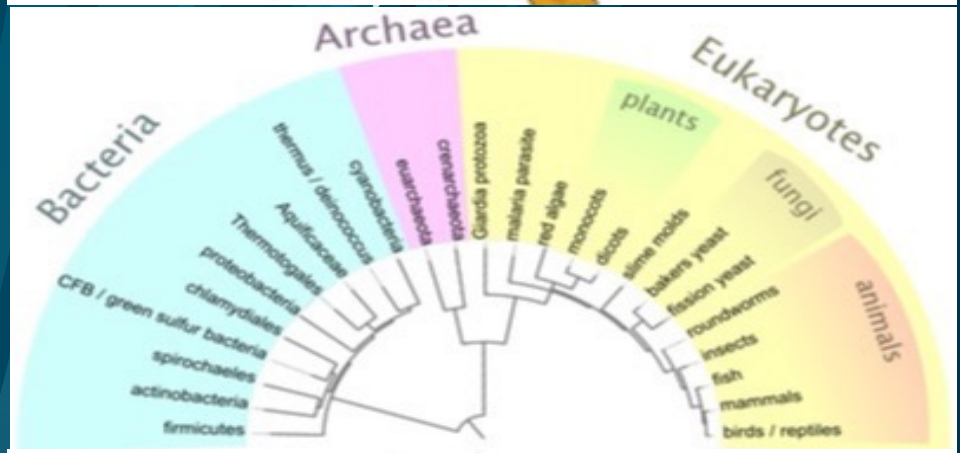
UiT The Arctic University of Norway

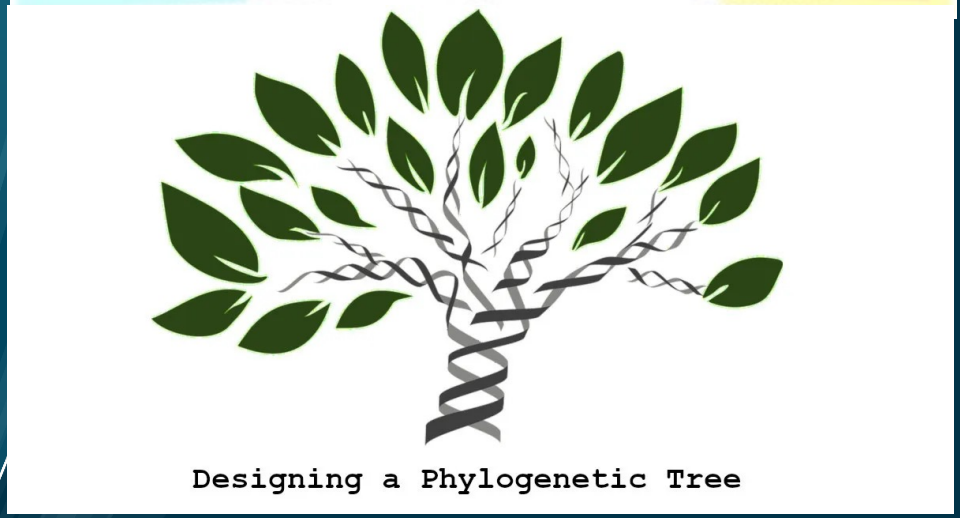
# Phylogenetic inferences

***Tree of life***

shripathi.bhat@uit.no

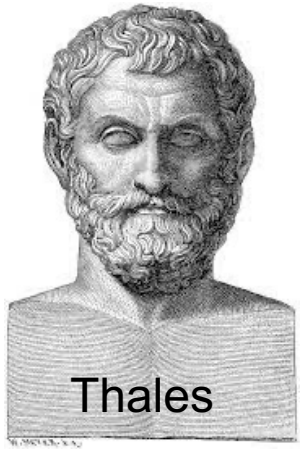
*Bioinformatician , RGG, NFH*



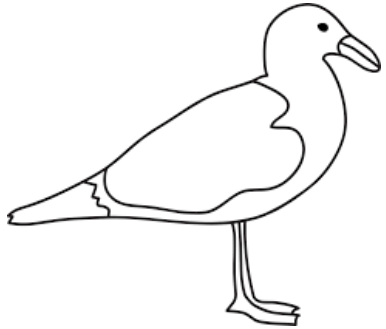

Google images

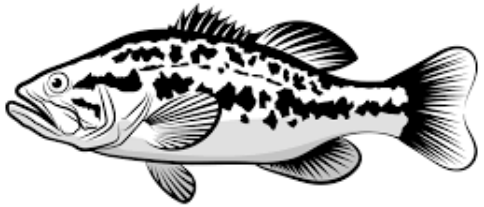
# Origin of phylogenetics

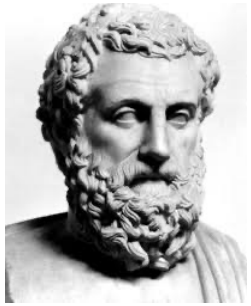



"*Birds in a way resemble fishes. For birds have their wings in the upper part of their bodies, and fishes have two fins in the front part of their bodies. Birds have feet on their under part, and most fishes have a second pair of fins in their under part…*"

**-Aristotle (384-322 B.C), *De Incessu Animalium*.**







Google images

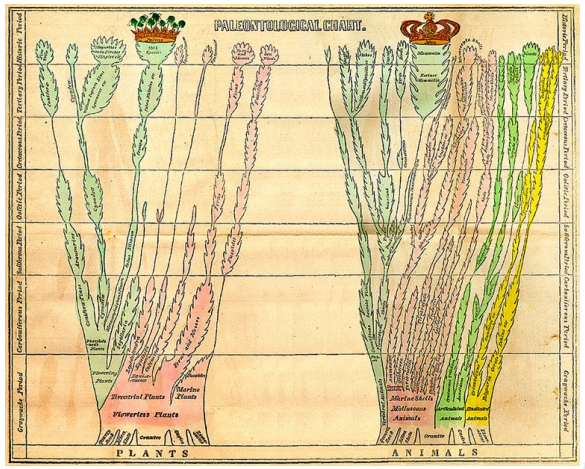
# Precursor for modern phylogenetics (Tree of life)

- Lamarck ’s works on branching tree (?) of anima

WRONG!

- Edward Hitchcocks’s  paleontology based tree of life

WRONG!

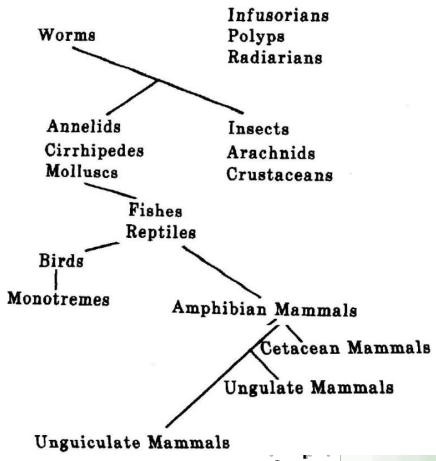



- Robert Chamber’s Vestiges of natural history of creation

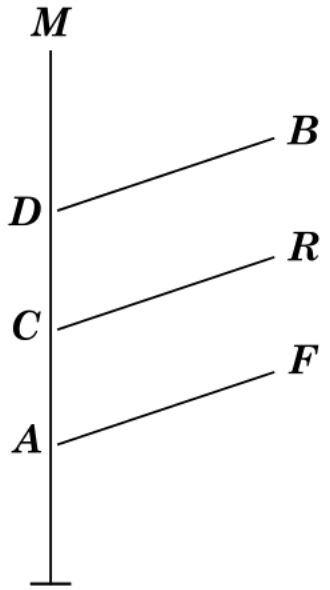
WRONG!

TABLE

SHOWING THE ORIGIN OF THE VARIOUS ANIMALS

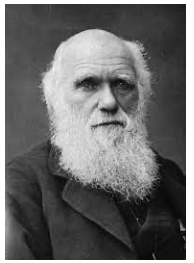


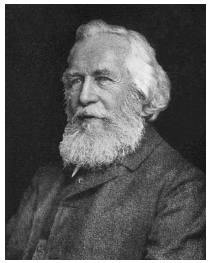

forms → Complex forms




Google images

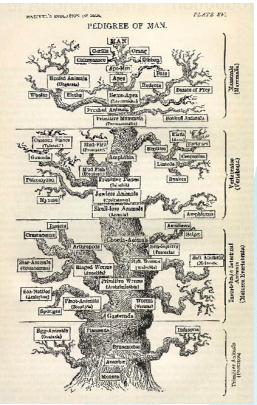
# Modern concept of tree of life





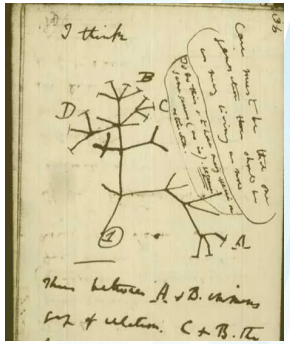
- Proposed by Charles Darwin and Ernest Haeckel

- Biodiversity rose due to the principle of descent with modification




- First time represented relationships between living organisms as trees

- On the Origin of species: principle of survival of the fittest




“ The affinities of all of the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during former years may represent the long succession of extinct species […] The limbs, divided into great branches, and these into lesser and lesser branches, were themselves once, when the tree was young, budding twigs, and this connection of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups”

-Charles Darwin in On the origin of species

Google images

# Evolution

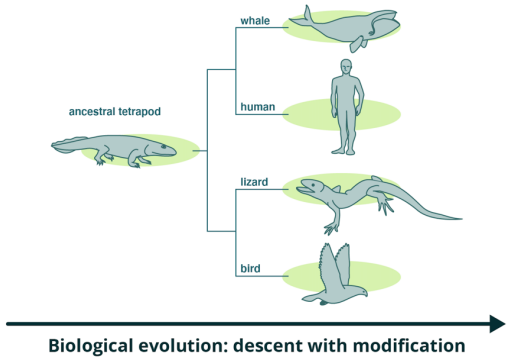
- Change over the time is not evolution.
  Caterpillar to moth, or trees loosing leaves is NOT evolution
- Then what is evolution?
- Phylogenetics is all about studying biological evolution

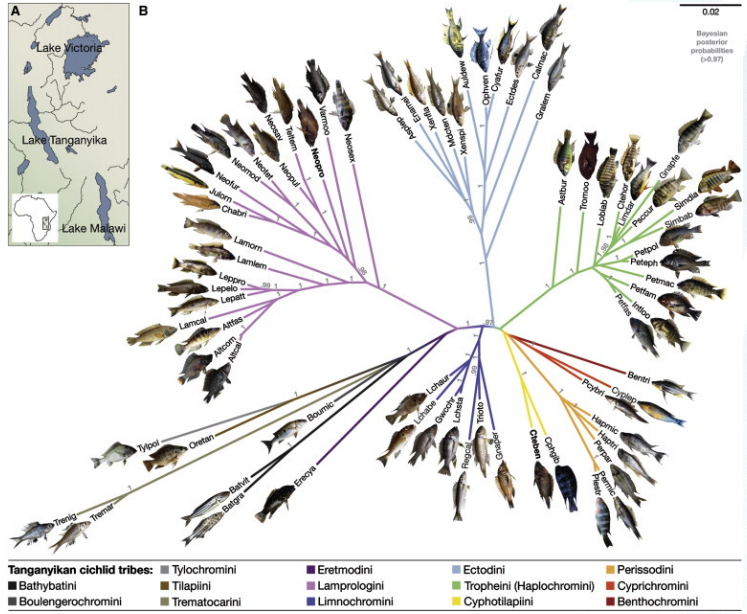
Change over time, not biological evolution

# Evolution: descent with inherited modification

- Biological evolution is simply a descent with inherited modification
- It can be as simple as change in allele frequency in a population from one generation to next generation = small scale evolution
- Or it can be the descent of different species from one common ancestor over many generations (millions of years) = large scale evolution
- We are all distant cousins

Humans, whales, Oaktree, sharks, seagulls




Muschick et al)

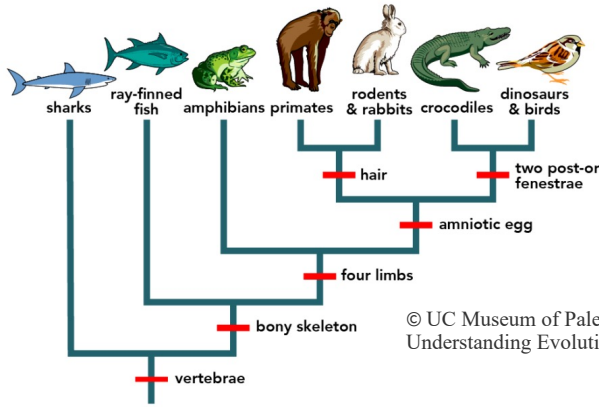
# Evolution

- The main idea of evolution is “life has a history”: which has changed over short (long) period of time, and different species share common ancestors

- The process of evolution produces a pattern of relationships between species

- Lineages evolve, split, modifications are inherited

  Evolutionary paths diverge (branching pattern) <-phylogeny



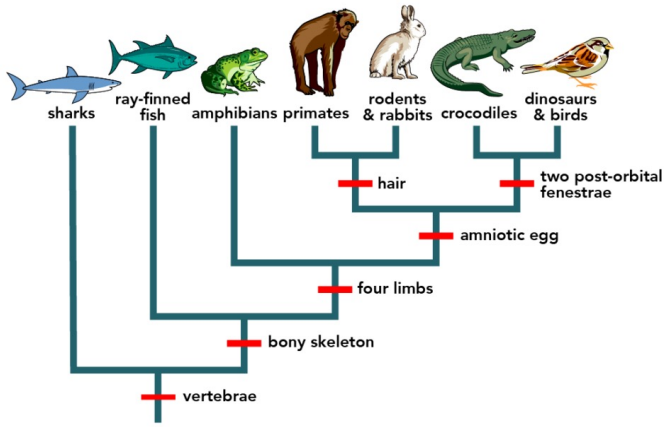
# Phylogeny

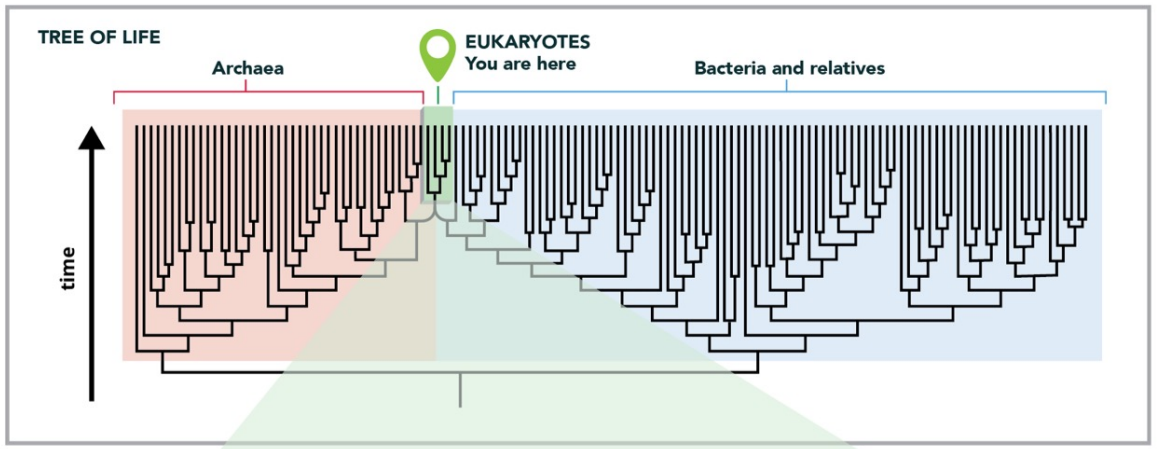
- Phylogen(y)etics is a reconstruction of evolutionary relationships/history among biological entities

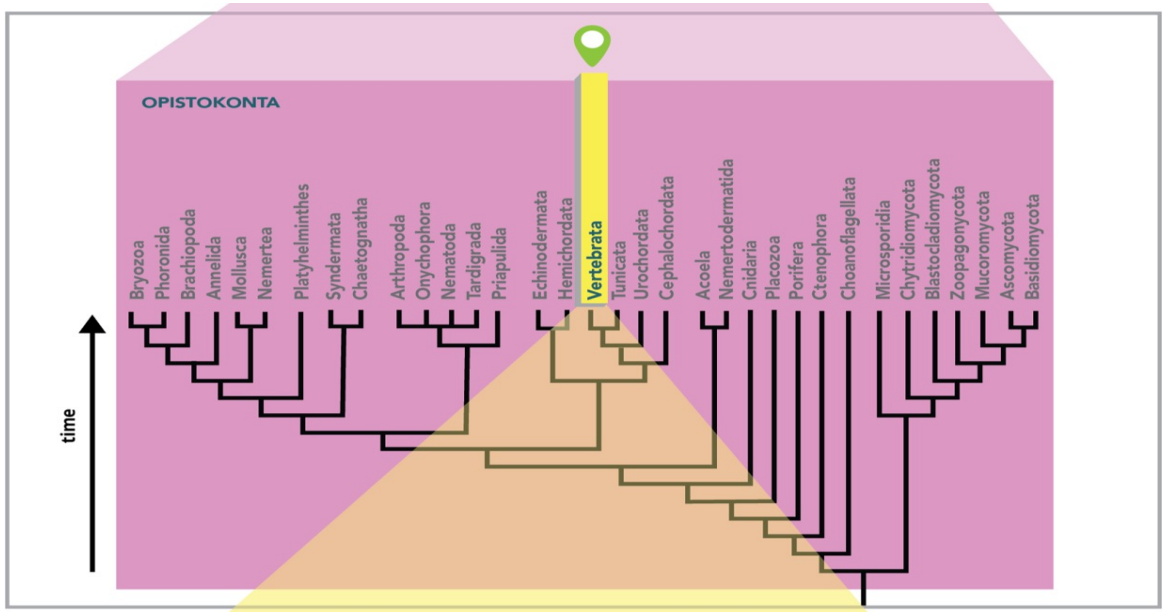
➢(Life) biological entities: species, individuals, characters, genes, proteins etc (anything you can measure in all species you are interested)

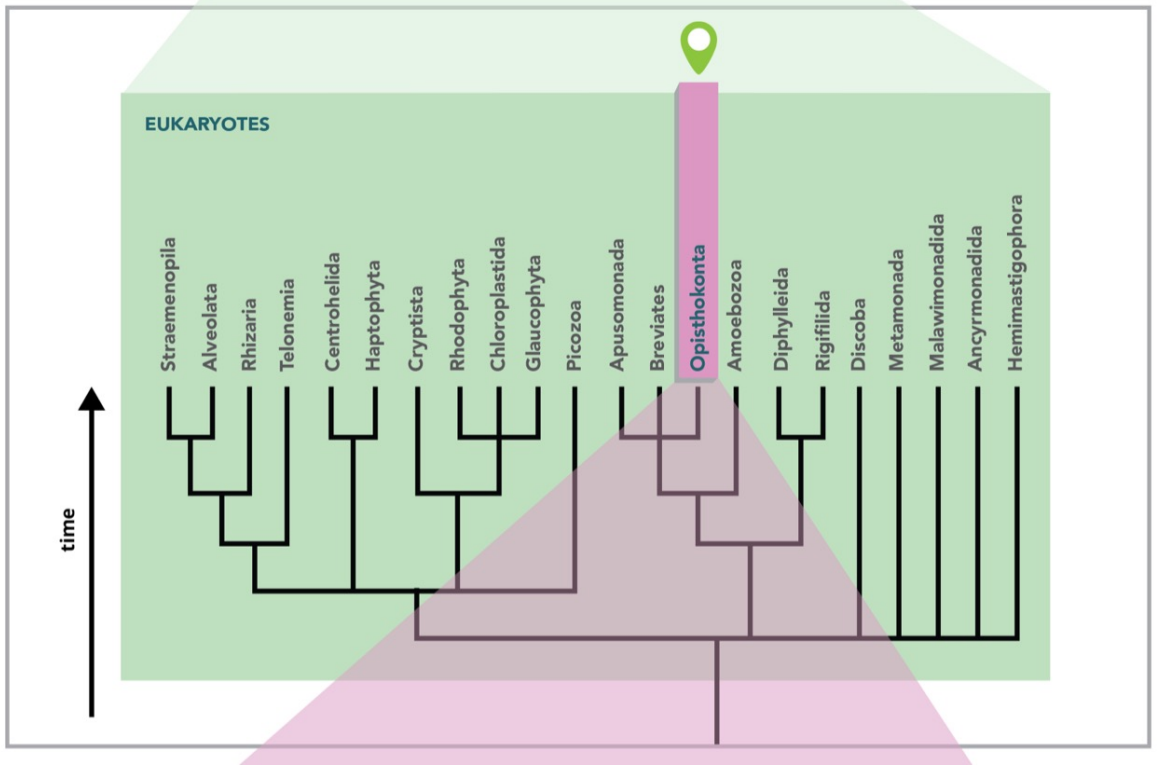
➢You need a hypothesis about the proposed relationship

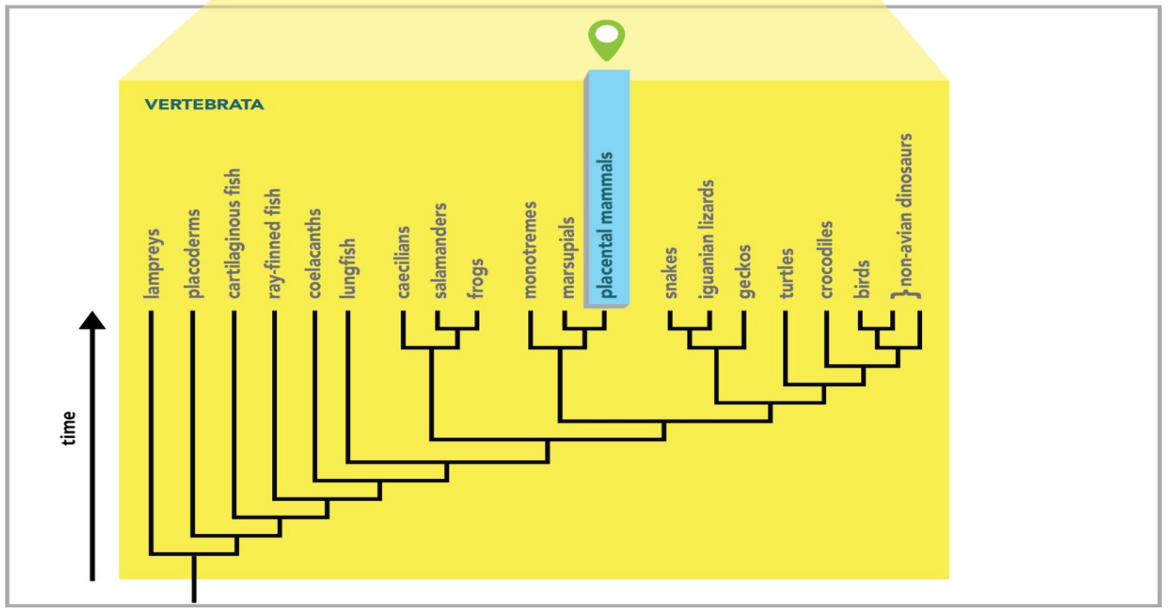
Molecular phylogenetics : when you use chemical molecules such as DNA or protein to make phylogenetic tree





# Molecular phylogenetics

- A set of techniques that enable the evolutionary relationships between DNA sequences to be inferred by making comparisons between those sequences

- **Objective of the molecular phylogenetics**:

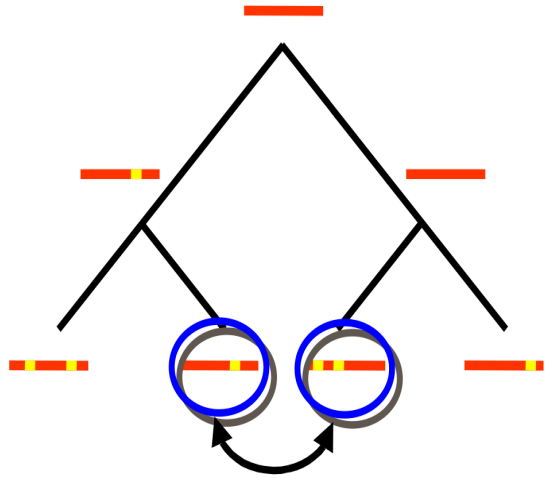
"IF GENOMES EVOLVE by the gradual accumulation of mutations, then the amount of difference in nucleotide sequence between a pair of genomes should indicate how recently those two genomes shared a common ancestor.

Two genomes that diverged in the recent past would be expected to have fewer differences than a pair of genomes whose common ancestor is more ancient. This means that by comparing three or more genomes it should be possible to work out the evolutionary relationships between them" -Genomes (2nd edition)

## Simple idea

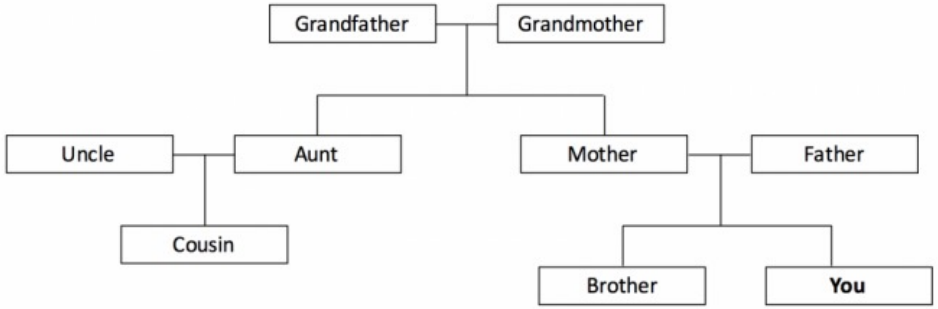
- From a common ancestor two DNA sequence diverged
- Each of these two sequences starts to accumulate nucleotide substitutions
- The number of mutations are used in molecular evolution analysis
- If two sequences of length N differ from each other at n s
  divergence between them is (n/N)*100
- Hence sequence analysis forms important part of

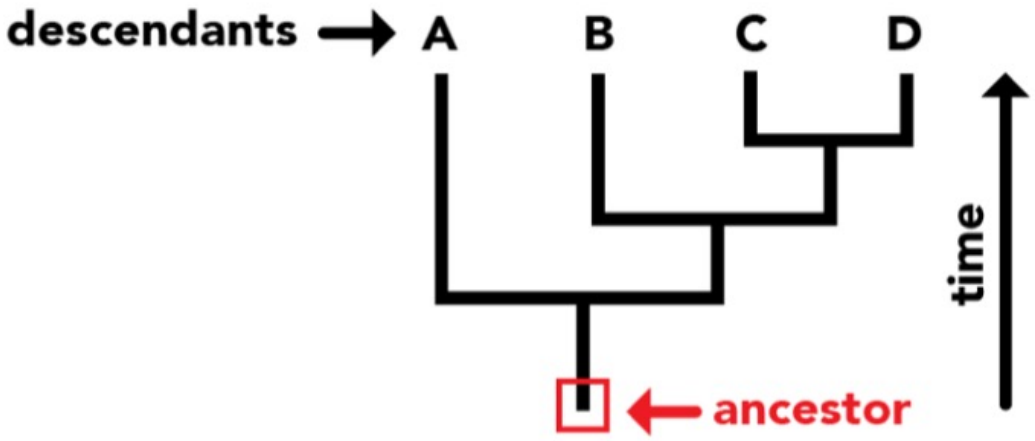
Phylogenetic tree construction

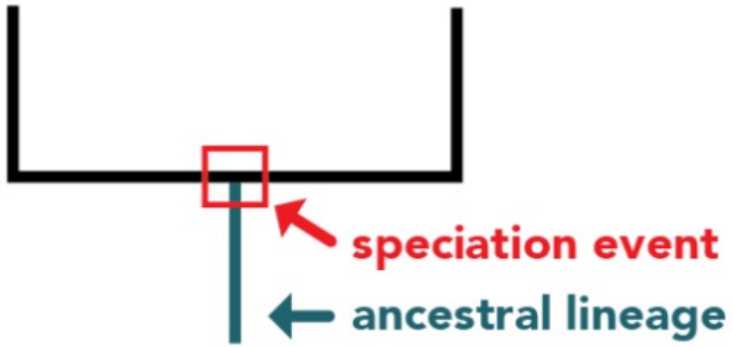


# Understanding phylogenies

- Phylogeny is like a family tree
- Root  represent ancestors
- Tips represent descendants
- Lineage split represent speciation(split) events



# Anatomy of phylogenetics (tree)

- An illustration of the evolutionary relationships among a group of organisms

## Anatomy of a tree

- Topology
- Branches
- Nodes
  - Tips/taxon
  - Internal nodes (ancestors)
  - Root
- Confidence

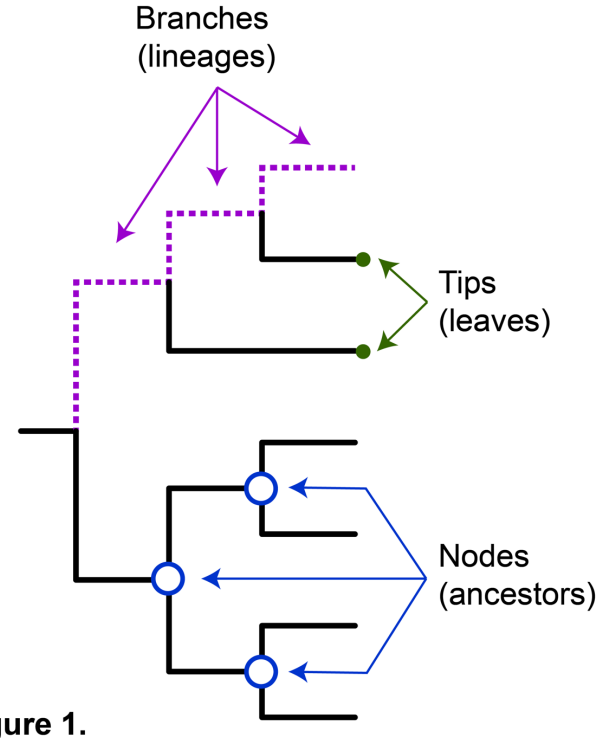
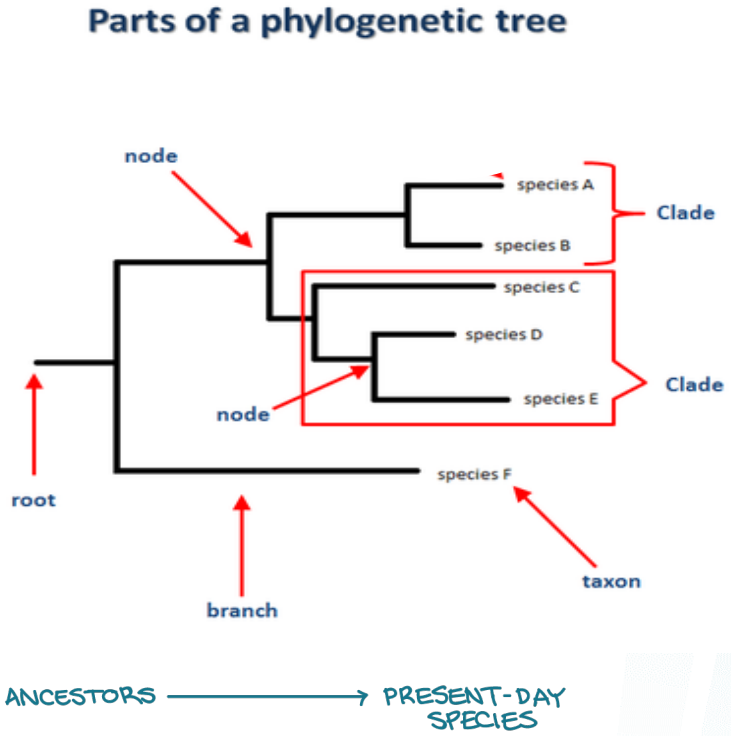
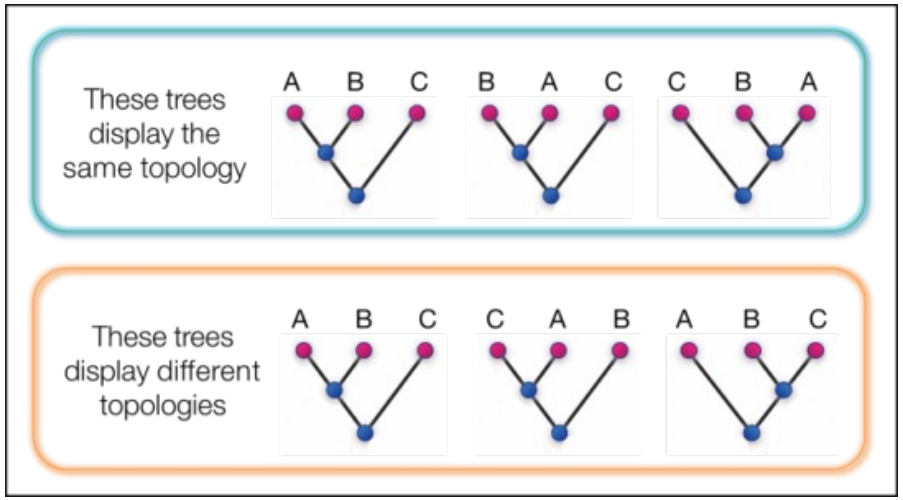


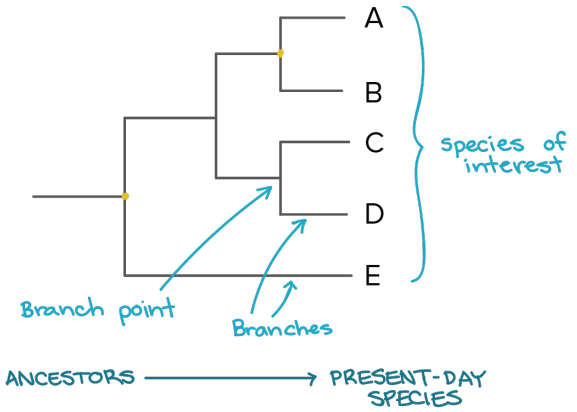

Figure 1.

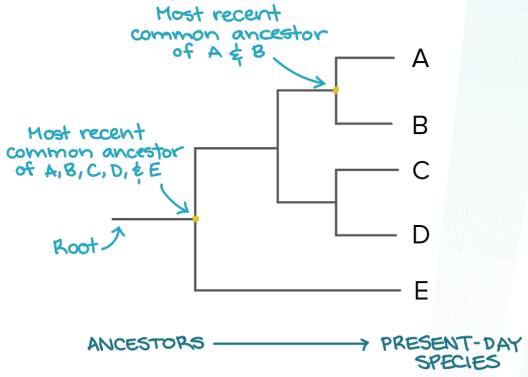


## Topology

- It is overall look of a tree: branching structure of the tree, representing hypothesis for set of species/individuals/characters/genes
- Similar sequences will be neighbours
- It is of particular biological significance



# Branches

- They show the path of transmission of genetic information from one generation to the next (through time)
- Branch length represent the genetic changes happened to reach next level in that branch (node)
- How to estimate genetic change or branch length
- External (recent diversions)
- Internal branches (old diversions)

| | |
|---|---|
| Human | ATGTTGACTC |
| Mouse | ATGCTGACTC |

1/10 =0.1 substitutions per site

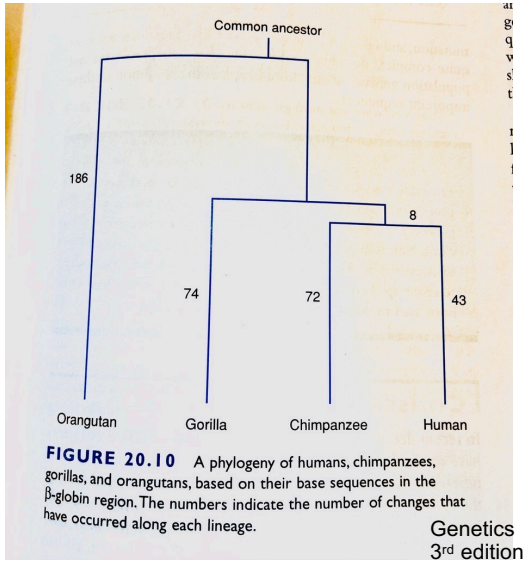
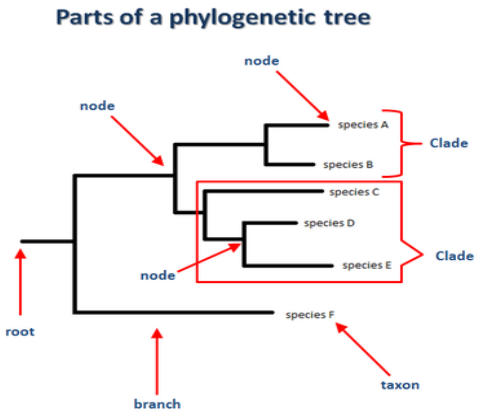
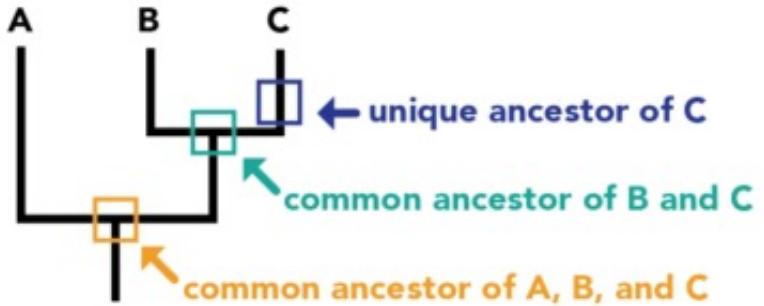
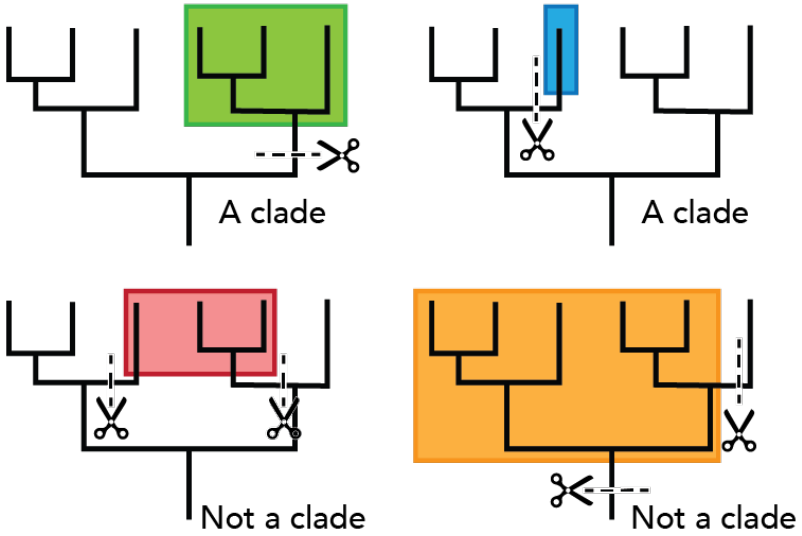


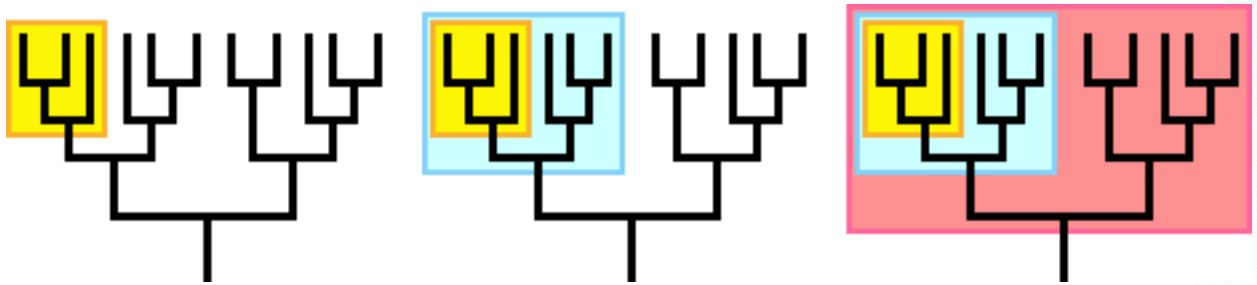

**FIGURE 20.10** A phylogeny of humans, chimpanzees, gorillas, and orangutans, based on their base sequences in the β-globin region. The numbers indicate the number of changes that have occurred along each lineage.

Genetics 3rd edition



How to read branches / trees

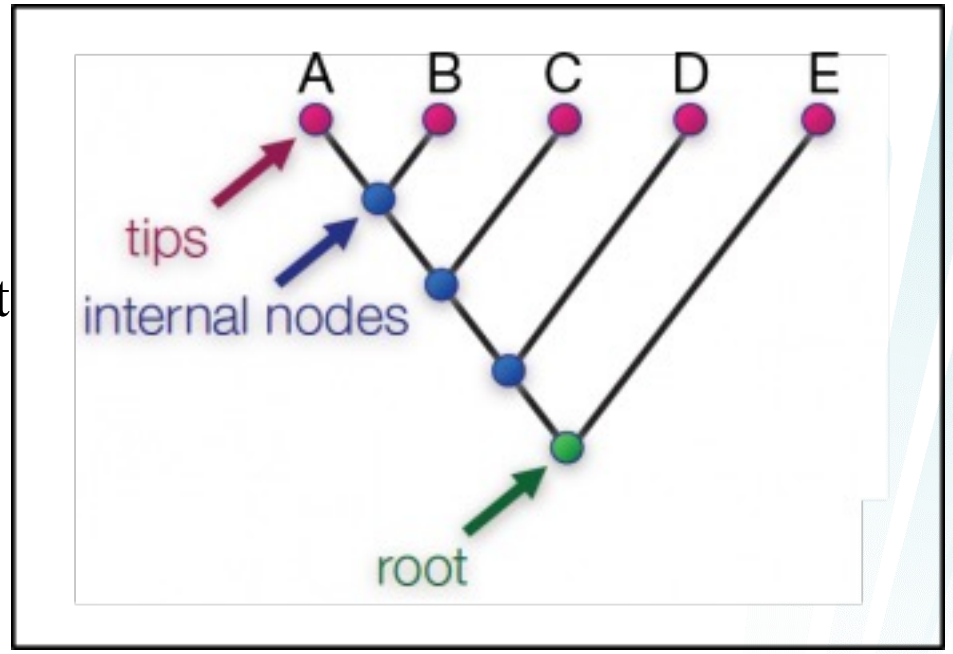
# Nodes

- Nodes : Points at the end of branches represent species /sequences/events at various points in a evolutionary history

- Represent a speciation /divergence event

- 2 nodes are connected by a branch

- The phylogenetic tree branch estimates # of changes that occurred from the time of separation




➢Tips /terminal nodes : extant samples (OTUs)

➢Internal nodes: inferred samples (ancestral) (ITUs)

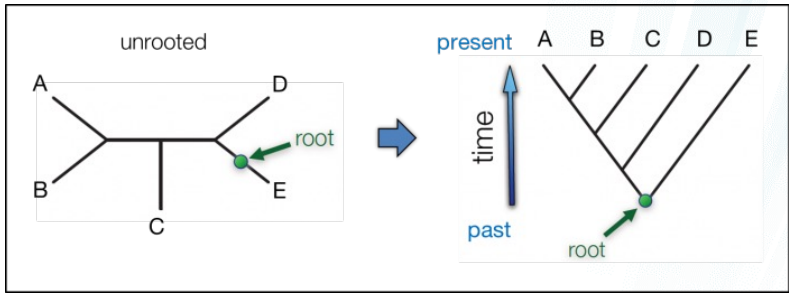
# Root

## Rooted tree

- The root of the phylogenetic tree is inferred to be the oldest point in the tree and corresponds to the theoretical last common ancestor of all taxonomic units included in the tree
- Gives directionality to evolution within the tree
- The paths from the root to the nodes correspond to evolutionary time

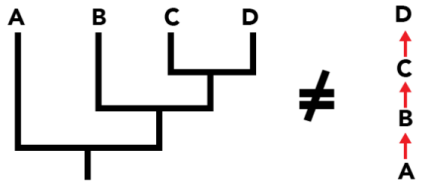
## Unrooted tree (also called phenogram some time)

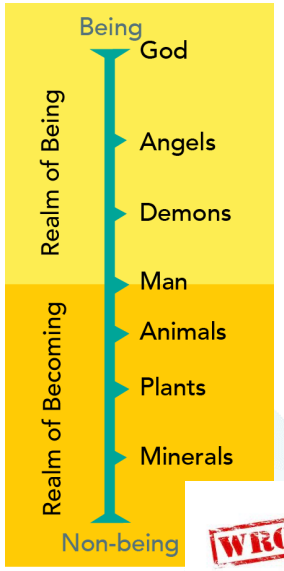
- “Root” less tree gives just relationship among different tips. No evolutionary path are explained
- Path between nodes don not represent evolutionary time in unrooted trees



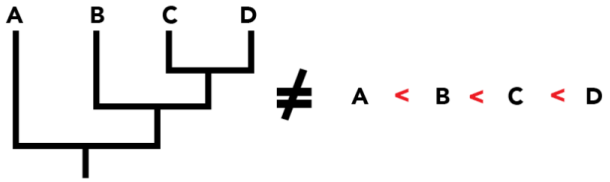
## Important things to remember about phylogenetic trees

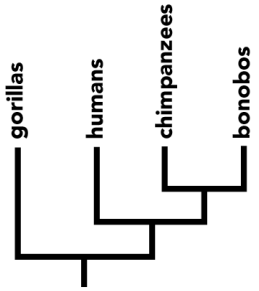
- Evolution produces a pattern of relationships among lineages that is tree like not ladder



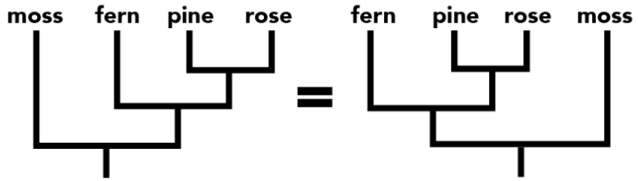




- There is no correlation with level of "advancement"



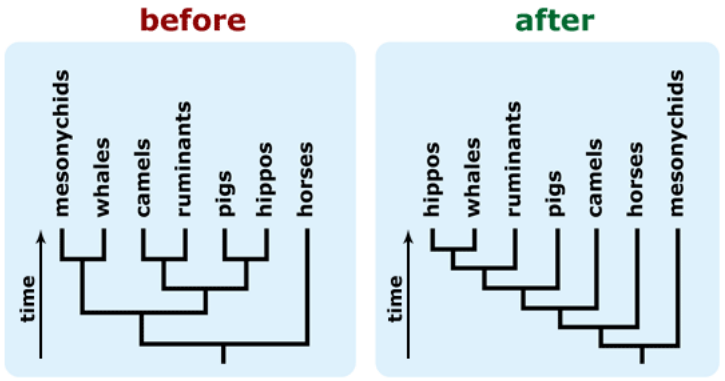




- You can change the orientation of the tree, but relations ships will not change

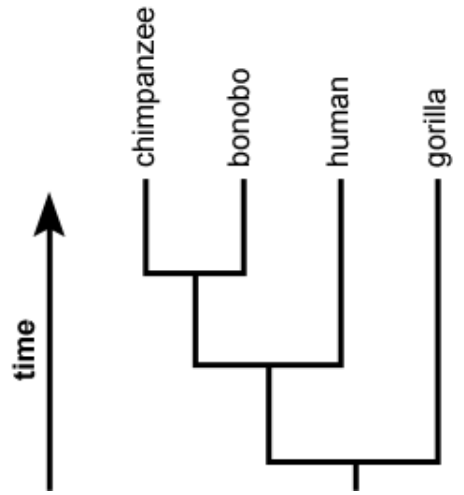


# How to build a phylogenetic tree

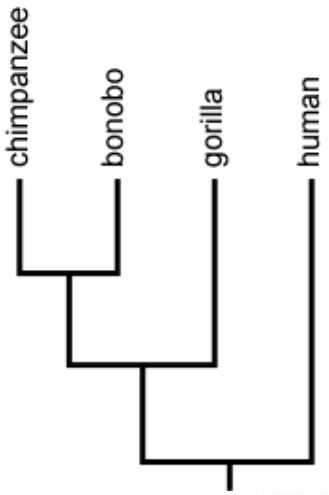
- Phylogenetic trees represent the evolutionary relationship among the lineages.
- Trees are hypotheses about the evolution of a group of organisms
- Test the hypothesis using available data.




$H_0$




$H_1$



## Construction of a phylogenetic tree: Steps

- Phylogenetics can not be observed directly rather, it is inferred (estimated). It is hypothesis, not a fact (we lack lot of intermediate species between present and last ancestors).
- **Steps to make a simple phylogenetic tree**

1. Choose taxa and outgroups (for a rooted tree) to be analysed
2. Choose characters to be used as evidence (**homologous**) or
3. Collect (the comparable DNA or AA sequences from samples) and align them
4. Right model of substitution (nucleotide or amino acid)
5. Convert (aligned sequences) or (other data) to a phylogenetic tree
6. Assess the accuracy of the reconstructed tree

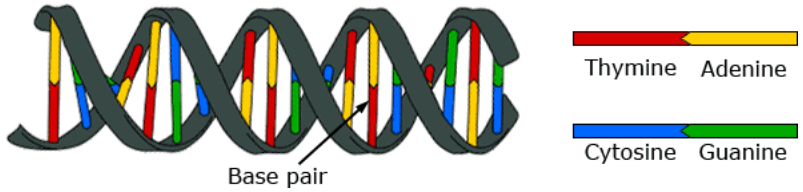
## Choose characters to be used as evidence

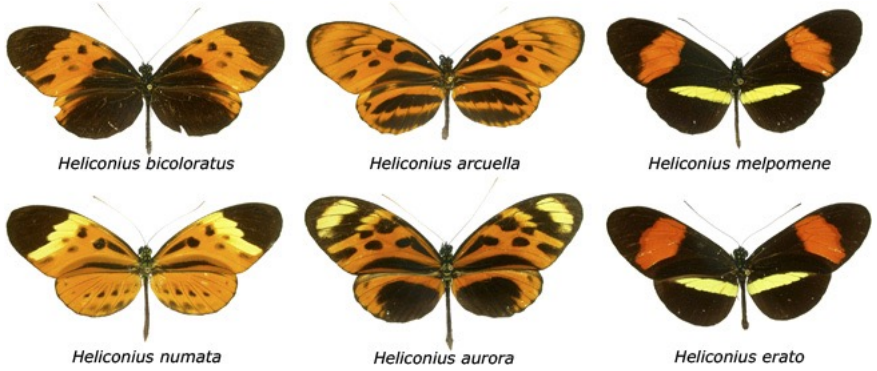
- Biologists need to collect data about the characters
- Characters need to be **heritable** and can be documented across study organisms

➢body morphology, nucleotide sequences, amino acid sequences

- Main goal in phylogeny is identifying less and less inclusive clades

  Shared ancestral traits and Shared derived traits




# Homologies and analogies

- Use homologous characters (similar characters in different organisms)

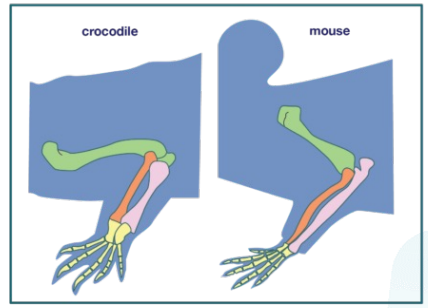
Homologous: Similar characters inherited from MRCA

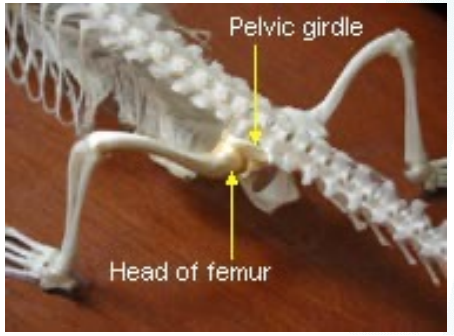
- Identifying homology:

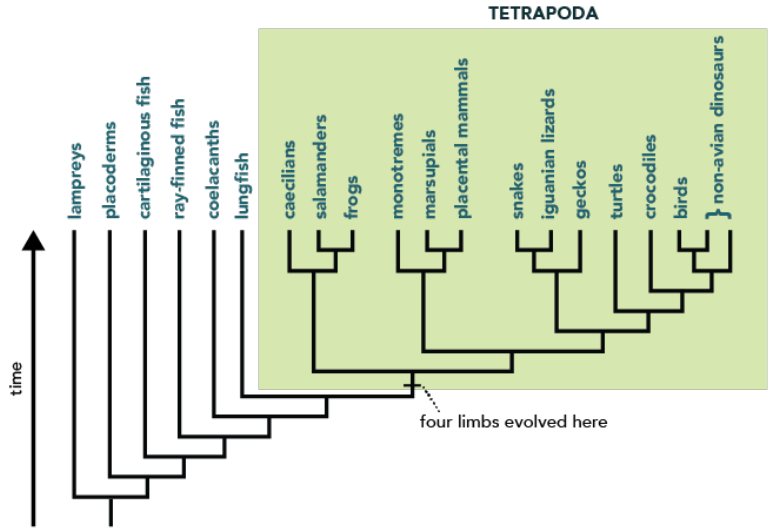
➢**Basic structure**

➢**Same relationship to other features**

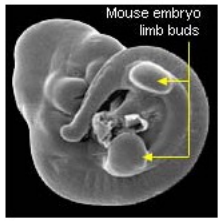
➢**Same development**



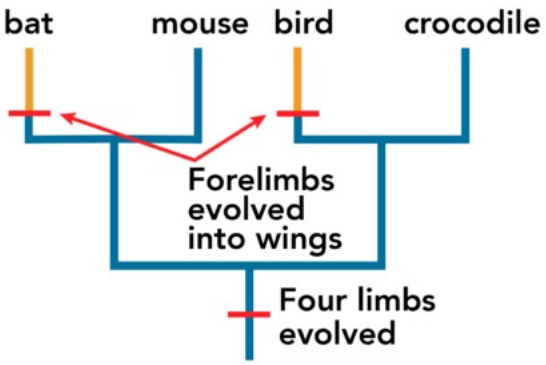



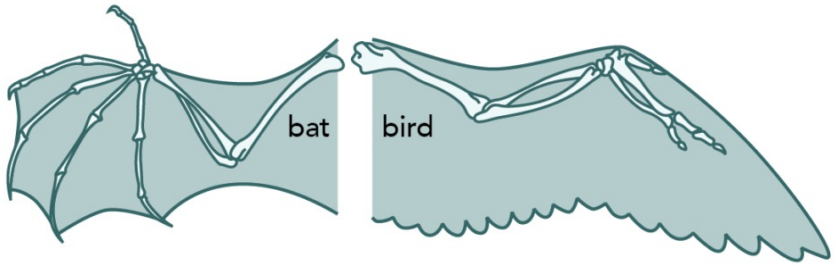

berkely.edu

# Analogous

- Not all similar-looking characters are homologies.
- Separate evolutionary origin and the result of convergent evolution.



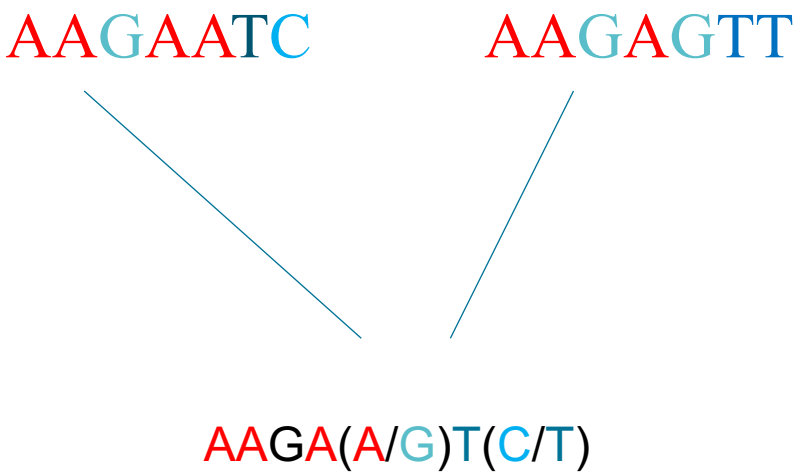



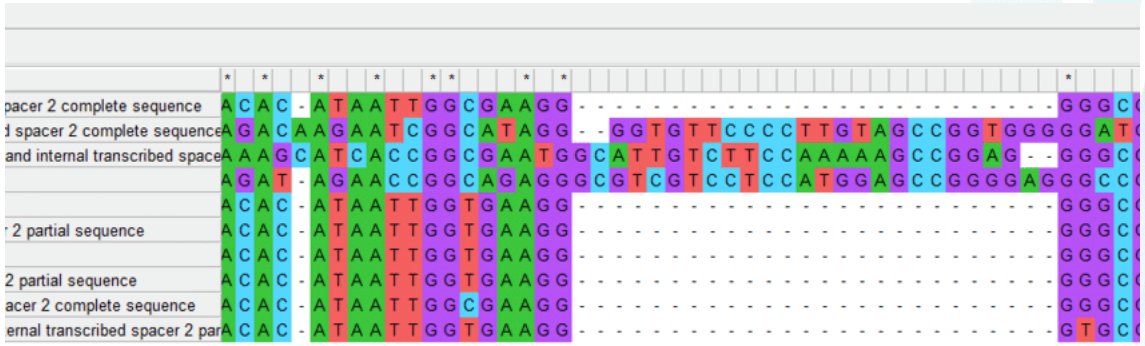
## 2. Collect the comparable sequences from samples and align them

- Orthologue(s) sequences are used for phylogenetic tree
- Are genes related by speciation events, means same genes in different species
- Different molecule type have different substitution rate
- Non-coding DNA regions have more substitution than coding regions
- Outgroup (reference to measure distances): which is closely related

## Relationship between phylogenetic analysis and sequence analysis

- When two sequences  found in 2 organisms are similar, we assume that they have one common ancestor
- The sequences alignment reveal which positions are conserved from the ancestor sequence

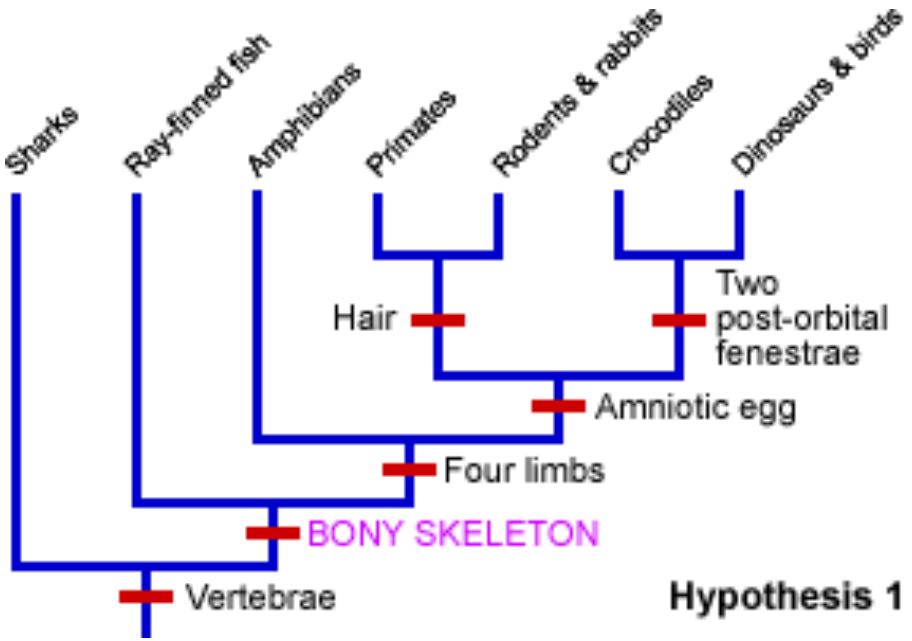


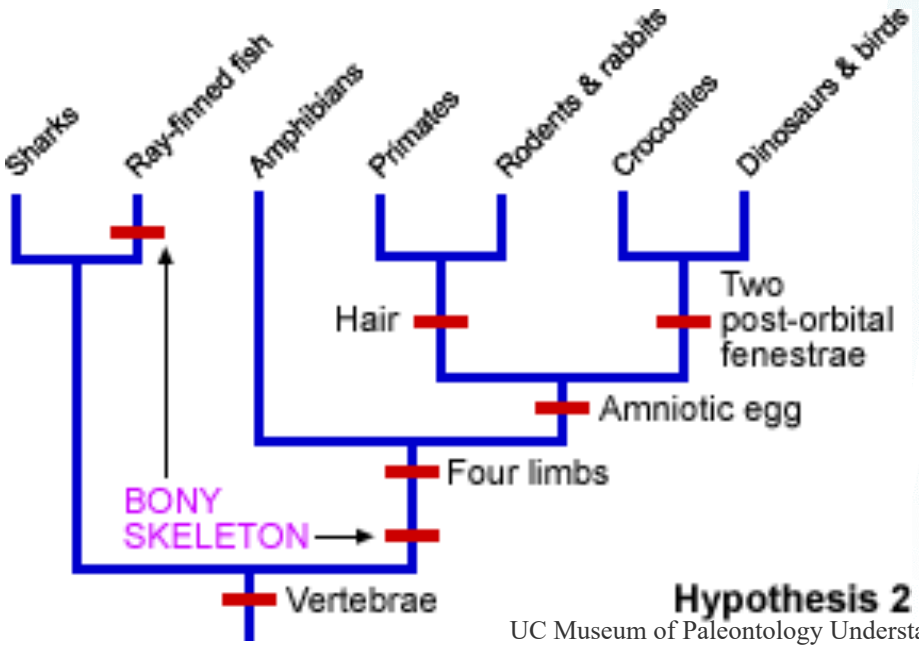



- The progressive multiple alignment of group of sequences: first aligns most similar ones and then moves to more divergent ones
- The alignment is influenced by the most similar pairs and arranged accordingly, but may not always represent the evolutionary history of the occurred changes
- Phylogenetic methods assume each position in a sequence can change independently from the other positions (substitution model)
- Gaps represent processes such as insertion, deletions or rearrangements

# 3. Convert (aligned sequences) or (other data) to a phylogenetic tree

- The basic method to all science  is the “parsimony” principle.
- Choose the simplest scientific explanation that fits the evidence
- All other things being equal, the best hypothesis is the one that requires the fewest evolutionary changes.
- Example:




UC Museum of Paleontology Understanding

# 4. Right nucleotide substitution model selection

- Model in science represent an abstraction of complex processes in order to make them mathematically tractable and hence useful to make reasonable predictions about the outcome of process or system under different scenarios through predictive formula.

  Weather forecasts are best know example for models

- Models of DNA sequence evolution : rates of change of fixed mutations among sequences
- In molecular phylogenetics, models are used to make predictions about the substitution process in molecular sequences (calculating their probabilities) along the branches of a tree or phylogeny.

# 5. Methods to build a phylogenetic tree

- Strong similarities between sequences

➢**Maximum parsimony**

- Weak similarities between sequences

➢**Distance matrix method** : Simple distance between every pair of sequence is calculated (as base differences)

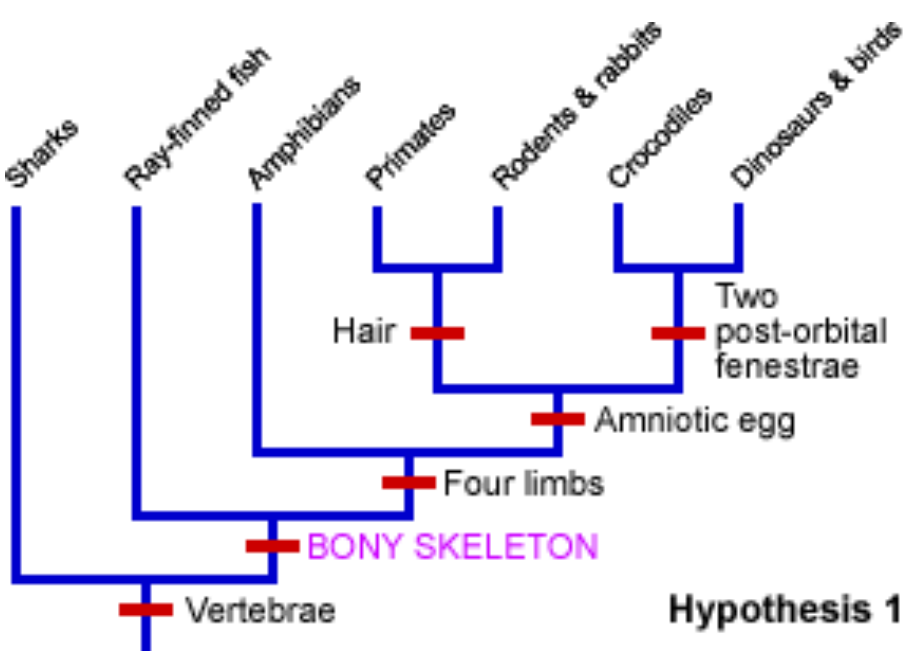
- Very weak similarities between sequences

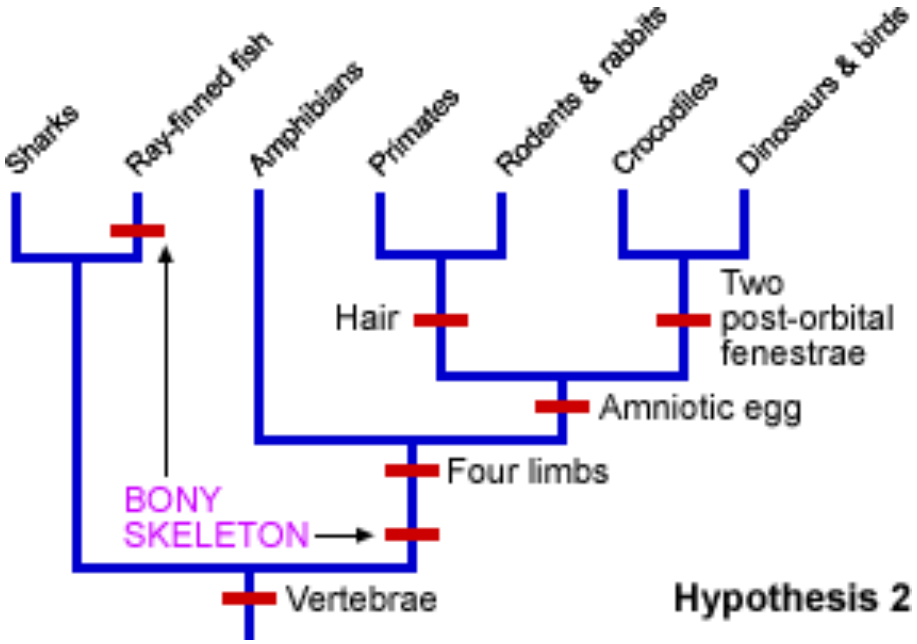
➢**Maximum Likelihood**: The method searches for the tree with the highest probability or likelihood

# Maximum Parsimony trees

- The principle of parsimony says the simplest explanation that describes the greatest number of observations is preferred over more complex explanations.
- **Character based method:** Relies on the raw characters themselves (ATGC or other). Don’t convert characters into statistics (see in distance trees).
- A tree with the fewest common ancestors is preferred (minimum net amount of evolution)
- Many trees are calculated and consensus tree is built combining all trees into a single approximated tree topology

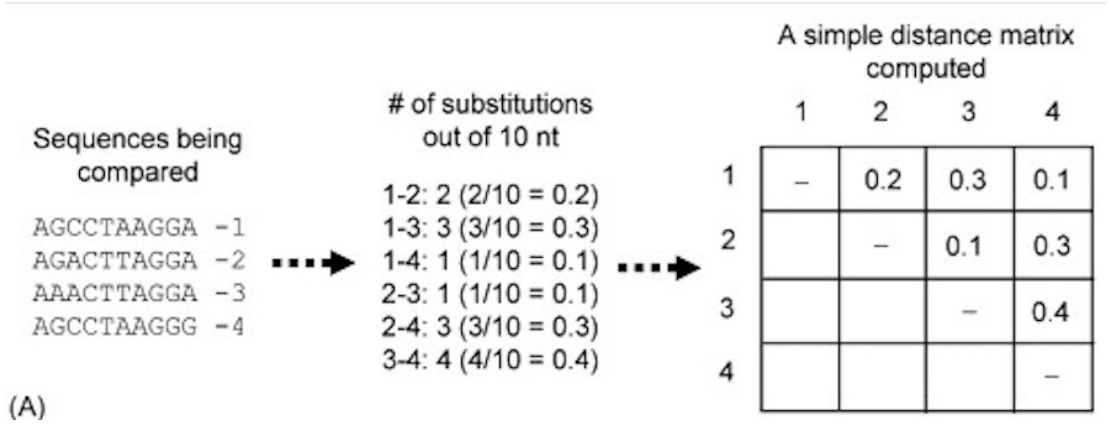
# Maximum Parsimony trees



# Distance trees

- Group representative taxa based on the number of nucleotide or amino acid substitution between sequences
- Simple distance between every pair of sequence is calculated (as base differences, p-distance)
- Base differences represent evolutionary distance (branch length)



# Distance method

- Distance-based phylogenetic reconstruction requires the calculation of the number of changes that have occurred in a given sequence
- Impossible to determine exact number of substitutions as we have no idea whether it is single or multiple times substitution
- **Remember**: Two sequences that are identical at a given position may not necessarily indicate the absence of a substitution, since it is possible that the sequence conservation at that site is due to an unobserved change followed by a reversion back to its original state

# Distance based trees

- We use substitution models to correct these nucleotide substitutions

❖**UPGMA** (unweighted pair group method with arithmetic mean) tree : a successive clustering approach

Assumption of a constant evolutionary rate (molecular clock) for all sequences

This approach is simple and consists of three simple steps: 1) Find the two organisms with least differences. 2) Group them together as one cluster and recalculate differences. 3) Repeat steps 1–2 until the tree is complete.

❖**Neighbour Joining tree** : Clusters closest sequences (neighbors) first

unequal rates of evolution

# Distance method tree

- **Neighbour joining tree**

➢ more realistic in terms of evolutionary model (unequal rates of evolution)

➢Builds unrooted tree

➢Reliable tree method

➢Fast and simple

- **UPGMA tree**

➢Builds rooted tree

➢branches tips come out equal

# Likelihood and Bayesian trees

- Like MPT, simultaneously compare all sequences (characters) (not pairwise)

➢Maximum likelihood method

➢Bayesian methods

# Maximum likelihood tree

- It evaluates a hypothesis about evolutionary history in terms of the probability that the proposed model and the hypothesized history would give rise to the observed data set.

- The supposition is that a history with a higher probability of reaching the observed state is preferred to a history with a lower probability.

- Considers more than one tree to reach the probable solution

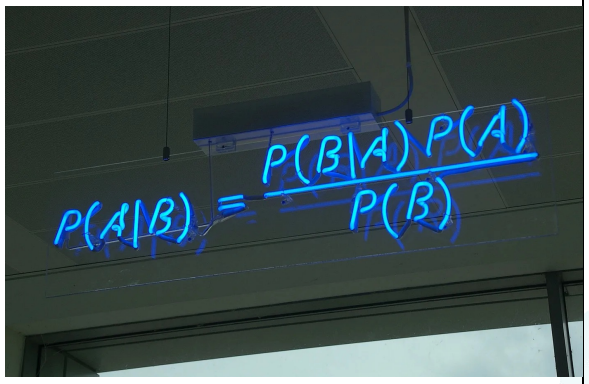
- Applies a model of evolution to make decision, more realistic

- But computationally very intensive, as it has to consider “all possible trees”

# The Bayesian Tree

- Method relies on Bayes' theorem : Probably of an event based on the prior knowledge of conditions that might be related to the event (conditional probabilities, posterior probabilities)

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$




how often A happens *given that B happens*, written **P(A|B): posterior**

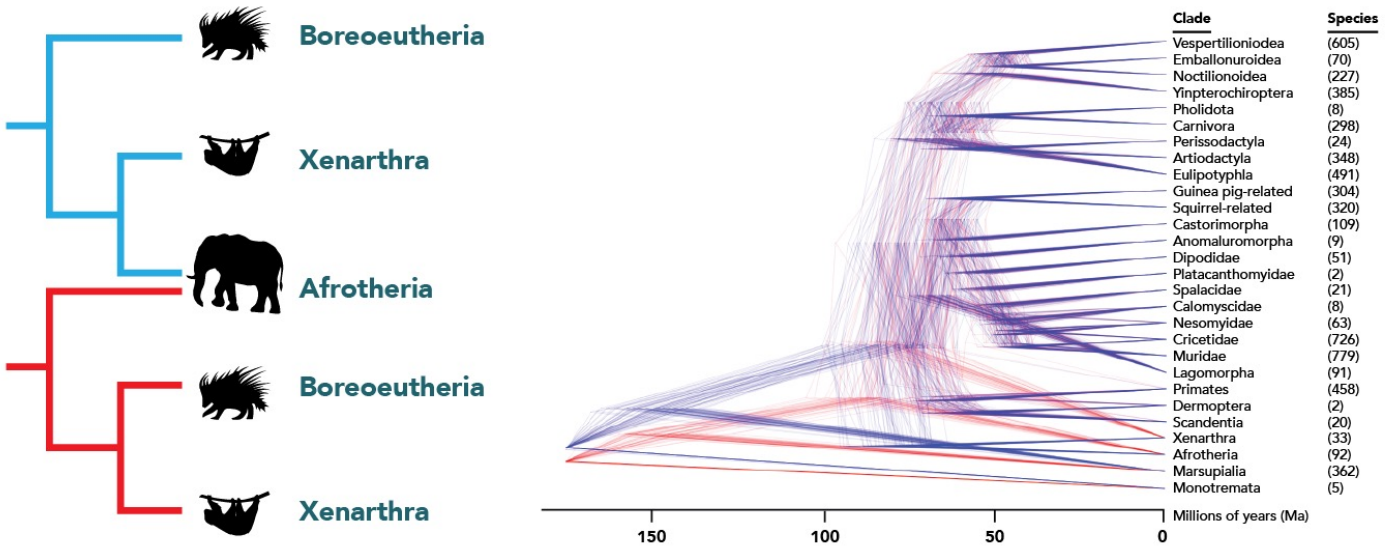
how often B happens *given that A happens*, written **P(B|A) : likelihood**

and how likely A is on its own, written **P(A) : prior**

and how likely B is on its own, written **P(B): marginal likelihood**

- MCMC approach: estimates PP for all possible trees
- The MCMC algorithm begins at multiple random locations in tree space, with each point (or chain) beginning to sample trees randomly. As each chain explores tree space, any tree sampled that has a higher PP than the previous tree serves as the starting point for the next sampling iteration, allowing the algorithm to sample those topologies with increasingly larger PP

- Bayesian tree produces lot of trees with pp and then initially algorithm eliminates all possible “bad” trees and then make a consensus tree from “good” trees
- Benefit with collection of trees (against single tree) is that it opens up alternative evolutionary history.



# Assess the accuracy (confidence) of the reconstructed tree

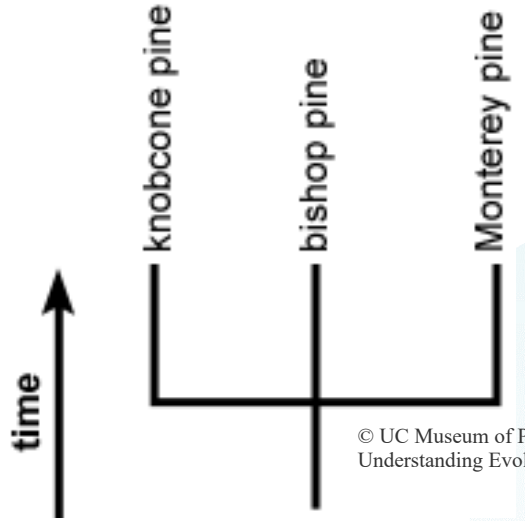
- Uncertainty in phylogenetics: Arises from lack of empirical evidence

- Is a method for testing how good a dataset fit the model
- Two way we can test relations on a tree

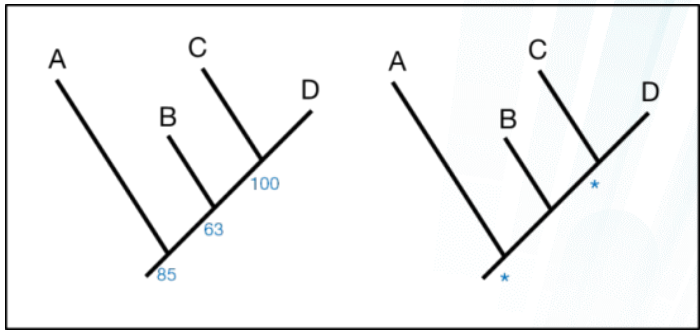
➢Looking into branching: polytomies (some time real)

➢Statistical method : branch support statistics

- Bootstrapping

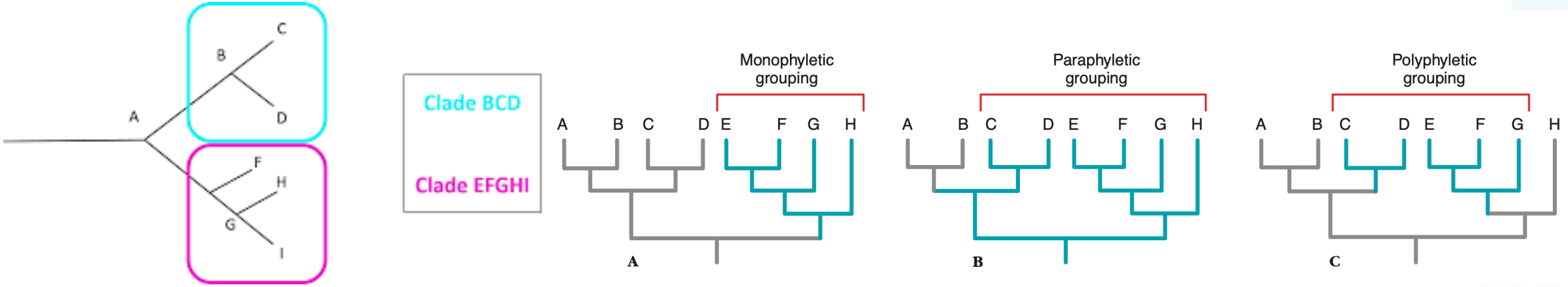


- **Bootstrapping**
- Build same tree leaving out some portion of evidence and check if same clades appear even after leaving out some data. Repeat the process leaving more data and asses the clade: bootstrap value (>95% is best)
- Part of the results will show the # of times a particular branch point occurred out of all the trees that were built.
- The higher the # - the more valid the branching point.



# So, what is phylogenetic inference?

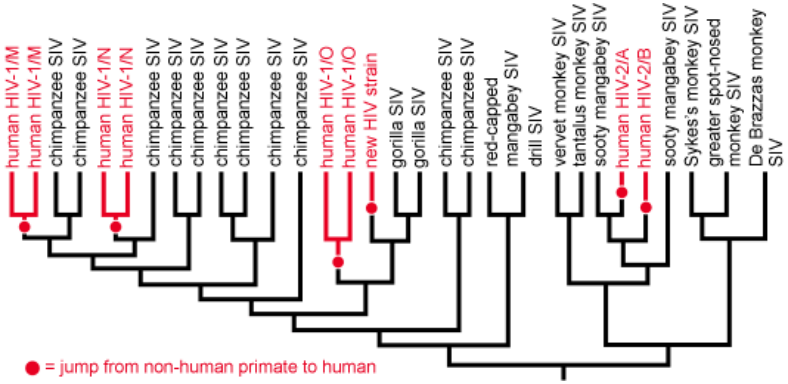
**Phylogenetic inference is the practice of reconstructing the evolutionary history of related species by grouping them in successively more inclusive sets based on shared ancestry.**



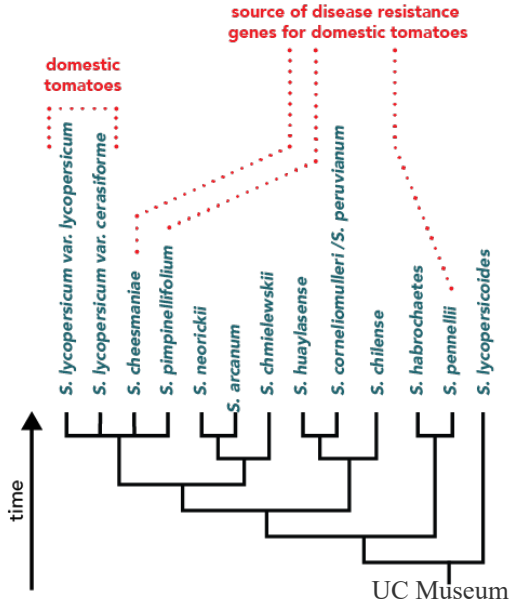
# Why phylogenetics?

➢ **Applications**

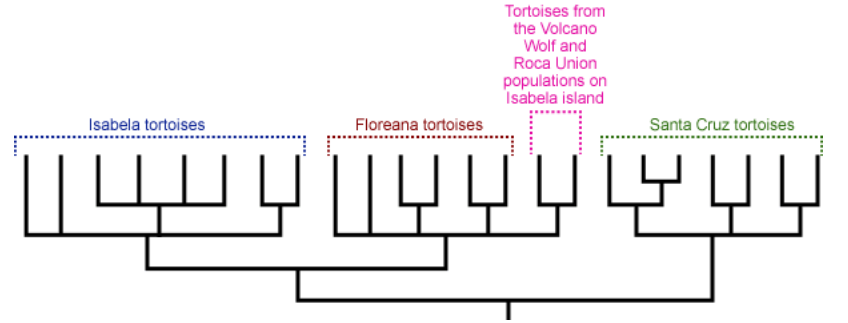
- **Identifying pathogens**

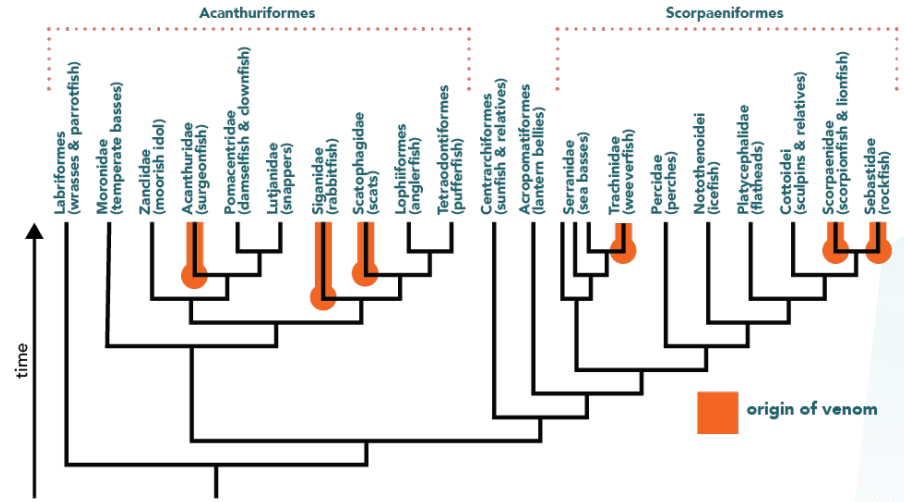


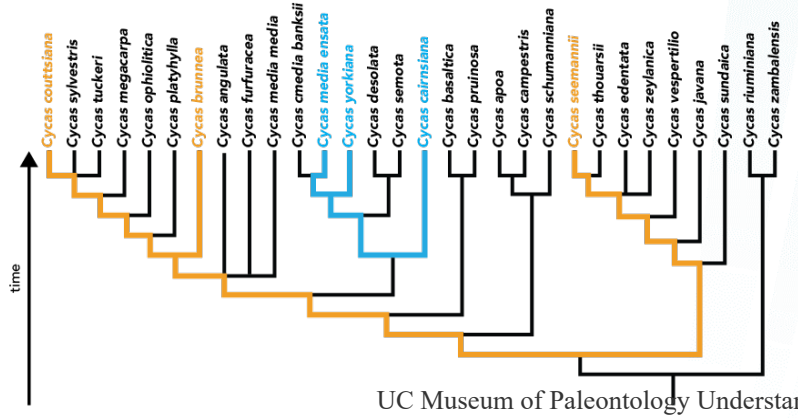

- **Selective breeding programs**



- Conservation genetics
- Drug discovery
- Defending diversity
- Tree as a forensic tool
- Classification



# We will see more of this in practical class

Some of the information and figures are taken from UC Museum of Paleontology Understanding Evolution
Genetics 3rd edition
https://blogg.vm.ntnu.no/evolusjon/en/2014/12/15/the-tree-of-life/