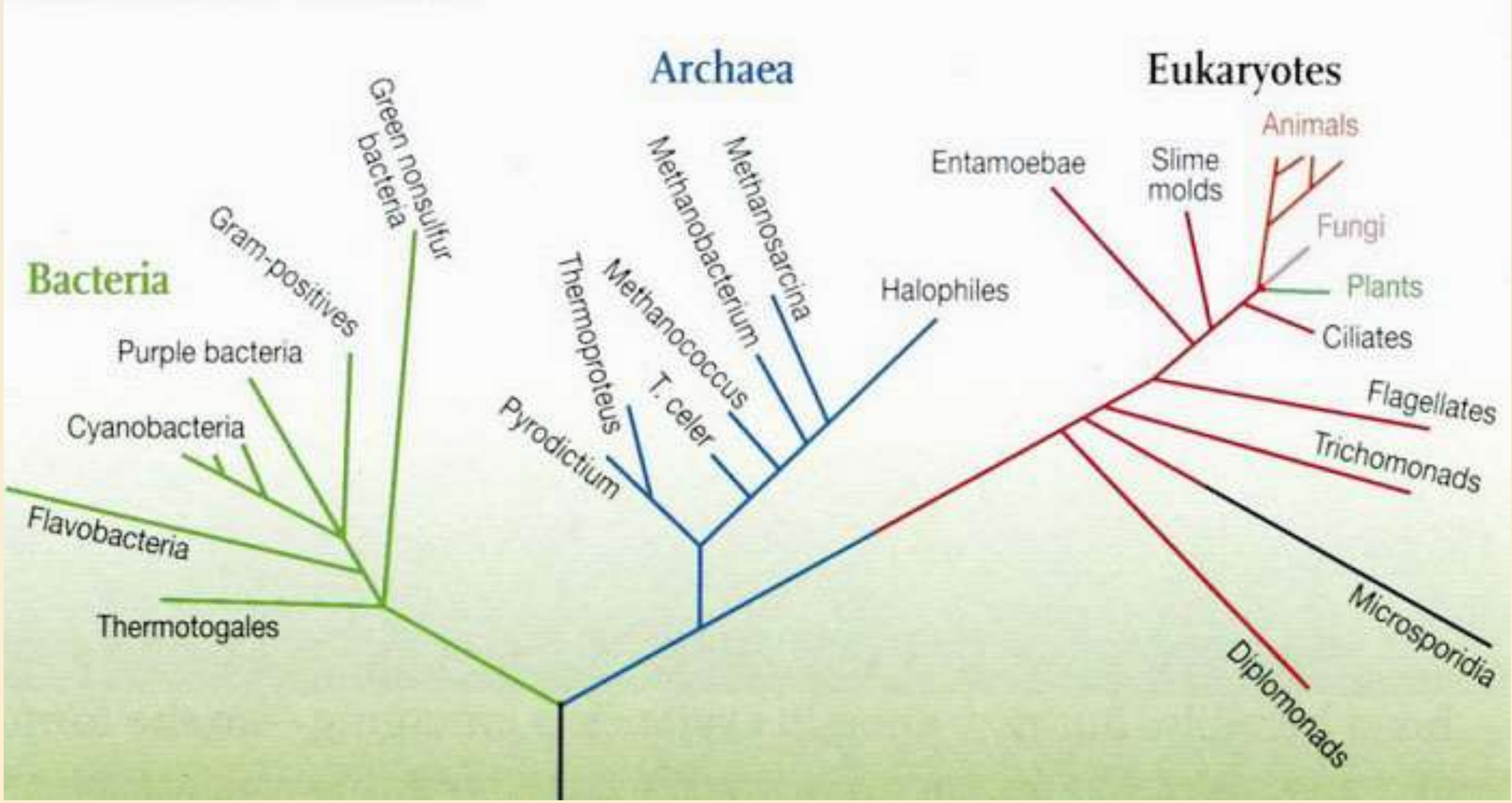
Reconstructing Phylogeny

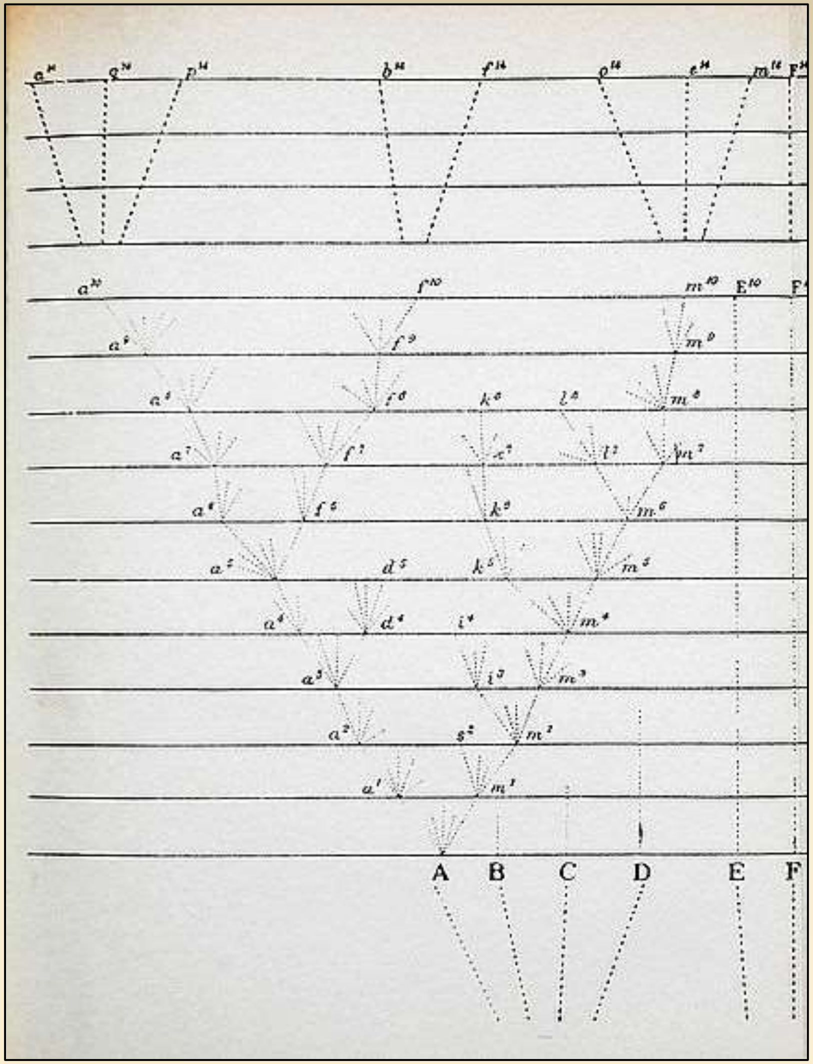
Classification
Phylogeny
Systematics



In *the Origin of Species*, Darwin included just one illustration — a "tree" depicting branching and extinction through time.

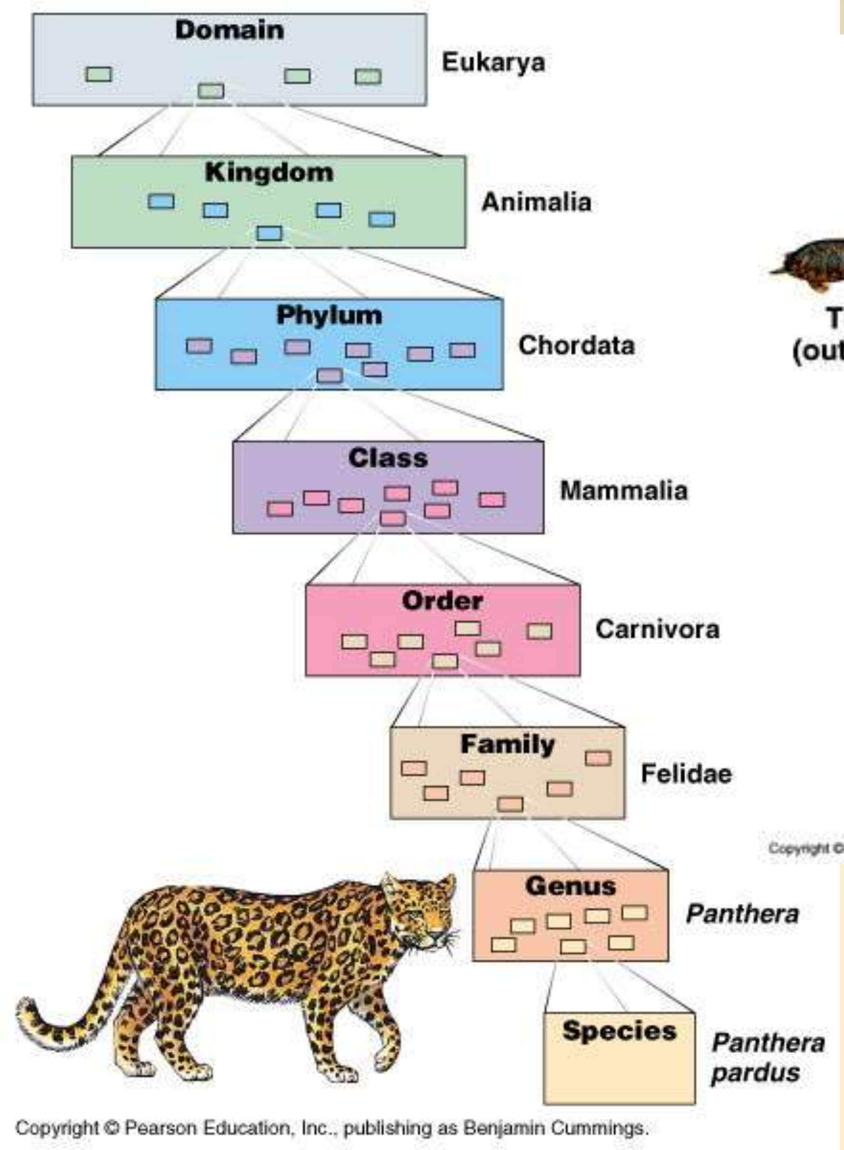
With this he crystallized the idea that species share common ancestors at various points back in time.

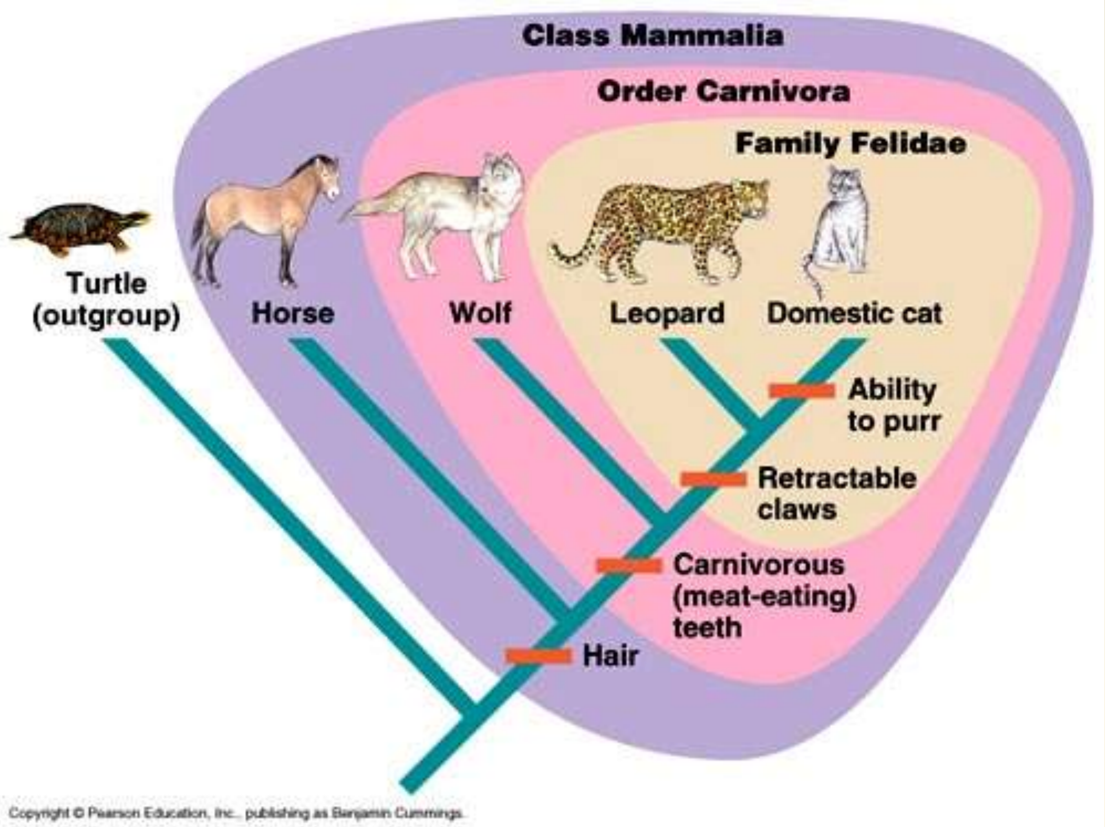
He referred to the genealogical relationships among all living things as **"the great Tree of Life."**



"The time will come, I believe, though I shall not live to see it, when we shall have fairly true genealogical trees of each great kingdom of Nature" - Charles Darwin

Nested box-within-box hierarchy is consistent with descent from a common ancestor, used as evidence by Darwin.





The Importance of Phylogenetic Trees

1. Increasing use of phylogenetic trees in the biological sciences.
2. Need to know what tree diagrams do and *do not* communicate.
3. Provide an efficient structure for organizing biodiversity info.
4. Develop accurate conception of totality of evolutionary history.
5. Important for aspiring biologists to develop this understanding.

Phylogenetic Tree of Life

Why is phylogeny important?

Understanding and classifying the diversity of life on Earth

Testing evolutionary hypotheses:

- test relationships
- 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!

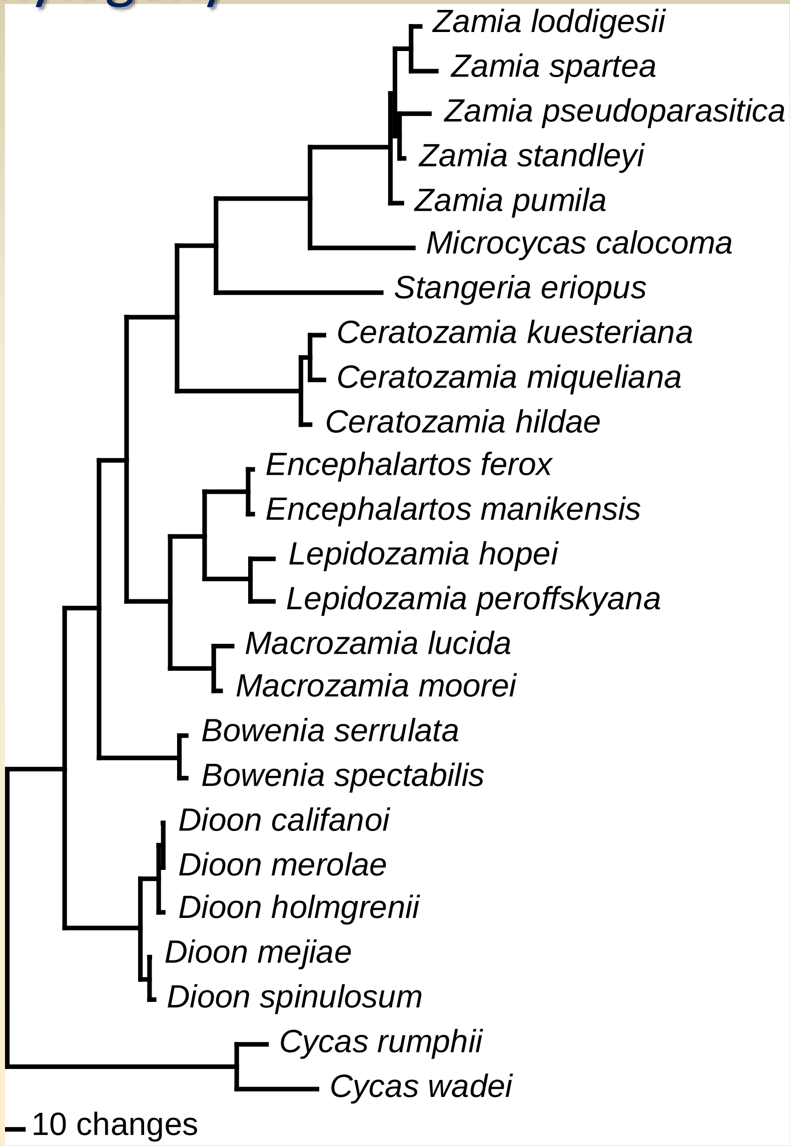
Uses of phylogenies: Taxonomy

e.g. Cycad Phylogeny

Similar organisms are grouped together

Clades share common evolutionary history

Phylogenetic classification names clades



Bogler & Francisco-Ortega. 2004. Bot. Rev.: 70.

ITS1 and ITS2

Strict Consensus

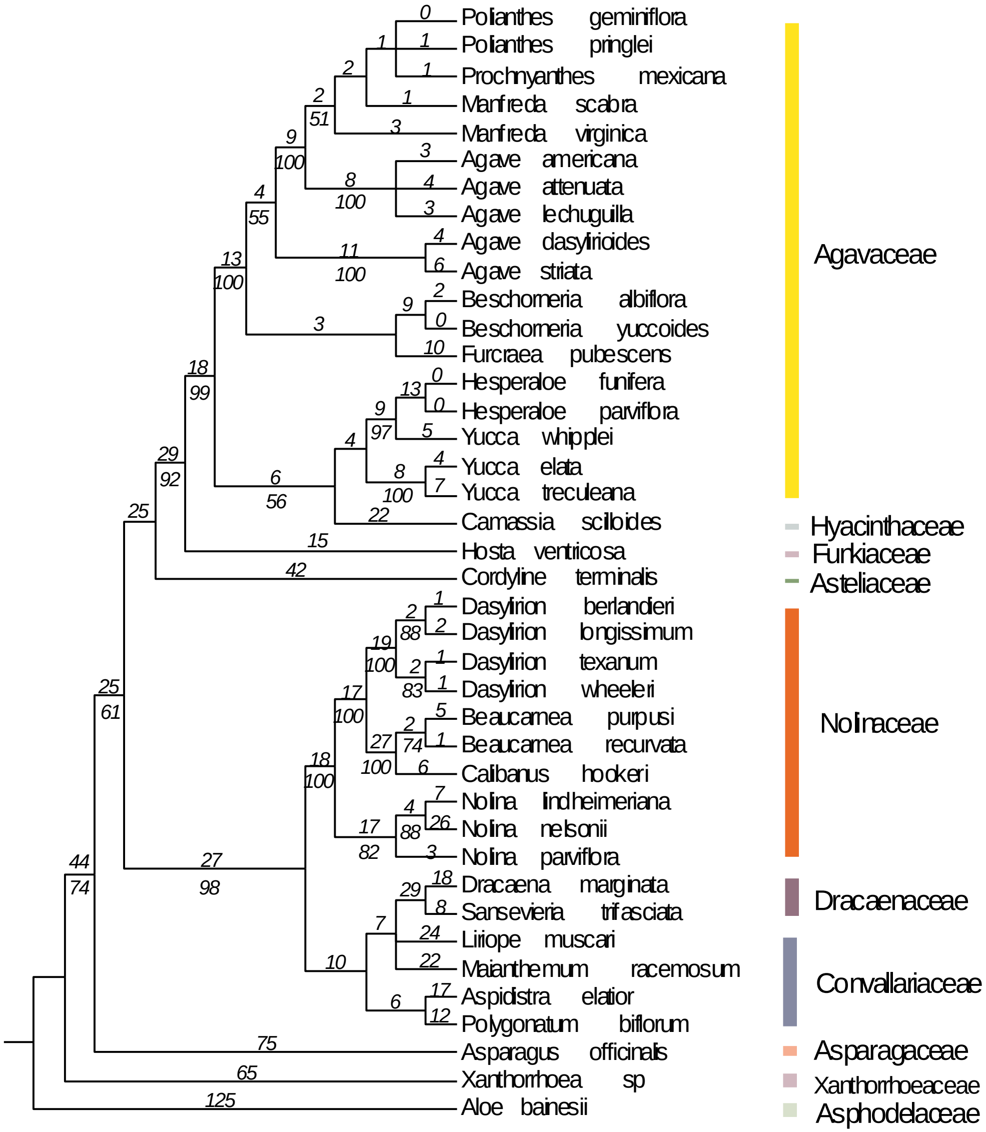
4 Trees

979 Steps

CI = 0.659

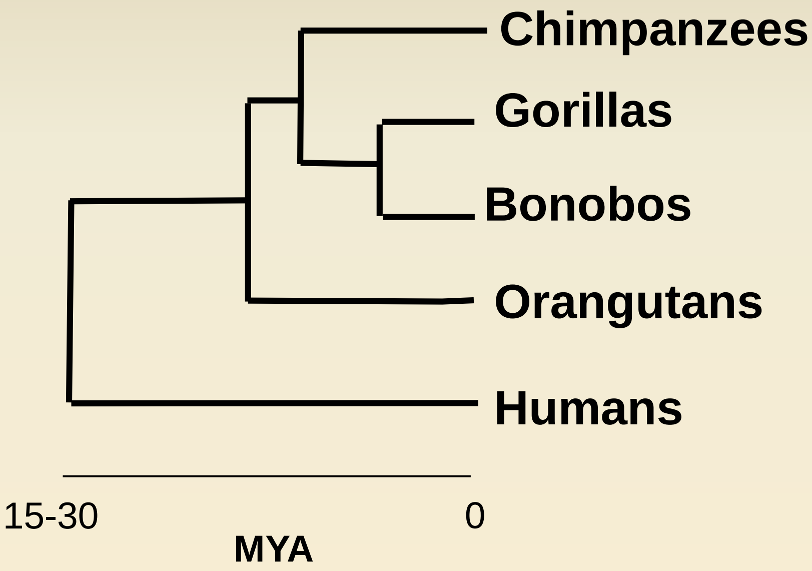
RI = 0.815

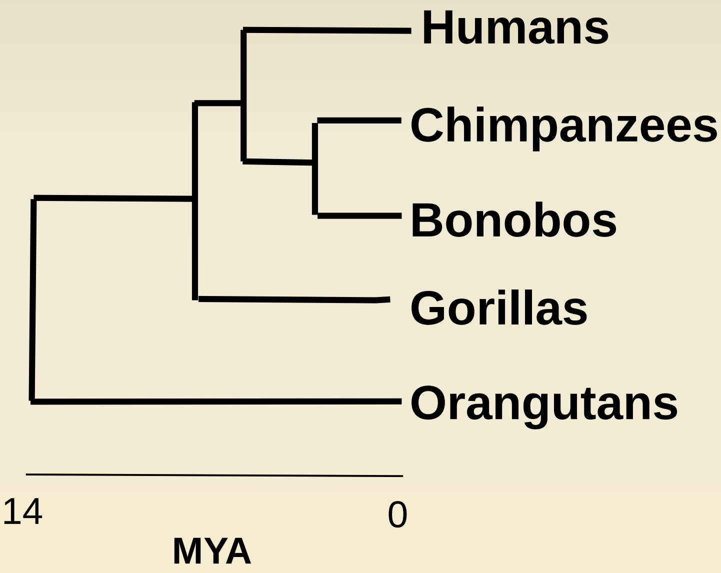
Combined Data: ITS1 and ITS2



Bogler and Simpson. 1996. AJB 83: 1225-1235.

Which species are the closest living relatives of modern humans?





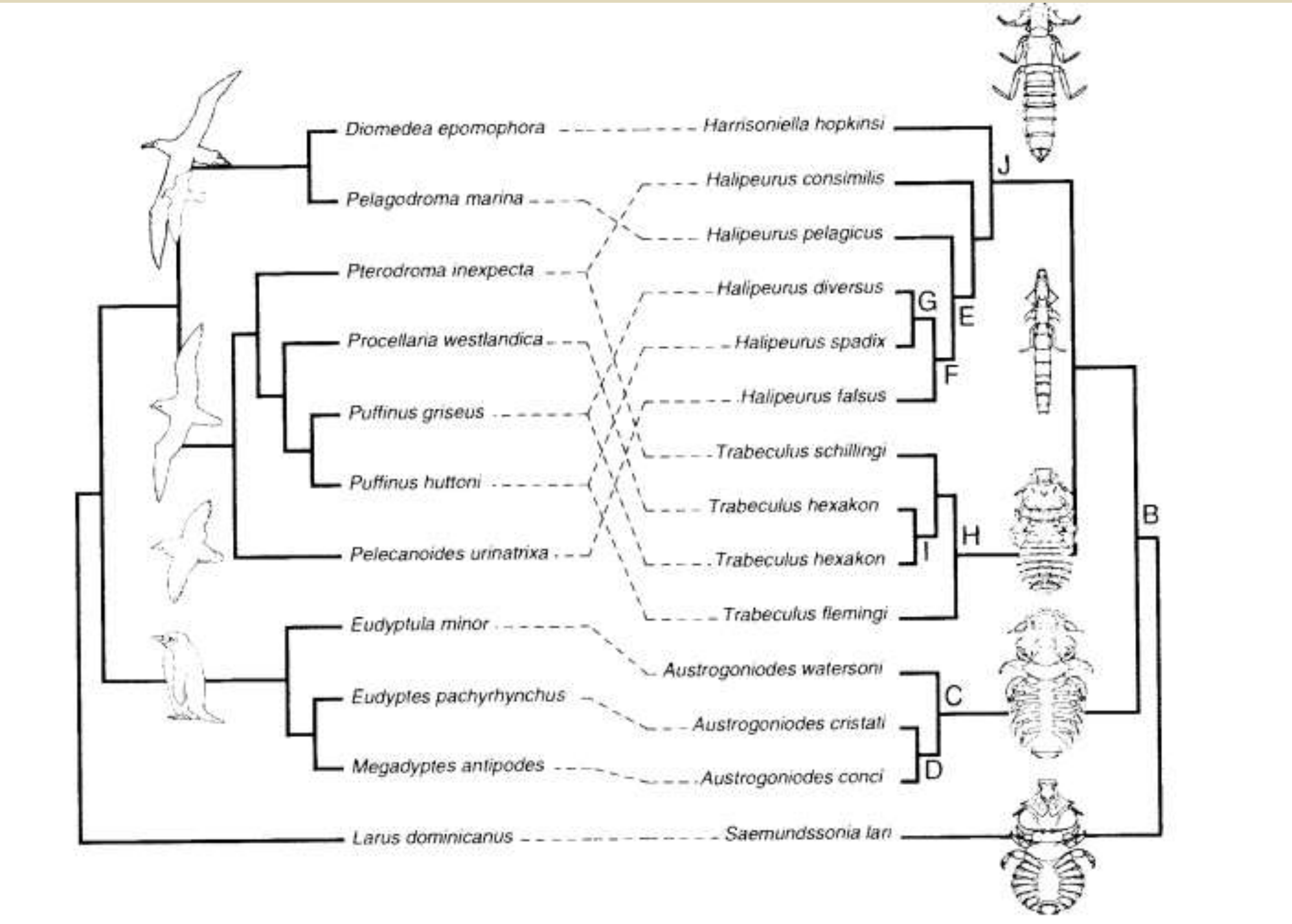
The pre-molecular view was that the great apes (chimpanzees, gorillas and orangutans) formed a clade separate from humans, and that humans diverged from the apes at least 15-30 MYA.

Mitochondrial DNA, most nuclear DNA-encoded genes, and DNA/DNA hybridization all show that bonobos and chimpanzees are related more closely to humans than either are to gorillas.

Uses of phylogenies: Co-evolution

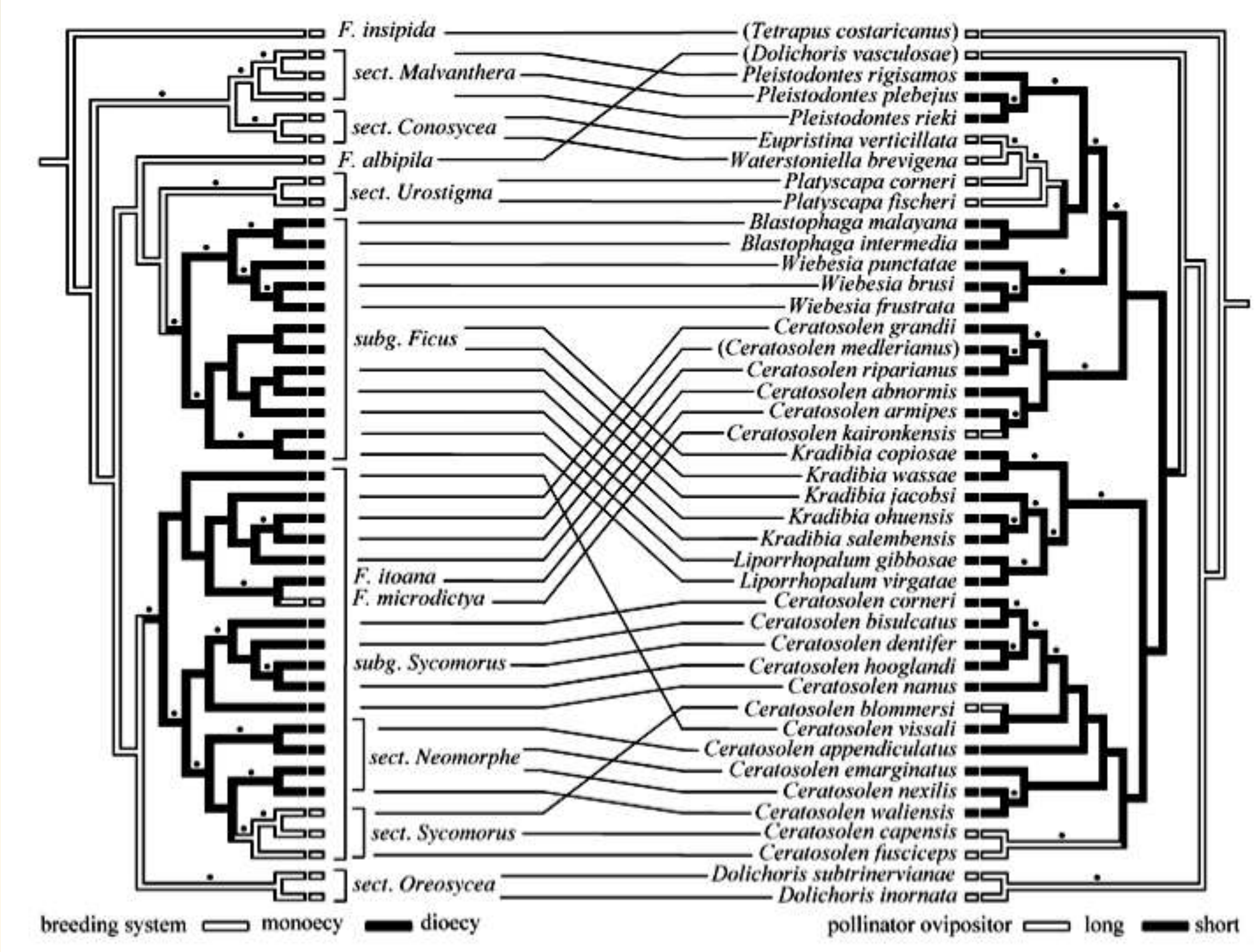
- Compare divergence patterns in two groups of tightly linked organisms (e.g. hosts and parasites or plants and obligate pollinators)
  - – Look at how similar the two phylogenies are
  - – Look at host switching
- Evolutionary arms races
  - – Traits in one group track traits in another group
    - e.g. toxin production and resistance in prey/predator or plant/herbivore systems, floral tube and proboscis length in pollination systems

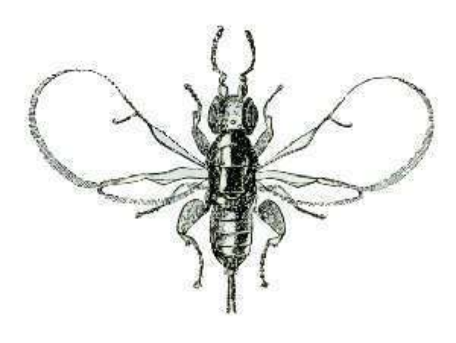
Example of host-parasite phylogeny
Seabirds and their Lice

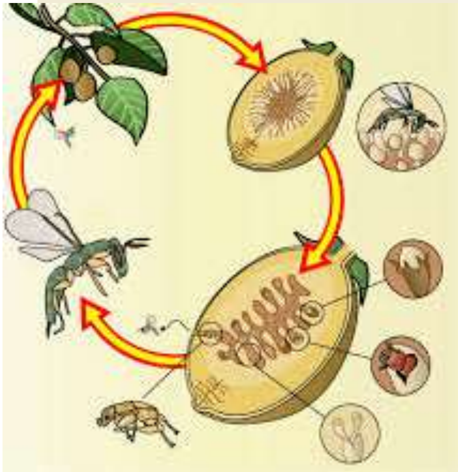


Source: Page, R.D.M., Cruickshank, R.H., Dickens, M., Furness, R.W., Kennedy, M., Palma, R.L., Smith, V.S. 2004. Phylogeny of "*Philoceanus* complex" seabird lice (Phthiraptera: Ischnocera) inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, 30: 633-652.

Example of plant-pollinator phylogeny
Figs and Fig Wasps

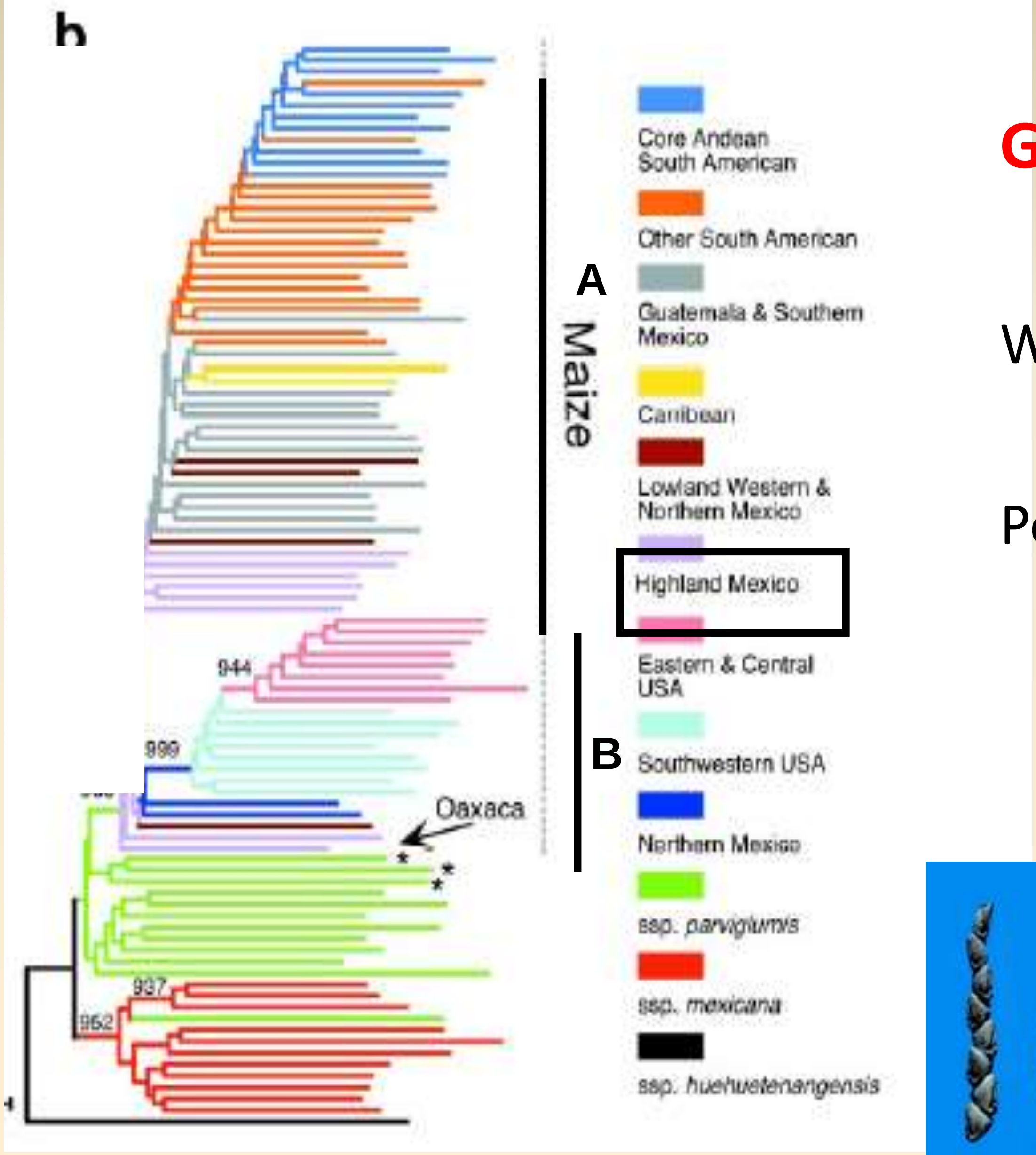






Source: George D Weiblen. Correlated Evolution in Fig Pollination. Systematic Biology 53(1): 128139, 2004.

Testing evolutionary hypotheses



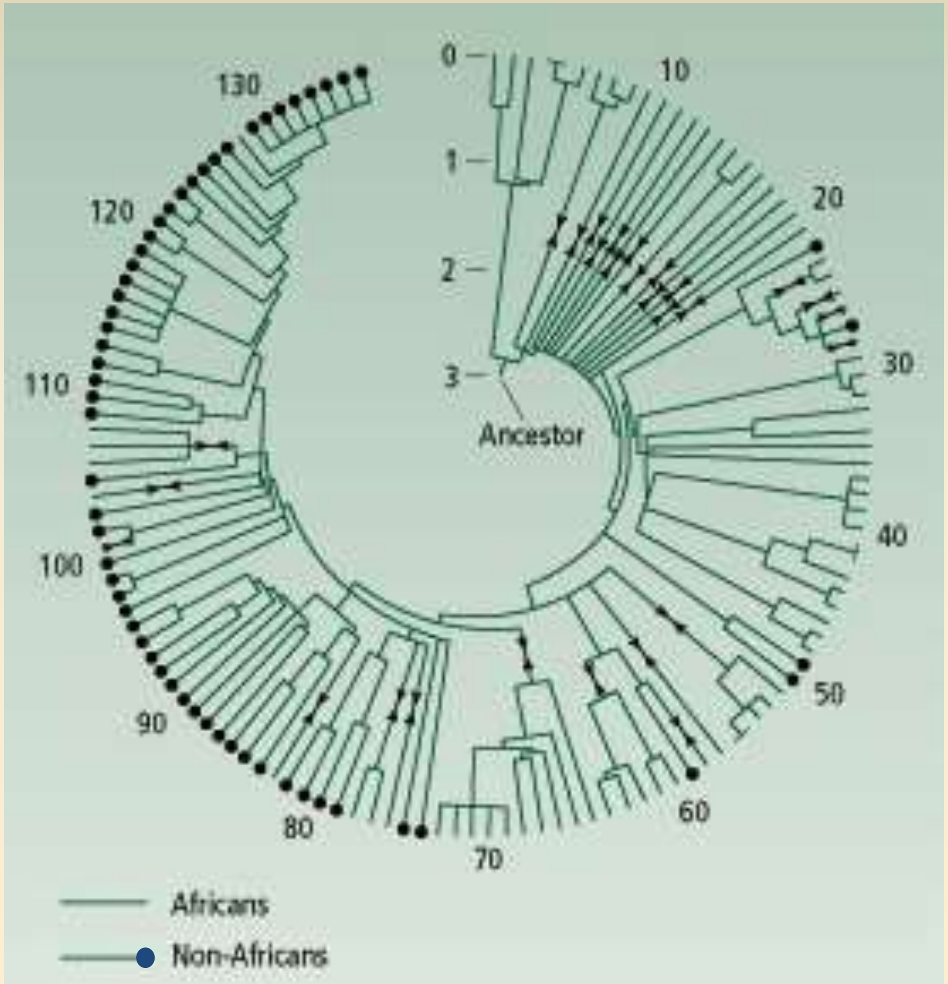
Matsuoka et al. (2002)

Geographic origins of Maize

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

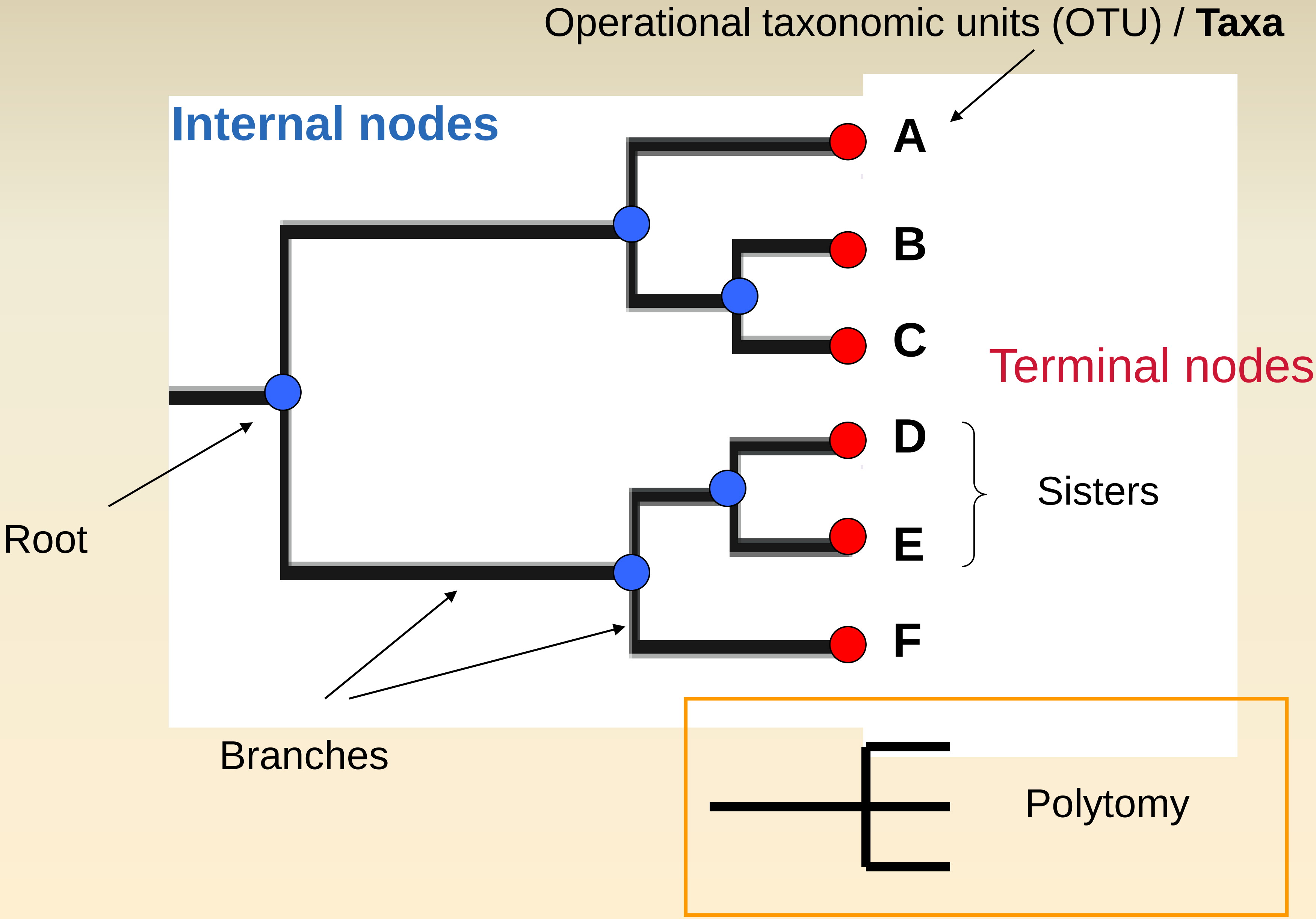
Where did humans originate?

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

Tree Terminology



Operational taxonomic units (OTU) / **Taxa**

Internal nodes

Terminal nodes

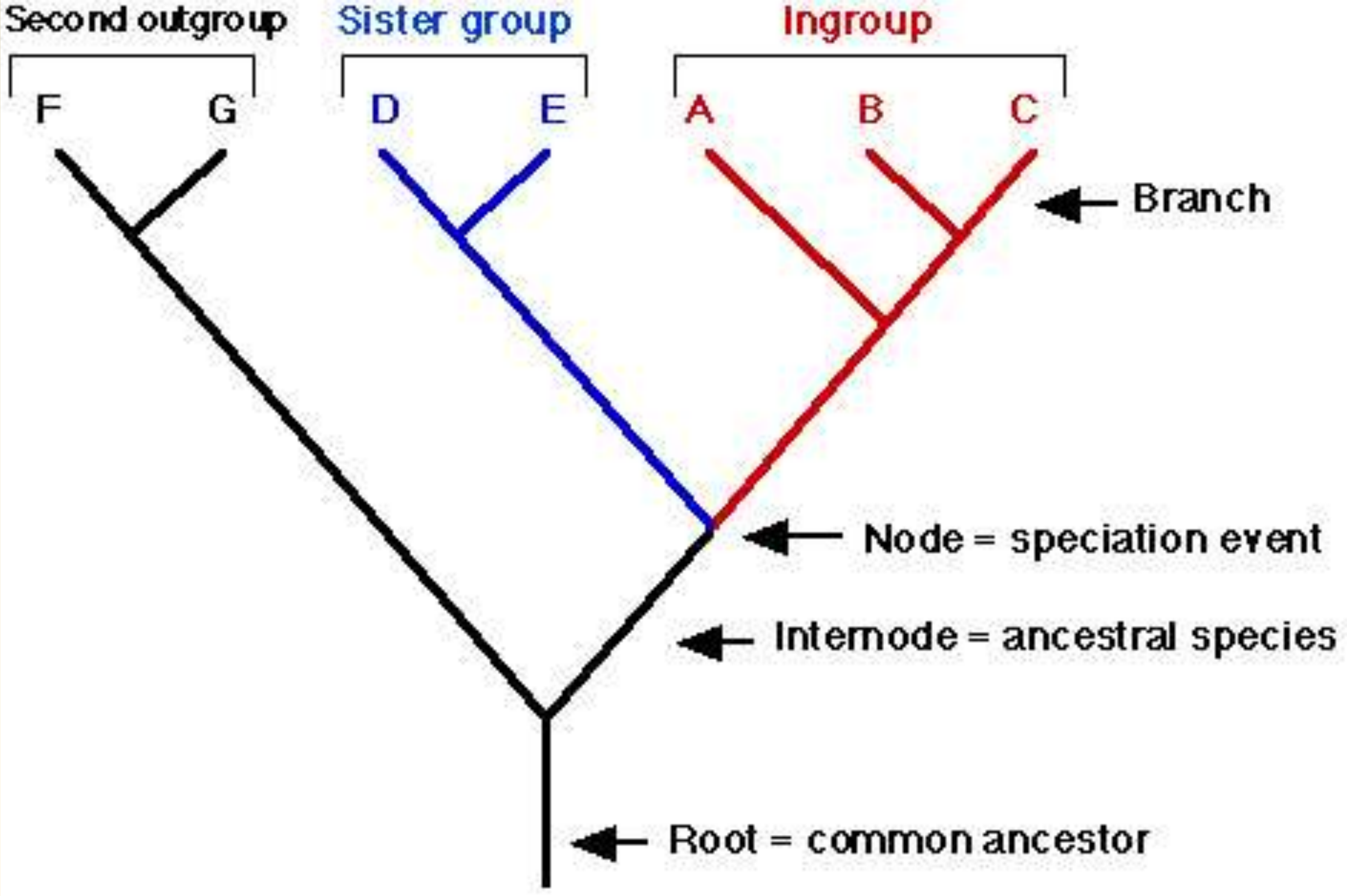
Sisters

Root

Branches

Polytomy

Tree Terminology



Sister Groups

A **Sister Group** is a pair of taxa that are most closely related to each other.

Humans are most closely related to chimpanzees, so humans & chimpanzees form a sister group.

Gorillas form a sister group to the clade containing humans and chimpanzees.

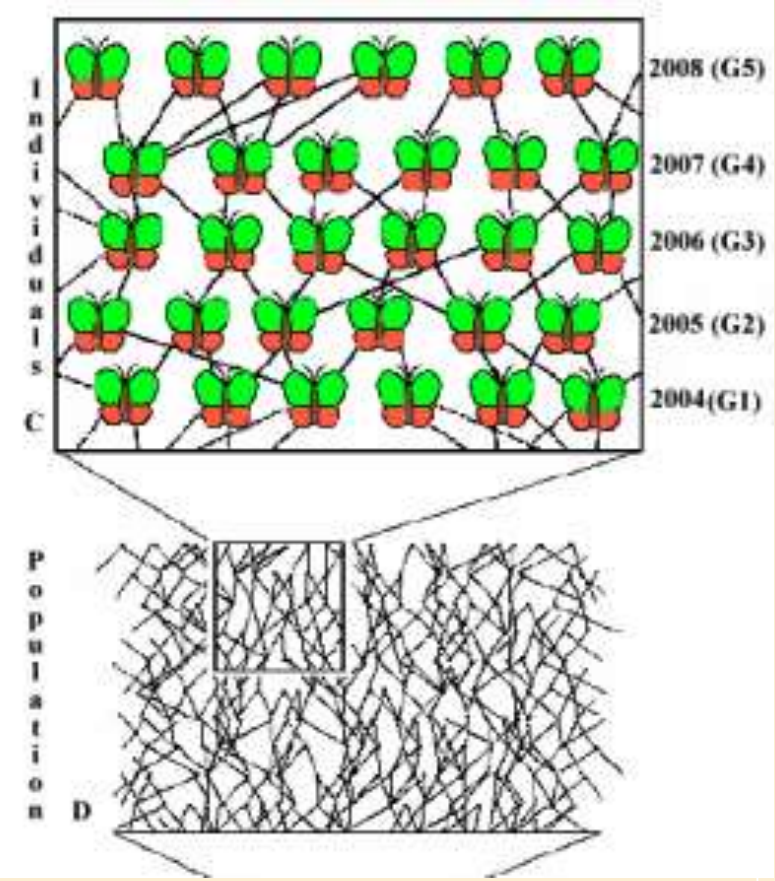
Ingroup — the group of organisms of primary interest.

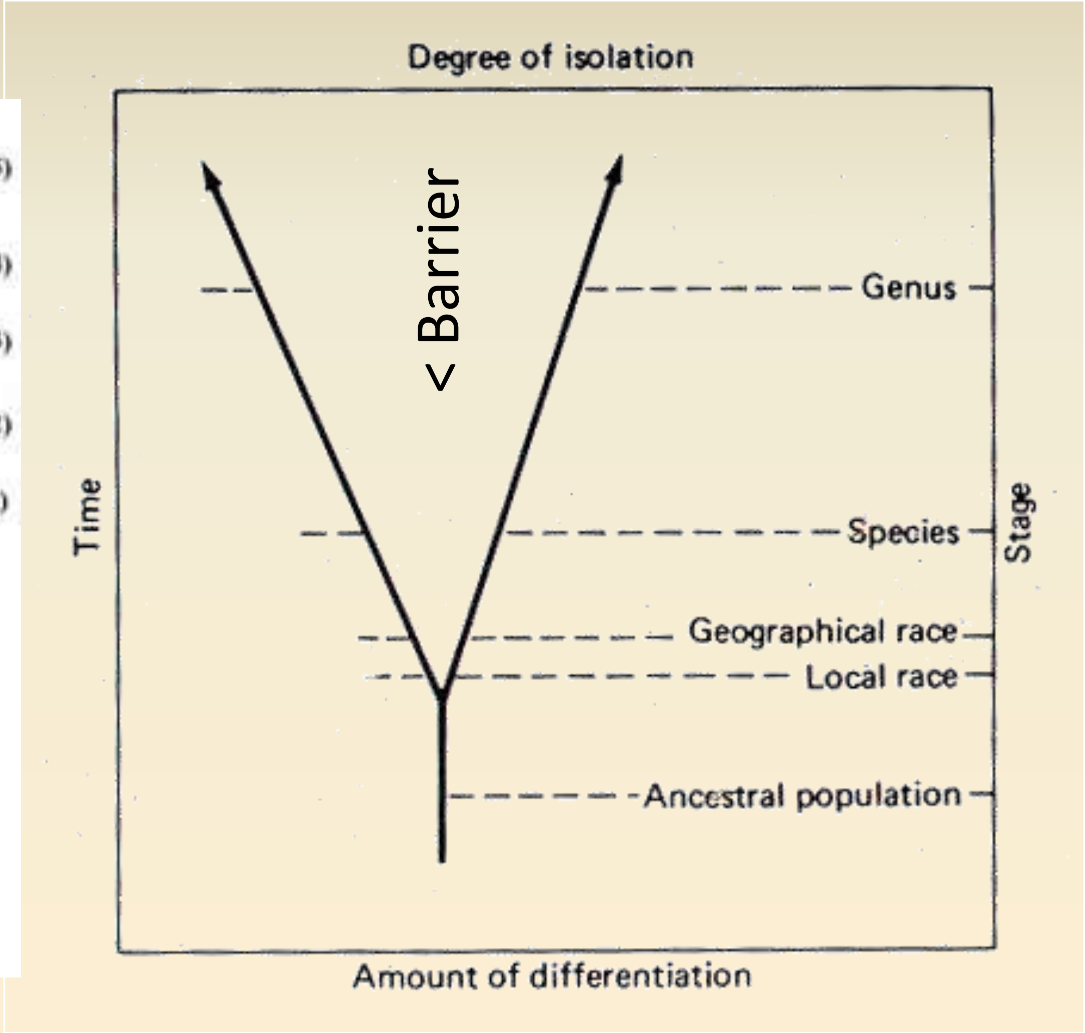
Outgroup — species or group known to be closely related to, but phylogenetically outside, the group of interest.

- Used to root the tree. Helps establish the direction of evolutionary change, the polarity of a character.

Tree Branches
populations interbreeding

Tree Nodes
Speciation Events





These trees depict equivalent relationships despite different styles

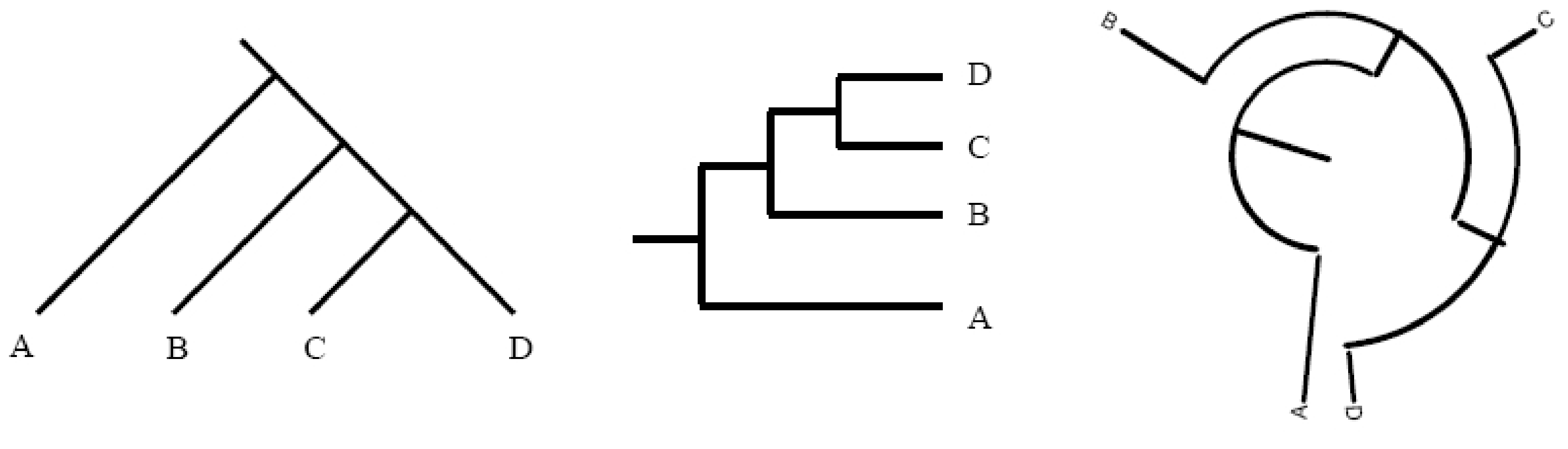
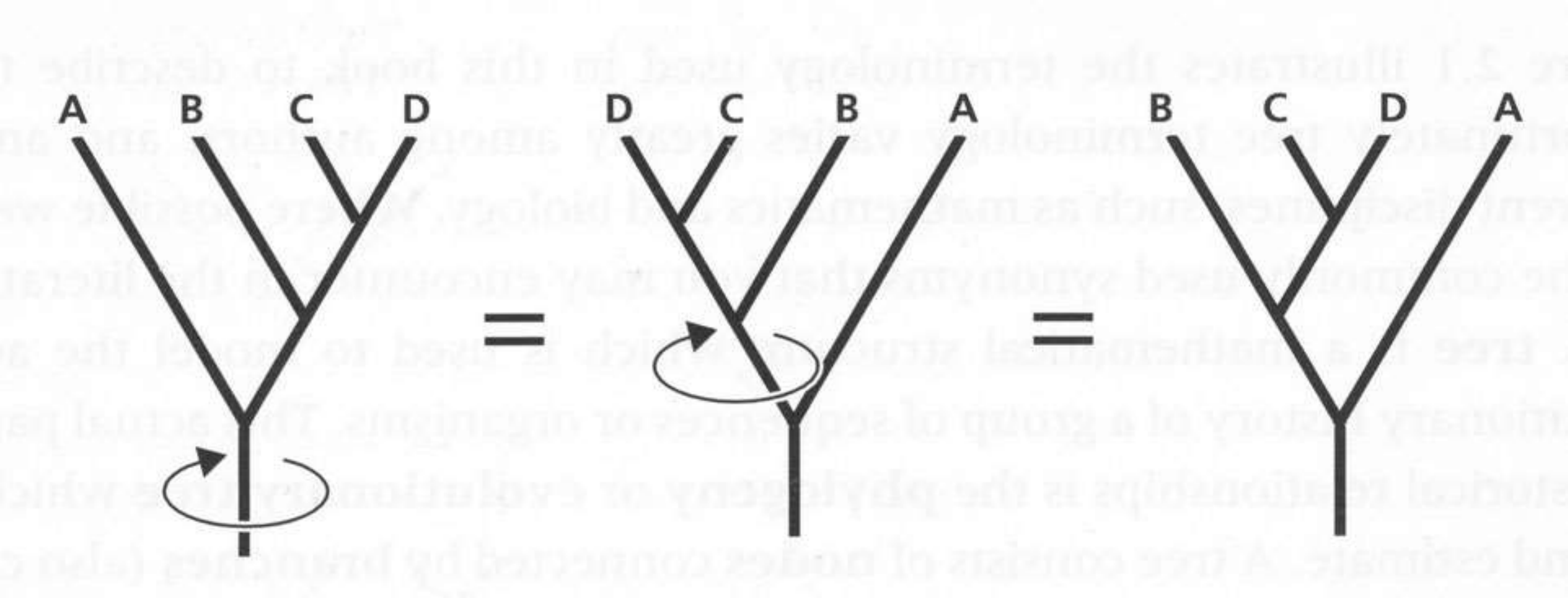


Figure 6 : These trees depict equivalent relationships despite being different in style.

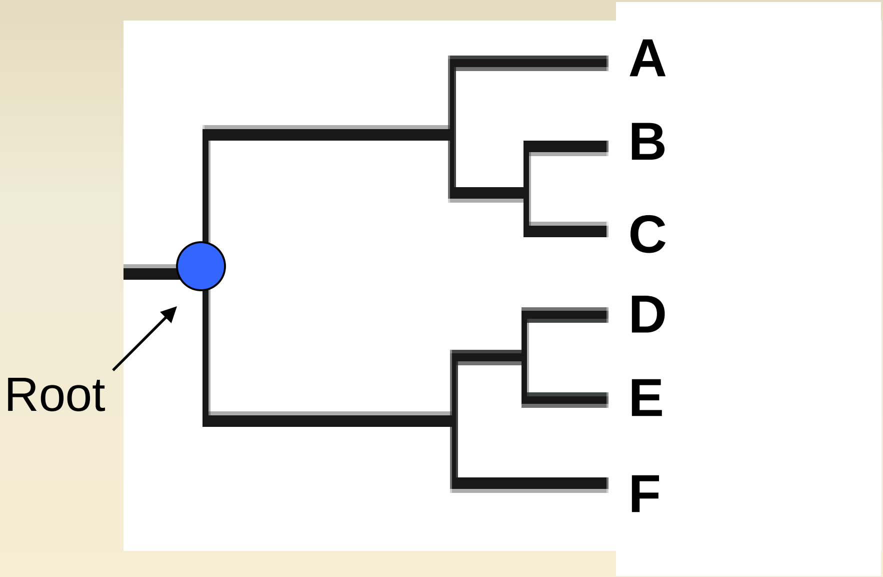
Branches can rotate around nodes…..

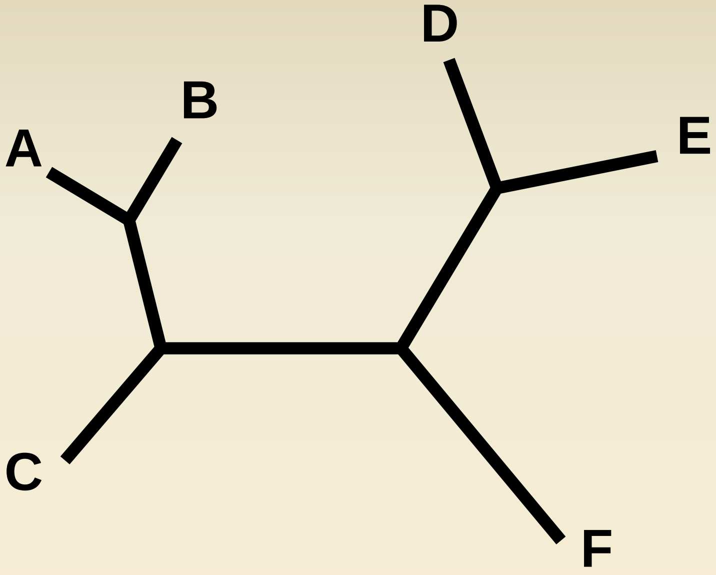


Three different representations of the same tree

Tree Terminology

Rooted vs. Unrooted trees



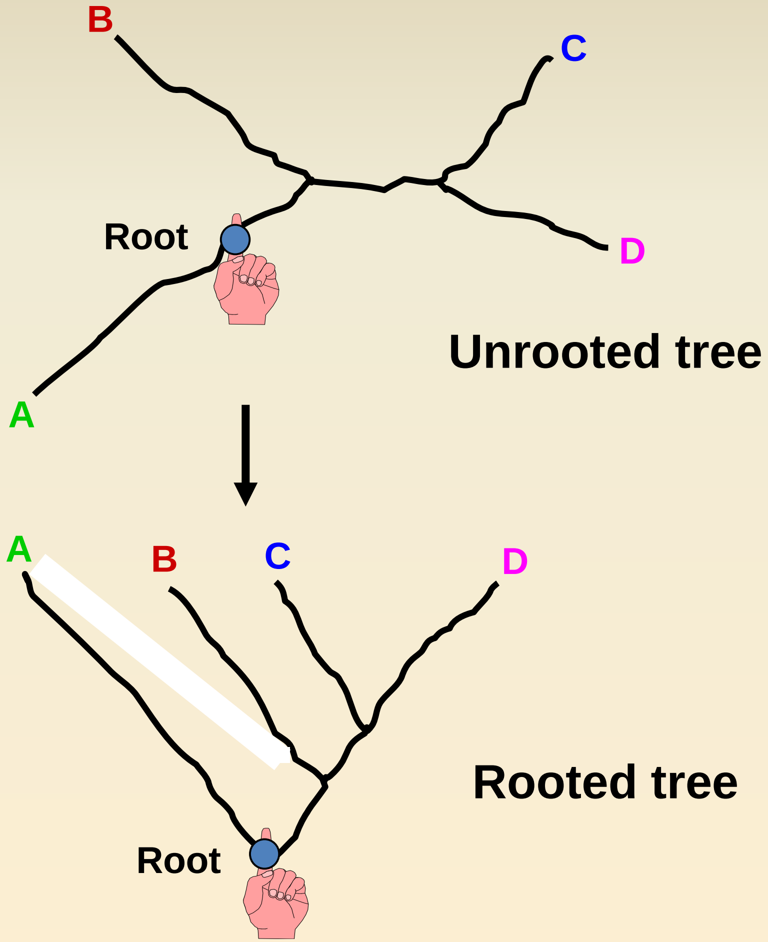


__Rooted trees:__ Has a root that denotes common ancestry

__Unrooted trees:__ Only specifies the degree of kinship among taxa but not the evolutionary path

Inferring evolutionary *relationships* between the taxa requires rooting the tree:

To root a tree mentally, imagine that the tree is made of string. ●Grab the string at the root and tug on it until the ends of the string (the taxa) fall opposite the root:

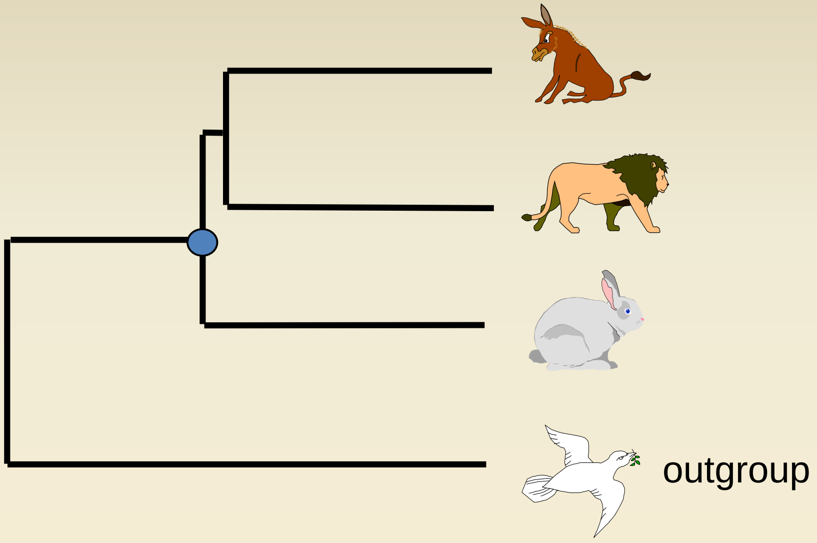


Note that in this rooted tree, taxon A is no more closely related to taxon B than it is to C or D.

There are two major ways to root trees:

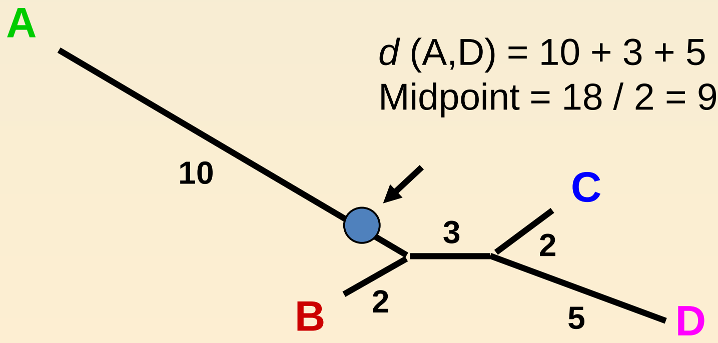
By outgroup:

Uses taxa (the “outgroup”) that are known to fall outside of the group of interest (the “ingroup”). Requires some prior knowledge about the relationships among the taxa. The outgroup can either be species (*e.g.*, birds to root a mammalian tree) or previous gene duplicates (*e.g.*, α-globins to root β-globins).



By midpoint or distance:

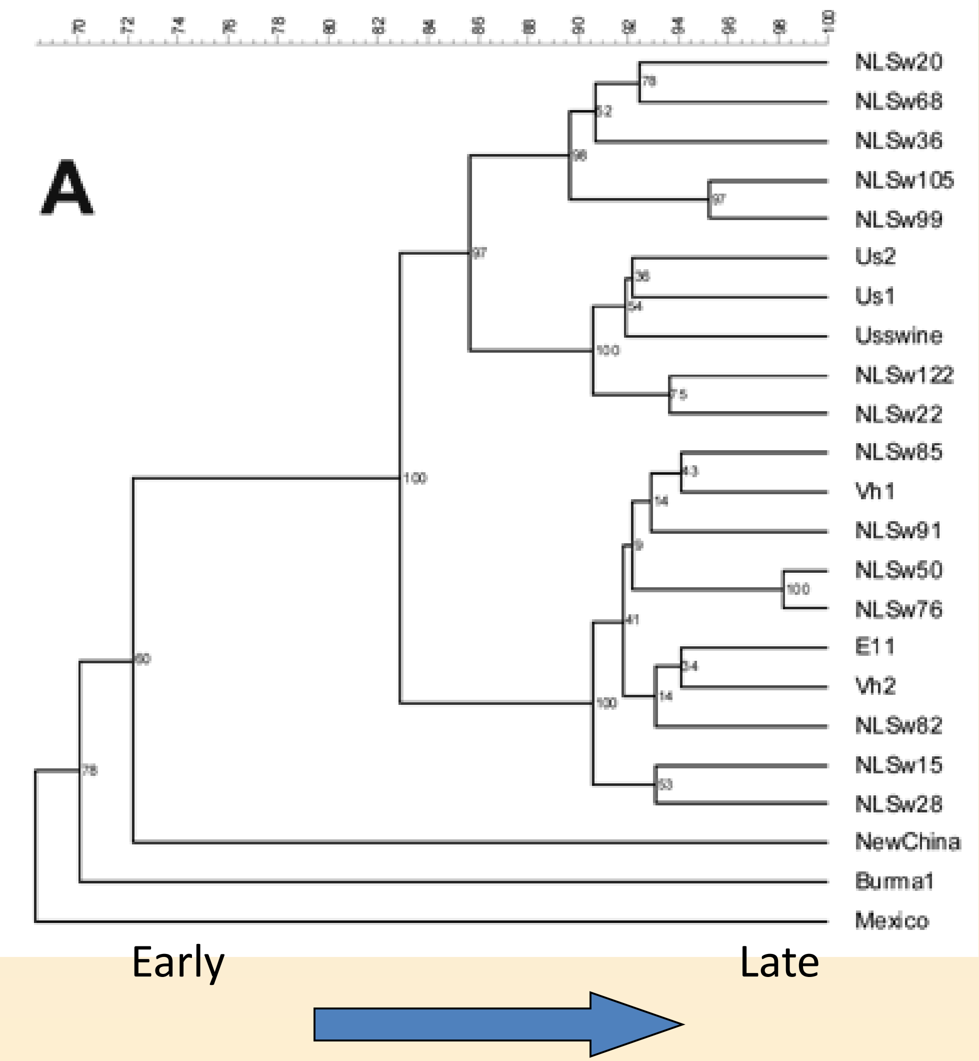
Roots the tree at the midway point between the two most distant taxa in the tree, as determined by branch lengths. Assumes that the taxa are evolving in a clock-like manner. This assumption is built into some of the distance-based tree building methods.



d (A,D) = 10 + 3 + 5 = 18
Midpoint = 18 / 2 = 9

Based on lectures by C-B Stewart,
and by Tal Pupko

Trees: rooted vs. unrooted



- A rooted tree has a single node (the root) that represents a point in time that is earlier than any other node in the tree.
- A rooted tree has directionality (nodes can be ordered in terms of “earlier” or “later”).
- In the rooted tree, distance between two nodes is represented along the time-axis only (the second axis just helps spread out the leafs)

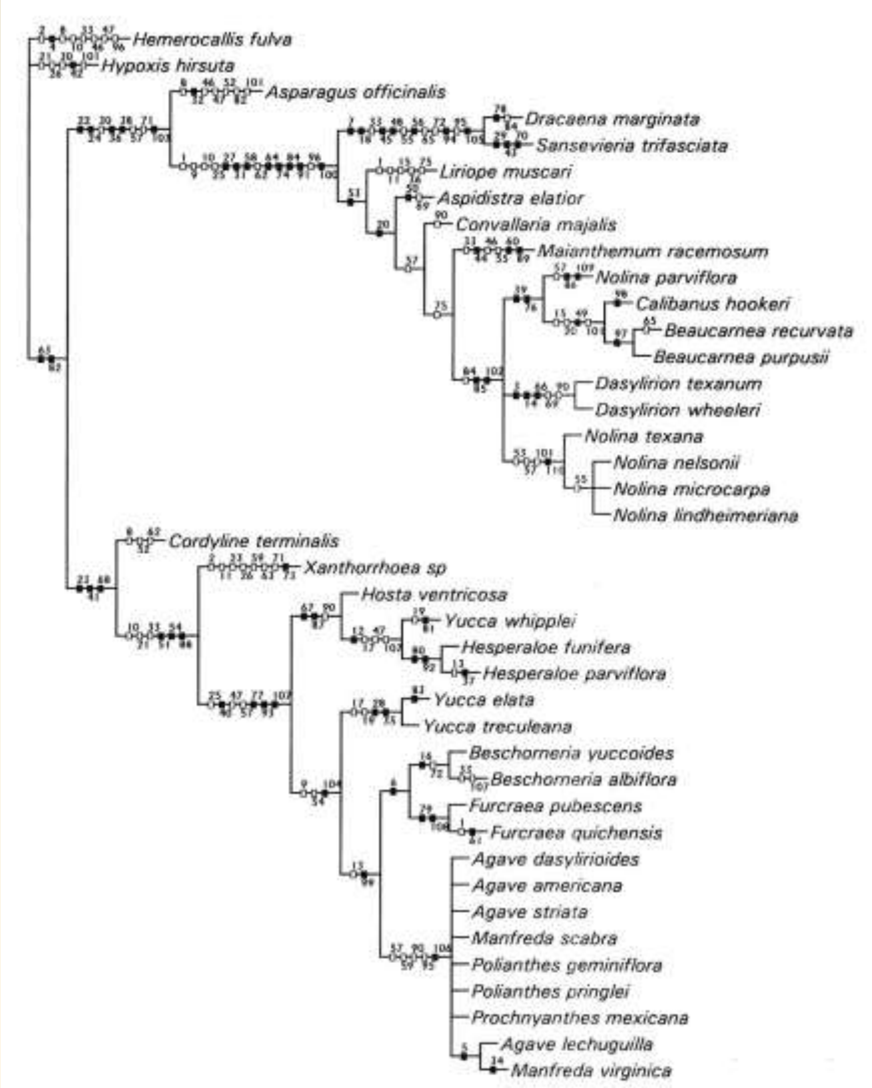
Dendrogram is a broad term for the diagrammatic representation of a phylogenetic tree.

Cladogram is a phylogenetic tree formed using cladistic methods. This type of tree only represents a *branching pattern*; i.e., its branch spans do not represent time or relative amount of character change.

Phylogram is a phylogenetic tree that has branch *spans proportional to the amount of character* change.

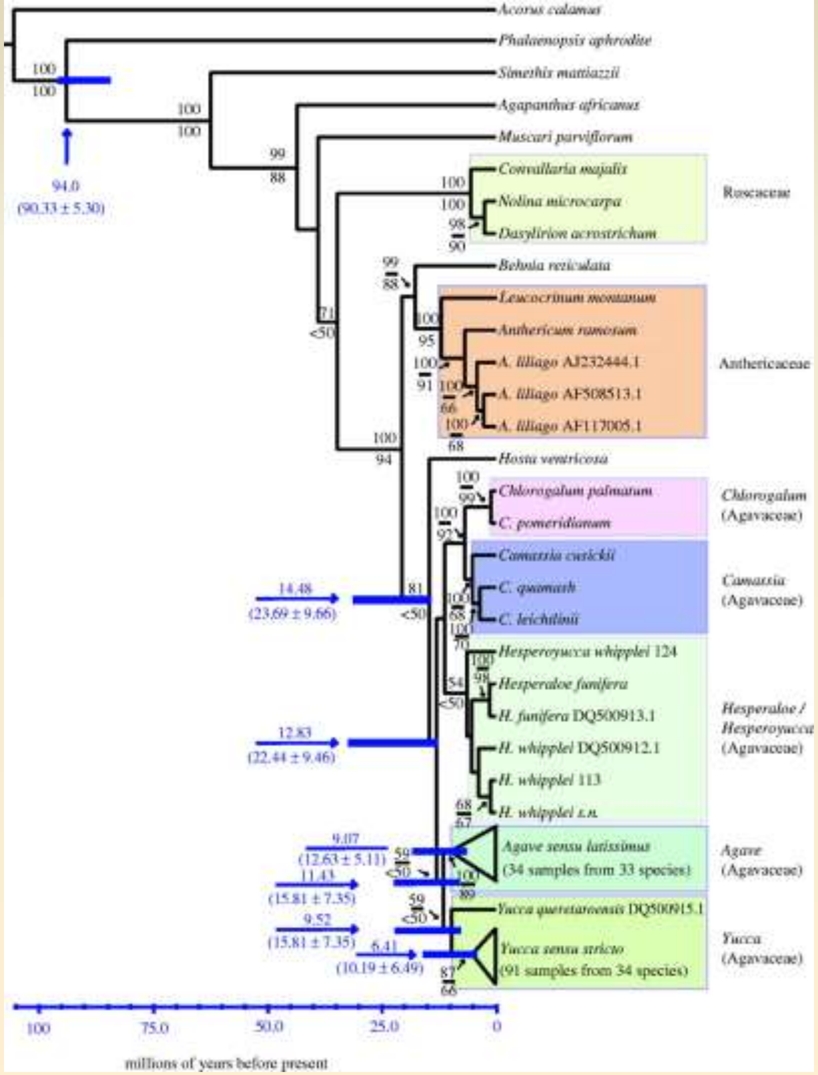
Chronogram is a phylogenetic tree that explicitly represents *evolutionary time* through its branch spans.

cpDNA Restriction Sites Phylogram



Bogler and Simpson. 1995. Syst. Bot. 20: 191

Chronogram of the Agavaceae based on the Bayes consensus tree derived from 153 cpDNA sequences from the trnL gene and the trnL–trnF intergenic spacer.



Smith C I et al. Proc. R. Soc. B 2008;275:249-258

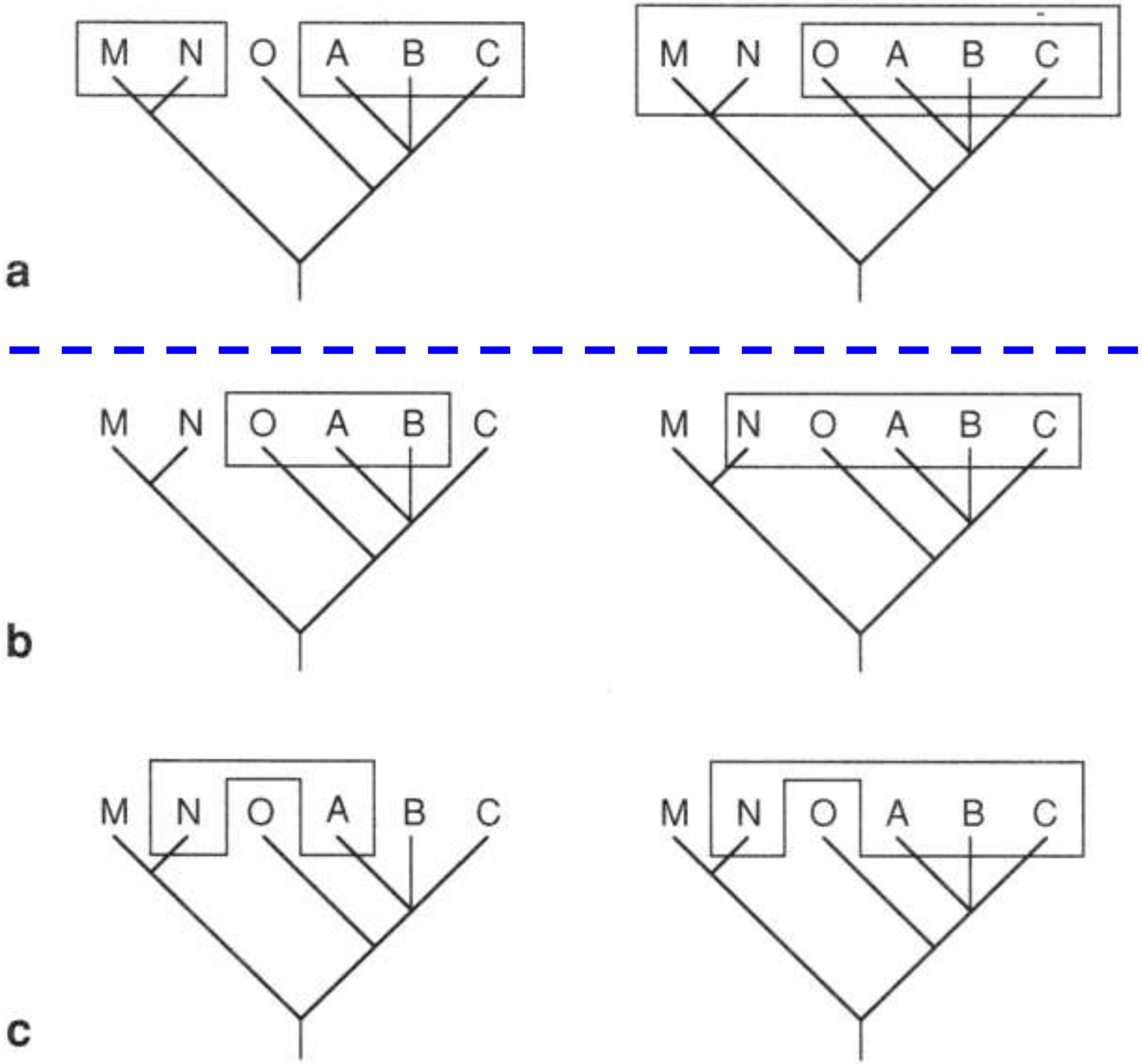


Fig. 1.1.—Examples of monophyletic (**a**), paraphyletic (**b**), and polyphyletic (**c**) groups.

Monophyly
(monophyletic)

Paraphyly
(paraphyletic)

Polyphyly
(polyphyletic)

Phylogeny and classification

Monophyly

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

Asteroidea

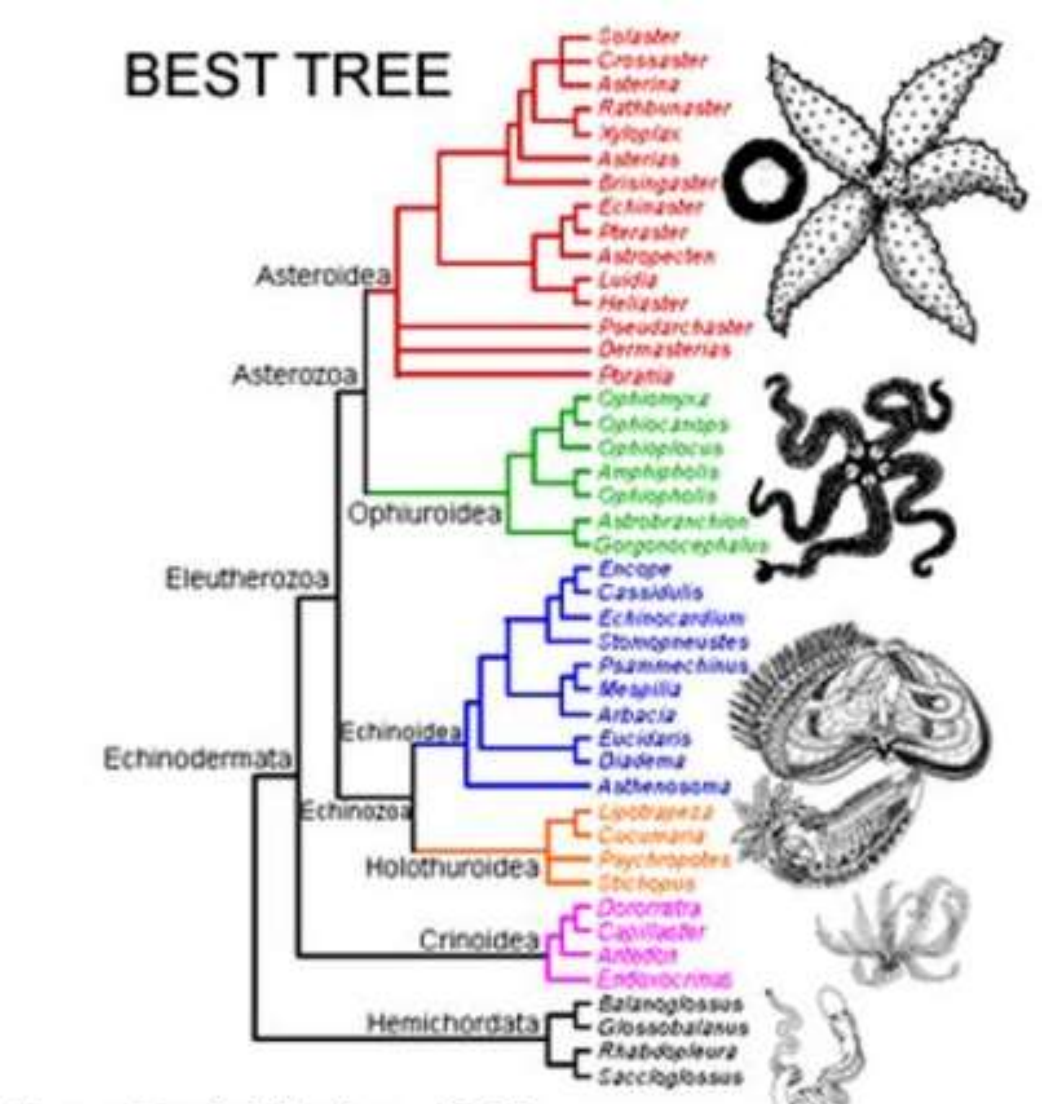
Ophiuroidea

Echinoidea

Holothuroidea

Crinoidea

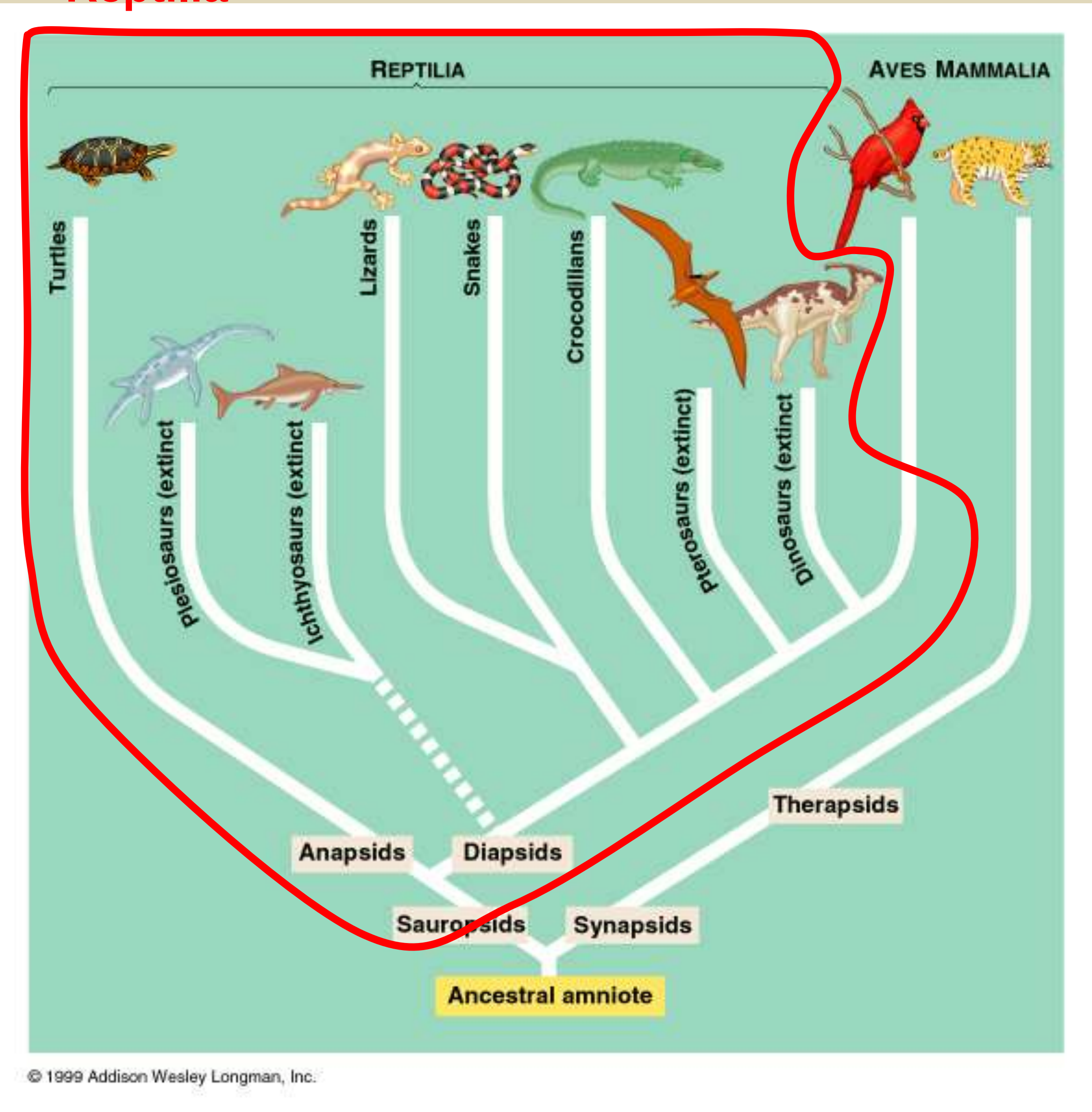
Each group shares a common ancestor that is not shared by any members of another group



From Daniel Janies. 2000.

Paraphyletic groups

Reptilia



Paraphyly

Birds are more closely related
to crocodilians than to other extant vertebrates

Archosauria = Birds + Crocs

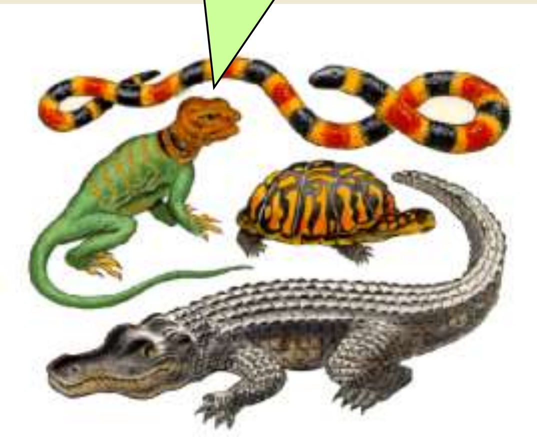
We think of reptiles as turtles,
lizards, snakes, and crocodiles

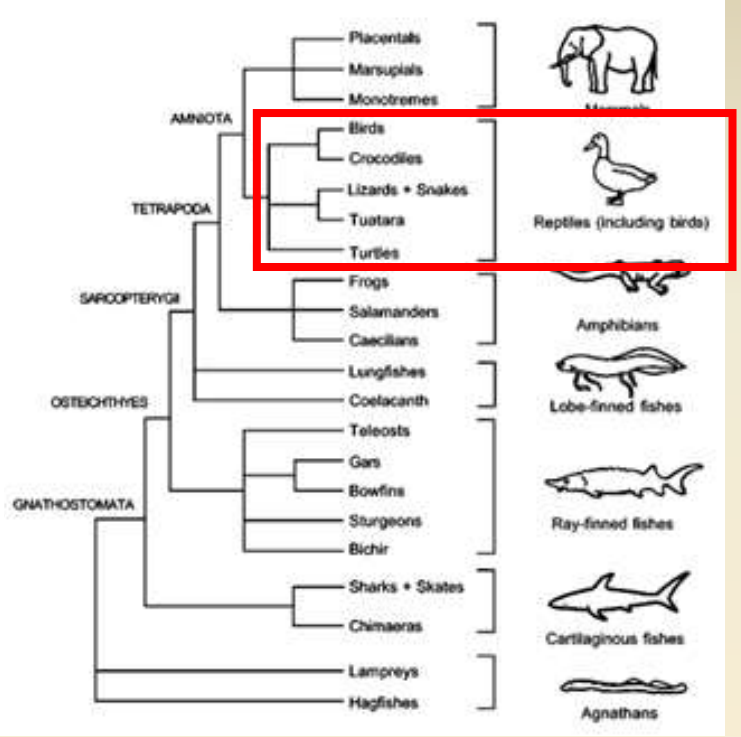
But Reptilia is a paraphyletic group unless it includes Aves

What does this mean?

It means that “reptiles” don’t exist!

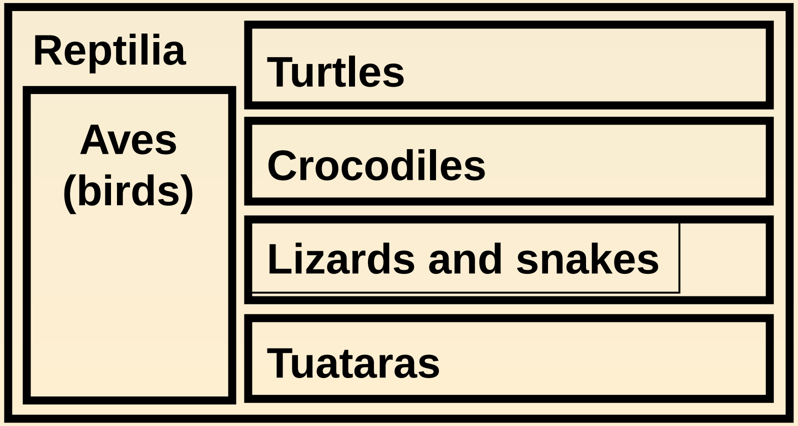
No, it means that you’re one of us!





What it means is that “reptile” is only a valid clade if it *includes* birds

Birds are still birds, but Aves cannot be considered a “Class” equivalent to Class Reptilia because it is evolutionarily nested *within* Reptilia



Reptilia

Aves (birds)

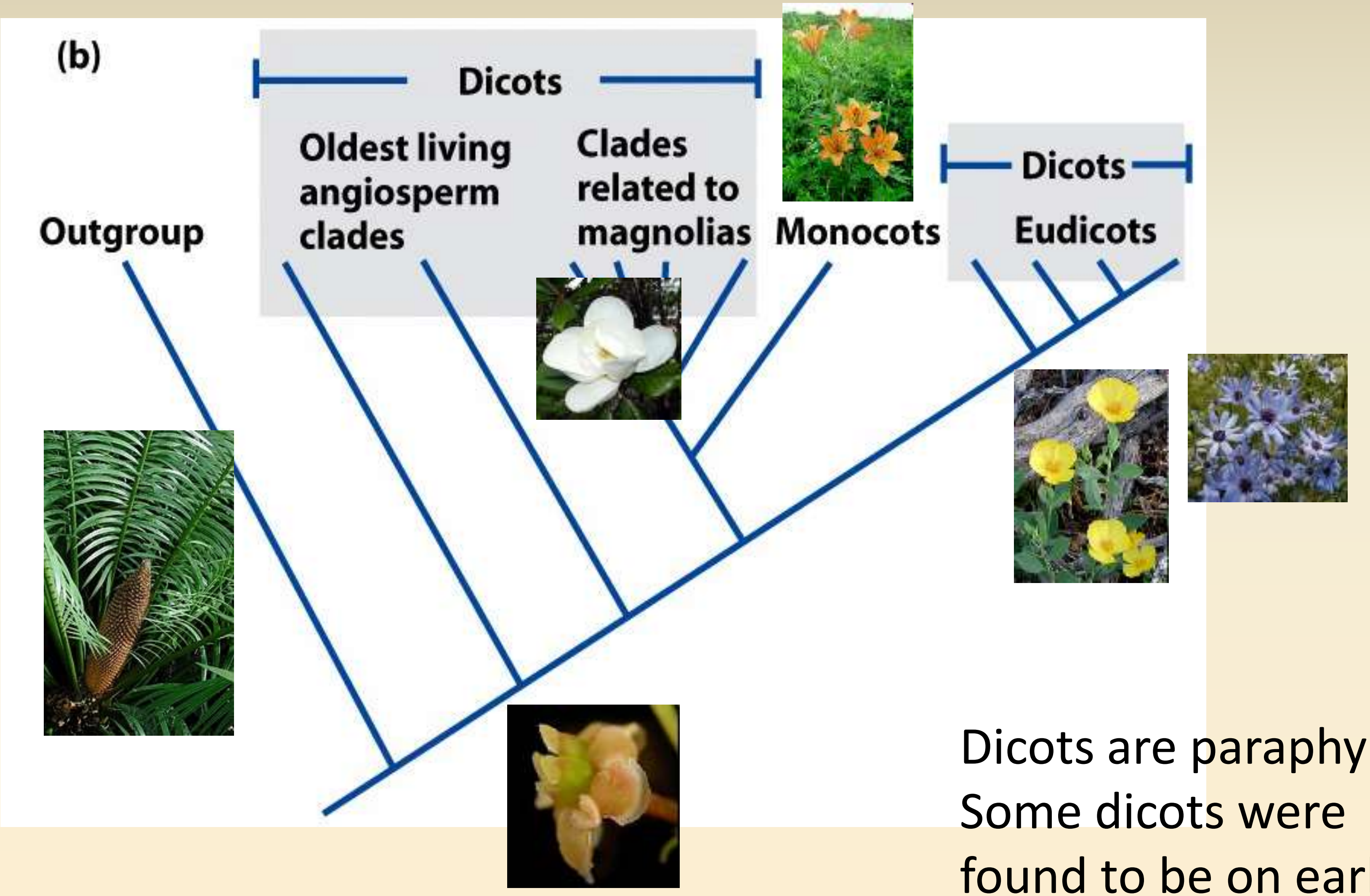
Turtles

Crocodiles

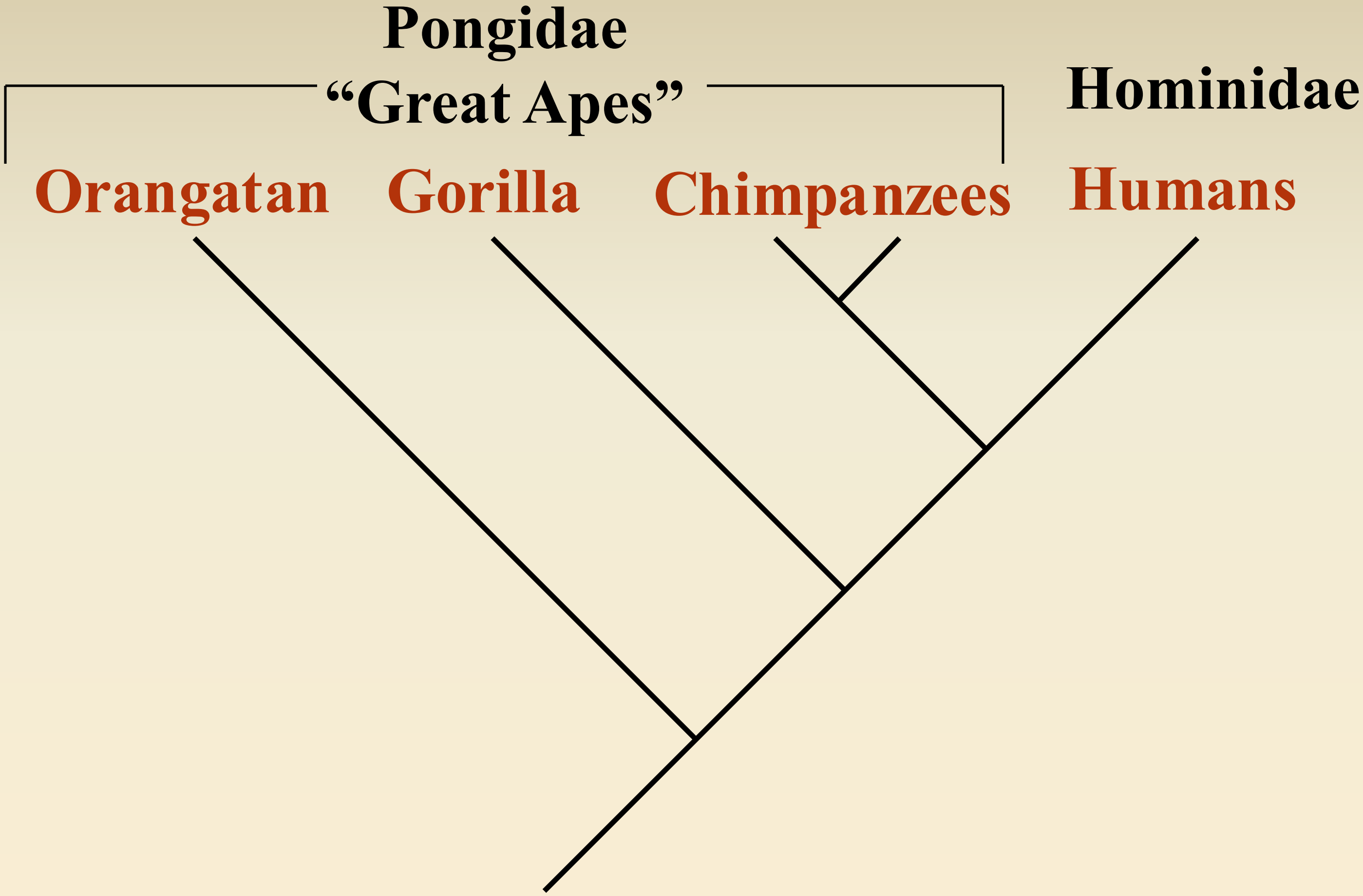
Lizards and snakes

Tuataras

Monophyly vs Paraphyly: Angiosperm



Dicots are paraphyletic. Some dicots were found to be on early branching clades



Pongidae

“Great Apes”

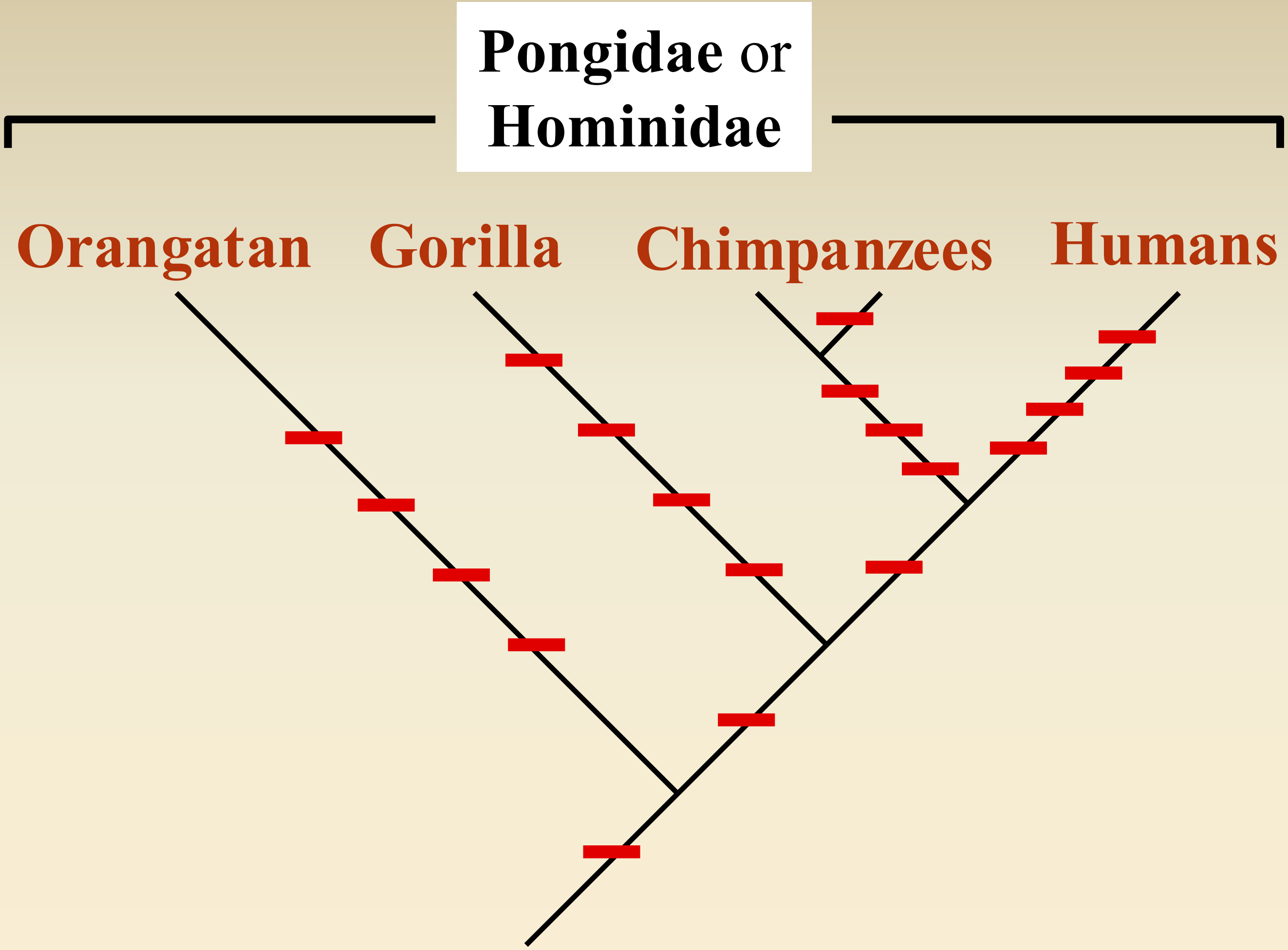
Hominidae

Orangatan

Gorilla

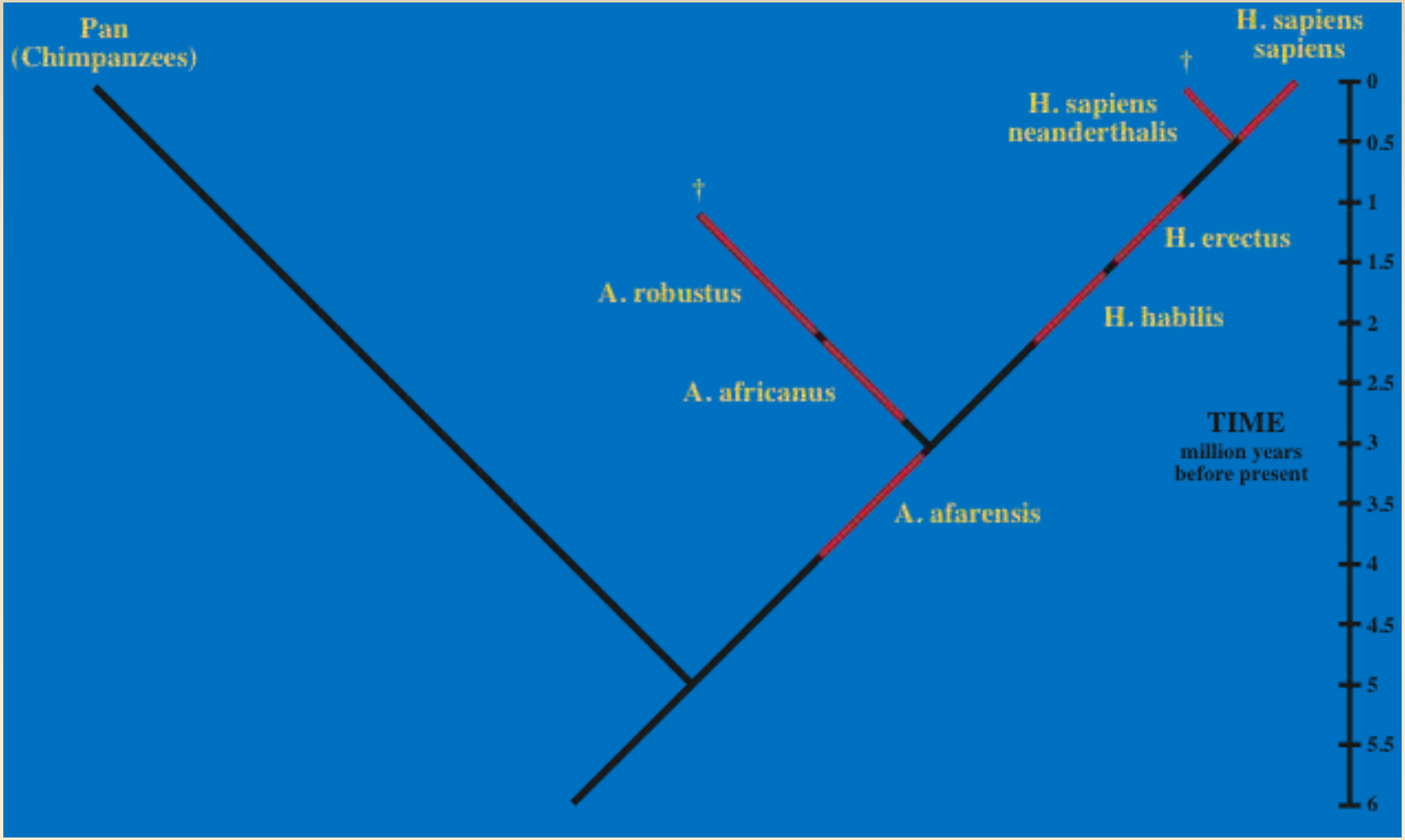
Chimpanzees

Humans



Pongidae or **Hominidae**

Orangatan **Gorilla** **Chimpanzees** **Humans**


Pan
(Chimpanzees)
H. sapiens
sapiens
†
H. sapiens
neanderthalis
†
A. robustus
A. africanus
H. erectus
H. habilis
A. afarensis
TIME
million years
before present
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6

We are human, but we are also apes.

We share unique human features.

We also share features with other apes (and with other animals, plants, fungi, bacteria, etc.).

Humans didn’t evolve from apes, humans are apes.

Questions - Methods

- What kinds of data do we use? Characters?
  - Morphology
  - Fossils
  - Behavior
  - Molecules (DNA)
- How do we make phylogenetic trees?
  - Similarity (distance, phenetics)
  - Cladistic methodology, Parsimony
- How do we decide among competing alternative trees?

Phylogeny is Reconstructed from Characters

Any character that is genetically determined can be used in a phylogenetic analysis.

Character - Heritable trait possessed by an organism; characters are usually described in terms of their **states**, for example: "hair present" vs. "hair absent," where "hair" is the character, and "present" and "absent" are its states.

Morphology—presence, size, shape, or other attributes of body parts, number lengths of legs, etc. The more discrete the better.

The fossil record is especially valuable, and the only option for many extinct taxa

Phylogenies of most extinct species depend almost exclusively on morphology.

Fossils provide evidence that helps distinguish ancestral from derived traits. The fossil record can also reveal when lineages diverged.

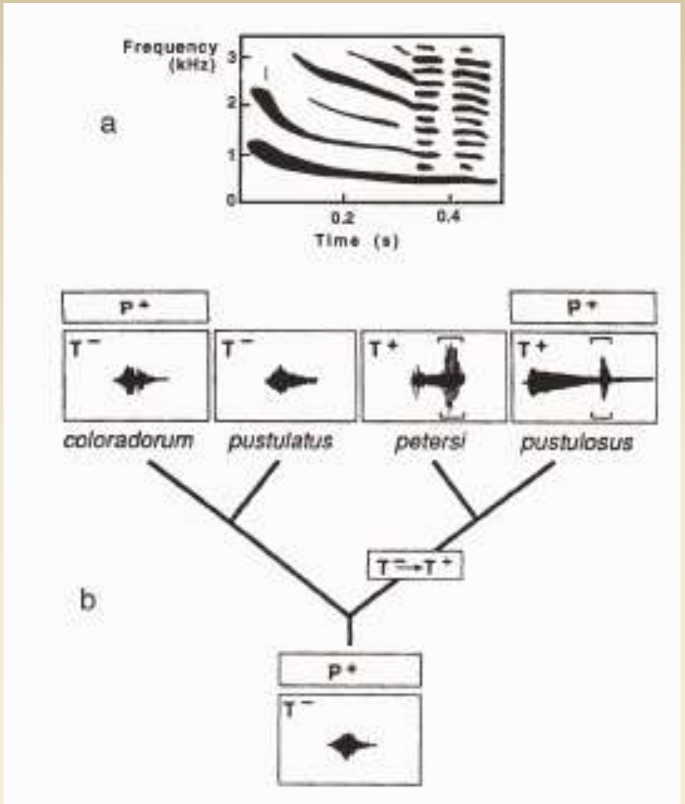
Ammonites

Limitations of using morphology:

- Some taxa show few morphological differences.
- It is difficult to compare distantly related species.
- Some morphological variation is caused by environment.
- Often determined by multiple genes, often not independent or discrete.
- Quantitative measures hard to deal with.

Behavior:

- Leks
- Parental Care
- Gregariousness
- Calls and Songs

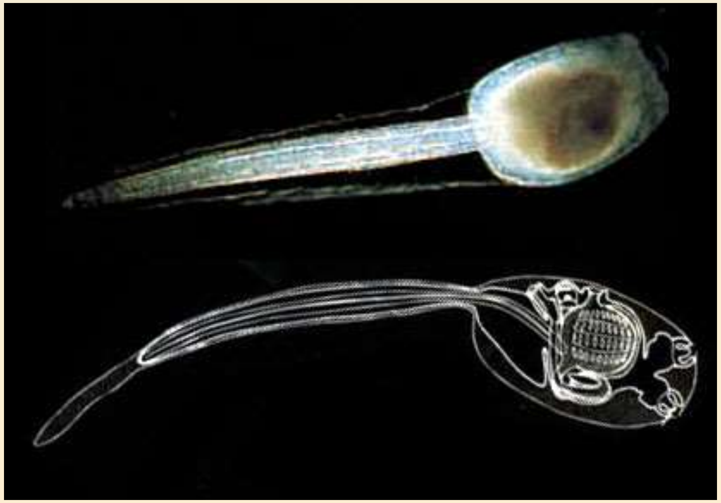


Development:

Similarities in developmental patterns may reveal evolutionary relationships.

Example:

The larvae of sea squirts has a notochord, which is also present in all vertebrates.



Larvae

Adult

Molecular data:

DNA sequences have become the most widely used data for constructing phylogenetic trees.

Nuclear, chloroplast, and mitochondrial DNA sequences are used.

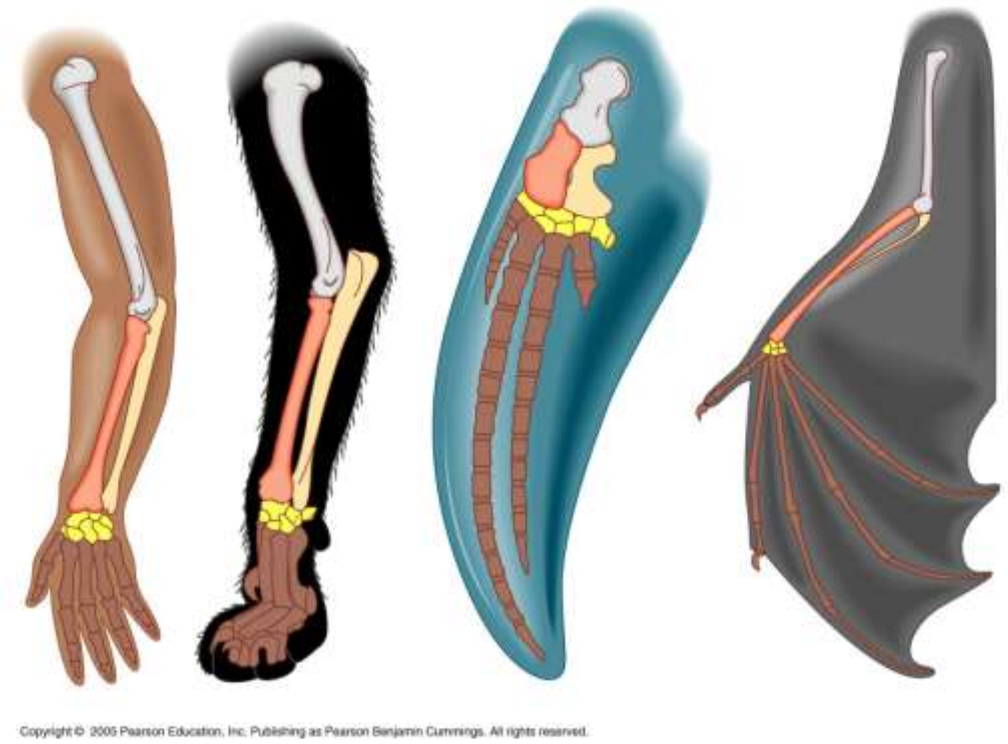
Information on gene products (such as amino acid sequences of proteins) are also used.

Homology: Characters are considered **homologous** when they are inherited from a common ancestor which possessed that feature.

Convergence: the *independent* (convergent) evolution of anatomical or functional similarity between unrelated or distantly related lineages or forms. The resulting similarities are only superficial, generally resulting from similar adaptation to similar environments and are NOT a result of common ancestry (and are therefore *NOT* homologies).

Homoplasy: A similar feature shared by two or more taxa that does not meet the criterion (or criteria) of homology. Homoplasies generally arise via **convergence**.

Homologous Characters – derived from common ancestor



Copyright © 2005 Pearson Education, Inc. Publishing as Pearson Benjamin Cummings. All rights reserved.

Homologous Characters:

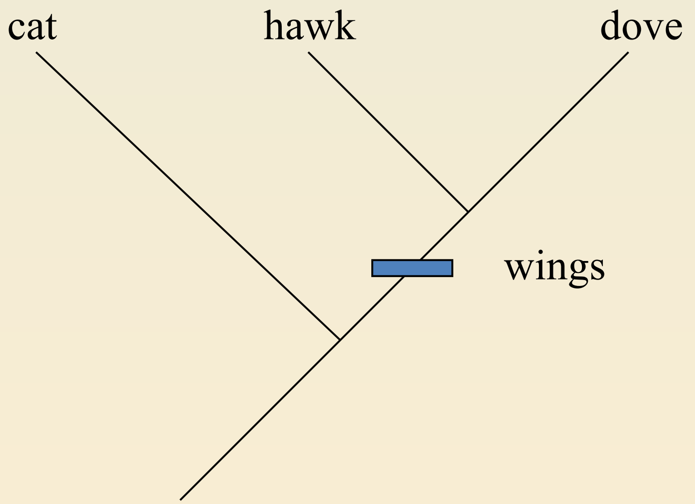
Shared by two or more species

Inherited from a common ancestor

They can be any heritable traits, including DNA sequences, protein structures, anatomical structures, and behavior patterns.

Homology

- A character is similar (or present) in two taxa because their common ancestor had that character:



- In this diagram, wings are homologous characters in hawks and doves because both inherited wings from their common winged ancestor

Each character of an organism evolves from one condition (the **ancestral trait**) to another condition (the **derived** trait).

Shared derived traits provide evidence of the common ancestry of a group and are called **synapomorphies**.

The vertebral column is a synapomorphy of the vertebrates. The ancestral trait was an undivided supporting rod.

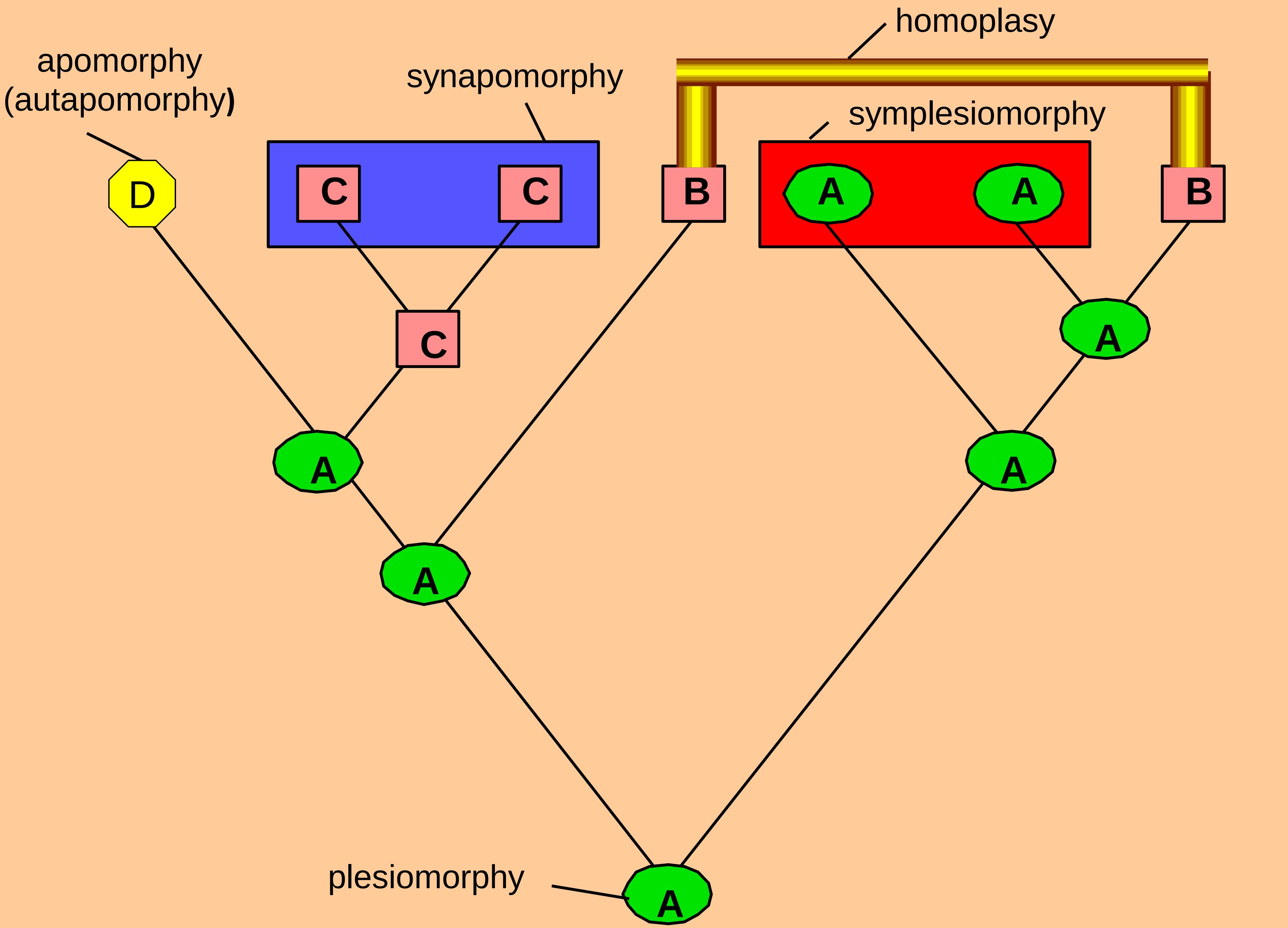
Terminology developed by Willi Hennig

PLESIOMORPHY: An ancestral or primitive character, often incorrectly used to group taxa.

APOMORPHY: a **derived** feature or character; derived from and differing from an ancestral (plesiomorphic) condition.

SYNAPOMORPHY: A shared, derived character (apomorphy) reflecting common ancestry used to group taxa. Hair is a synapomorphy of mammals.

SYMPLESIOMORPHY: A plesiomorphy shared by two or more taxa.


homoplasy
apomorphy
(autapomorphy)
synapomorphy
symplesiomorphy
D
C
C
B
A
A
B
C
A
A
A
A
plesiomorphy
A

Synapomorphies reveal the relationships among tetrapods

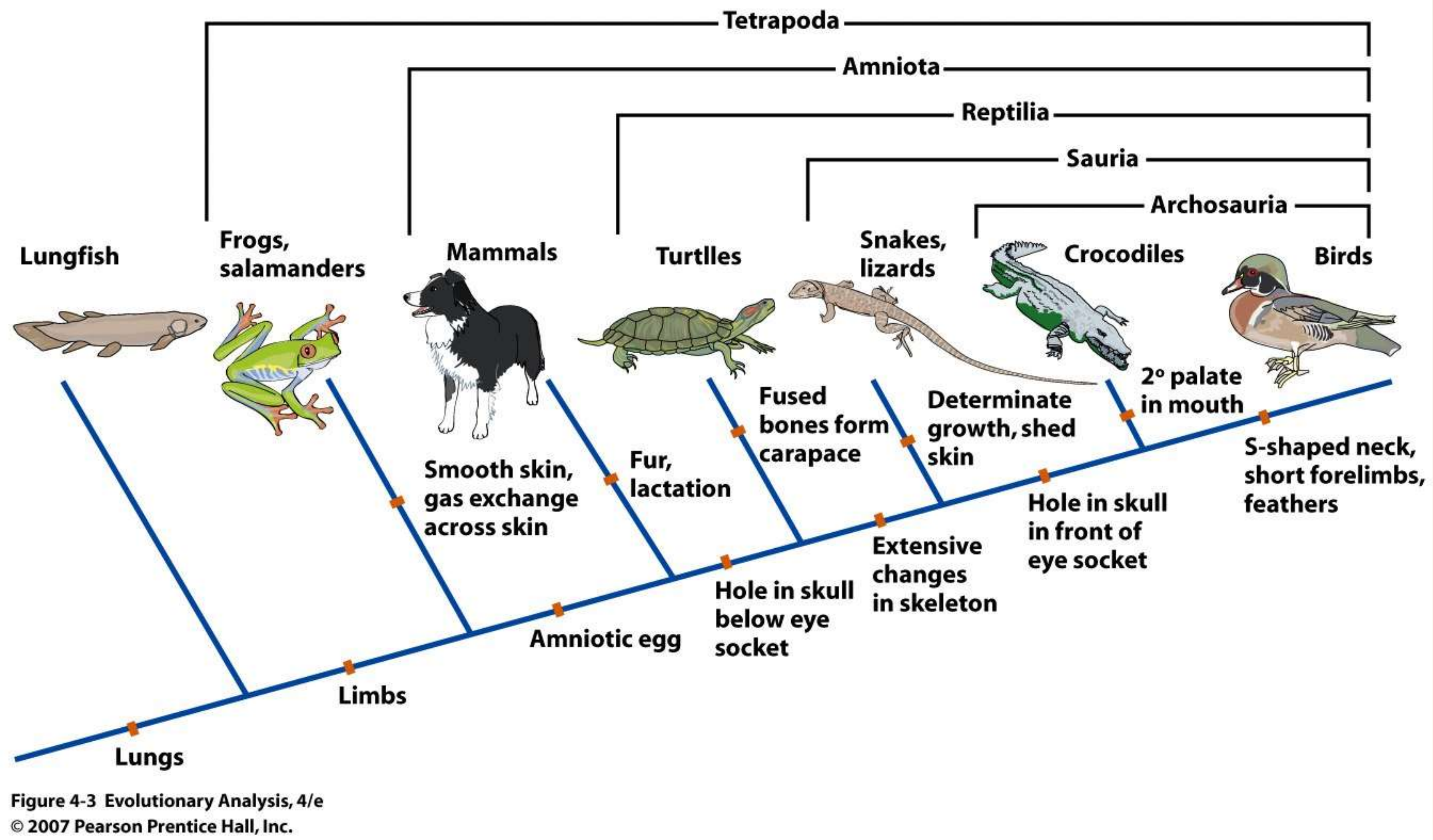
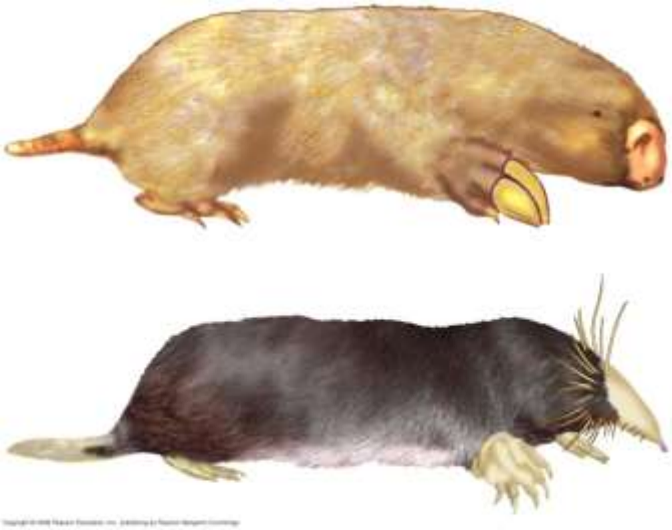

Tetrapoda
Amniota
Reptilia
Sauria
Archosauria
Lungfish
Frogs, salamanders
Mammals
Turtlles
Snakes, lizards
Crocodiles
Birds
Lungs
Limbs
Smooth skin, gas exchange across skin
Amniotic egg
Fur, lactation
Hole in skull below eye socket
Fused bones form carapace
Extensive changes in skeleton
Determinate growth, shed skin
Hole in skull in front of eye socket
2° palate in mouth
S-shaped neck, short forelimbs, feathers

Figure 4-3 Evolutionary Analysis, 4/e
© 2007 Pearson Prentice Hall, Inc.

Trees built from synapomorphies = cladograms

Convergent Evolution

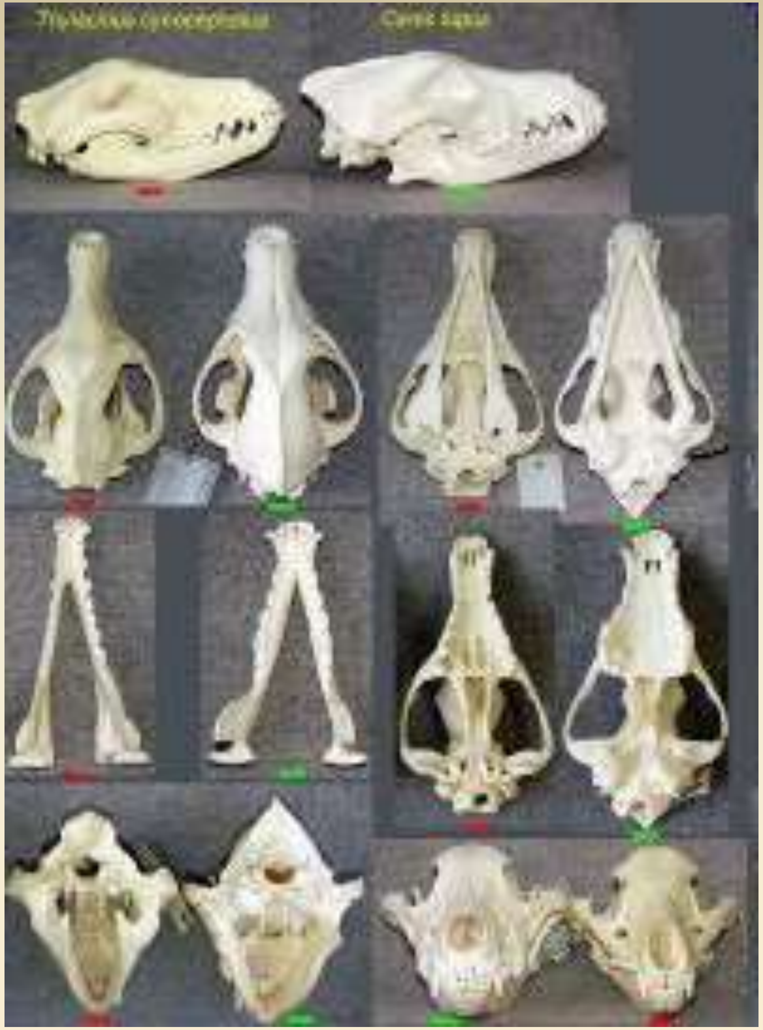
Similarity between species that is caused by a similar but evolutionarily independent response to similar selection pressures Ancestors are different in appearance, but the two descendants now look alike for that trait.



Convergent evolution: Australian "mole" and N. Am. "mole"

Convergent evolution within mammals

Marsupial Tasmanian wolf



Grey Wolf

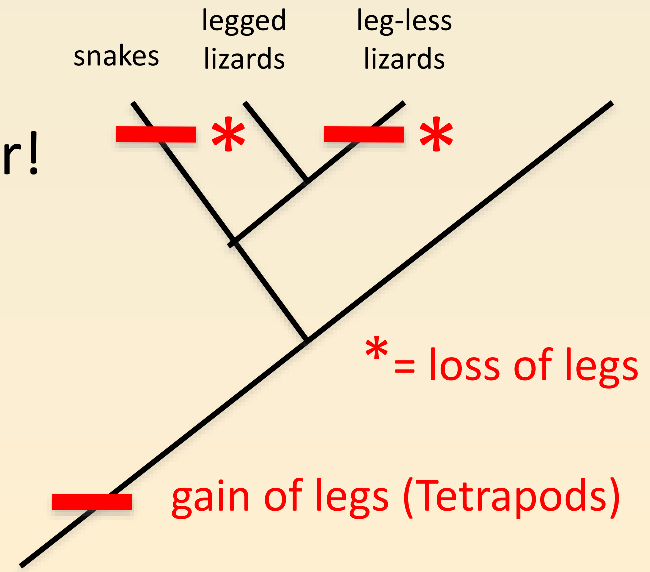
The skulls of the Thylacine (left) and the Grey Wolf, *Canis lupus*, are almost identical, although the species are only very distantly related (different infraclasses). The skull shape of the Red Fox, *Vulpes vulpes*, is even closer to that of the Thylacine.

Leg-less lizards

Snake

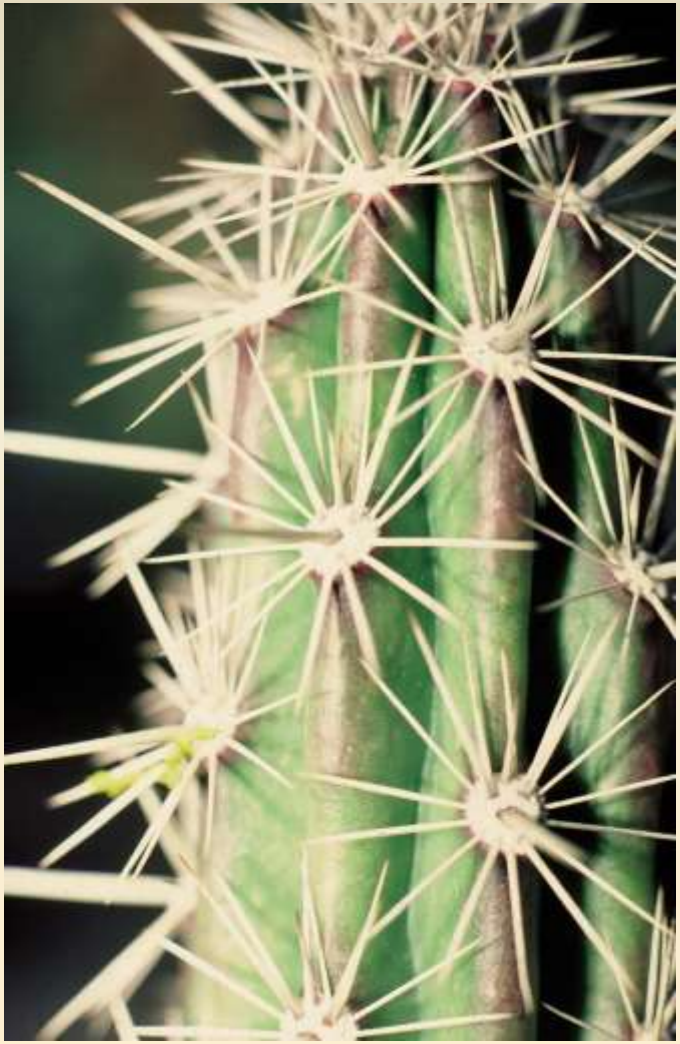
Both examples of **reversal** within Tetrapods:
loss of a derived feature – forelimbs.

Example of **convergence** relative to one another!
Independently evolved.

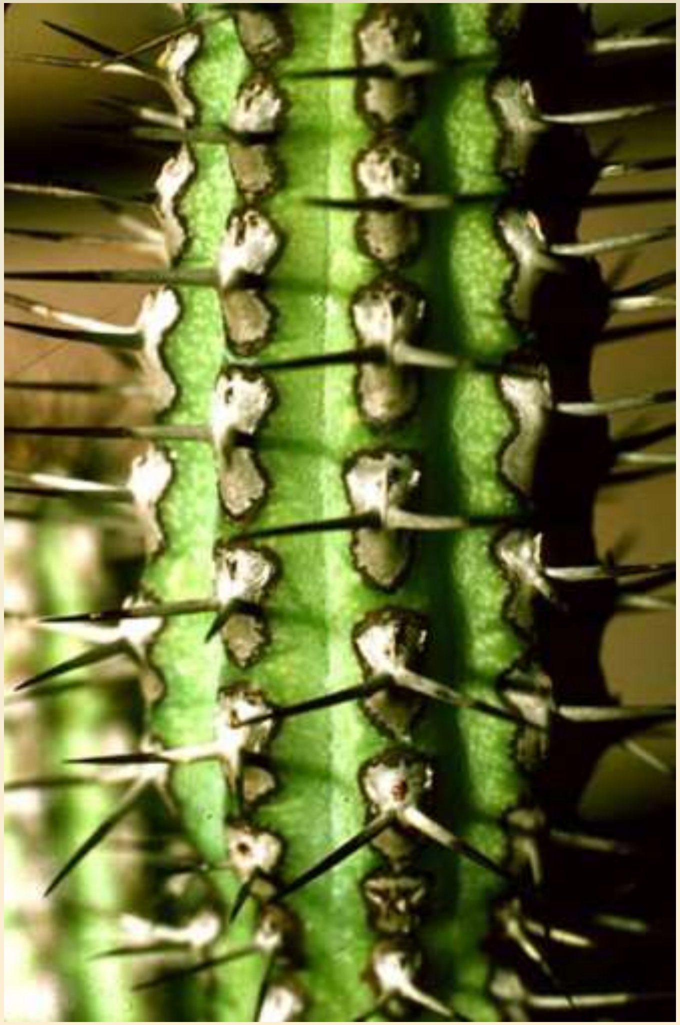


Convergent evolution:

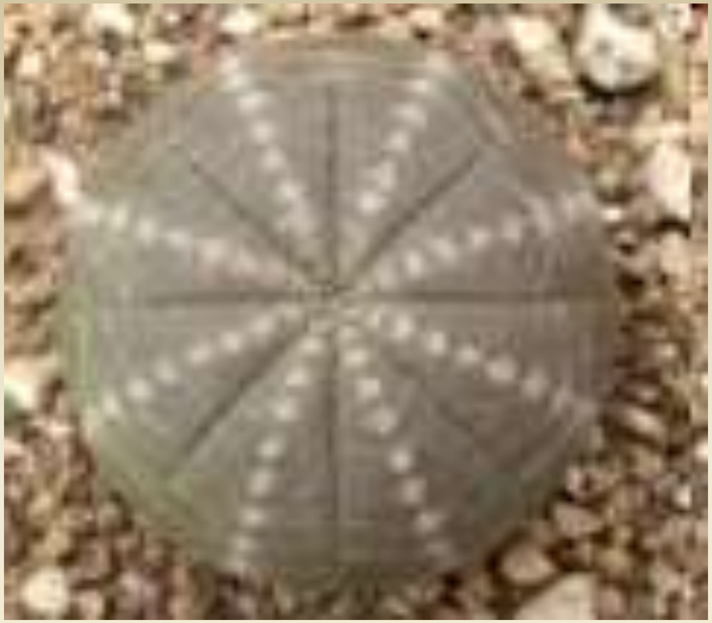
spines of cacti & euphorbs

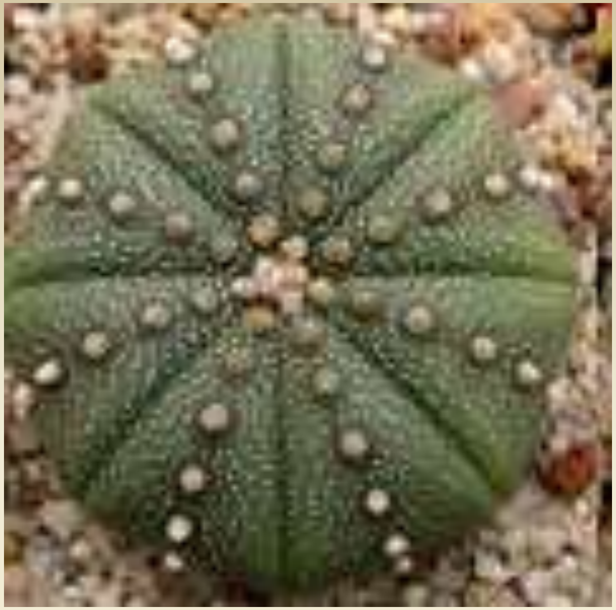


Cactus



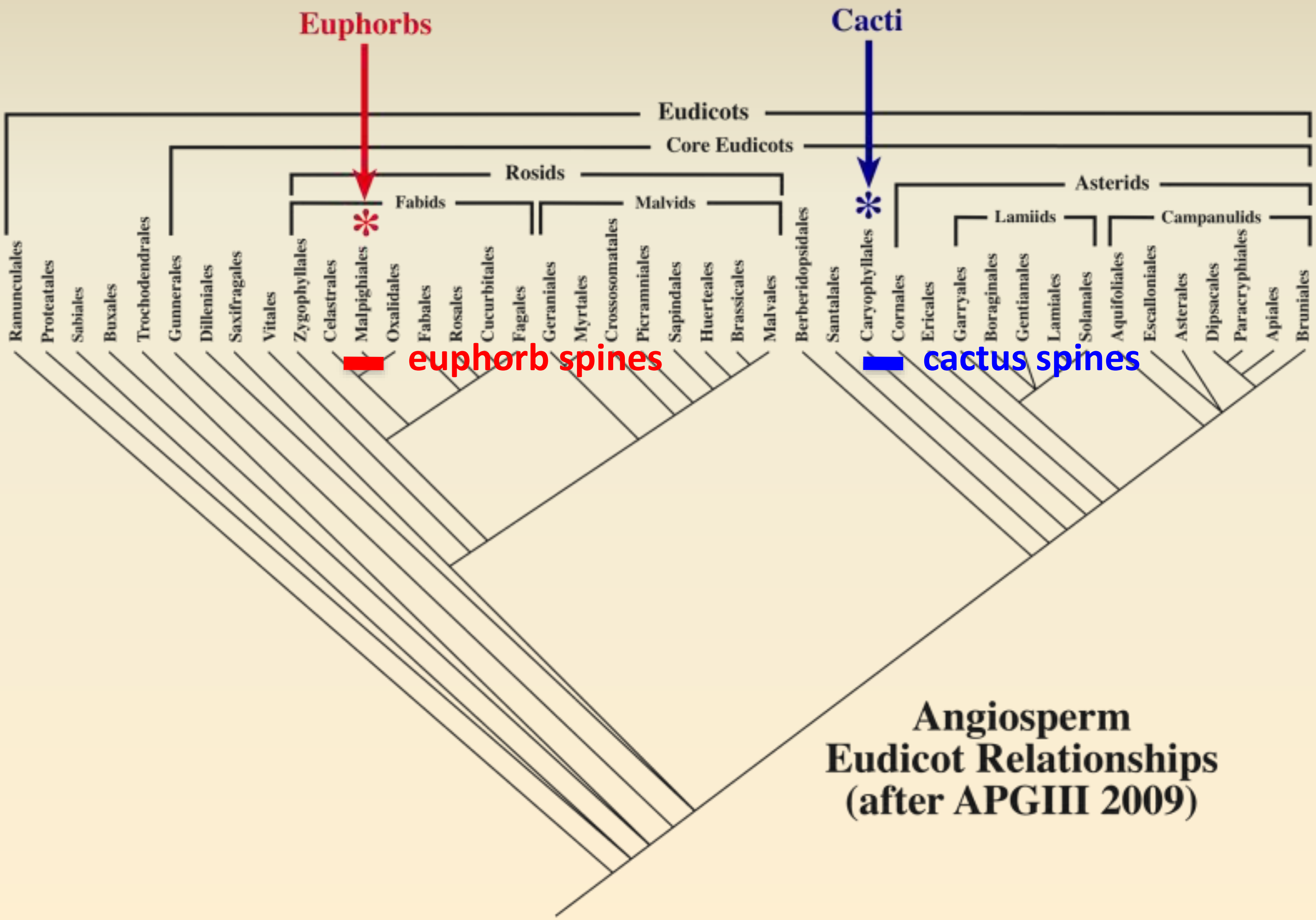
Euphorb





Convergent evolution of succulence: Euphorbiaceae left, Cactaceae right
The trait succulence is a homoplasy arising from convergent evolution

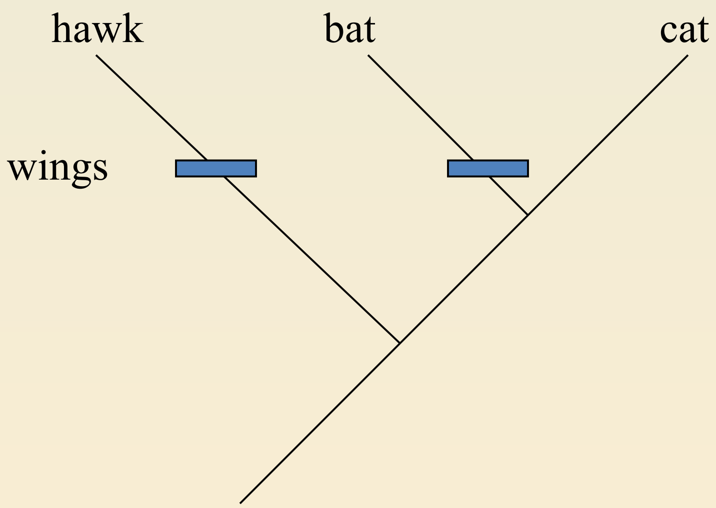
Convergent evolution: spines of cacti & euphorbs



Angiosperm Eudicot Relationships (after APGIII 2009)

Homoplasy

- A character is similar (or present) in two taxa because of independent evolutionary origin (i.e., the similarity does not derive from common ancestry):



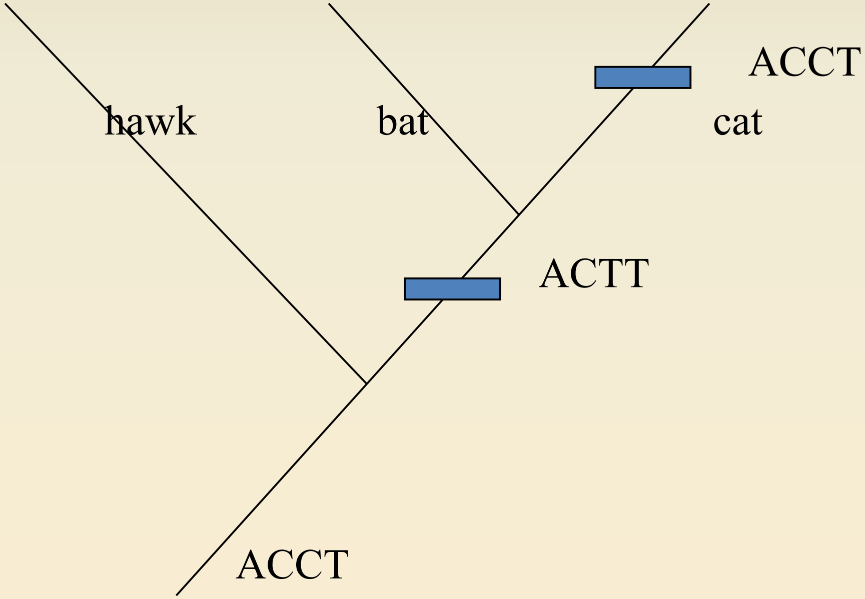
- In this diagram, wings are a homoplasy in hawks and bats because their common ancestor was an un-winged tetrapod reptile. Bird wings and bat wings evolved independently.

Types of homoplasy

- Convergence
  - Independent evolution of similar traits in distantly related taxa — streamlined shape, dorsal fins, etc. in sharks and dolphins
- Parallelism
  - Independent evolution of similar traits in closely related taxa — evolution of blindness in different cave populations of the same fish species
- Reversal
  - A character in one taxon reverts to an earlier state (not present in its immediate ancestor)

Reversal

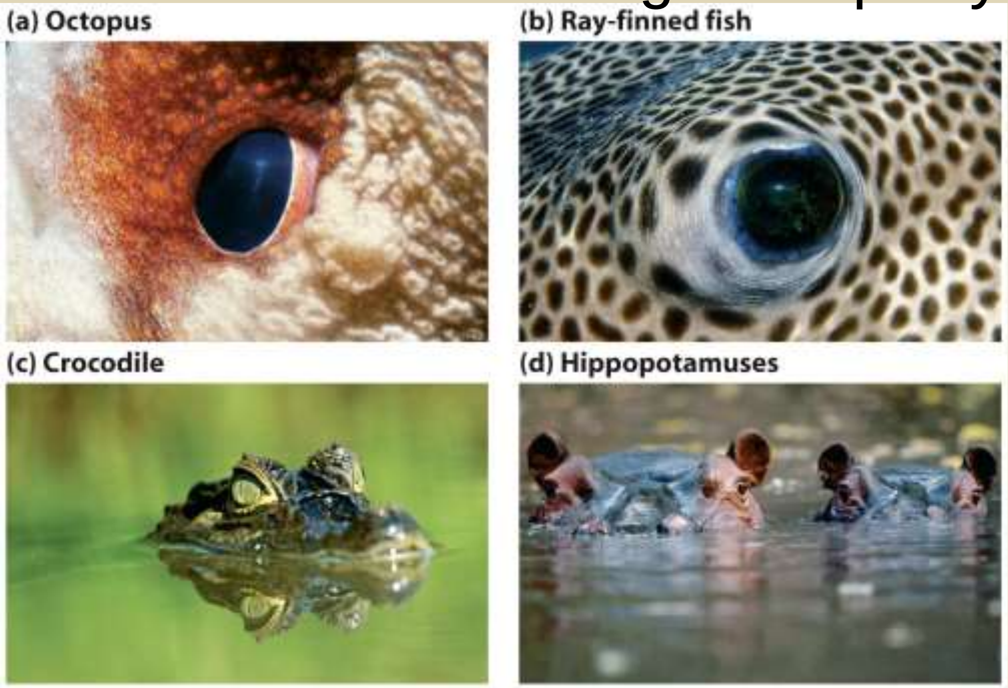
- A character is similar (or present) in two taxa because a reversal to an earlier state occurred in the lineage leading to one of the taxa:



- In this diagram, hawks and cats share the ancestral nucleotide sequence ACCT, but this is due to a reversal on the lineage leading to cats

V. The Right Traits

The importance of recognizing and using homologous traits versus shared traits reflecting homoplasy



(a) Octopus

(b) Ray-finned fish

(c) Crocodile

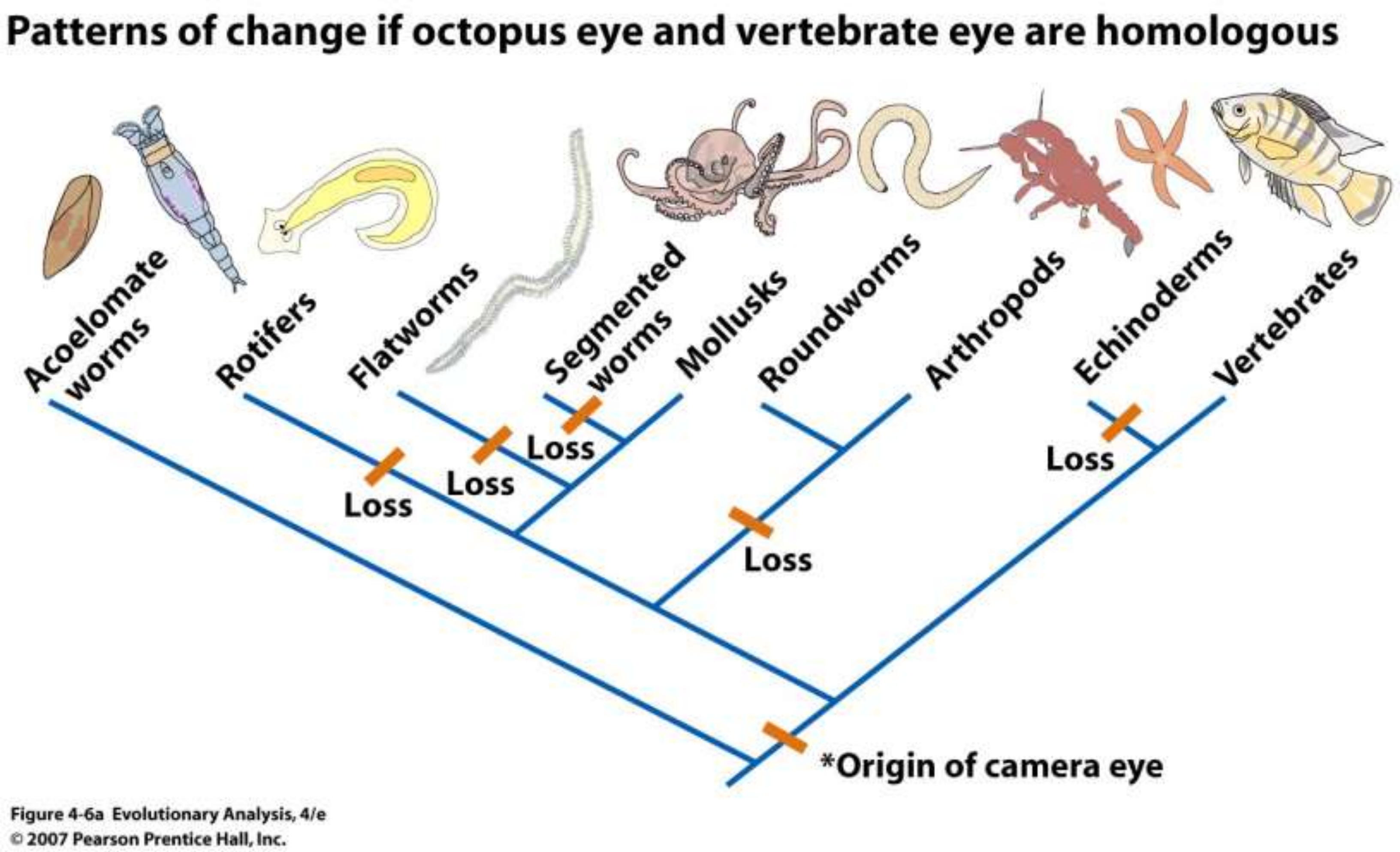
(d) Hippopotamuses

Figure 4-4 Evolutionary Analysis, 4/e
© 2007 Pearson Prentice Hall, Inc.

Homology: A trait that is similar between two species because of inheritance of that trait from a common ancestor

Homoplasy: A trait that is similar between two species because of convergent evolution, parallelism or reversal, but not because of shared ancestry

VI. Parsimony: least number of steps to construct a phylogeny



Using parsimony to distinguish homology from homoplasy

(Tree made from DNA synapomorphies) (also development)

Pattern of change if octopus eye and vertebrate eye are convergent

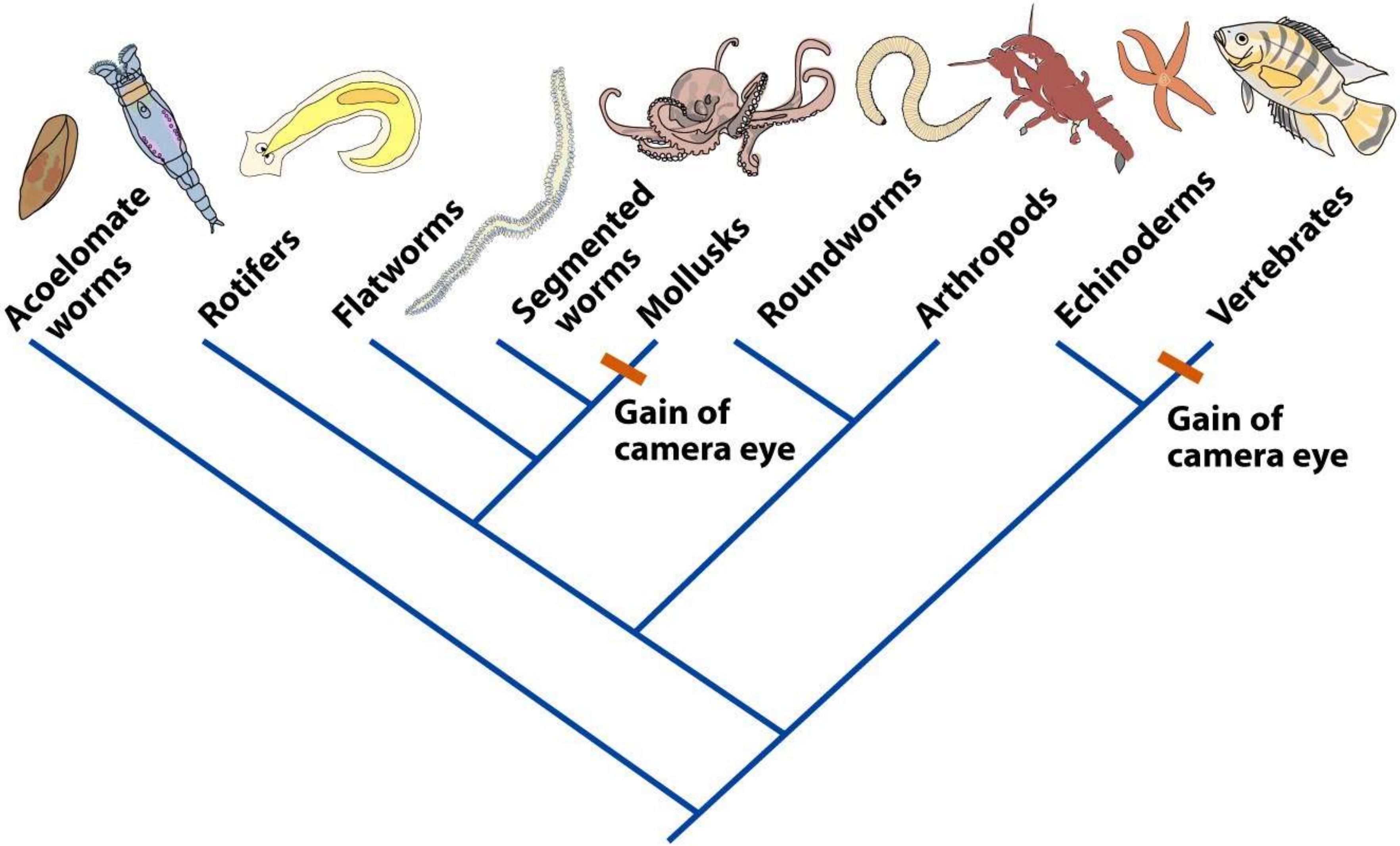
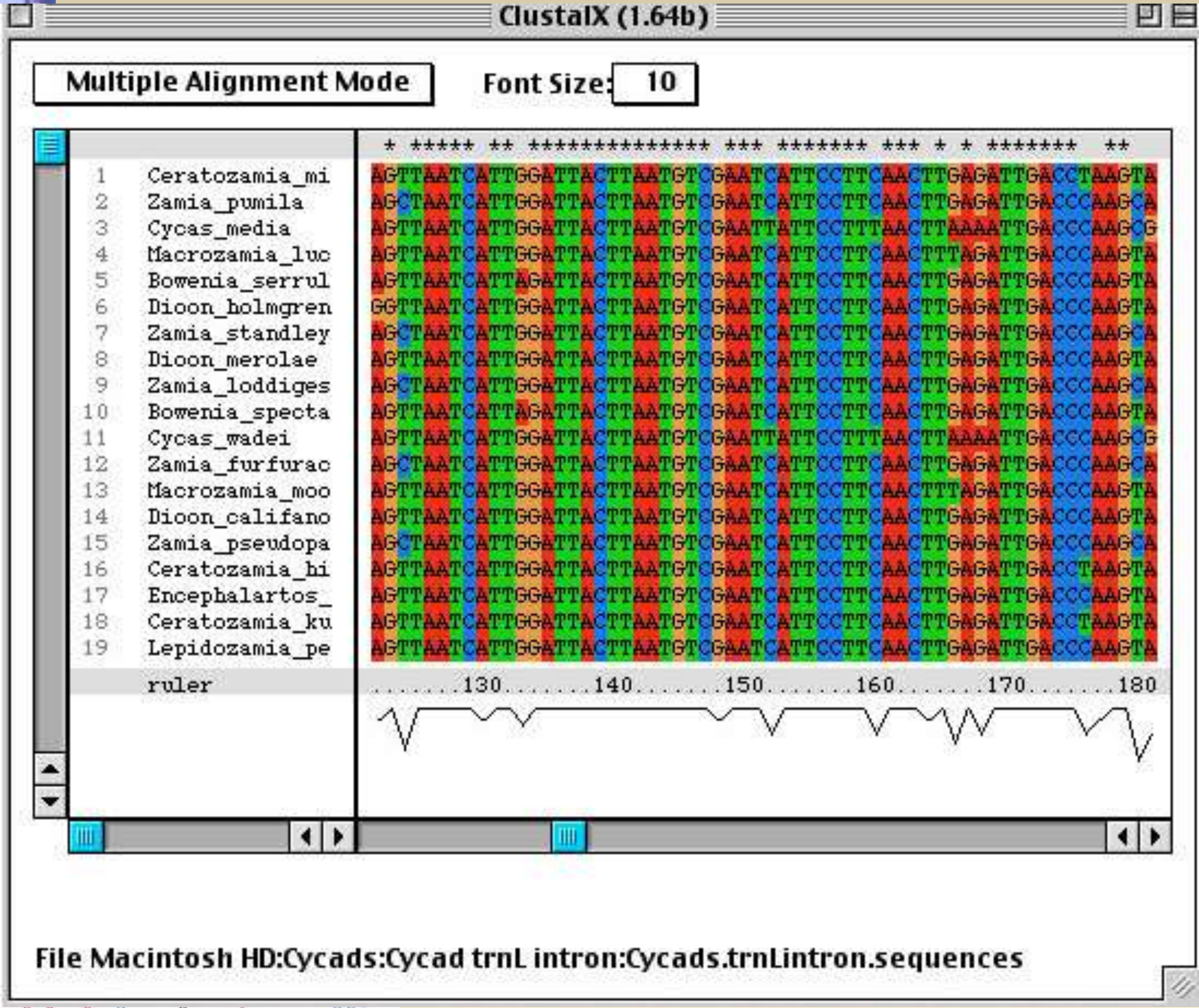


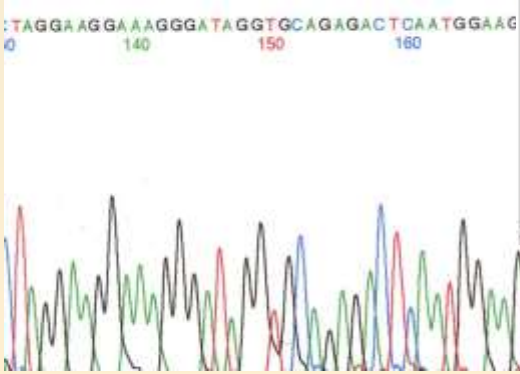
Figure 4-6b Evolutionary Analysis, 4/e

DNA Sequencing



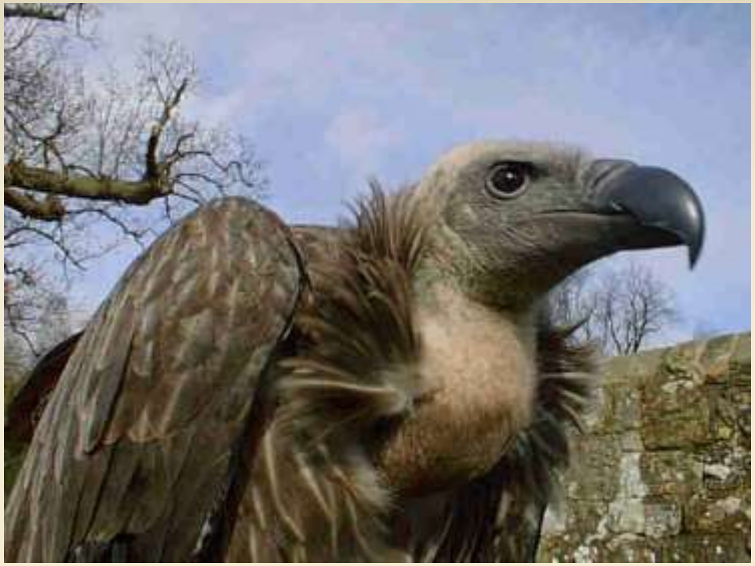
```
                         * ***** ** ************** *** ******* *** * * *******   **
 1   Ceratozamia_mi      AGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCTAAGTA
 2   Zamia_pumila        AGCTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGCA
 3   Cycas_media         AGTTAATCATTGGATTACTTAATGTCGAATTATTCCTTTAACTTAAAATTGACCCAAGCG
 4   Macrozamia_luc      AGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTTAGATTGACCCAAGTA
 5   Bowenia_serrul      AGTTAATCATTAGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGTA
 6   Dioon_holmgren      GGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGTA
 7   Zamia_standley      AGCTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGCA
 8   Dioon_merolae       AGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGTA
 9   Zamia_loddiges      AGCTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGCA
10   Bowenia_specta      AGTTAATCATTAGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGTA
11   Cycas_wadei         AGTTAATCATTGGATTACTTAATGTCGAATTATTCCTTTAACTTAAAATTGACCCAAGCG
12   Zamia_furfurac      AGCTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGCA
13   Macrozamia_moo      AGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTTAGATTGACCCAAGTA
14   Dioon_califano      AGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGTA
15   Zamia_pseudopa      AGCTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGCA
16   Ceratozamia_hi      AGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCTAAGTA
17   Encephalartos_      AGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGTA
18   Ceratozamia_ku      AGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCTAAGTA
19   Lepidozamia_pe      AGTTAATCATTGGATTACTTAATGTCGAATCATTCCTTCAACTTGAGATTGACCCAAGTA
     ruler               .......130.......140.......150.......160.......170.......180
```

File Macintosh HD:Cycads:Cycad trnL intron:Cycads.trnLintron.sequences

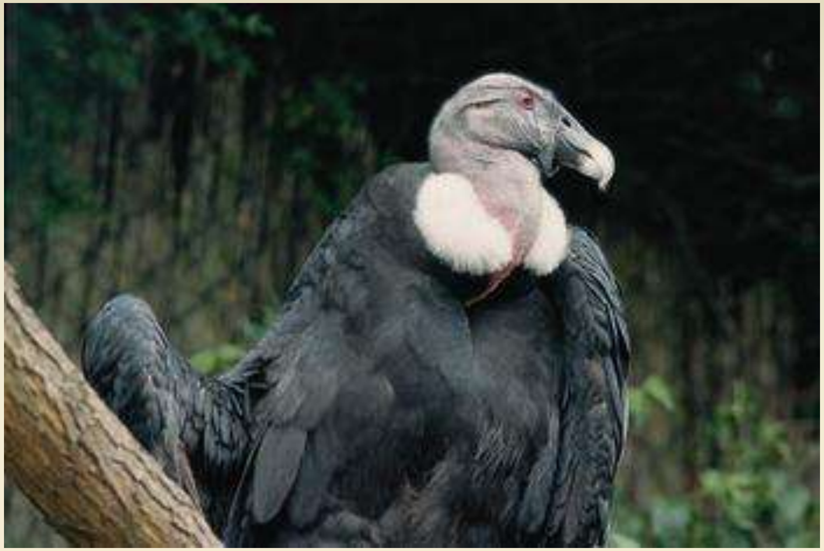


Sequence Alignment

Morphology vs. molecular data



African white-backed vulture
(old world vulture)



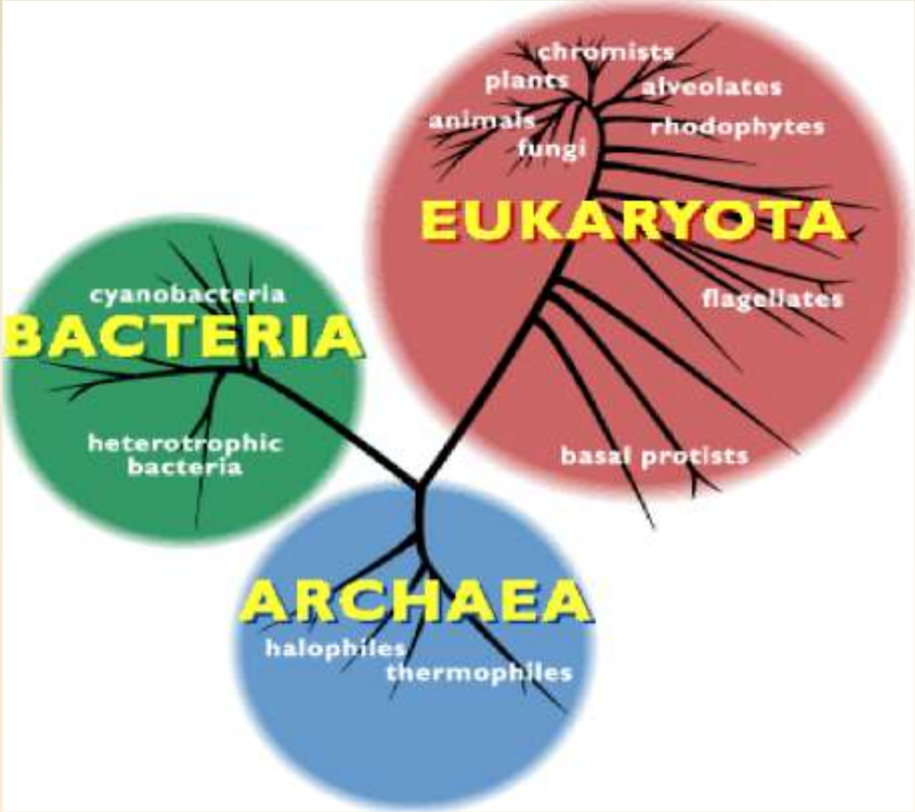
Andean condor
(new world vulture)

New and old world vultures seem to be closely related based on morphology.

Molecular data indicates that old world vultures are related to birds of prey (falcons, hawks, etc.) while new world vultures are more closely related to storks

Similar features presumably the result of convergent evolution

Molecular data: single-celled organisms



Molecular data useful for analyzing single-celled organisms (which have only few prominent morphological features).

Molecular Data

Many more molecular characters available for analysis than morphological ones.

Identity is easier to define: ATCG vs. whether a flower color is pink or white.

Nonetheless, molecular data are still subject to homoplasy: reversals and convergence as well as long branch attraction (errors due to mutation rate being fast and number of characters small: leads to wrong phylogenetic tree appearing to be correct.

In spite of the pitfalls, DNA sequence data are now overwhelmingly the tool of choice for generating phylogenetic hypotheses.

Methods

How do we infer phylogeny?

Three “schools” of phylogenetic thought:

1. Evolutionary systematics

2. Phenetics - (Distance)

3. Cladistics/phylogenetics

1. Evolutionary systematics

Arose during the Modern Synthesis of Evolution (Ernst Mayr, Theodosius Dobzhansky, G.G. Simpson)

Tried to be synonymous with evolutionary biology & “Neo-Darwinism”

Goal: Think of relationships among organisms as how Natural Selection made them.

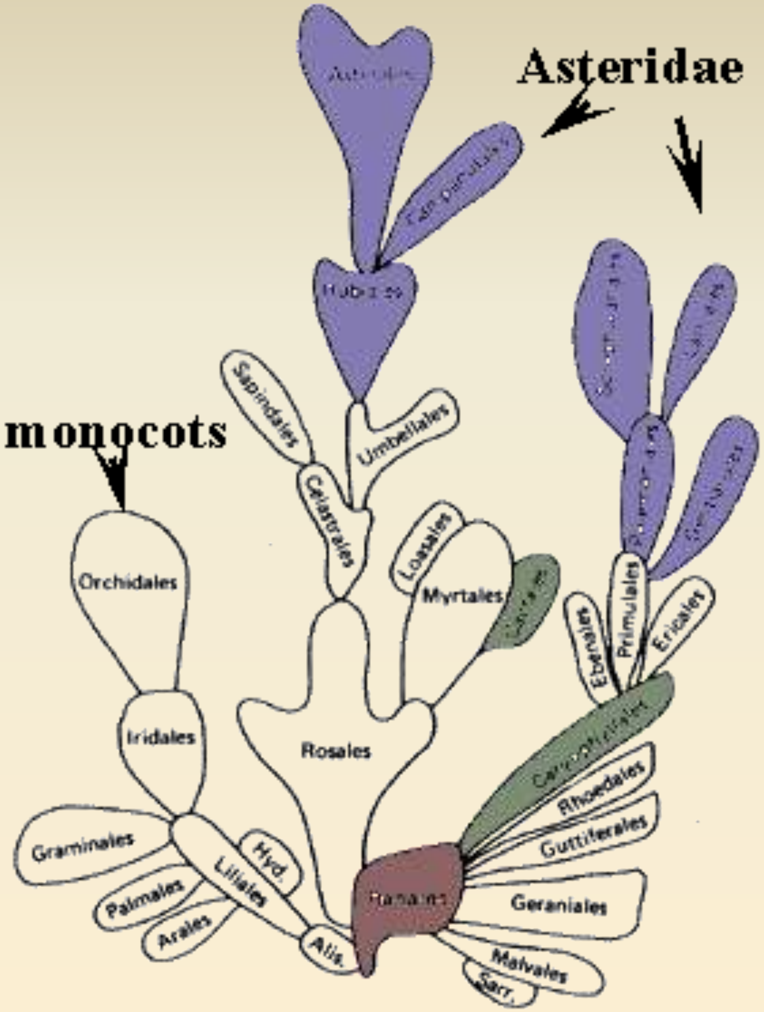
Very little (if any) methodology or “operationalism”. Construct scenarios, but no formal system of theories.

Difficult to formulate testable hypotheses.

Often only classifications, with little attempt to depict relationships as “trees” (phylogenies).

”Trust the experts”

Bessey’s Cactus of angiosperms



Evolution of the Horse

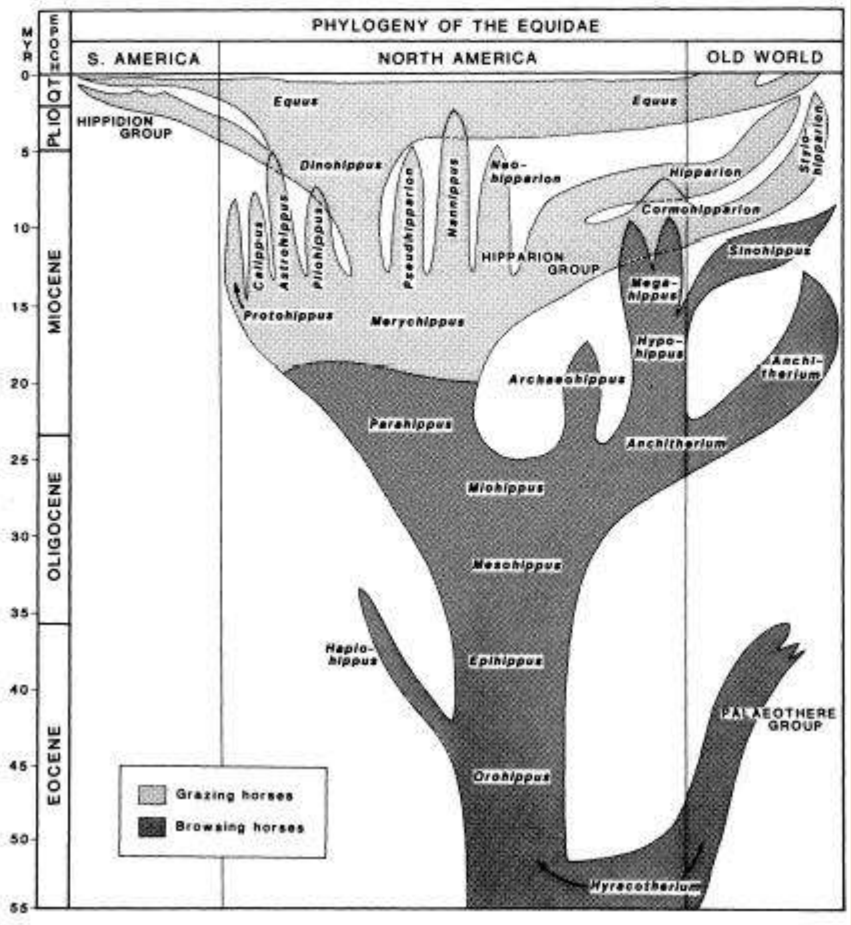


FIGURE 1. Current phylogeny of the Equidae, with particular emphasis on the North American taxa.

2. Phenetic classification – Distance, Similarity

The basic idea of phylogenetic reconstruction is simple:

Taxa that are closely related (descended from a relatively recent common ancestor) should be more similar to each other than taxa that are more distantly related.

So, all we need to do is build trees that put similar taxa on nearby branches.

This is the phenetic approach to tree building

Type of Data

- **Character**-based
  - Examine each character (e.g., residue) separately

- **Distance**-based
  - Input is a matrix of distances between species
  - percent similarity
  - fraction of residue they disagree on, or alignment score between them

Types of data used in phylogenetic inference:

Character-based methods: Use the aligned characters, such as DNA or protein sequences, directly during tree inference.

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

Distance-based methods: Transform the sequence data into pairwise distances (dissimilarities), and then use the matrix during tree building.

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

Example 1: Uncorrected "p" distance (=observed percent sequence difference)

Example 2: Kimura 2-parameter distance (estimate of the true number of substitutions between taxa)

Based on lectures by C-B Stewart, and by Tal Pupko

2. Phenetic classification – Distance, Similarity

Based on overall similarity.

Those organisms most similar are classified more “closely” together.

Steps:

1. Calculate pairwise distances (similarities) for all taxa.
2. Make distance matrix (table of pairwise distances).
3. Calculate tree from distance matrix.

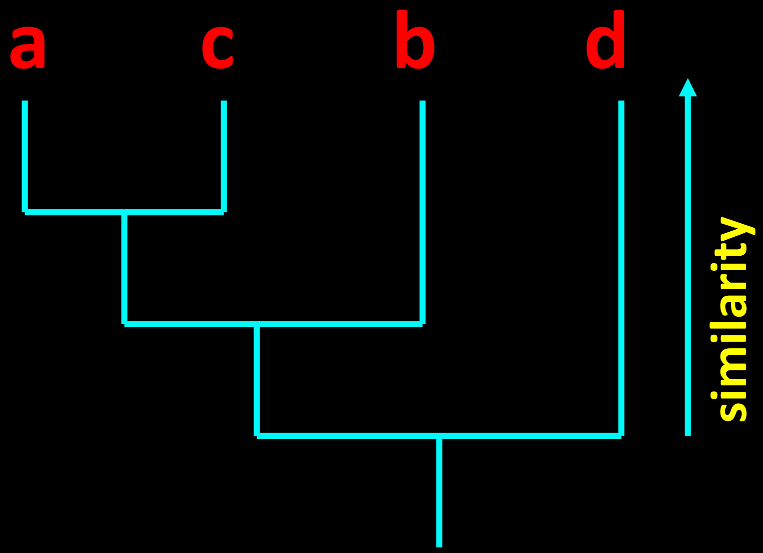
Phenetics - Simple Data

Character	a	b	c	d
1	1	1	0	0
2	1	1	0	0
3	1	1	1	0
4	1	1	1	0
5	1	1	1	0
6	1	1	1	0
7	0	1	0	0
8	0	1	0	0
9	0	1	0	0

Phenetics

Similarity Matrix

	a	b	c	d
a	-	6	7	3
b		-	4	0
c			-	5
d				-



B = (BA + BC)/2

B = 5

D = 0

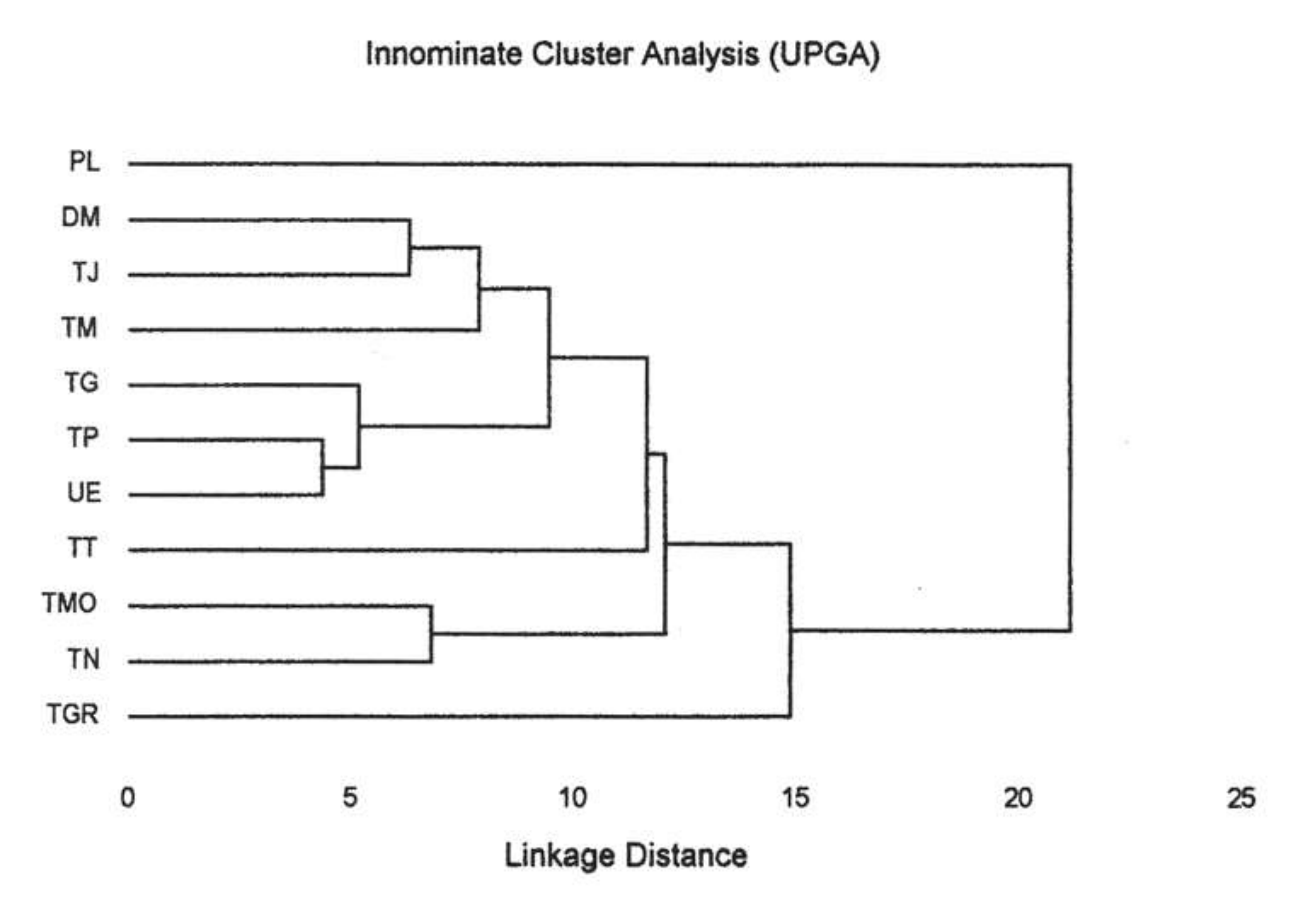
Phenetics

Phenograms do not necessarily represent phylogenetic relationships

Similarity - number of character states 2 species share

Relationship - how recently they diverged from a common ancestor

Phenetics: “phenograms”



Unweighted Pair Group Method using Arithmetic Averages (UPGMA)

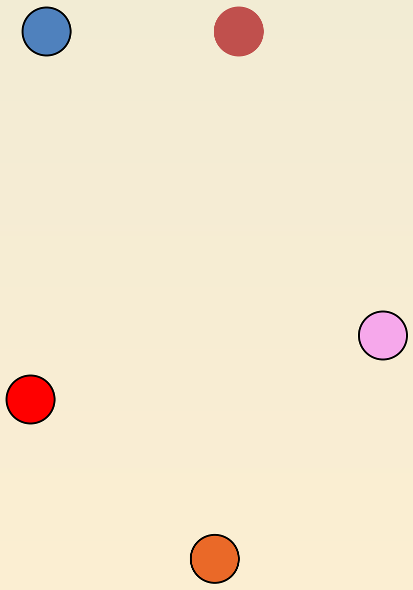
UPGMA is a type of **Distance-Based** algorithm.

Despite its formidable acronym, the method is simple and intuitively appealing.

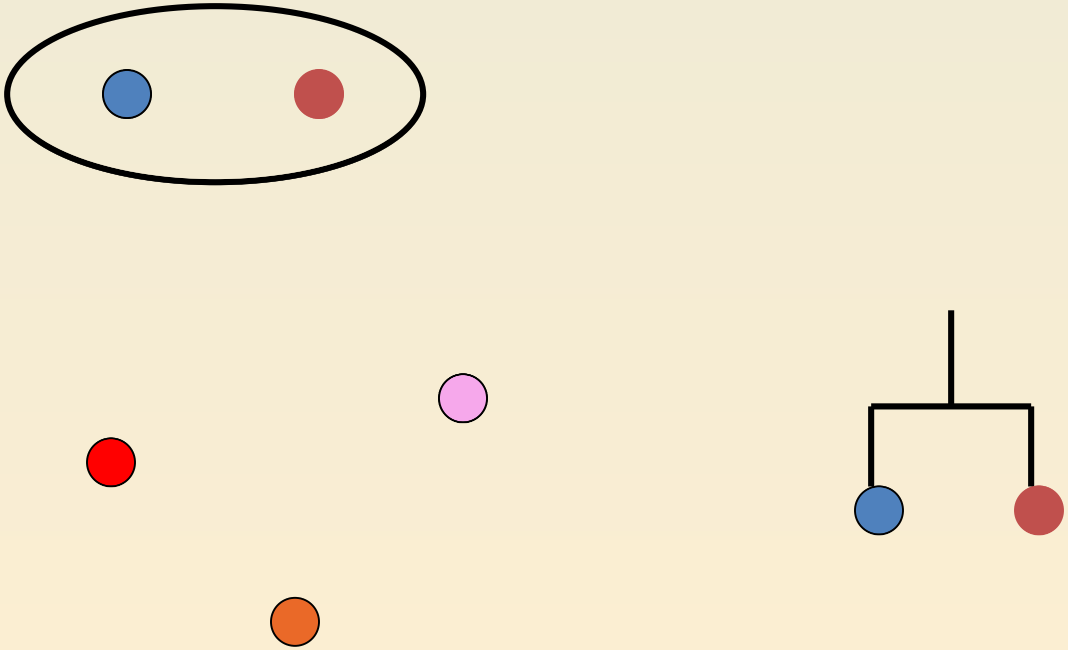
It works by clustering the sequences, at each stage amalgamating two clusters and, at the same time, creating a new node on the tree.

Thus, the tree can be imagined as being assembled upwards, each node being added above the others, and the edge lengths being determined by the difference in the heights of the nodes at the top and bottom of an edge.

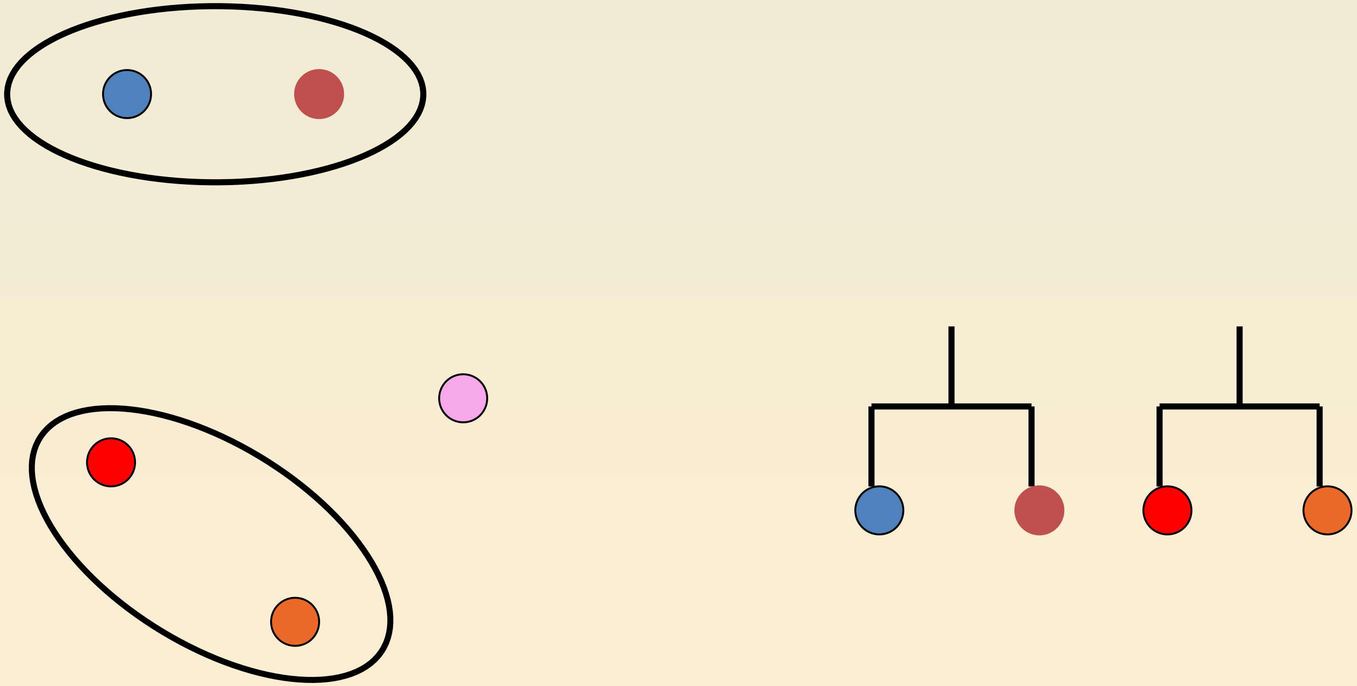
An example showing how UPGMA produces a rooted phylogenetic tree



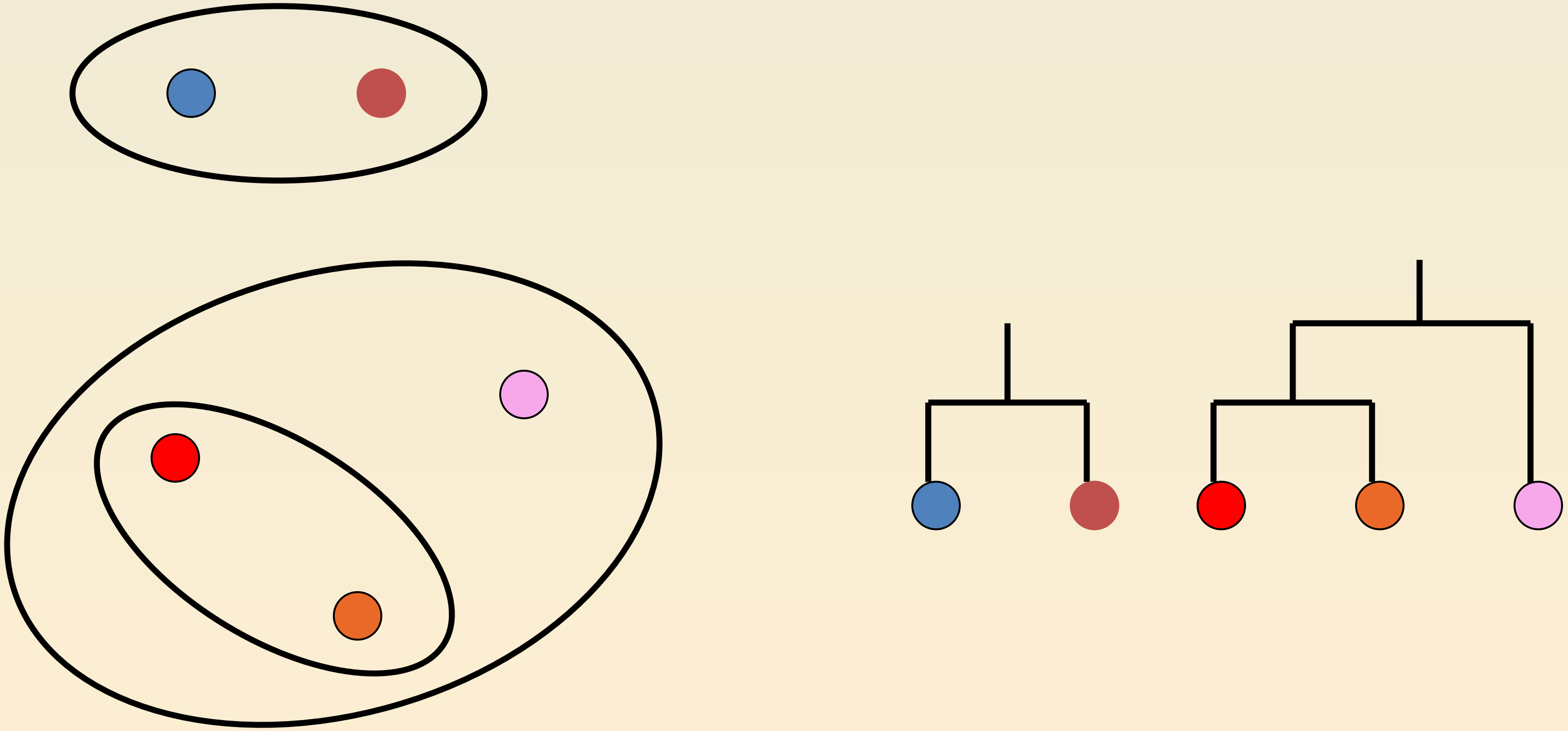
An example showing how UPGMA produces a rooted phylogenetic tree



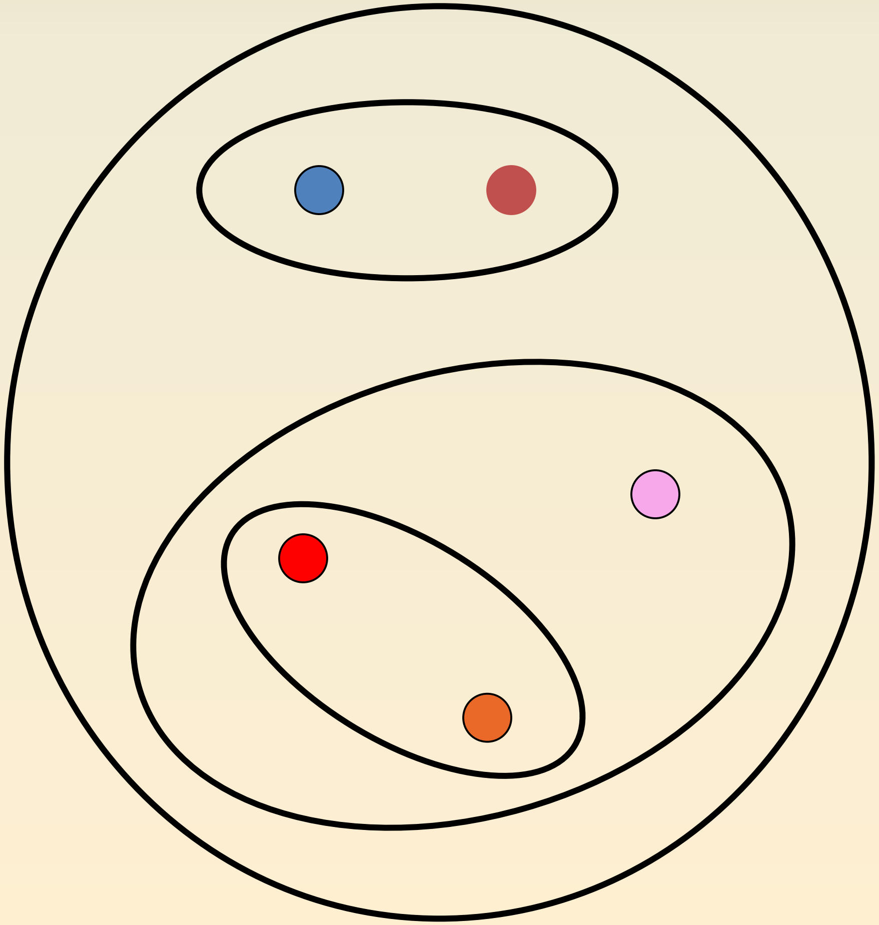
An example showing how UPGMA produces a rooted phylogenetic tree

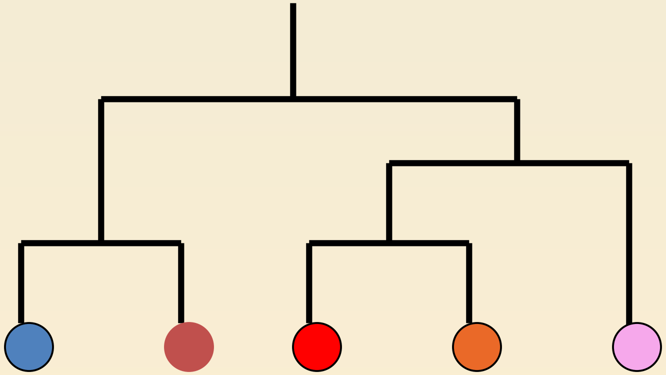


An example showing how UPGMA produces a rooted phylogenetic tree



An example showing how UPGMA produces a rooted phylogenetic tree





Methods of tree estimation

- Distance based - Phenetics
  - Minimum distance
    - Shortest summed branch lengths
- Character based
  - Maximum parsimony (MP)
    - Fewest character changes
  - Maximum likelihood (ML)
    - Highest probability of observing data, given a model
  - Bayesian
    - Similar to ML, but incorporates prior knowledge

Phenetics

Emphasizes the **overall similarity** of phenotypes in grouping and classifying taxa.

Maintains principles of Neo-Darwinism, but includes no estimation of processes.

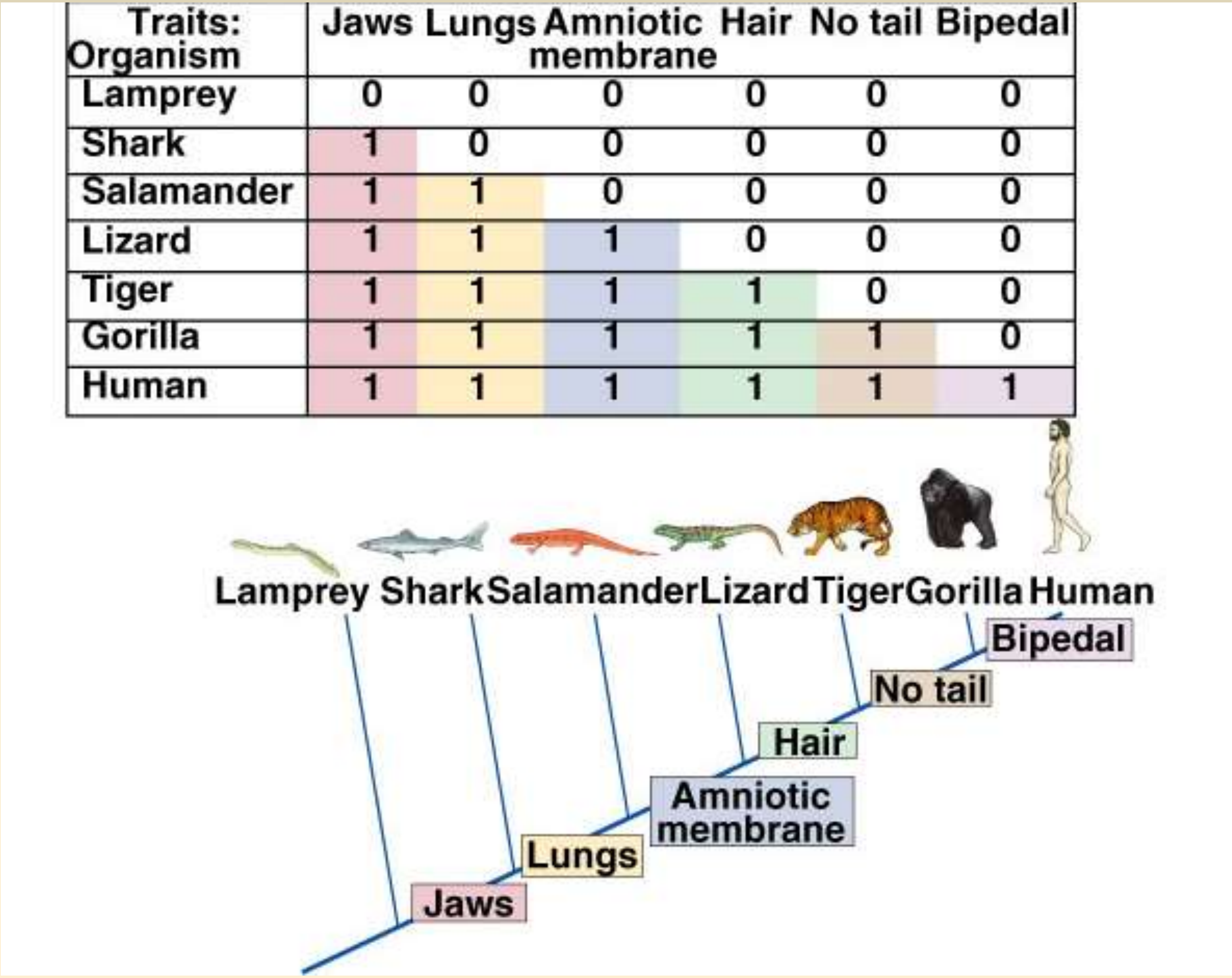
Largely methodological/operational. No philosophical basis.

Uses any and all data, as long as it can be quantified.

Resulting “trees” called “Phenograms.”

Statements of similarity only. Useful for summarizing resemblance

3. Cladistics (Phylogenetics) - Sequentially group taxa by shared derived character states (apomorphies)

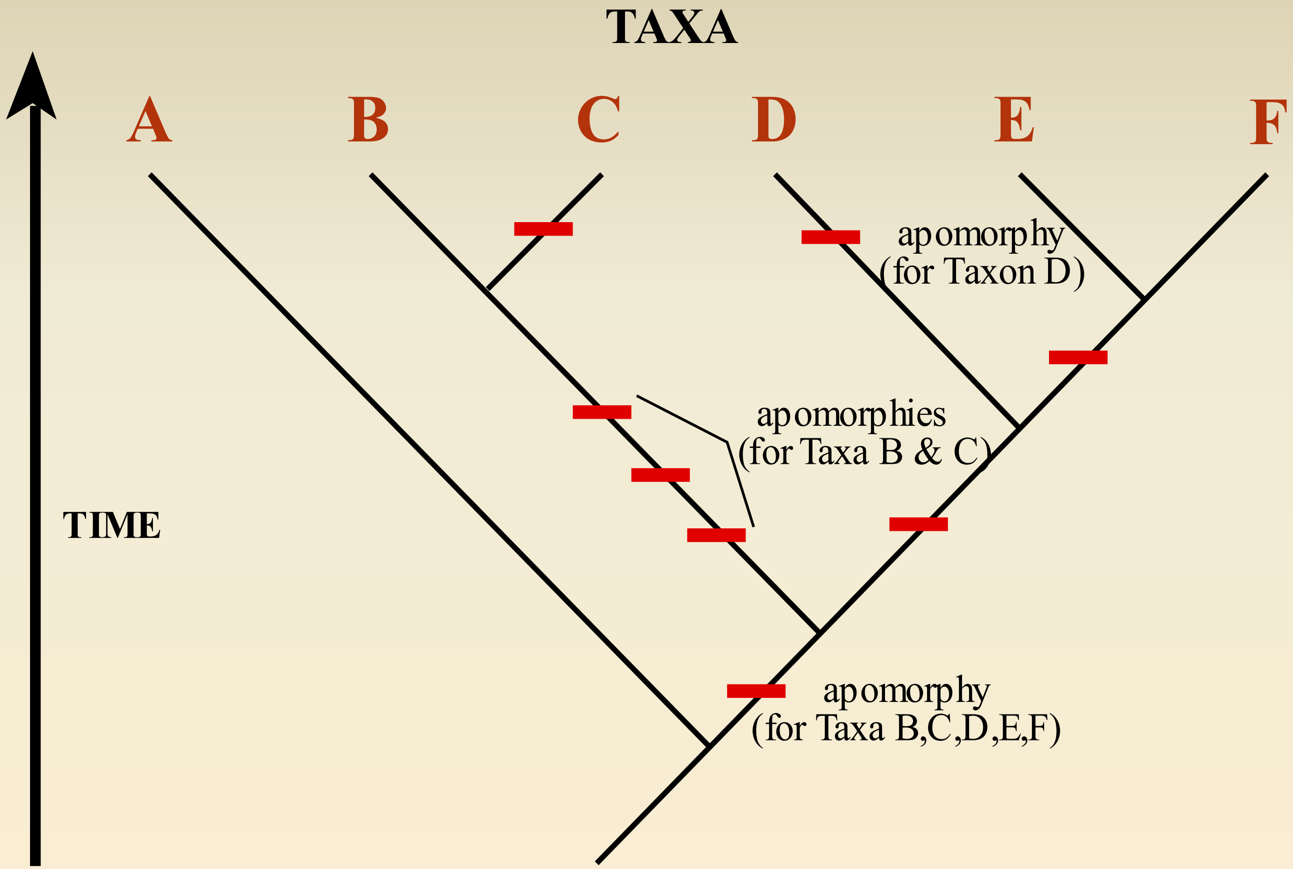


Traits: Organism	Jaws	Lungs	Amniotic membrane	Hair	No tail	Bipedal
Lamprey	0	0	0	0	0	0
Shark	1	0	0	0	0	0
Salamander	1	1	0	0	0	0
Lizard	1	1	1	0	0	0
Tiger	1	1	1	1	0	0
Gorilla	1	1	1	1	1	0
Human	1	1	1	1	1	1

Apomorphies are the result of evolution.

Taxa are grouped by shared apomorphies

Taxa sharing apomorphies underwent the same evolutionary history and should be grouped together.



apomorphy
(for Taxon D)

apomorphies
(for Taxa B & C)

apomorphy
(for Taxa B,C,D,E,F)

Cladogram or Phylogenetic Tree

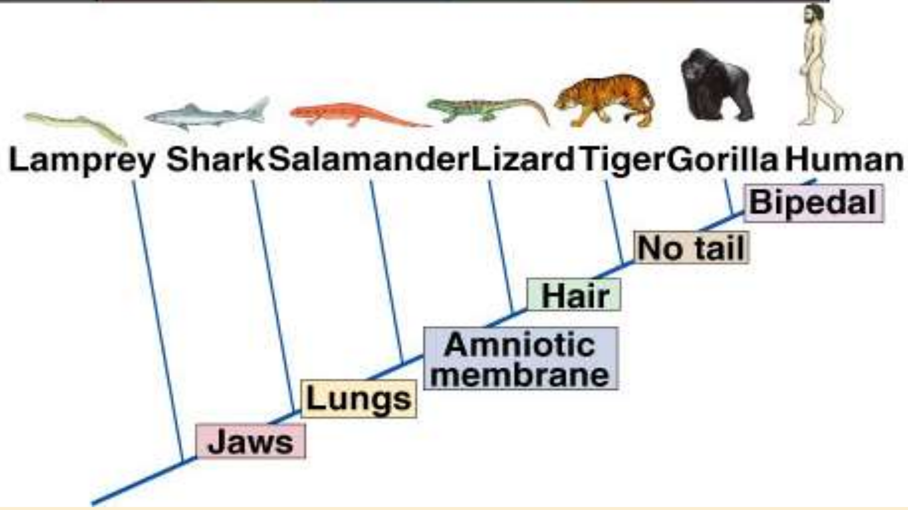
Outgroup comparison

A species or group of species closely related to, but not a member of, the group under study is designated an outgroup.

Character states exhibited by the outgroup are assumed ancestral, and other states are considered derived.

In this case, the Lamprey, a jawless fish, is the outgroup

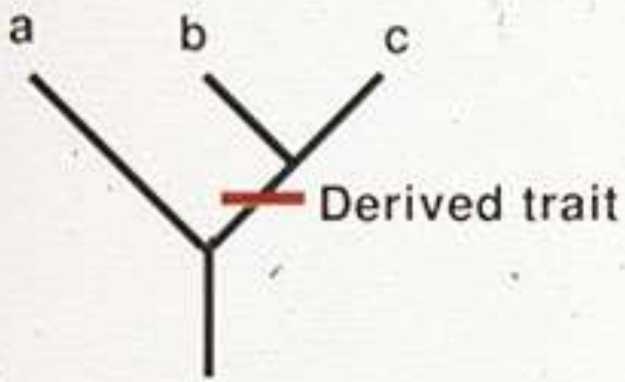
Traits: Organism	Jaws	Lungs	Amniotic membrane	Hair	No tail	Bipedal
Lamprey	0	0	0	0	0	0
Shark	1	0	0	0	0	0
Salamander	1	1	0	0	0	0
Lizard	1	1	1	0	0	0
Tiger	1	1	1	1	0	0
Gorilla	1	1	1	1	1	0
Human	1	1	1	1	1	1



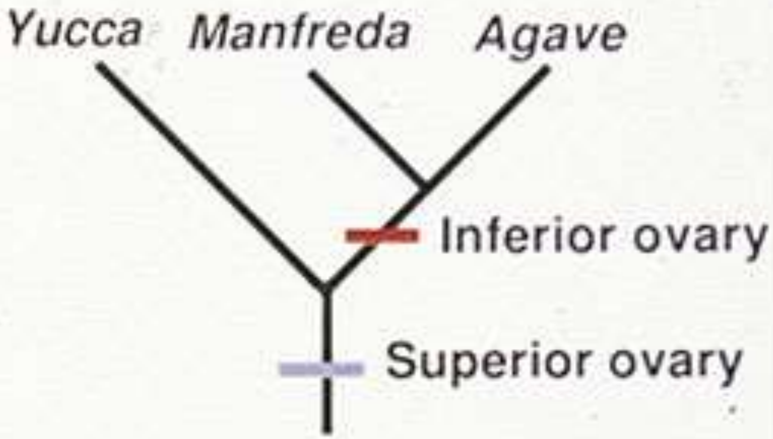
Cladistics

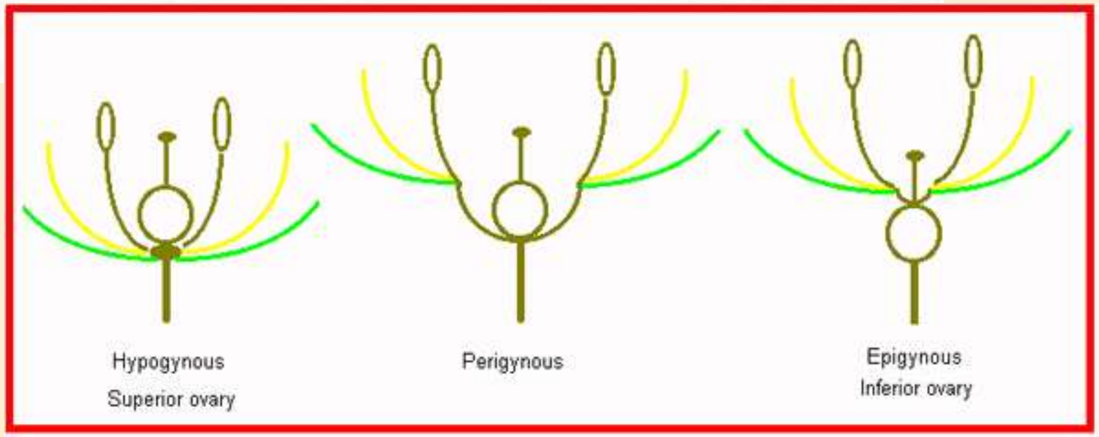
Phylogeny reconstruction

Shared derived characters



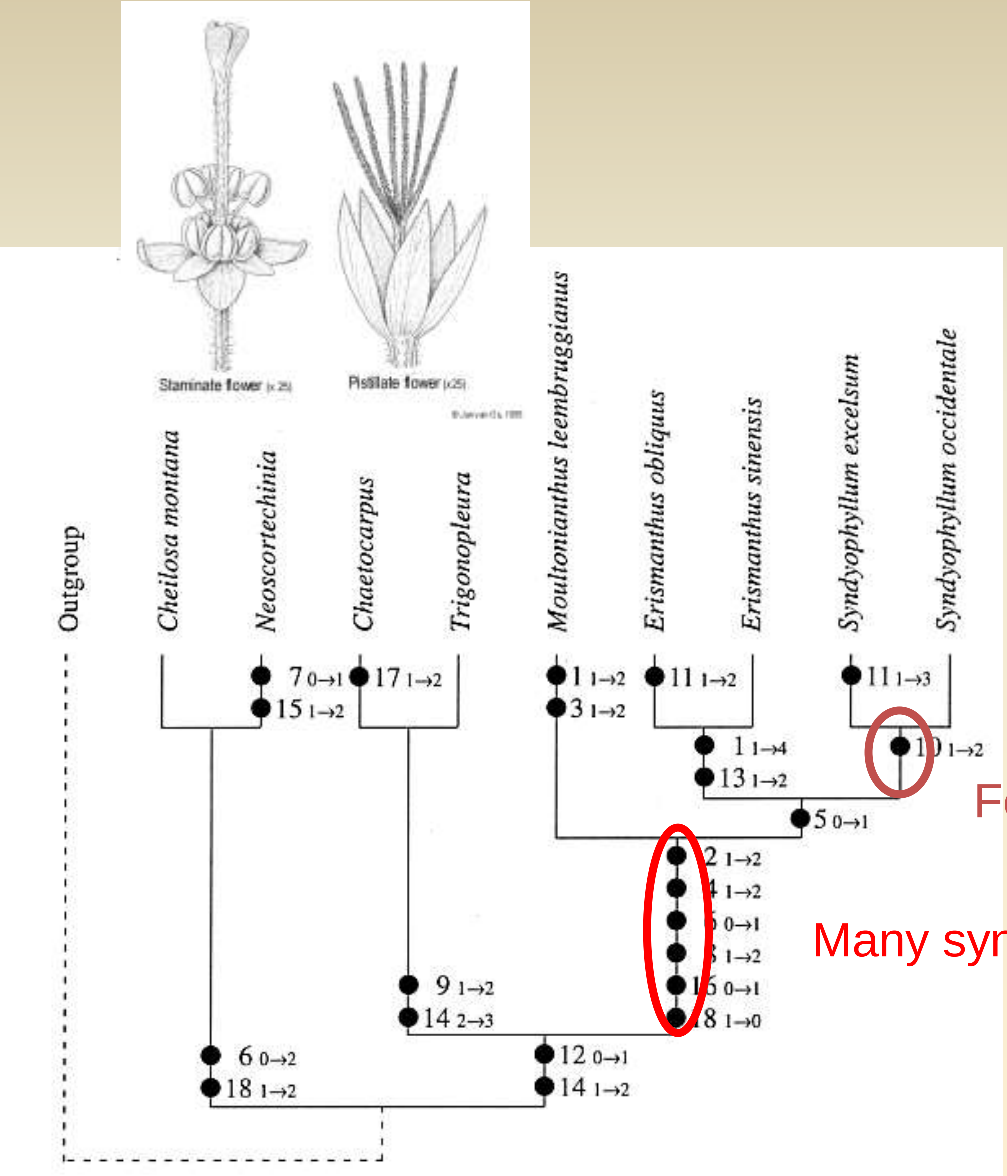
Common ancestor
Ancestral Condition





How can we tell how well a clade is supported?

In part, by the number of synapomorphies



Few synapomorphies = weaker support

Many synapomorphies = stronger support

Fig. 1. The phylogeny of the tribe Erismantheae (genera *Erismanthus*, *Moultonianthus*, and *Syndyophyllum*) in relation to the tribes Chaetocarpeae (*Chaetocarpus* and *Trigonopleura*) and Cheiloseae (*Cheilosa* and *Neoscortechinia*). Delimitation of tribes after Webster (1994).

Cladistics

- By definition, homology indicates evolutionary relationship — when we see a shared homologous character in two species, we know that they share a common ancestor
- Build phylogenetic trees by analyzing shared homologous characters
- Of course, we still have the problem of deciding which shared similarities are homologies and which are homoplasies.

Steps for General Parsimony Analysis:

Study specimens. Gather data.

Analyze characters, establish polarity if possible, or choose outgroup.

Create data matrix (Excel, Mesquite), taxa on one axis, characters on the other.

Score each characters (0,1) for all taxa.

Use computer program to find most parsimonious tree.

Common programs: Mega, PAUP, TNT

Maximum Parsimony Analysis

All possible trees are determined for each position of the sequence alignment

Each tree is given a score based on the number of evolutionary step needed to produce said tree

The most parsimonious tree is the one that has the fewest evolutionary changes for all sequences to be derived from a common ancestor

Usually several equally parsimonious trees result from a single run.

Finding optimal trees - heuristics

The number of possible trees increases exponentially with the number of taxa making exhaustive searches impractical for many data sets (an NP complete problem)

Heuristic methods are used to search tree space for most parsimonious trees by building or selecting an initial tree and swapping branches to search for better ones

The trees found are not guaranteed to be the most parsimonious - they are best guesses

How many possible trees?

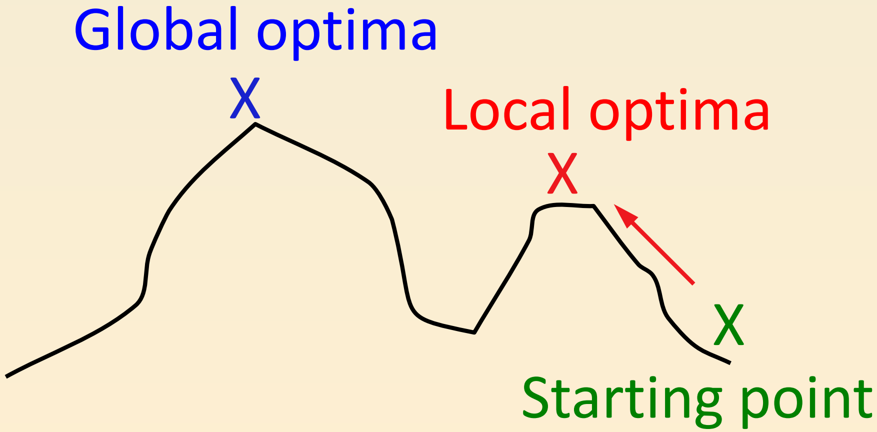
Ingroup taxa	Number of trees
1	1
2	1
3	3
5	105
10	34,459,425
50	2.75292×10^{76}

Reducing the time for searching “tree space”

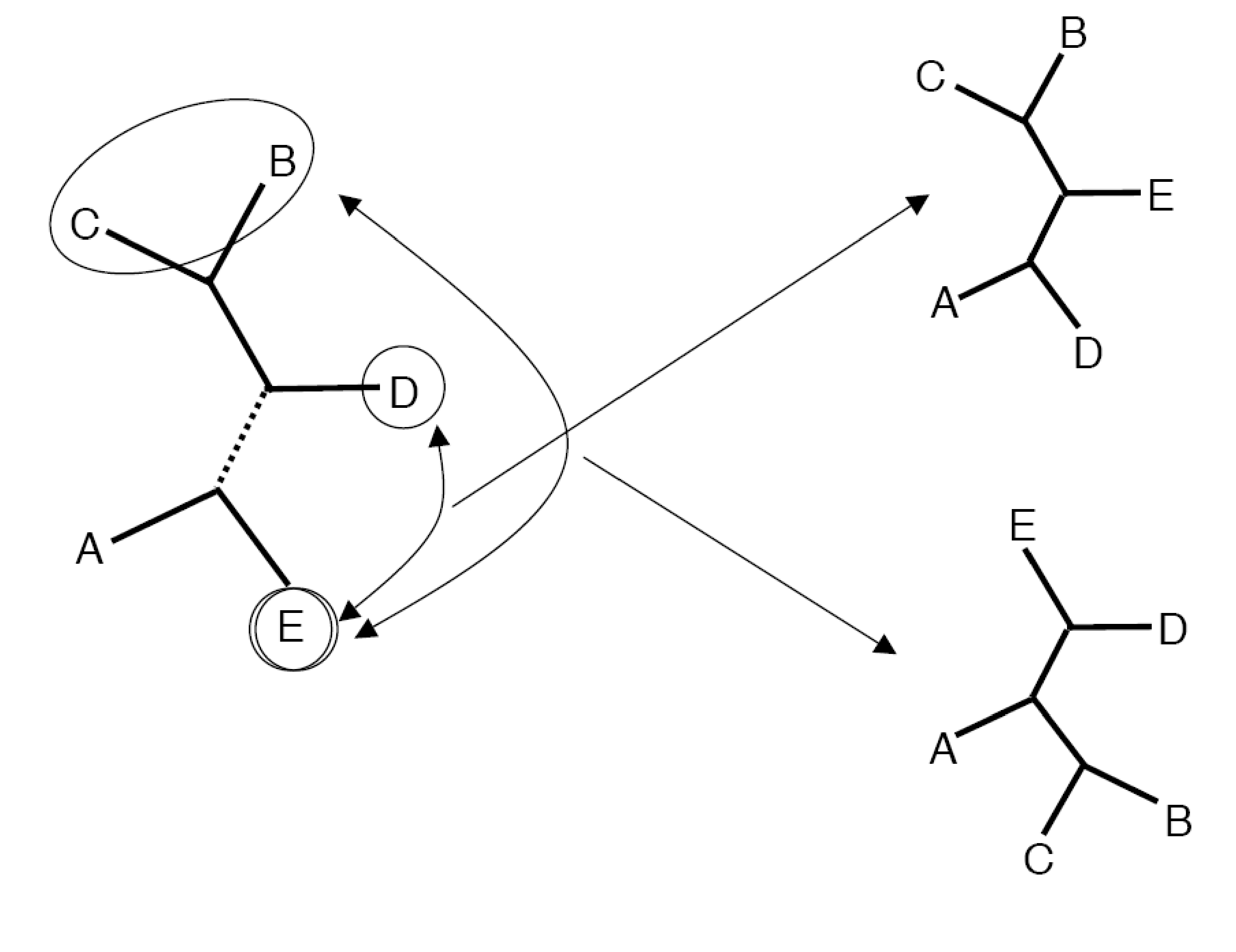
Heuristic search

Find an initial tree, and move within near-by tree-space, discarding worse alternatives

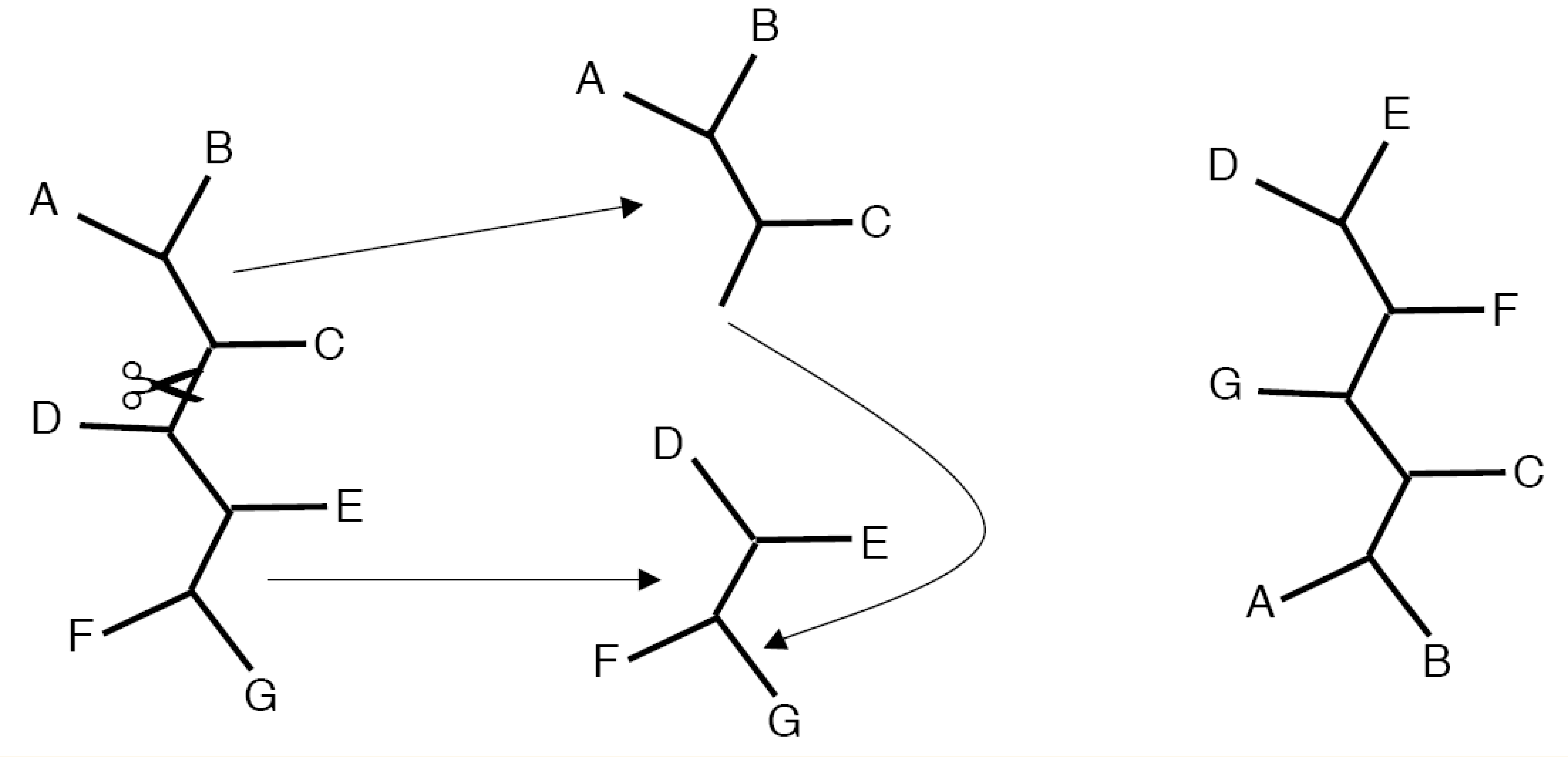
Only a small amount of tree-space is searched and there is no guarantee of finding the optimal tree - can be trapped in local maxima



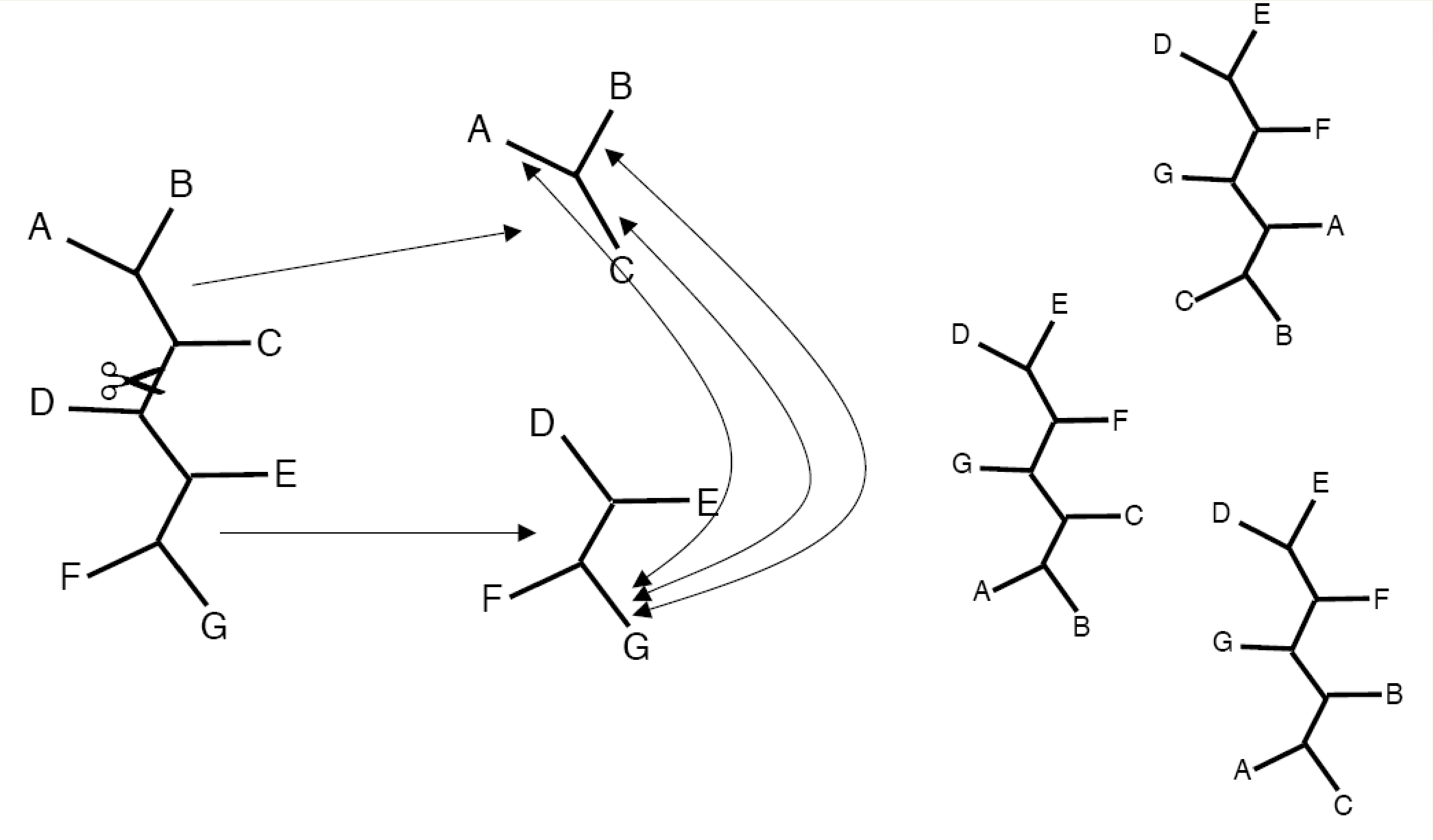
Branch Swapping: Nearest-Neighbor Interchange



Branch Swapping: Subtree Pruning and Regrafting



Branch Swapping: Tree Bisection and Reconnection



Principle of Parsimony

That cladogram (tree) having the fewest number of “steps” (evolutionary changes) is the one accepted.

The ‘most-parsimonious’ tree is the one that requires the fewest number of evolutionary events (*e.g.,* nucleotide substitutions, amino acid replacements) to explain the sequences.

Okham’s razor: the simplest explanation is the best.

Results of parsimony analysis

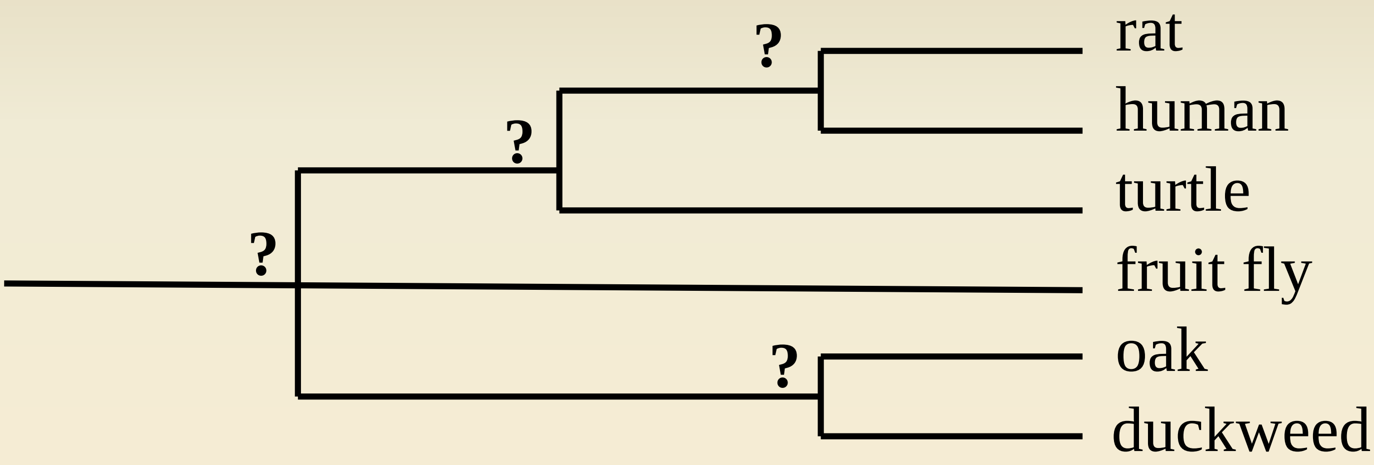
- One or more most parsimonious trees.
- Hypotheses of character evolution associated with each tree (where and how changes have occurred).
- Branch lengths (amounts of change associated with branches)
- Various tree and character statistics describing the fit between tree and data
- Suboptimal trees - optional

Consistency index

- ***Homoplasy:*** Multiple emergence of the same state in a phylogeny
- Perfect fit (= compatible characters) $\Rightarrow$ no homoplasy
- Let m_i = min #(steps possible for site *i*) and s_i = min #(steps for site *i* given the tree)

  Minimum # steps divided by actual number of steps
- The ***consistency index*** is C.I. = $\sum m_i / \sum s_i$ ($0 < \mathrm{CI} \leq 1$)
- **CI measures amount of homoplasy in tree**

How confident are we about the inferred phylogeny?



Bootstrap support values

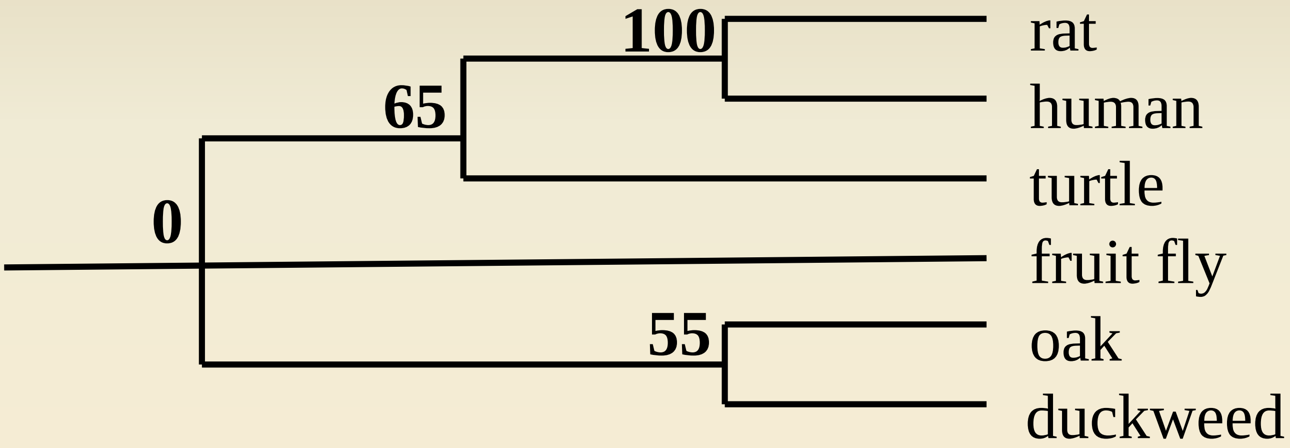
Characters are **resampled with replacement** to create many bootstrap replicate data sets (pseudosamples)

Each bootstrap replicate data set is analyzed.

Process is replicated 100x, 1000x, or more

Frequency of occurrence of a group (bootstrap proportions) is a measure of support for the group

Bootstrap values



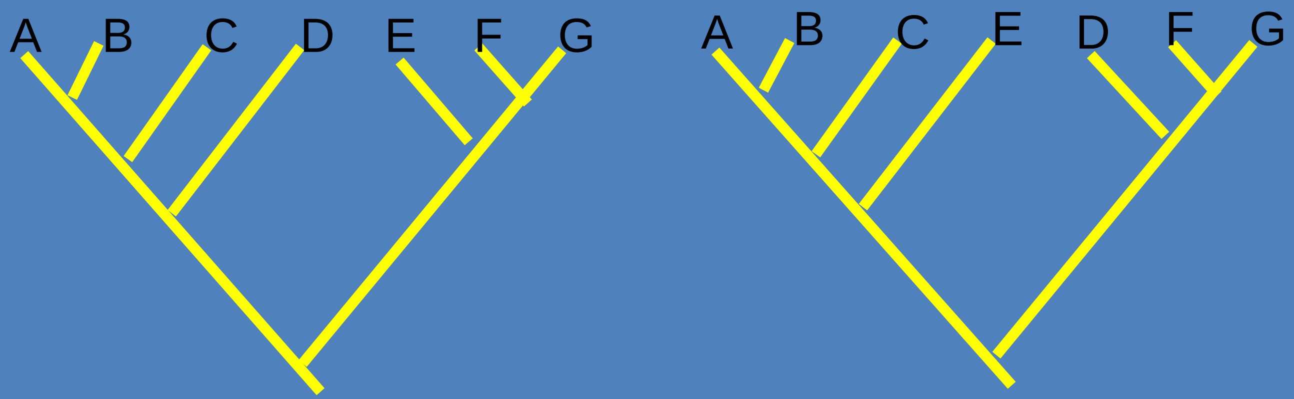
- Values are in percentages
- Conventional practice: only values 60-100% are shown

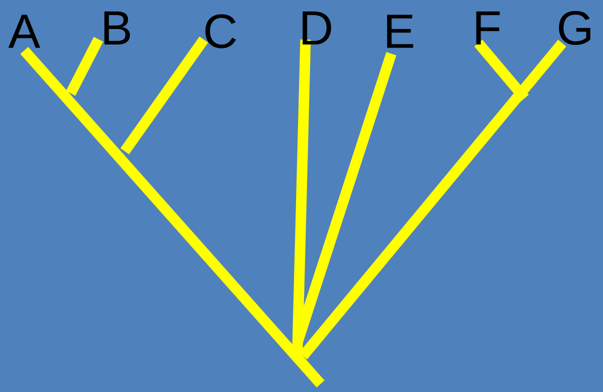
Multiple optimal trees

- Many methods can yield multiple equally optimal trees
- We can further select among these trees with additional criteria, but
- Typically, relationships common to all the optimal trees are summarised with ***consensus trees***

Strict consensus methods

Two Equally Parsimonious Tree





Collapse the two nodes that are different

STRICT CONSENSUS TREE

Parsimony - advantages

- is a simple method - easily understood operation
- does not seem to depend on an explicit model of evolution
- gives both trees and associated hypotheses of character evolution
- should give reliable results if the data is well structured and homoplasy is either rare or widely (randomly) distributed on the tree

Parsimony - disadvantages

- May give misleading results if homoplasy is common or concentrated in particular parts of the tree, e.g:
  - base composition biases
  - long branch attraction
- Underestimates branch lengths
- Parsimony often justified on purely philosophical grounds - We prefer the simplest hypotheses

  - But this is not always the case.

Maximum Likelihood: The explanation that makes the observed outcome the most likely

$$L = \Pr(D|H)$$

Probability of the data, given an hypothesis

The hypothesis is a tree topology, its branch-lengths and a model under which the data evolved

First use in phylogenetics: Cavalli-Sforza and Edwards (1967) for gene frequency data; Felsenstein (1981) for DNA sequences

Maximum Likelihood

Creates all possible trees like Maximum Parsimony method but instead of retaining trees with shortest evolutionary steps……

Employs a **model of evolution** whereby different rates of transition/transversion (A->T, G->C) ration can be used

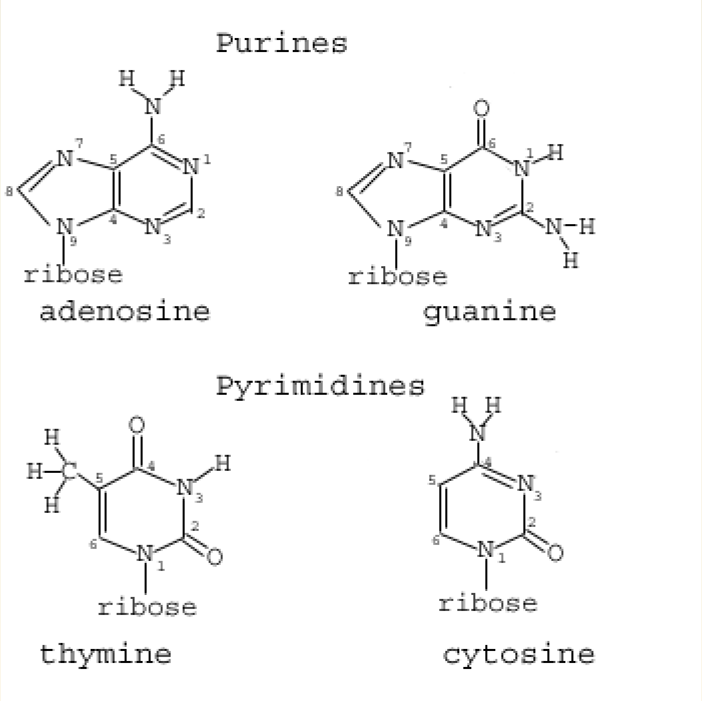
Each tree generated is calculated for the probability that it reflects each position of the sequence data.

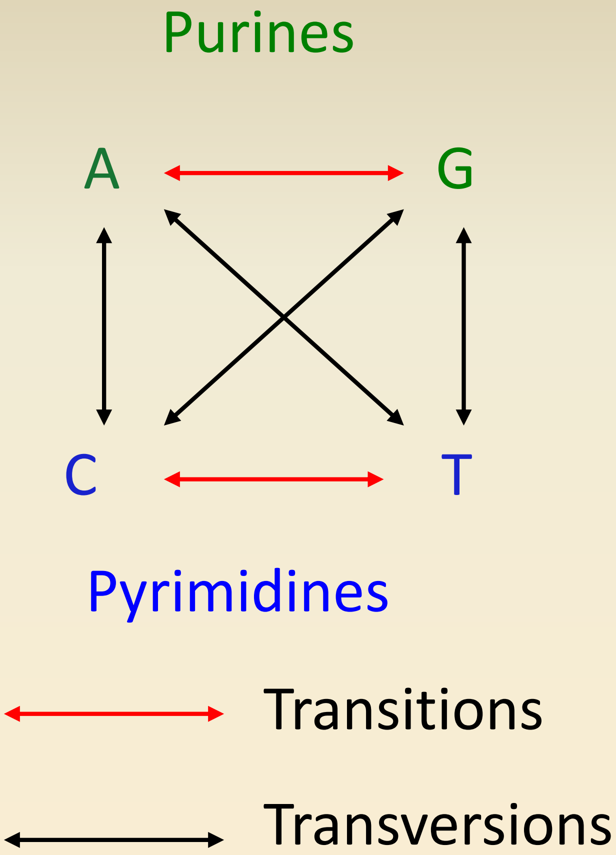
Calculation is repeated for all nucleotide sites

Finally, the tree with the best probability is shown as the maximum likelihood tree - usually only a single tree remains

It is a more realistic tree estimation because it does not assume equal transition-transversion ratio for all branches.

Transitions and Transversions





Purines

A G

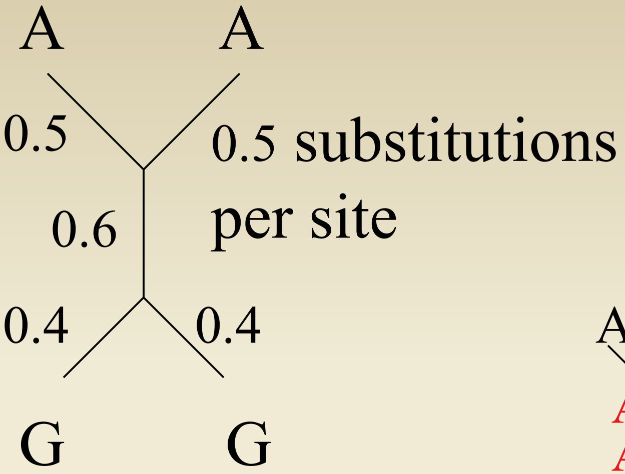
C T

Pyrimidines

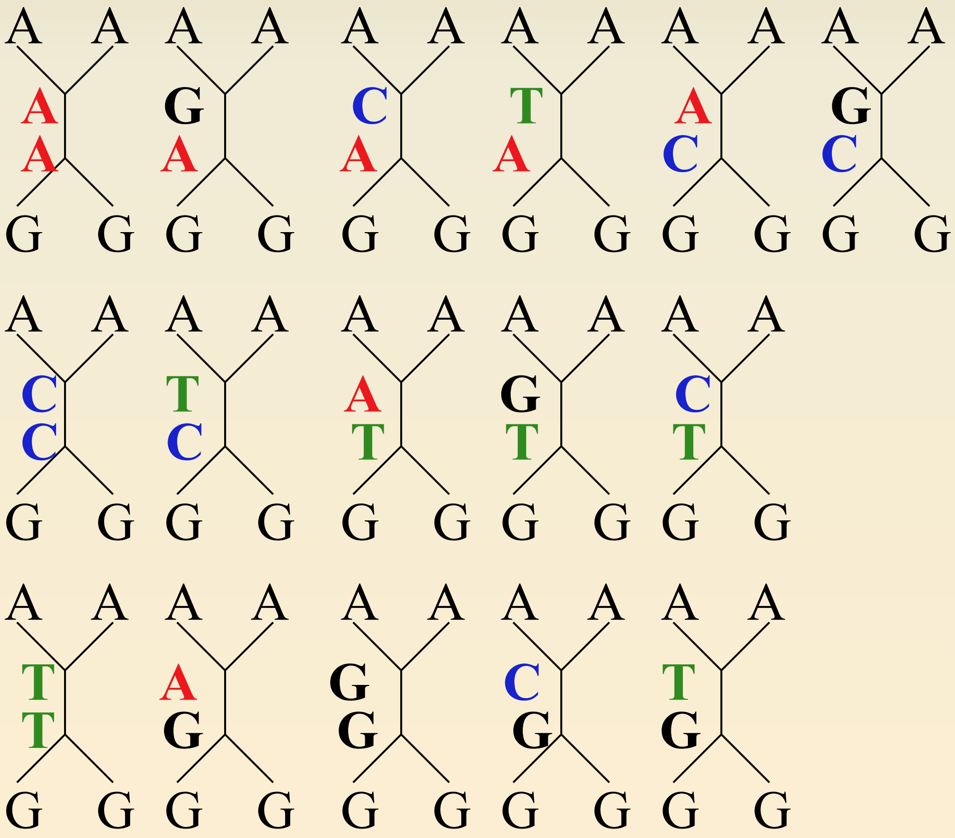
Transitions

Transversions

Substitution is more likely to change from A to G, than A to T



Model of rate change e.g. Kishino-Hasegawa (1985): 4 base frequencies, transition/transversion (ti/tv ratio)



Sum the probabilities for each of the 16 internal node combinations to get the likelihood for this single nucleotide site

The likelihood of a tree is the product of the site likelihoods. Taken as natural logs, the site likelihoods can be summed to give the log likelihood:

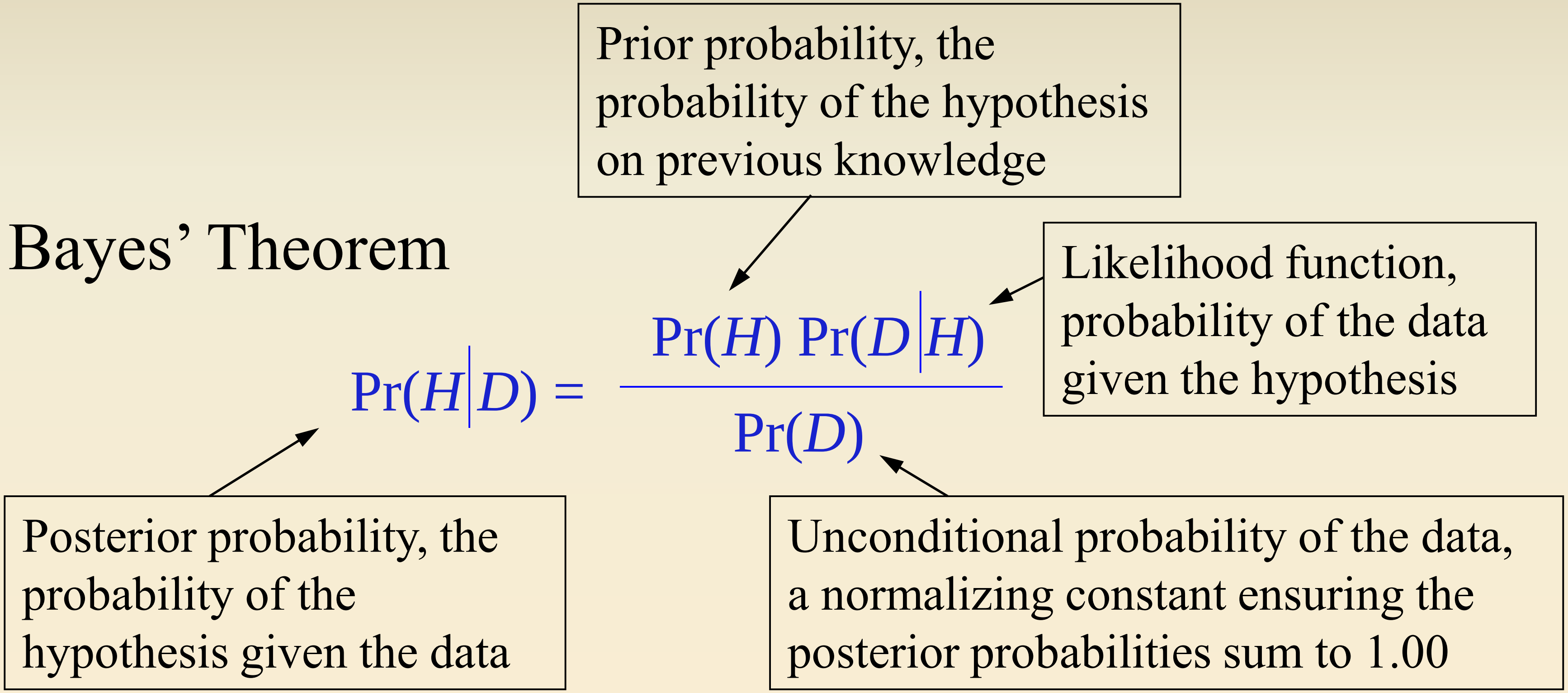
The tree with the highest –ln*L* is the ML tree

- ML is computationally intensive (slow)

- If branch-lengths are long, such that substitutions occur multiple times along the same branch for the same site, ML will be more consistent than MP – if the evolutionary process is sufficiently well modelled.

Bayesian Inference: The explanation with the highest posterior probability

Bayes' Theorem



$$\Pr(H|D) = \frac{\Pr(H)\ \Pr(D|H)}{\Pr(D)}$$

Prior probability, the probability of the hypothesis on previous knowledge

Likelihood function, probability of the data given the hypothesis

Posterior probability, the probability of the hypothesis given the data

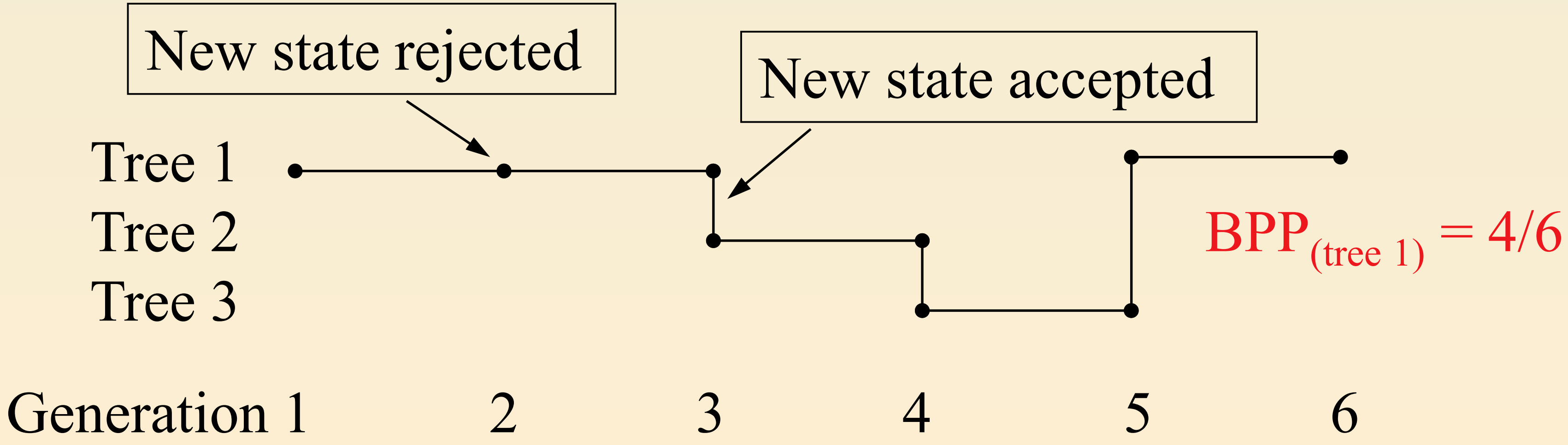
Unconditional probability of the data, a normalizing constant ensuring the posterior probabilities sum to 1.00

First use in phylogenetics: Li (1996, PhD thesis), Rannala and Yang (1996)

Bayesian inference in phylogenetics is essentially a likelihood method, but may more closely reflect the way humans think.

- It is Informed by prior knowledge (e.g. fossil data)
- emphasis is placed on $\Pr(H|D)$ instead of $\Pr(D|H)$

Markov chain Monte Carlo (MCMC) is used to approximate Bayesian posterior probabilities *(BPP) over 1,000s – 1,000,000s of generations



New state rejected

New state accepted

Tree 1

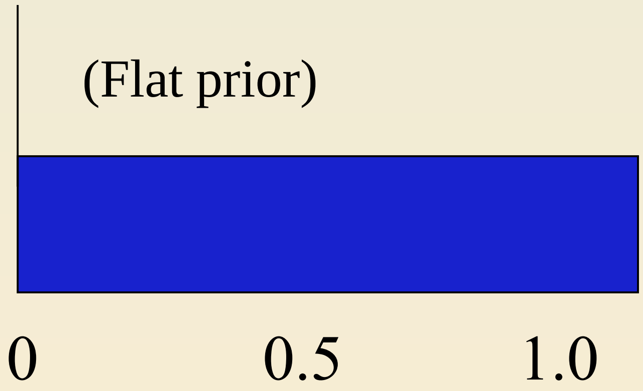
Tree 2

Tree 3

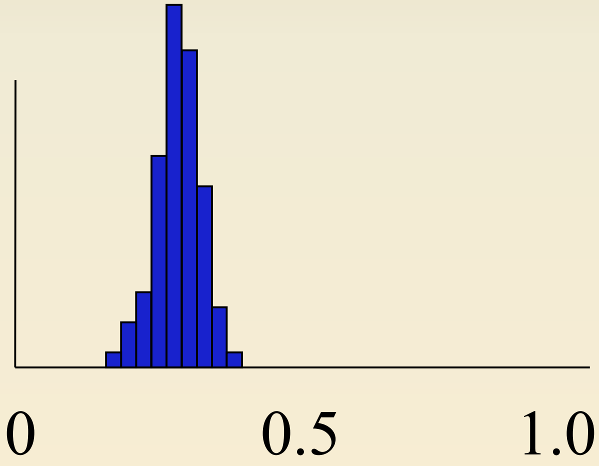
$\text{BPP}_{(\text{tree 1})} = 4/6$

Generation 1 2 3 4 5 6

Posterior probabilities are integrated over all trees in the posterior distribution – providing density distributions rather than the optimization of likelihood



Prior for a parameter value (e.g. proportion of invariant sites)



Posterior for the proportion of invariant sites

End