram***: A tree that depicts inferred historical branching (evolutionary) relationships among species.

A B C D E

How are these 5 species related to each other?

Can you identify some characters?

A B C D E




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **A** | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| **B** | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| **C** | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| **D** | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| **E** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

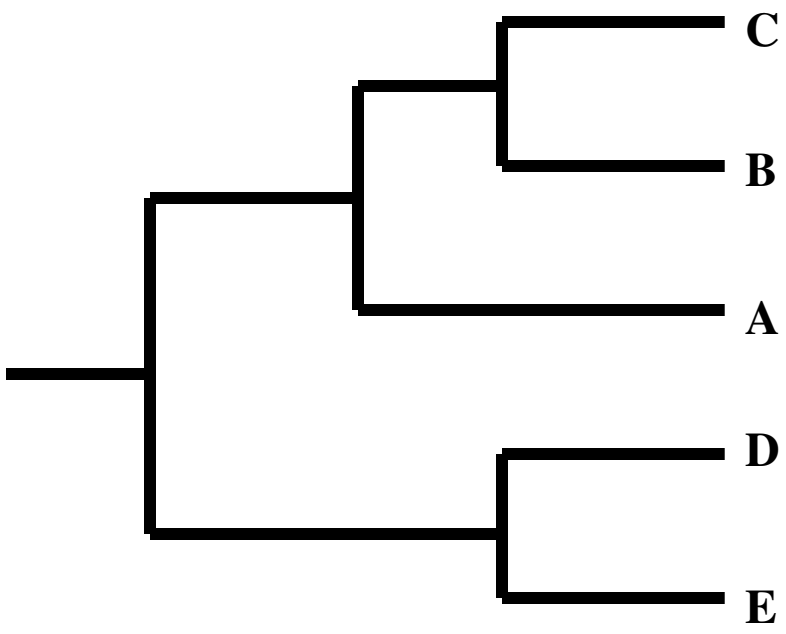
# Phenetic approach: tree based on overall similarity/dissimilarity

***Phenogram***: A branching diagram that links species by estimates of overall similarity or dissimilarity.
UPGMA tree

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **A** | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| **B** | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| **C** | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| **D** | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| **E** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

| | A | B | C | D | E |
|---|---|---|---|---|---|
| **A** | – | | | | |
| **B** | 3 | – | | | |
| **C** | 4 | 1 | – | | |
| **D** | 7 | 6 | 5 | – | |
| **E** | 7 | 4 | 3 | 2 | – |

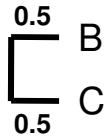
dissimilarity matrix



# Unweighted paired group method with arithmetic mean (UPGMA) *Sokal & Michner 1958*

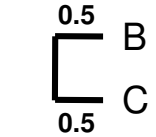
| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | – | | | | |
| B | 3 | – | | | |
| C | 4 | 1 | – | | |
| D | 7 | 6 | 5 | – | |
| E | 7 | 4 | 3 | 2 | – |

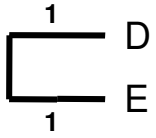
dissimilarity matrix




$d_{BC\text{-}A} = (d_{BA}+d_{CA})/2 = (3+4)/2 = 7/2 = 3.5$
$d_{BC\text{-}D} = (d_{BD}+d_{CD})/2 = (6+5)/2 = 11/2 = 5.5$
$d_{BC\text{-}E} = (d_{BE}+d_{CE})/2 = (4+3)/2 = 7/2 = 3.5$

| | BC | A | D | E |
|---|---|---|---|---|
| BC | – | | | |
| A | 3.5 | – | | |
| D | 5.5 | 7 | – | |
| E | 3.5 | 7 | 2 | – |



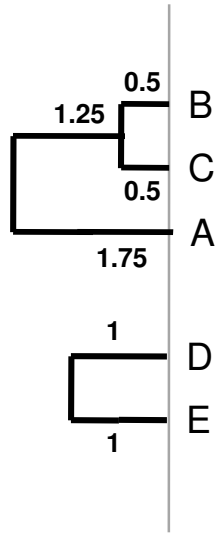




| | BC | A | DE |
|---|---|---|---|
| BC | – | | |
| A | 3.5 | – | |
| DE | 4.5 | 7 | – |

$d_{DE\text{-}BC} = (d_{DB}+ d_{DC}+d_{EB}+d_{EC})/4 =$
$(6+5+4+3)/4 = 18/4 = 4.5$
$d_{DE\text{-}A} = (d_{DA}+d_{EA})/2 = (7+7)/2 = 14/2 = 7$

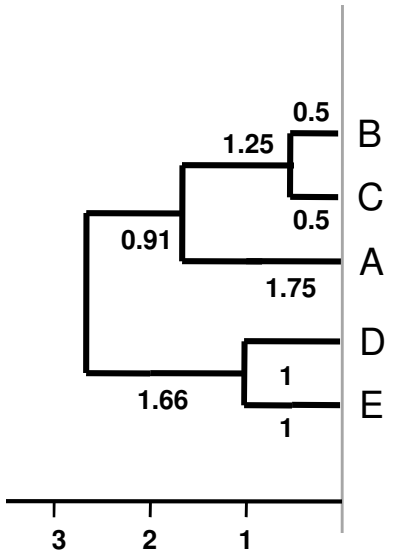
| | BC | A | DE |
|---|---|---|---|
| BC | – | | |
| A | 3.5 | – | |
| DE | 4.5 | 7 | – |

$d_{BC\text{-}A}$ = 3.5 [3.5/2=1.75]



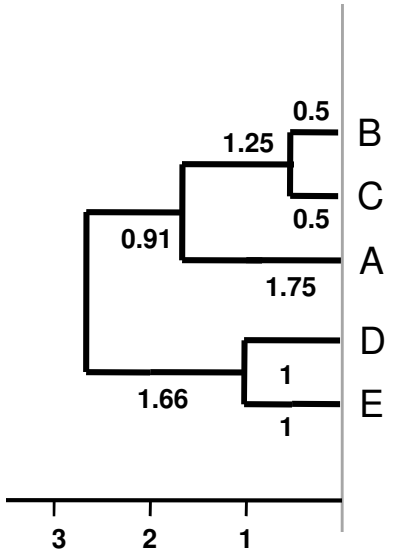

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | – | | | | |
| B | 3 | – | | | |
| C | 4 | 1 | – | | |
| D | 7 | 6 | 5 | – | |
| E | 7 | 4 | 3 | 2 | – |

$d_{ABC\text{-}DE}$ = (AD+AE+BD+BE+CD+CE)/6
(7+7+6+4+5+3)/6 = 5.33 [5.33/2=2.66]



Observed distance

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | – | | | | |
| B | 3 | – | | | |
| C | 4 | 1 | – | | |
| D | 7 | 6 | 5 | – | |
| E | 7 | 4 | 3 | 2 | – |



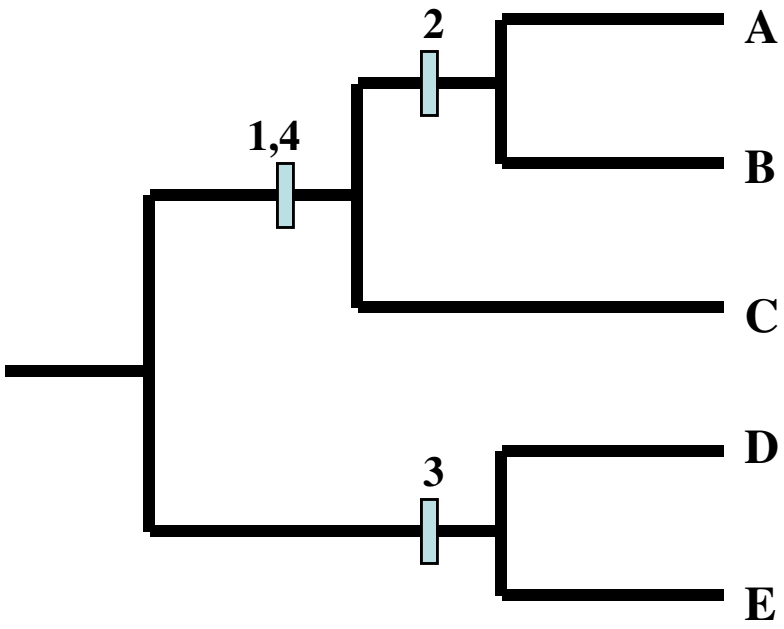

Tree distance

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | – | | | | |
| B | 3.5 | – | | | |
| C | 3.5 | 1 | – | | |
| D | 5.3 | 5.3 | 5.3 | – | |
| E | 5.3 | 5.3 | 5.3 | 2 | – |

***Phenogram***: A branching diagram that links species by estimates of overall similarity or dissimilarity. Generated by UPGMA method.

**Cladistics**: Tree based on shared derived characters

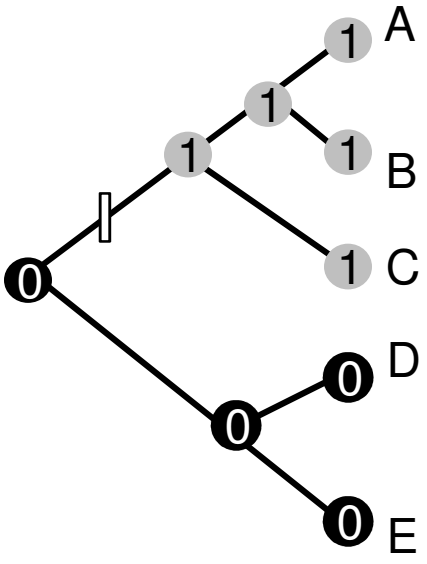
**Synapomorphies (shared derived characters) are characters inherited from the most recent common ancestor. (evolutionary homologies) Similarity due to common ancestry**



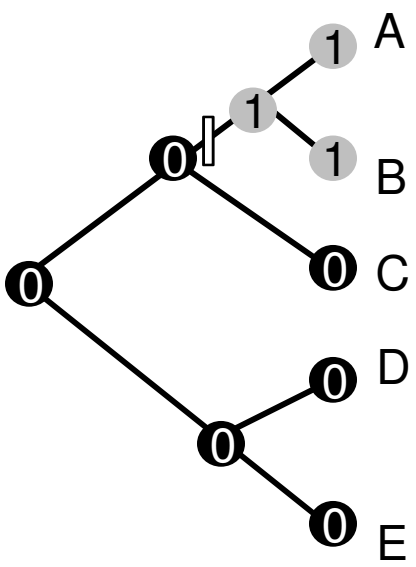

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| B | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| C | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| D | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| E | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **A** | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| **B** | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| **C** | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| **D** | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| **E** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

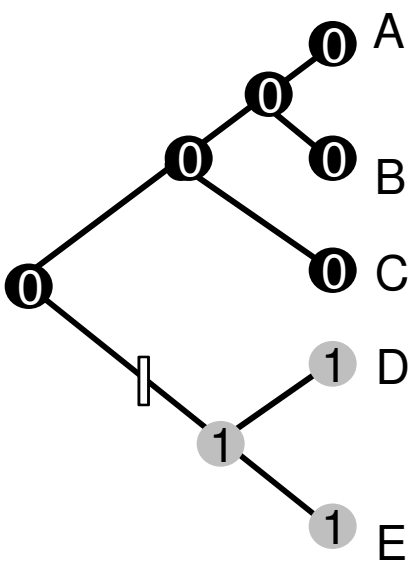
**Evolutionary homologies:**
**Similarity due to common ancestry**



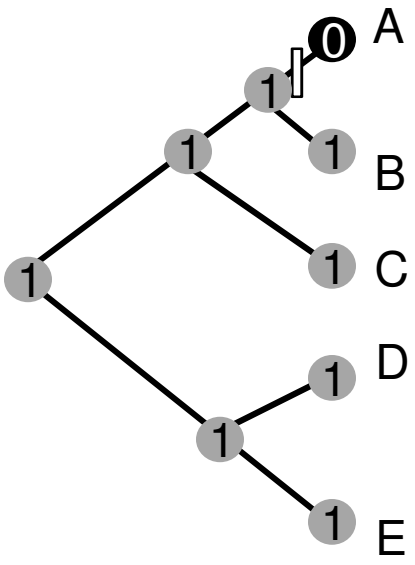

1,4) Synapomorphies (A+B+C)




2) Synapomorphy (A+B)

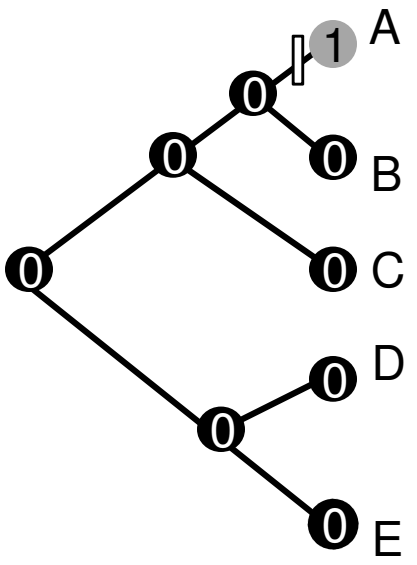



3) Synapomorphy (D+E)

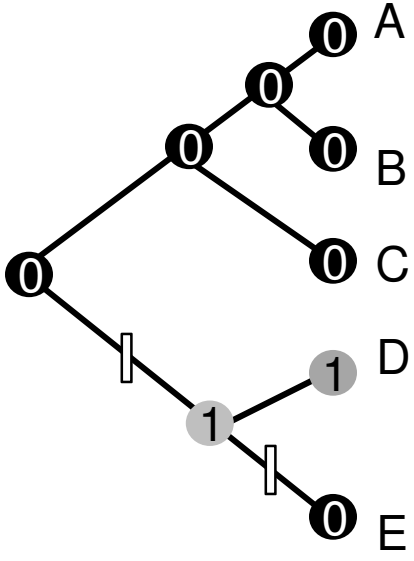



5) Loss of character (A)

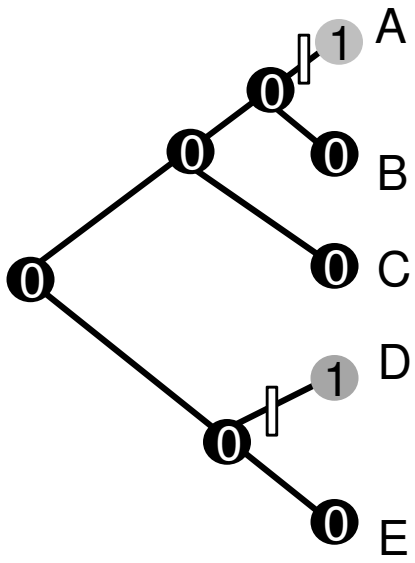
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **A** | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| **B** | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| **C** | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| **D** | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| **E** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |




8) Unique character (A)




6) Reversal (E)




7) Convergence (A and D)

*Homoplasy*

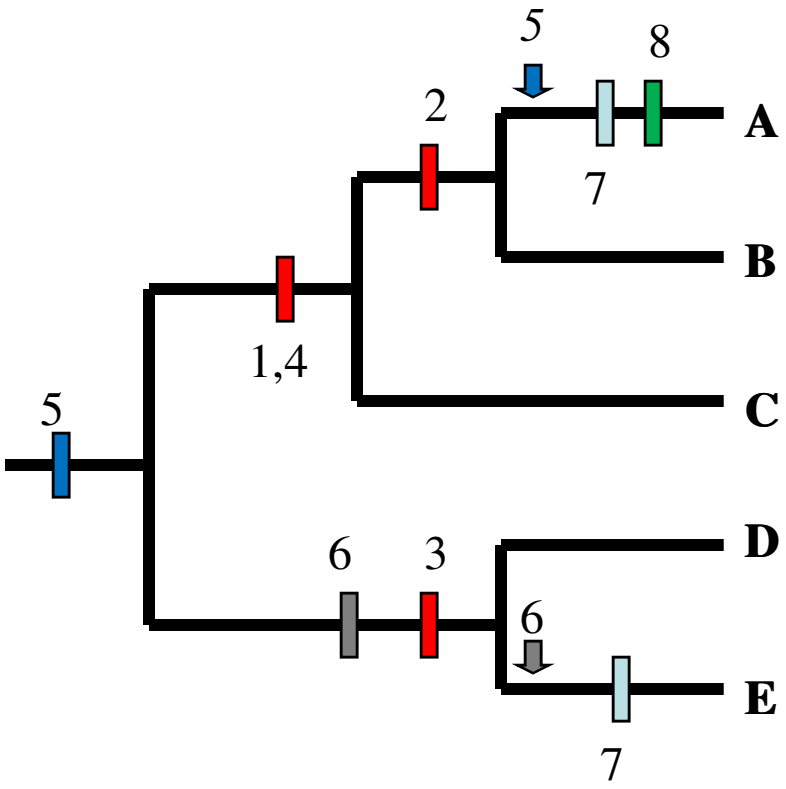
Reversal

Convergence

Character loss

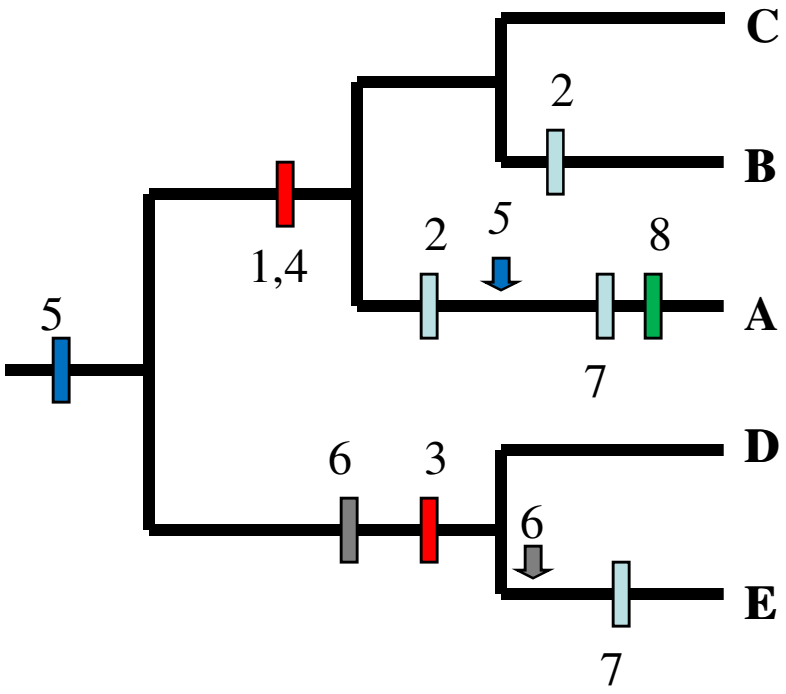
Unique

Shared derived (**homologies**)



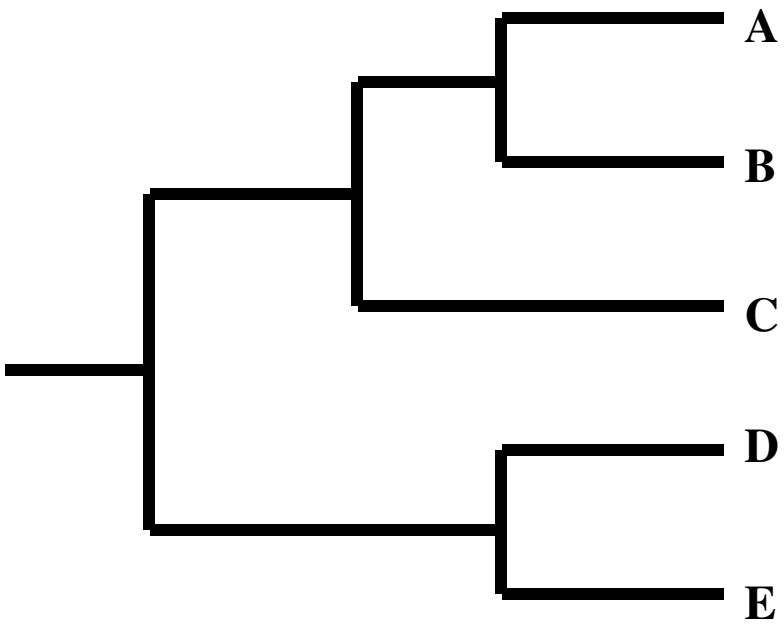

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **A** | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| **B** | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| **C** | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| **D** | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| **E** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

Principle of parsimony:
Parsimony: the principle of endorsing the simplest explanation that covers a case optimal cladogram, tree requiring the minimum number of steps to explain the character distribution.

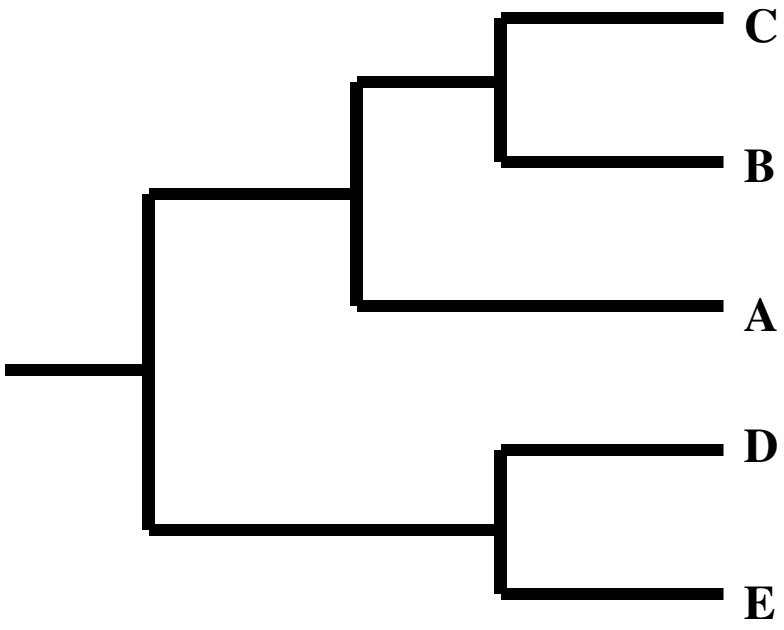



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| B | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| C | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| D | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| E | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

Cladogram=evolutionary tree=phylogeny
Phenogram is usually a good indicator of evolutionary relationships,
but strictly speaking is not a phylogeny

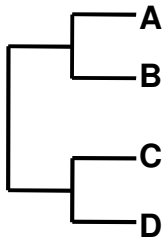



11 steps

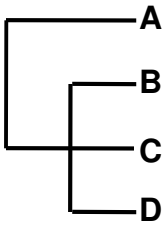



12 steps

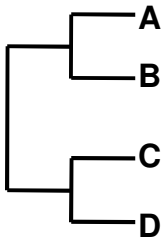
| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 1 | 1 | 0 | 1 |
| B | 1 | 0 | 0 | 0 |
| C | 0 | 0 | 1 | 0 |
| D | 0 | 0 | 1 | 1 |



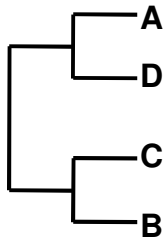

*Site 1*



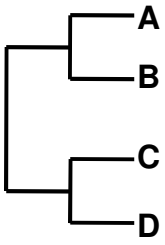

*Site 2*



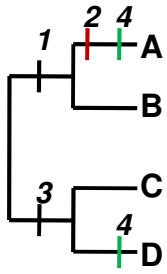

*Site 3*

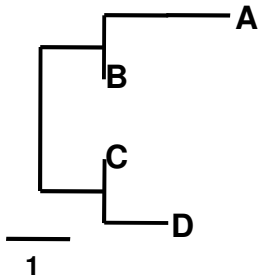



*Site 4*



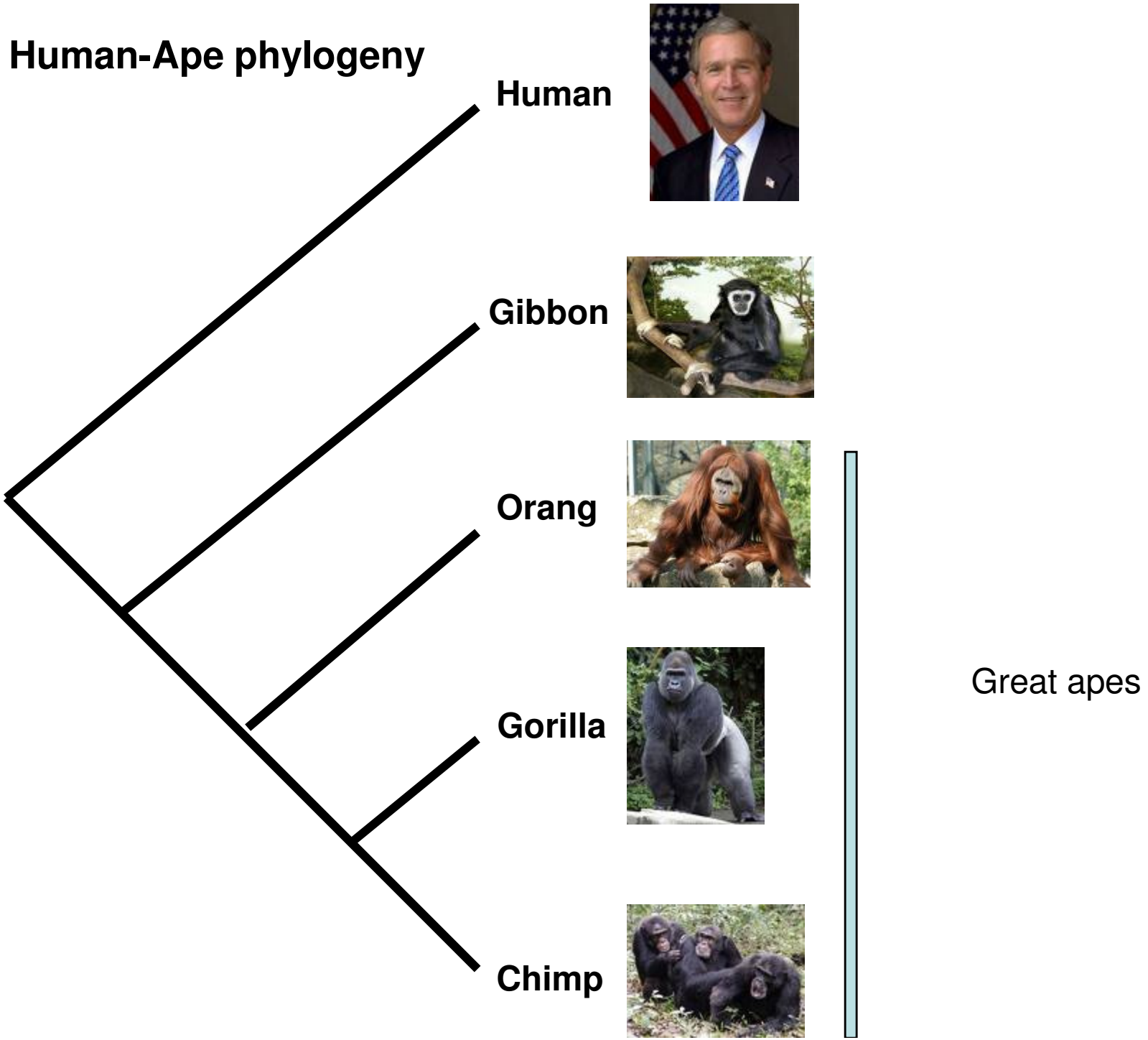

Cladogram

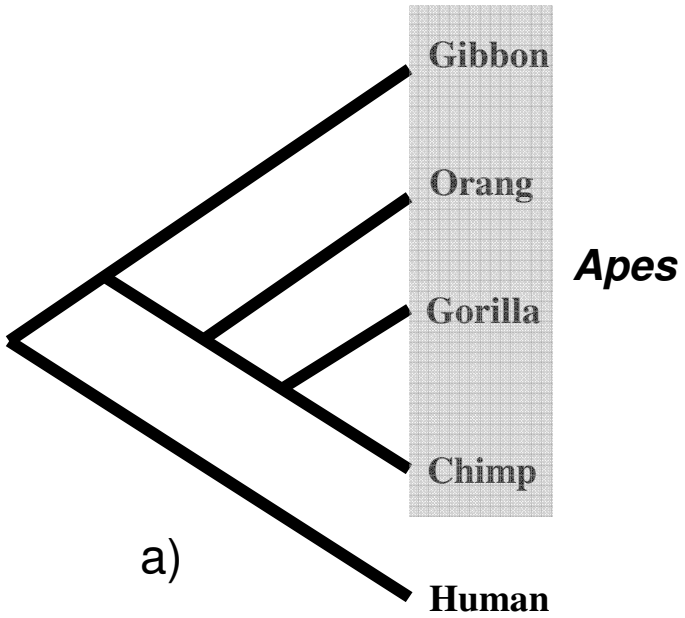


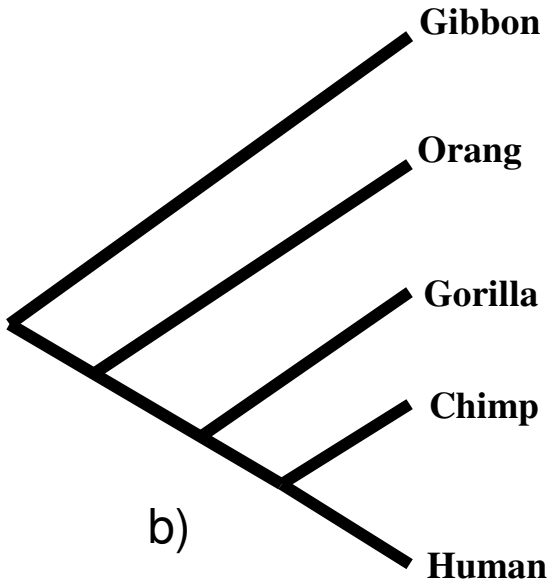




Phylogram

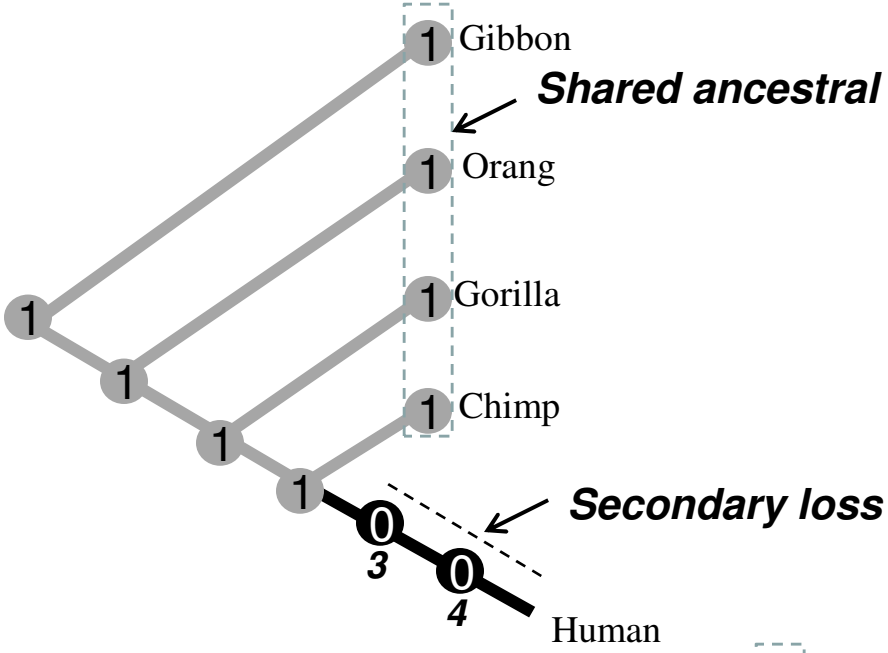
## Human-Ape phylogeny

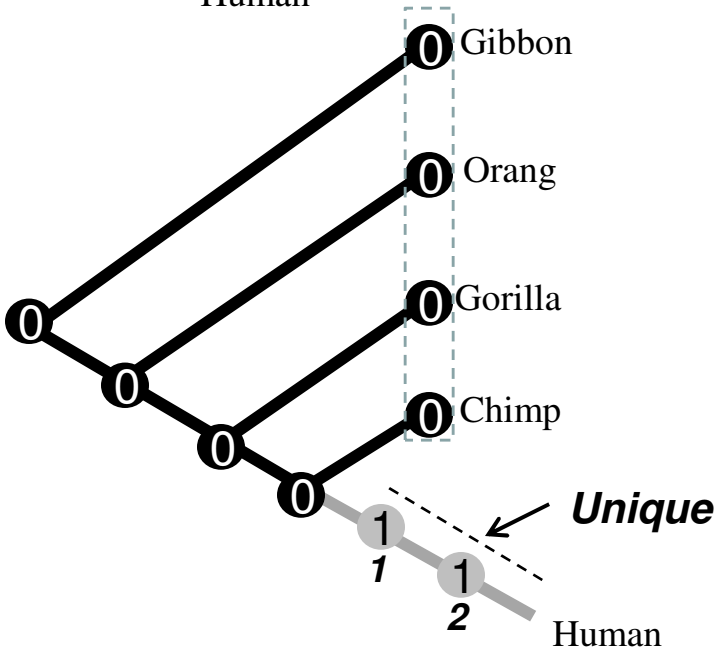



Gibbon
Orang
Gorilla
Chimp
Apes
a)
Human


Gibbon
Orang
Gorilla
Chimp
b)
Human

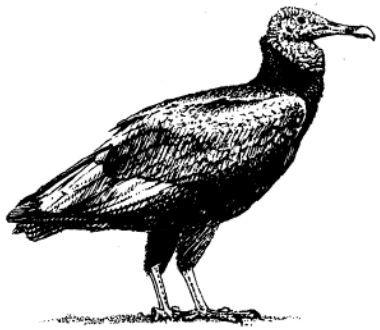
| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Gibbon | 0 | 0 | 1 | 1 |
| Orang | 0 | 0 | 1 | 1 |
| Gorilla | 0 | 0 | 1 | 1 |
| Chimp | 0 | 0 | 1 | 1 |
| Human | 1 | 1 | 0 | 0 |



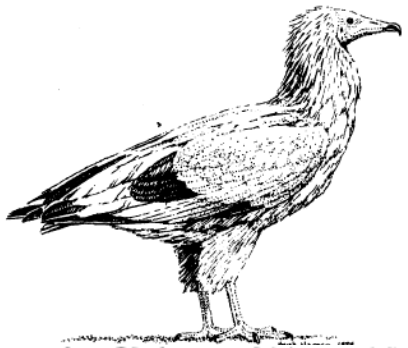




Cladistic vs. Phenetic
Homoplasies
Shared ancestral
Unique characters

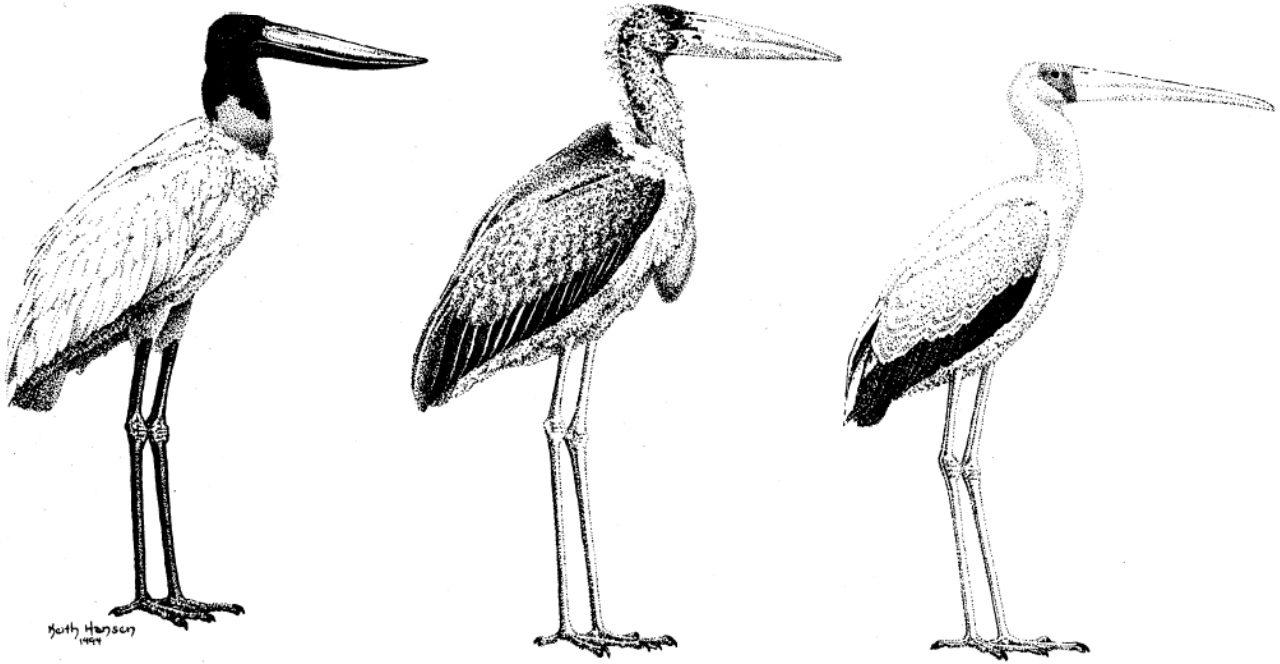
# Molecular systematics vs Morphology based classification



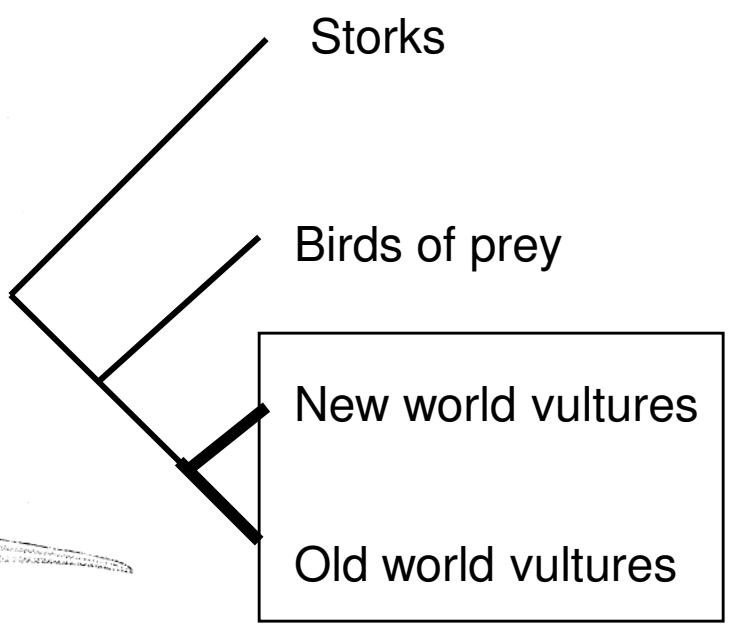
Black Vulture (New World)

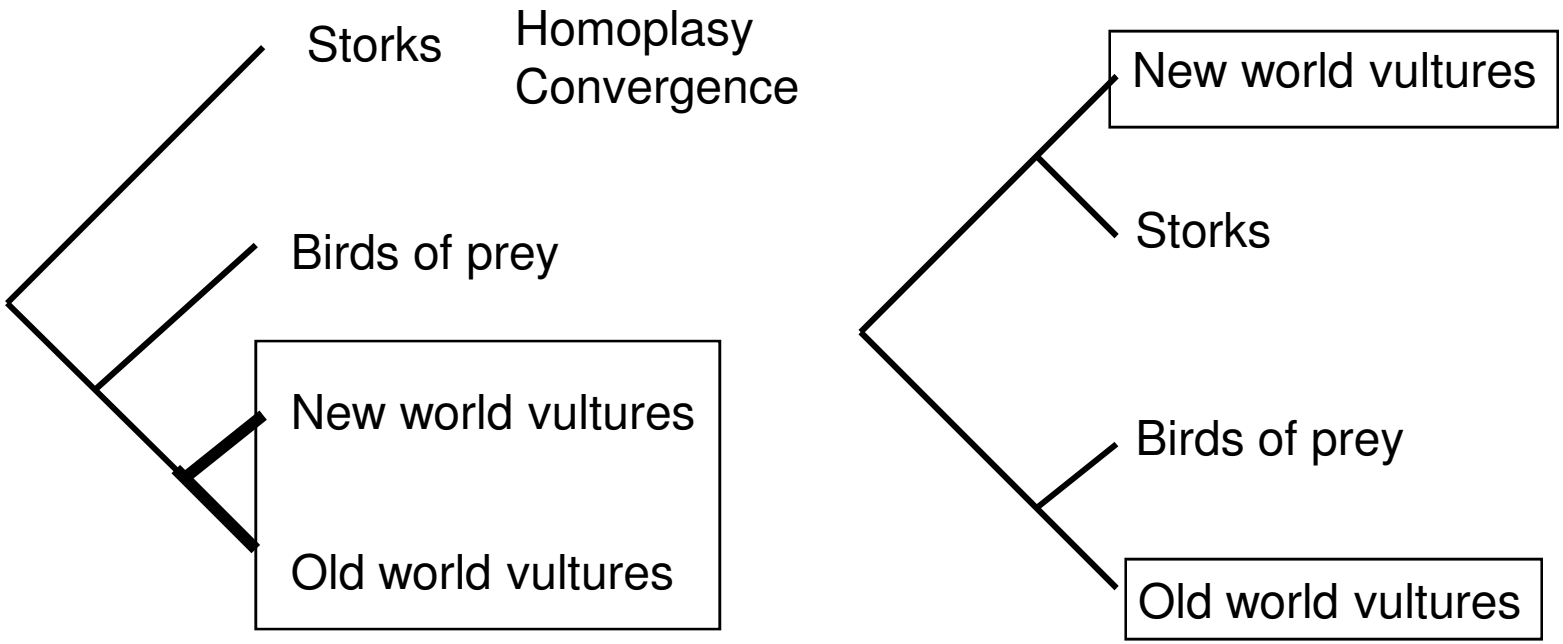


Egyptian Vulture (Old World)

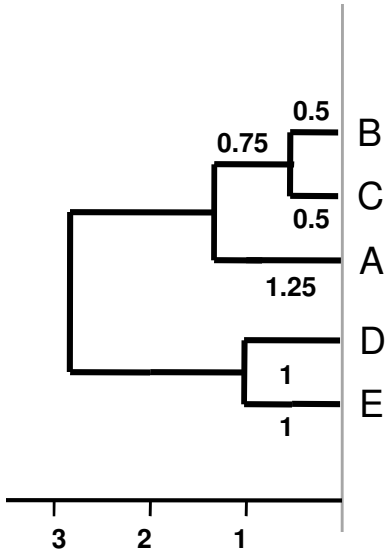


storks



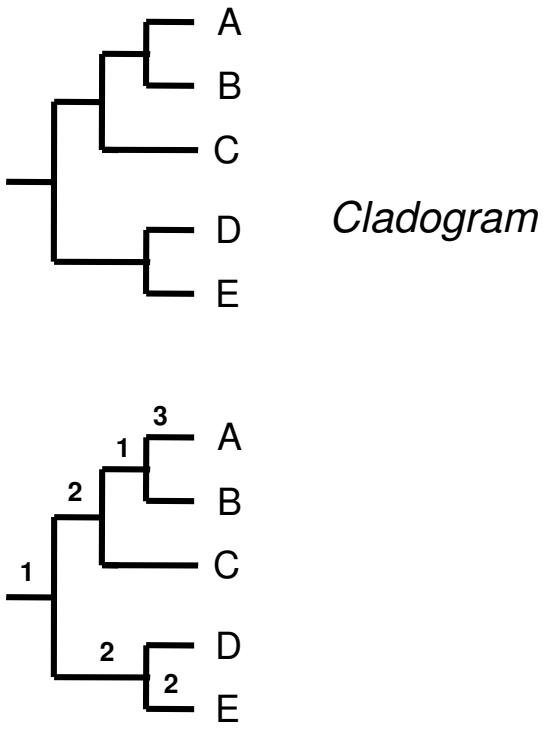

Storks
Homoplasy
Convergence
Birds of prey
New world vultures
Old world vultures
New world vultures
Storks
Birds of prey
Old world vultures

## Phenetic approach (UPGMA tree)

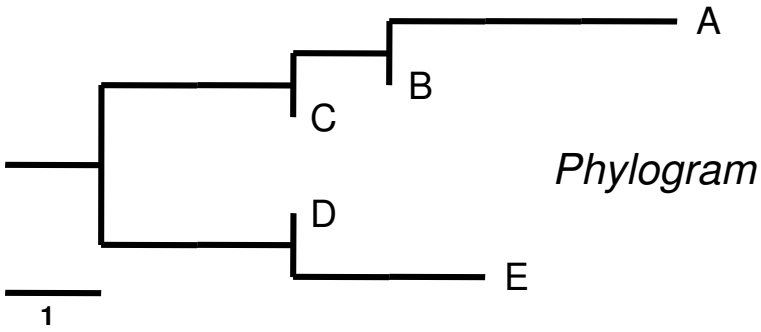



*Phenogram (dendrogram)*: A branching diagram that links species by estimates of overall similarity or dissimilarity.

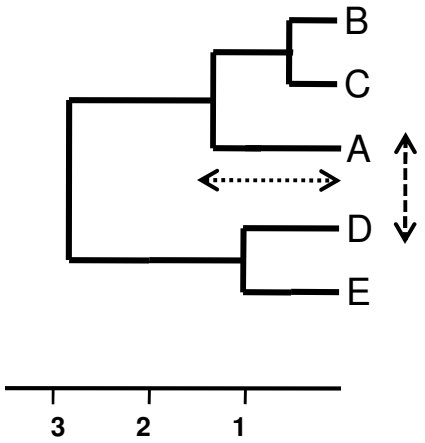
## Cladistic approach (parsimony)



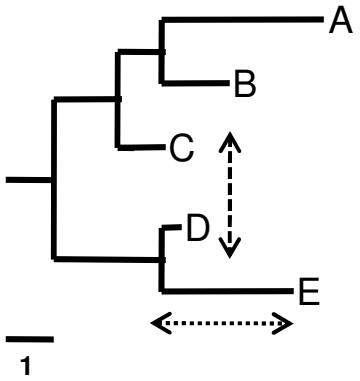

*Tree length=11*



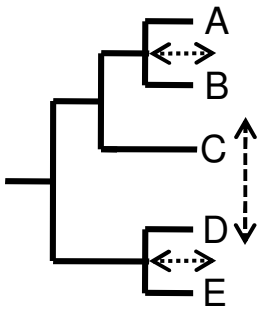
# Vertical vs. Horizontal distance

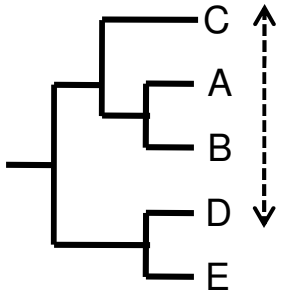



*Phenogram*




*Phylogram*



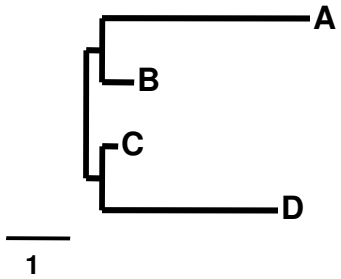



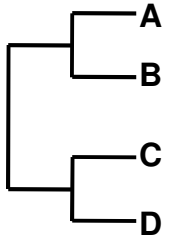

*Cladogram*



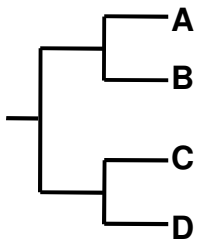
Horizontal distance has
no meaning in cladogram
But useful in phylogram/phenogarm

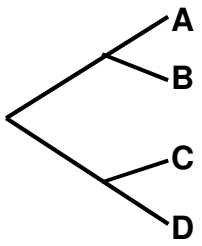
Vertical distance has no
meaning in cladogram and
phylogram/phenogarm

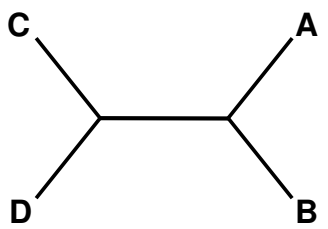


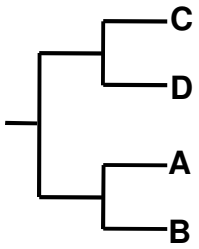
**Rooted tree** **Unrooted tree**

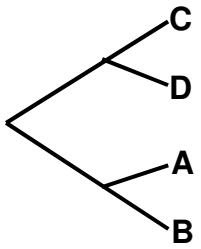


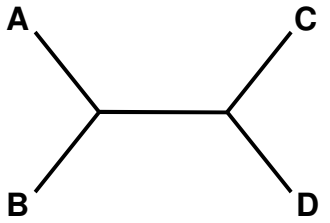



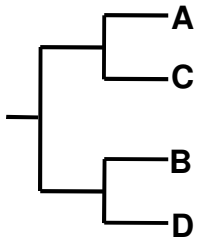

*Rectangular cladogram* *Slant cladogram*

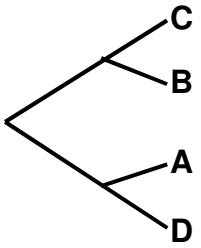


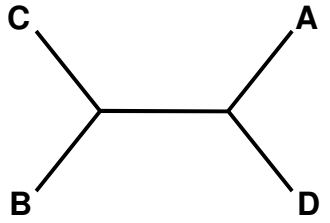



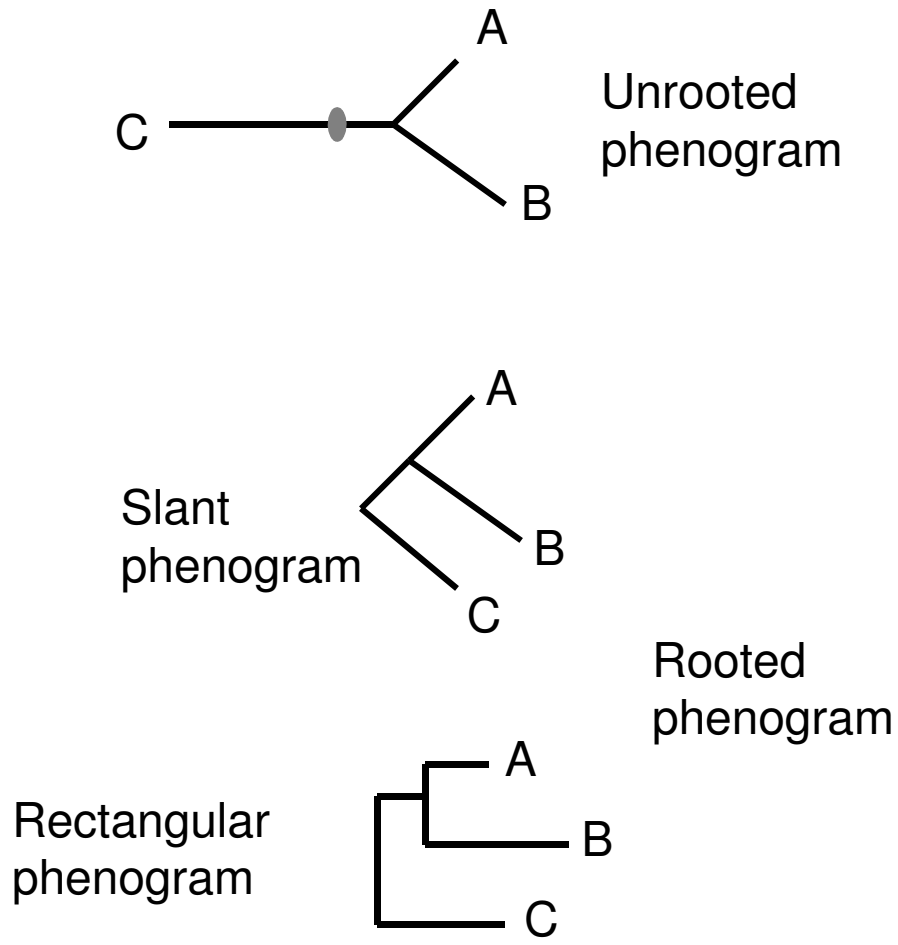




*Tree topology*



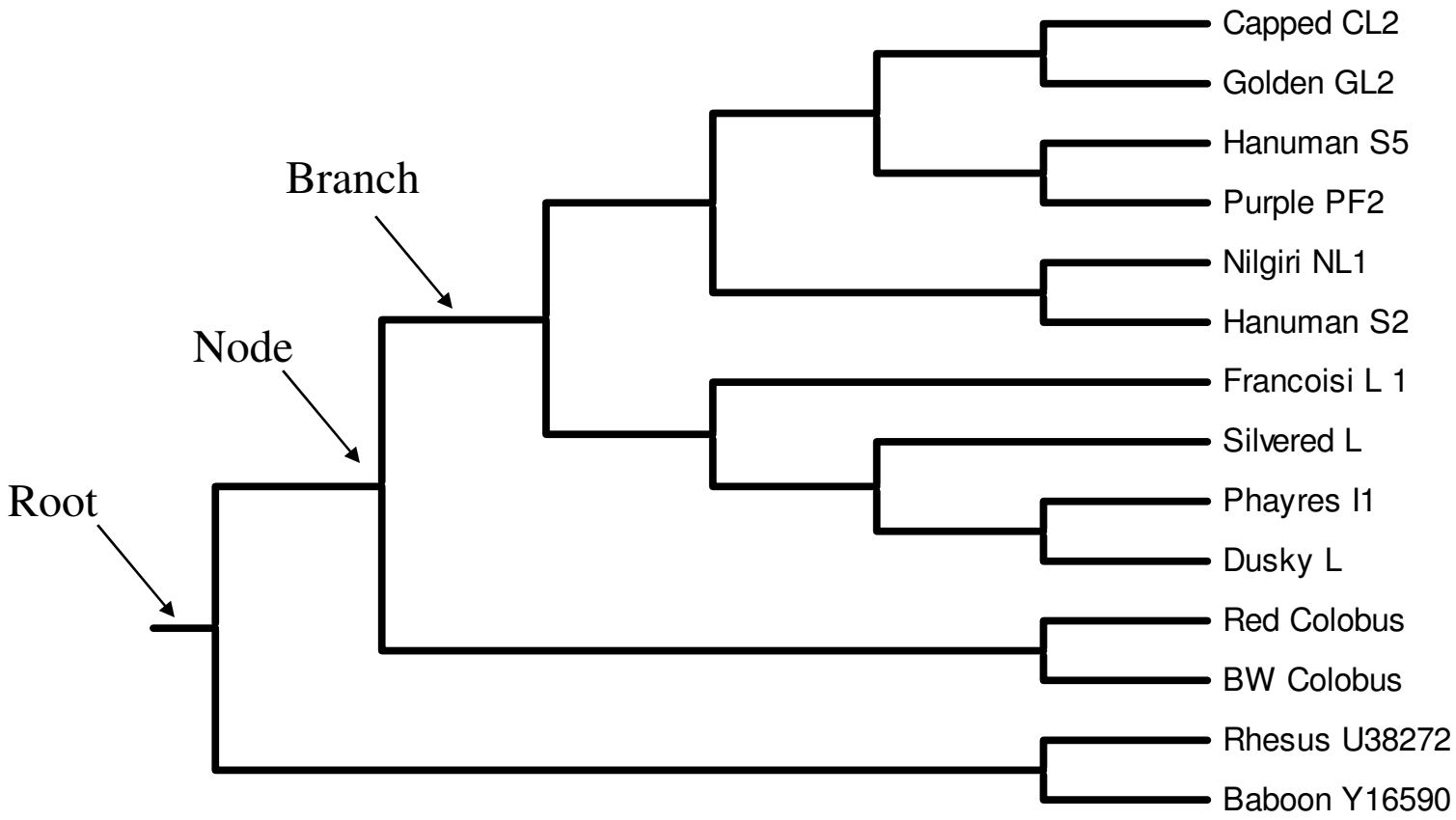

A
C
B
Unrooted
phenogram
A
Slant
phenogram
B
C
Rooted
phenogram
A
Rectangular
phenogram
B
C
Scale: number of changes

**Nodes**

External (operational taxonomic units OTU= extant species)

Internal (common ancestor)

Internal branch : between 2 nodes. External branch : between a node and a taxa

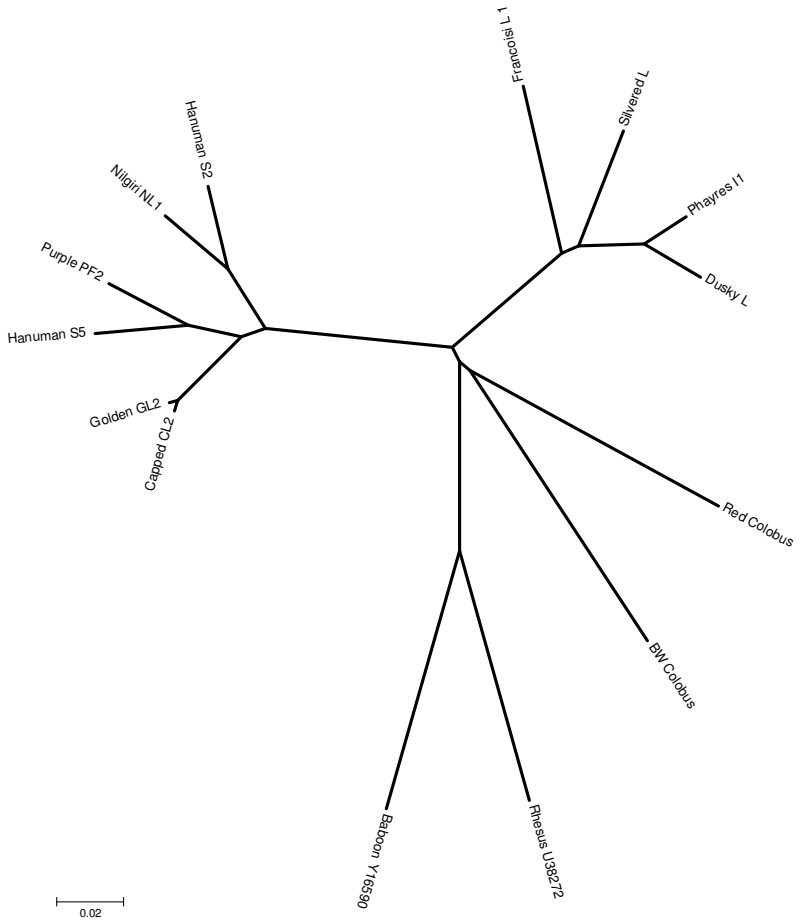


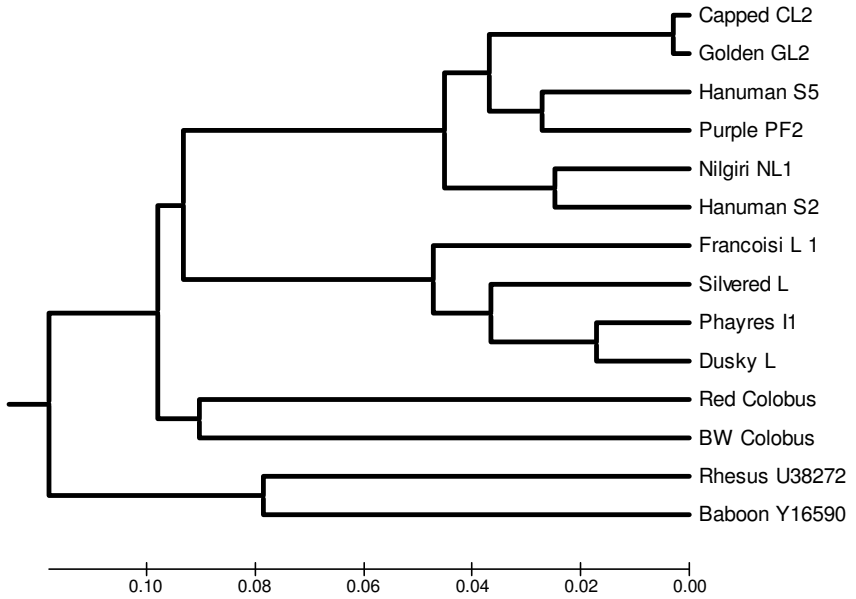

Tree Topology = shape of tree = branching order between nodes

Nodes=common ancestors; branches=lineages

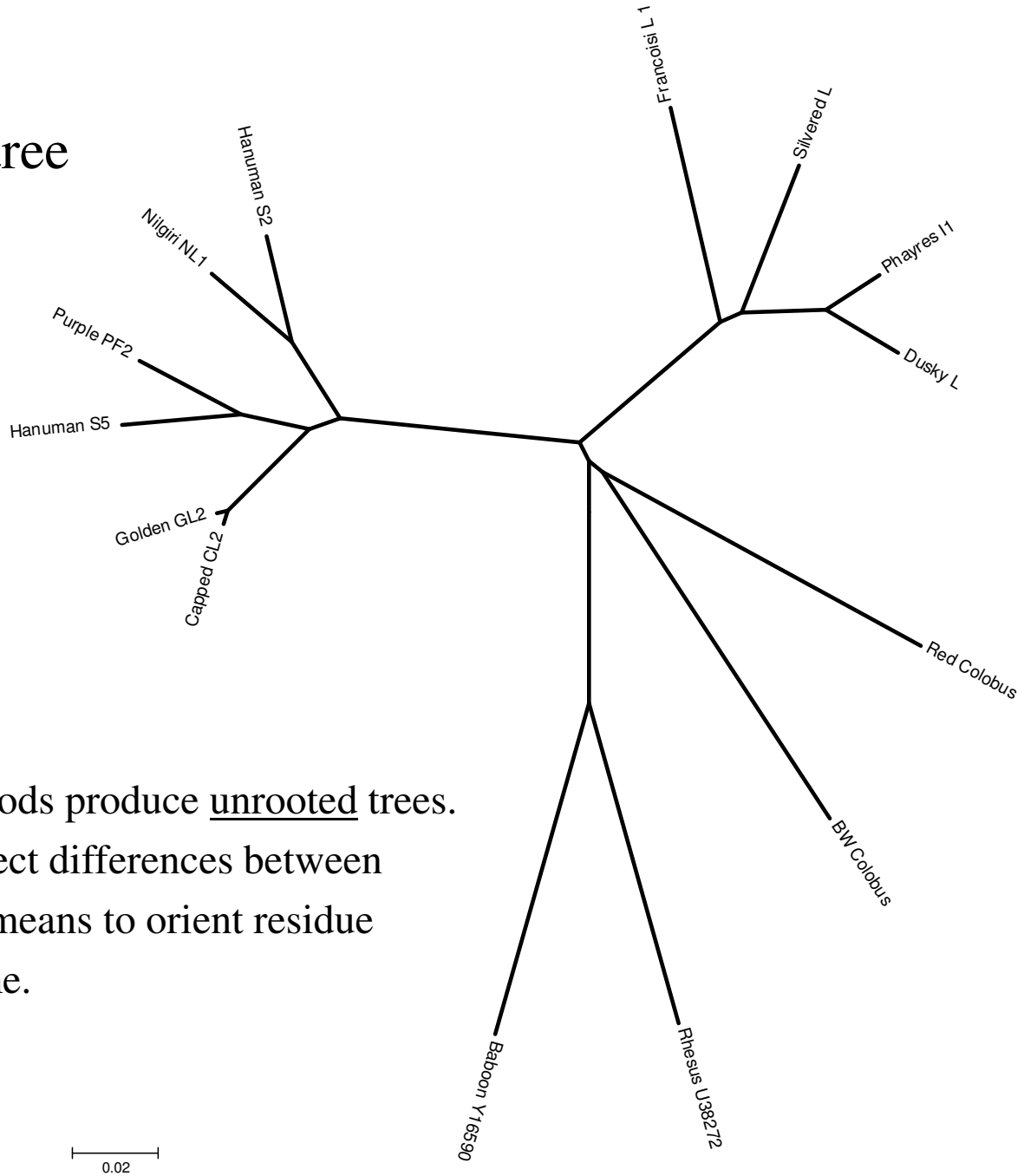
# Rooted vs. unrooted trees

The point on the tree that represents the earliest time in the evolutionary history of those sequences.



# Unrooted NJ tree




Most phylogenetic methods produce <u>unrooted</u> trees. This is because they detect differences between sequences, but have no means to orient residue changes relatively to time.

Why root trees?
—including root improves the estimate of the phylogeny.
Gives stability to the phylogeny.
—direction of evolution is known in an rooted tree
(facilitates the study of morphological and behavioural characters).
Useful for reconstruction of character evolution.

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.

–Mid-point method (molecular clock assumption) : 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 external nodes.

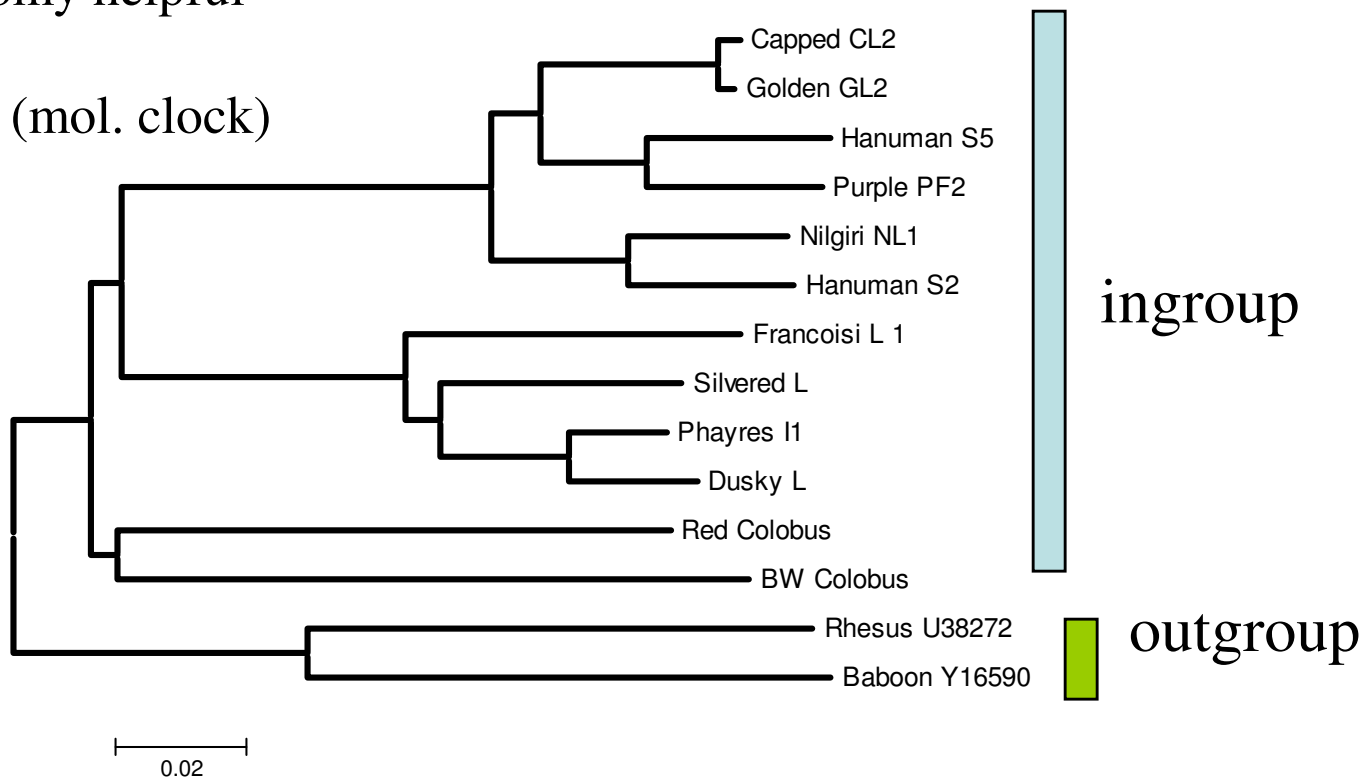
# Choice of out group

Should not be very distantly related
Might lead to serious topological errors due to saturation/multiple hits

Should not be too closely related
It might be one of the ingroup species

Traditional taxonomy helpful

Mid point rooting (mol. clock)

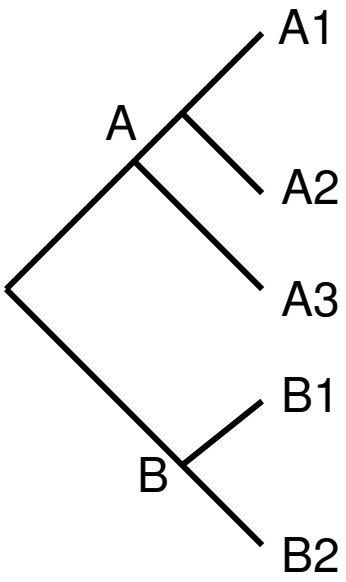


## Monophyly, paraphyly, polyphyly

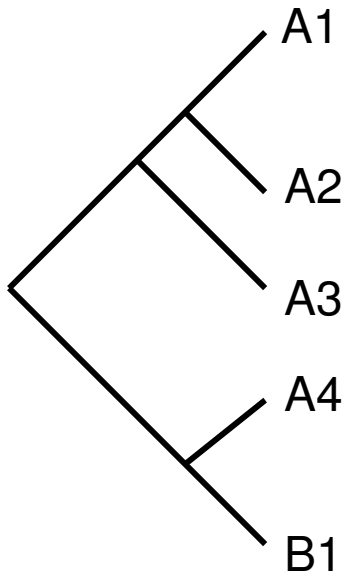
Monophyletic group consists of all descendants of ancestral taxa

Clade = monophyletic group

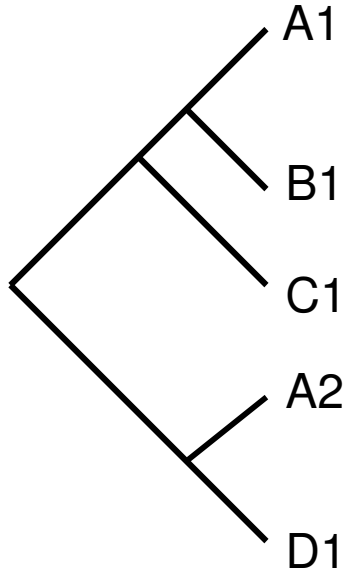
Genus A monophyletic
Genus B monophyletic




A is paraphyletic with respect to B

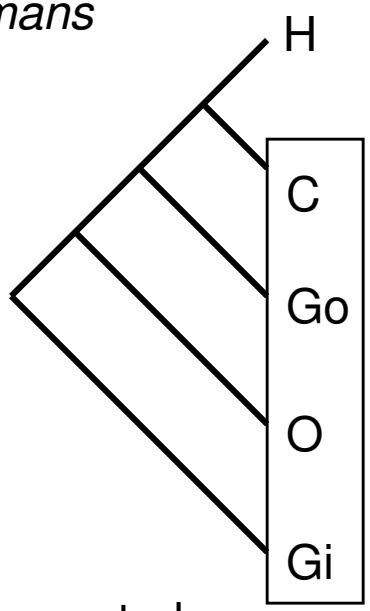



A is polyphyletic

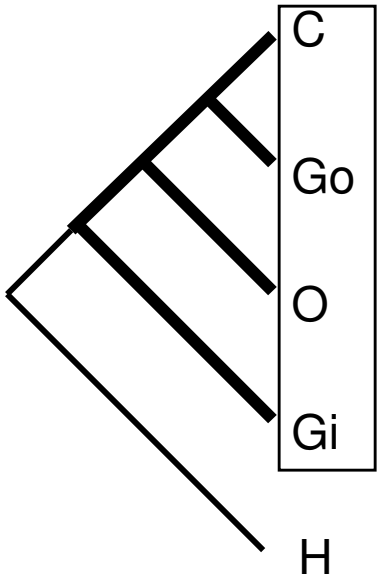


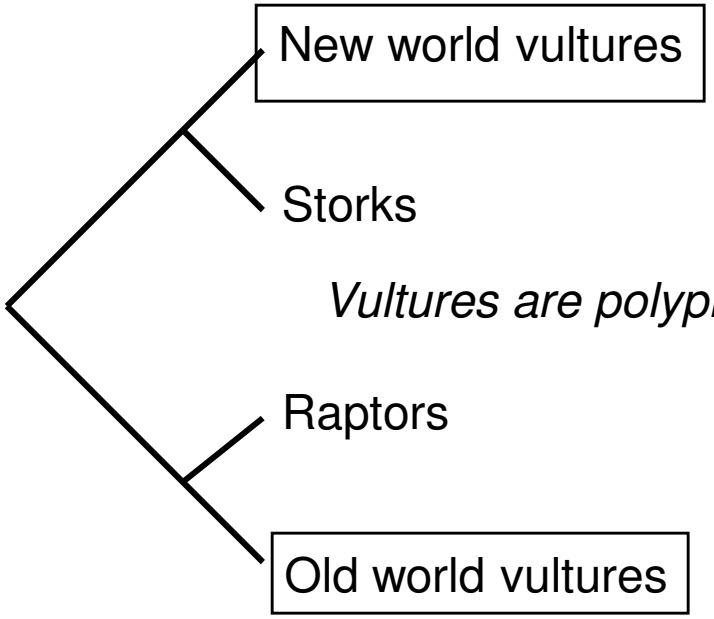

Are paraphyly and polyphyly artifacts of our biased taxonomy?

*Apes are paraphyletic to humans*

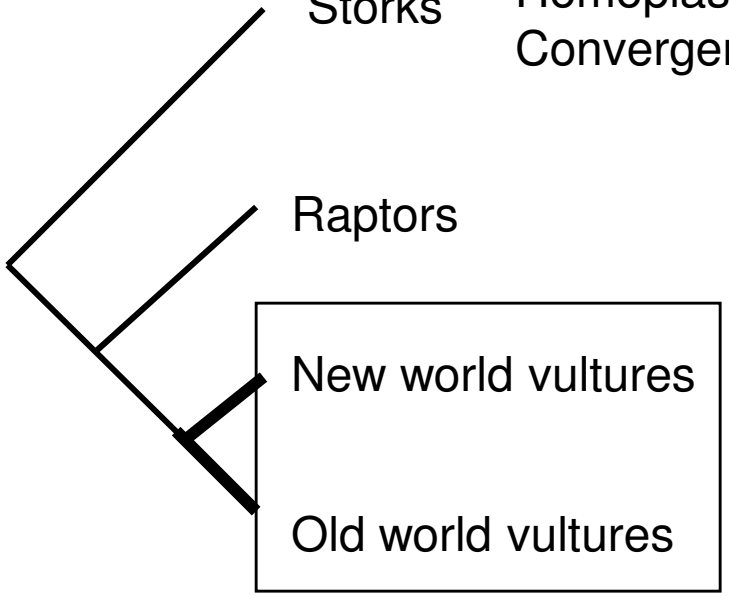



Unique
Shared ancestral



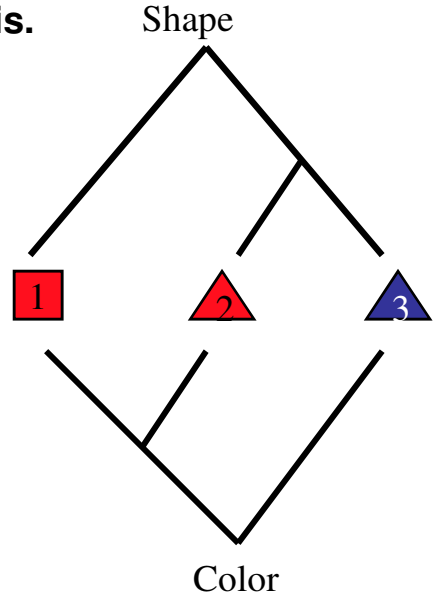

*Vultures are polyphyletic*




Homoplasy
Convergence

# Morphological (non-molecular) vs. molecular data

***Morphological data***
**What data to be recorded (soft tissue, skeletal, coat color, behaviour)**
**How those data are to be coded**
**Whether certain data points are to be included or excluded from analysis.**
**Lack of variation**
**Need experts**
**Advantage: Usually inexpensive**



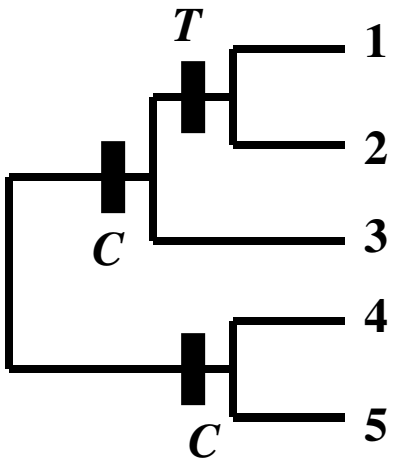

***Advantage of molecular data***
**Large number of characters**
**Easy to code data**
**High levels of variation**
**Molecular data is comparable across taxa…insects vs. vertebrates**
**Data acquisition technique identical across taxa**
**Less prone to errors and individual biases**
**Disadvantage: very expensive**

**So lets all do molecular systematics and abandon morphology based taxonomy?**

**If one makes the assumption that morphology based taxonomy is a good estimate of evolutionary relationship, and goes on to test certain hypothesis, then it might be problematic.**

**Phylogeny: Evolutionary tree that shows how different species are related to each other**

**Molecular phylogeny: phylogeny based on molecular data (DNA sequence)**




DNA sequence

| | |
|---|---|
| Species 1 | AAATCGGT |
| Species 2 | AAATCGGT |
| Species 3 | AAATCGGA |
| Species 4 | AACTGGGA |
| Species 5 | AACTGGGA |

Distance approach
UPGMA, NJ (PHYLIP, PAUP, MEGA)

Character state approach
Parsimony, Maximum likelihood, Bayesian
(PAUP, MEGA, Mr. Bayes)