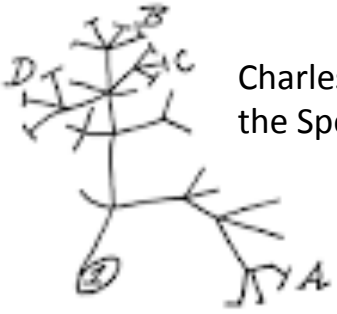
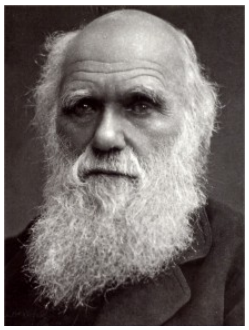
"…the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever-branching and beautiful ramifications."



Charles Darwin, Origin of the Species (1859)



Theory of Evolution

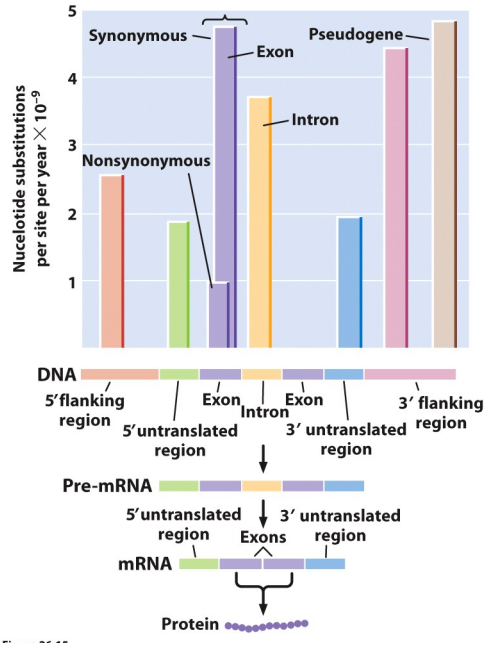
- **Evolution** is the change in the *inherited characteristics* of biological populations over *successive generations*.

Genetic variation + selection

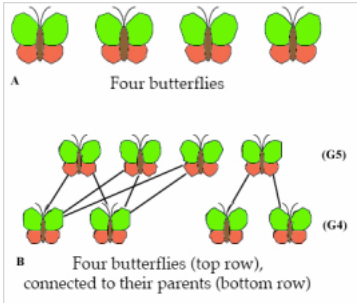
- **Principle of common ancestry** - Any two species share a (possibly distant) common ancestor

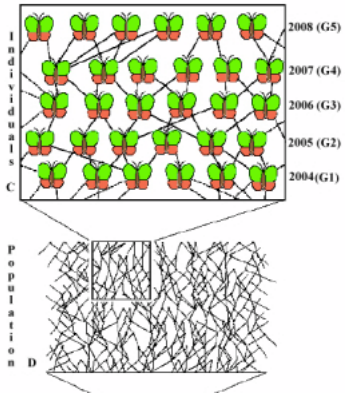
But it is important to realize that:

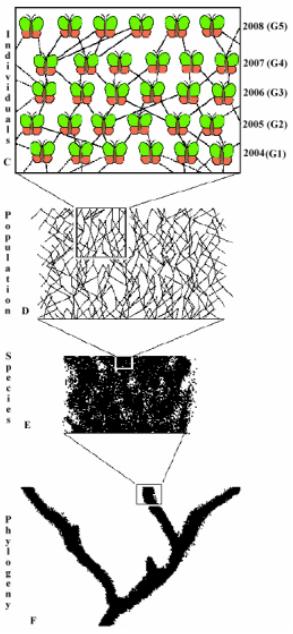
Different parts of genome (or gene) evolve at different rates because each region has a different effect on the organism's fitness (the gene's function).



What an Evolutionary Tree Represents







Zoooooming out…

Branching pattern of four species

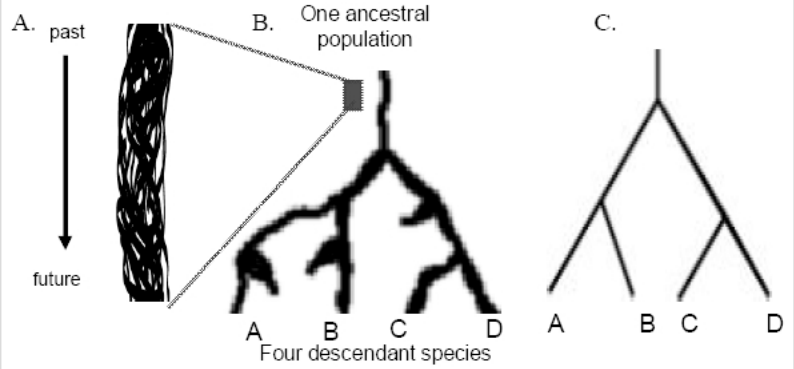
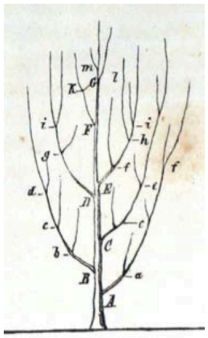


Figure 2 : Branching pattern of four species.

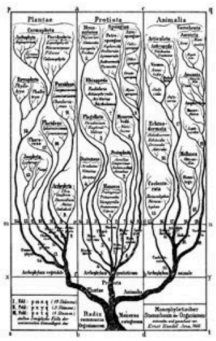
What is phylogeny?

Phylogeny is the pattern of evolutionary relationships among species, describing their descent from a common ancestor.

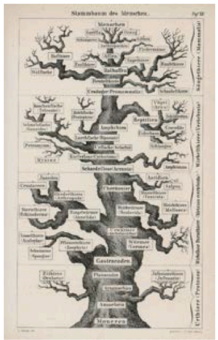
Augustin Augier, 1801



Heinrich Bronn, 1858

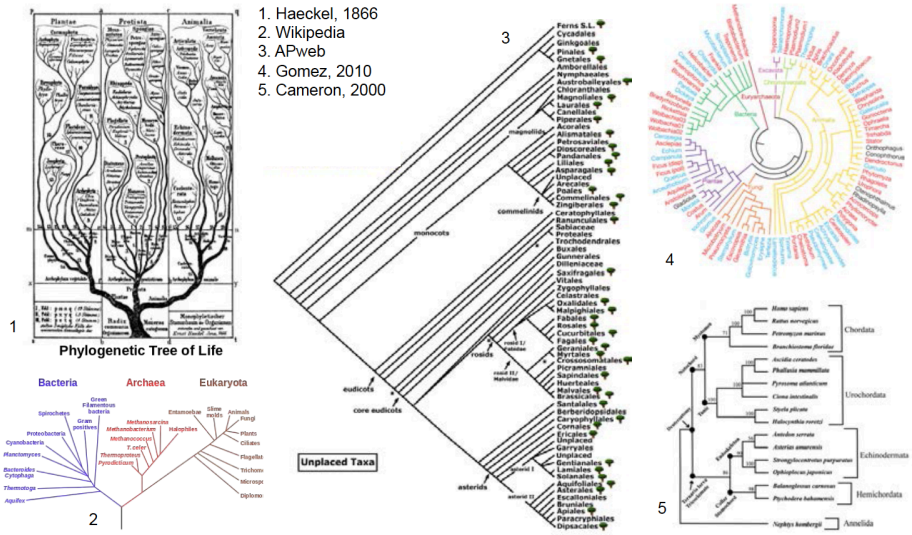


Haeckel, 1866

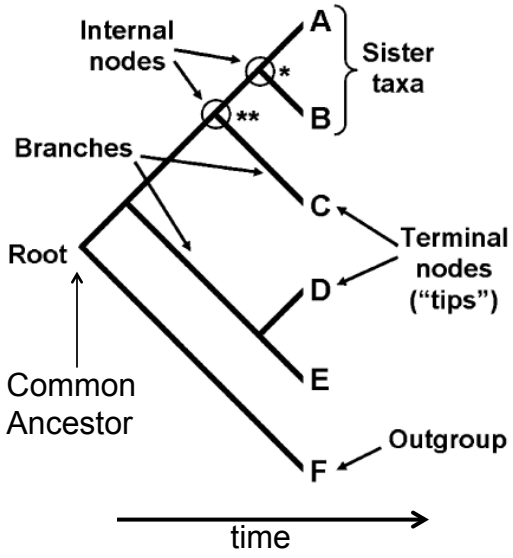


Haeckel, 1874

A phylogenetic tree is a diagram that describes the phylogeny



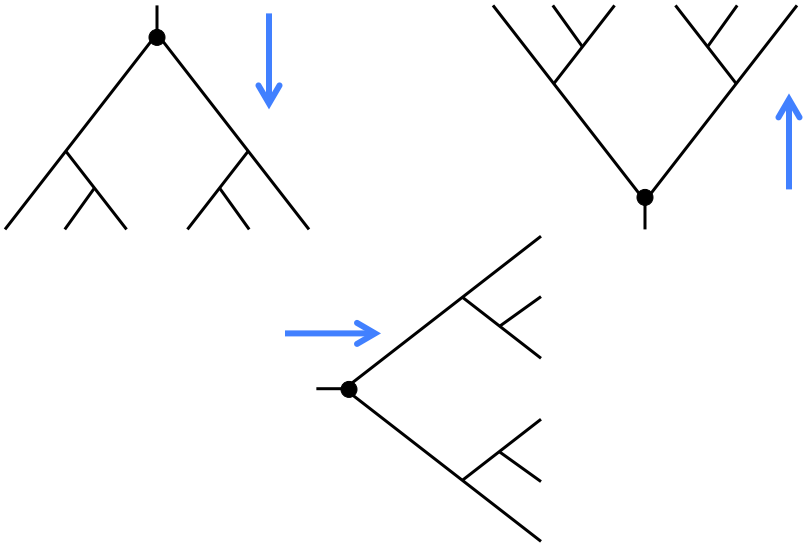
Anatomy of a phylogenic tree



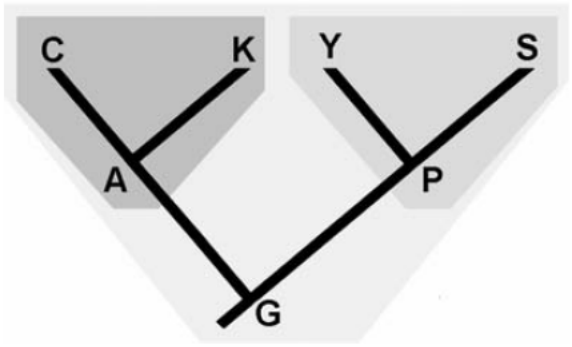
All trees represent the historical pattern of ancestry, divergence and descent.

Topology refers to the general pattern of branching within the tree.

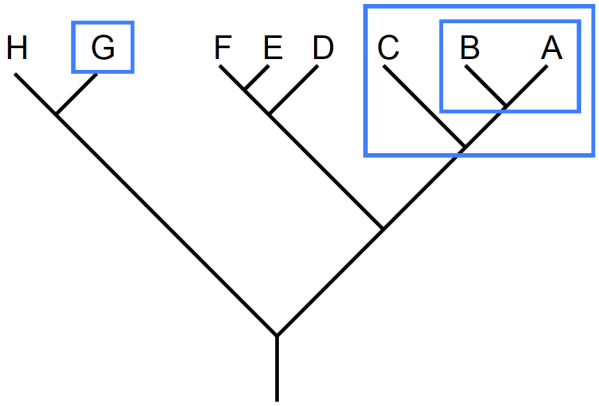
Rooted trees can be re-drawn in several orientations



Phylogenies indicate both relatedness and historical descent

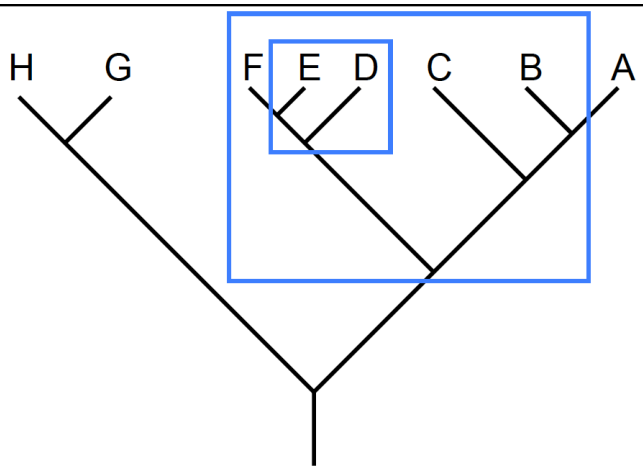


A **Clade** is a subtree (in grey) and shows all descendants from a common ancestor.



Monophyletic group (or **clade**) = a single lineage; a group composed of a common ancestor and all of its descendants.

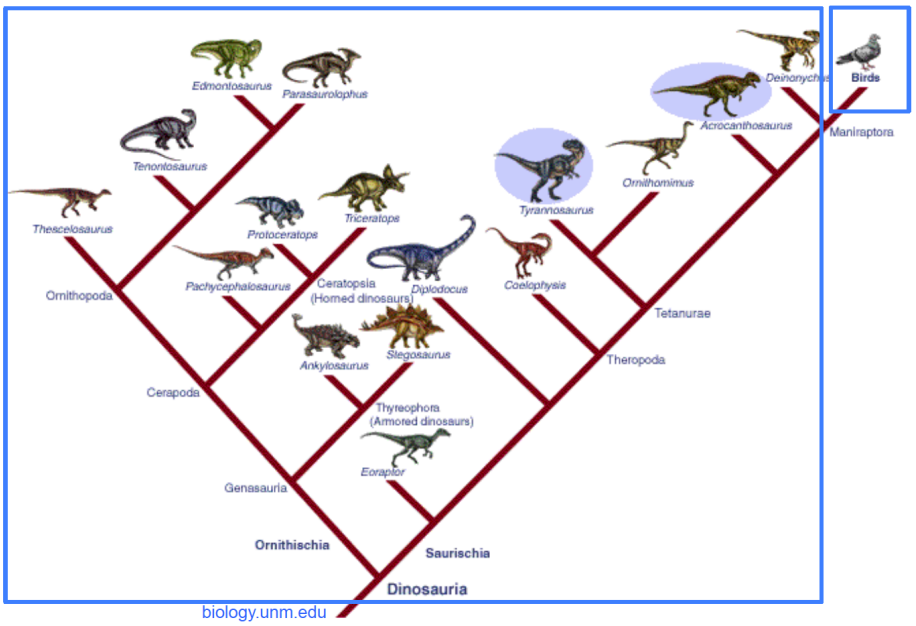
mono = one, phylum = tribe



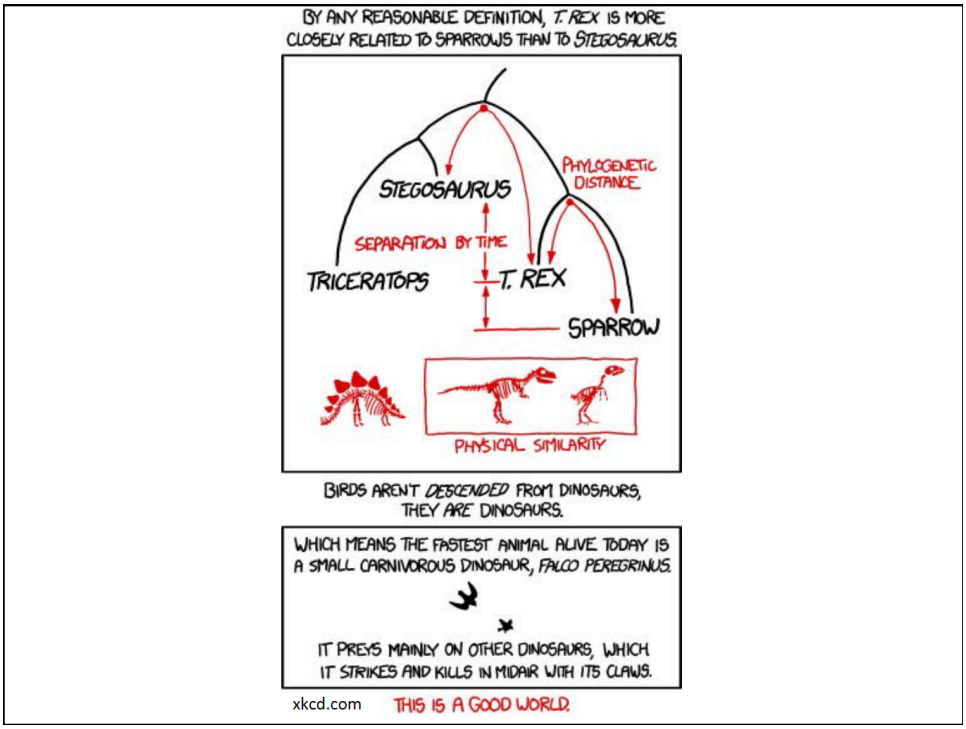
Paraphyletic group = a group containing a common ancestor and some, but not all, of its descendants.

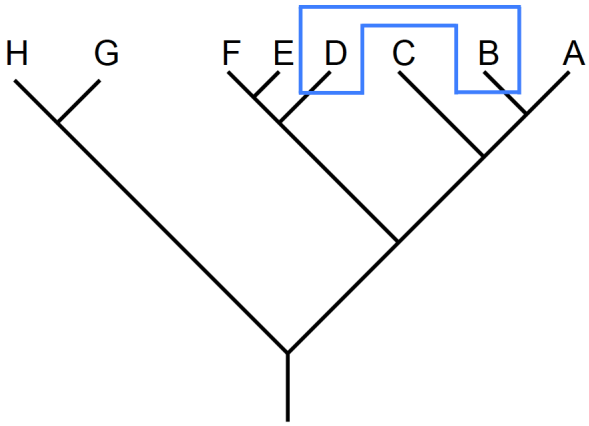
para = near, “not quite”, phylum = tribe

Reptiles, dinosaurs: examples of **paraphyletic** groups



biology.unm.edu

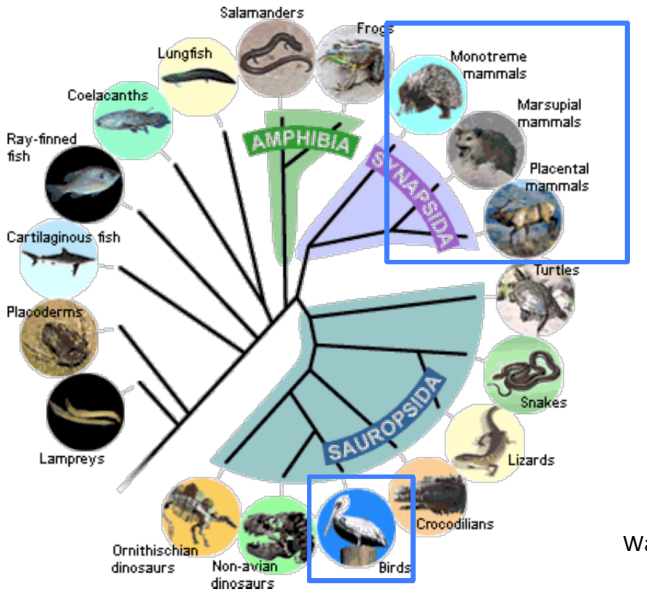




Polyphyletic group = multiple lineages; a group that does not contain the common ancestor of its members.

poly = many, phylum = tribe

Homeothermia: an example of a **polyphyletic** group

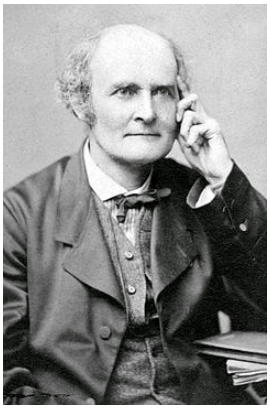


Warm blooded

evolution.berkeley.edu

The Newick format

In computer programs, trees are represented in a linear form by a string of nested parentheses, enclosing taxon names (and possibly also branch lengths and bootstrap values), and separated by commas. This type of representation is called the **Newick format**. The originator of this format in mathematics was Arthur Cayley.

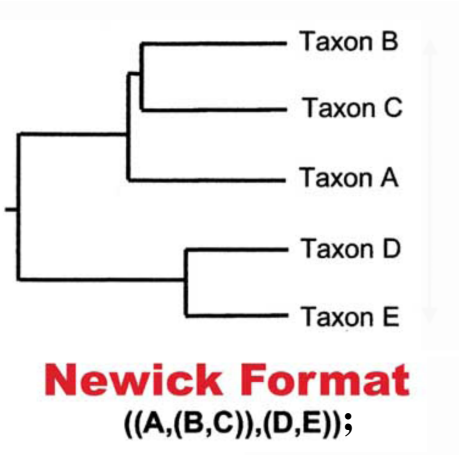


The Newick format

The Newick format for phylogenetic trees was adopted on June 26, 1986 at an informal meeting at *Newick's Lobster House* in Dover, New Hampshire. The Newick format currently serves as the *de facto* standard for representing phylogenetic tree and is employed by almost all phylogenetic software tools. Unfortunately, it has never been described in a formal publication; the first time it is mentioned in a publication is in 1992.

The Newick format

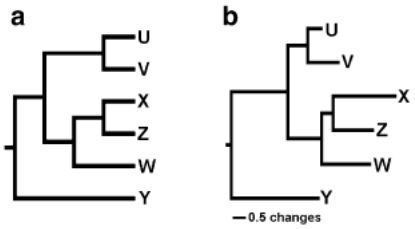
In the Newick format, the pattern of the parentheses indicates the topology of the tree by having each pair of parentheses enclose all members of a monophyletic group. A phylogenetic tree in the Newick format always ends in a semicolon (;).



The Newick format

One can use the Newick format to write down rooted trees, unrooted trees, multifurcations, branch lengths, and bootstrap values.

Cladograms versus phylograms

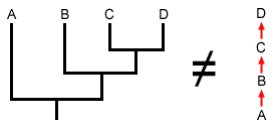


A. Cladograms only depict branching pattern, regardless of how it is presented.

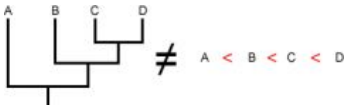
MOST TREES! Essentially present a particular "clustering".

B. Phylograms present branch lengths as proportional to some measure of divergence, MUST provide a scale!

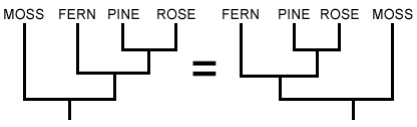
So when reading a phylogeny, it is important to keep three things in mind:



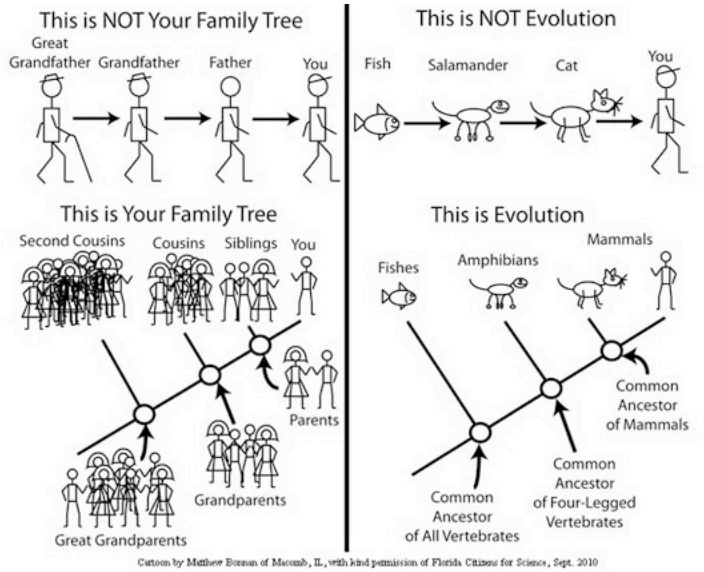
1. Evolution produces a pattern of relationships A B C D among lineages that is tree-like, not ladder-like.



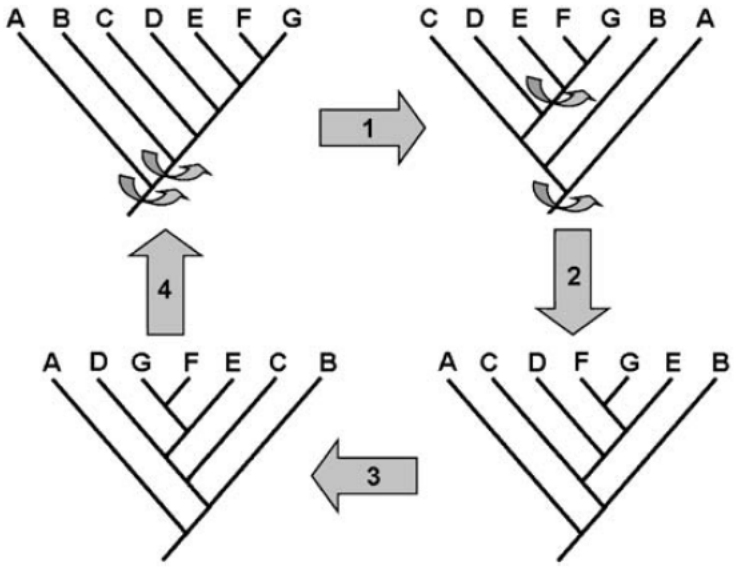
2. Just because we tend to read phylogenies from left to right, there is no correlation with level of "advancement."



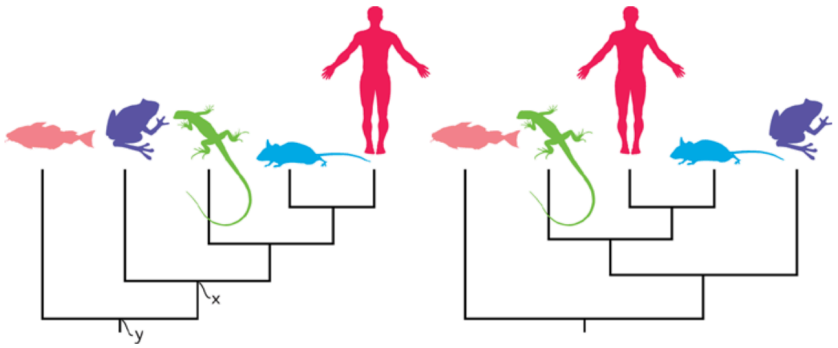
3. For any speciation event on a phylogeny, the choice of which lineage goes to the right and which goes to the left is arbitrary. The phylogenies at left are equivalent



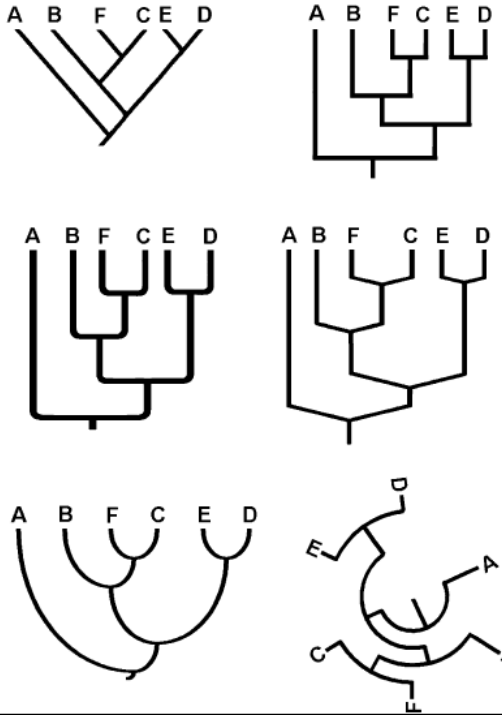
All internal nodes can be rotated without changing the topology



So are these trees the same?



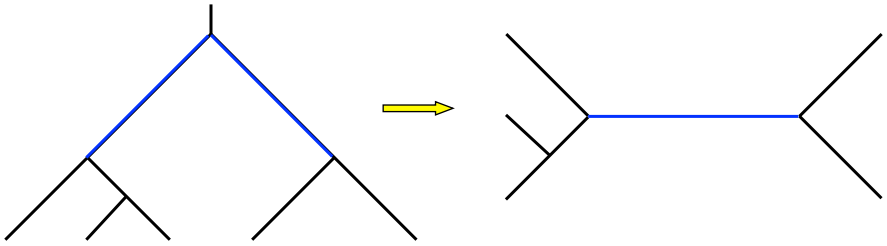
The same tree can be drawn in many different styles.



Note the branching pattern is the same in all six trees!

Some trees lack a root!

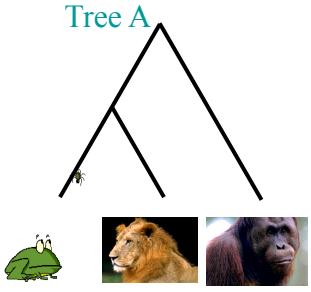
Unrooted tree represents the same phylogeny without the root node



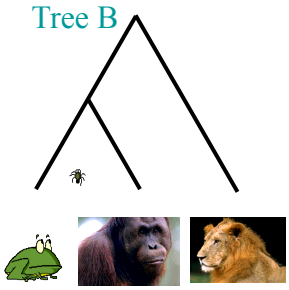
Depending on the model, data from current day species does not distinguish between different placements of the root.

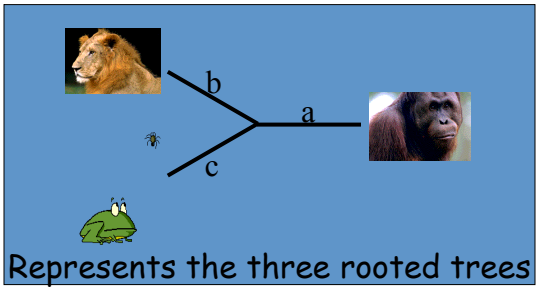
Rooted versus unrooted trees

Tree A

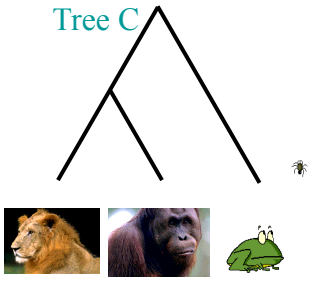


Tree B





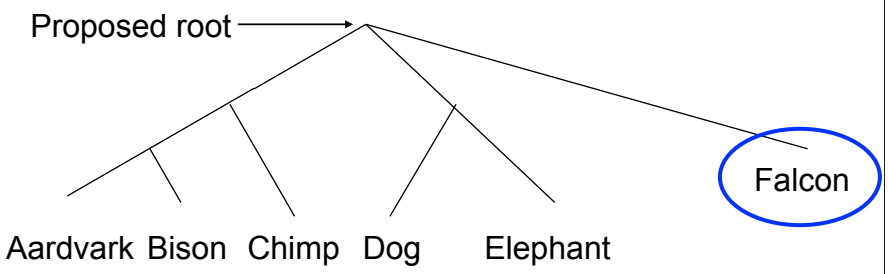
Tree C



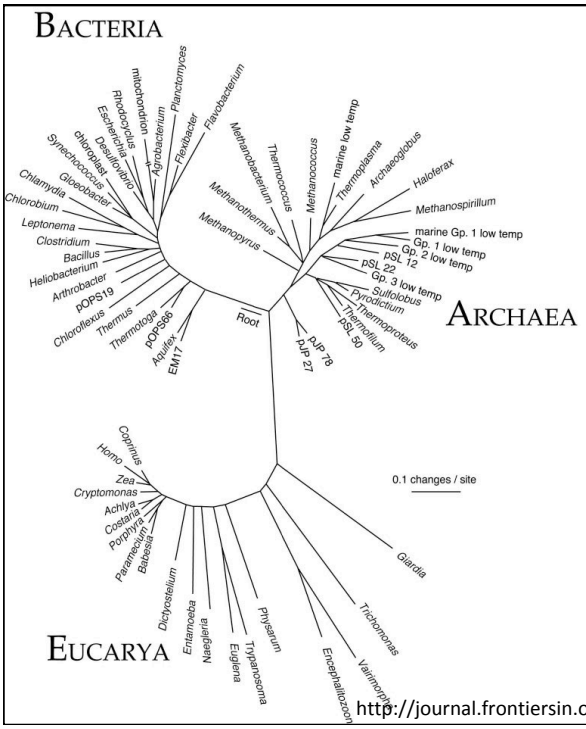
Positioning Roots in Unrooted Trees

We can estimate the position of the root by introducing an **outgroup**:

- a set of species that are definitely distant from all the species of interest



Ingroup – lineage under consideration (everybody not circled)
Outgroup – lineage that is not part of the ingroup (circled in blue)



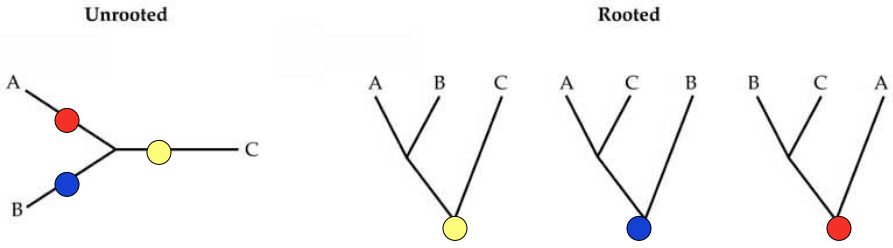
So then how do we root the tree of life?

This rooting is supported by phylogenetic analyses of protein paralogs (elongation factors and V/F types ATPase subunits) that originated by duplication before LUCA.

http://journal.frontiersin.org/article/10.3389/fmicb.2015.00717/full

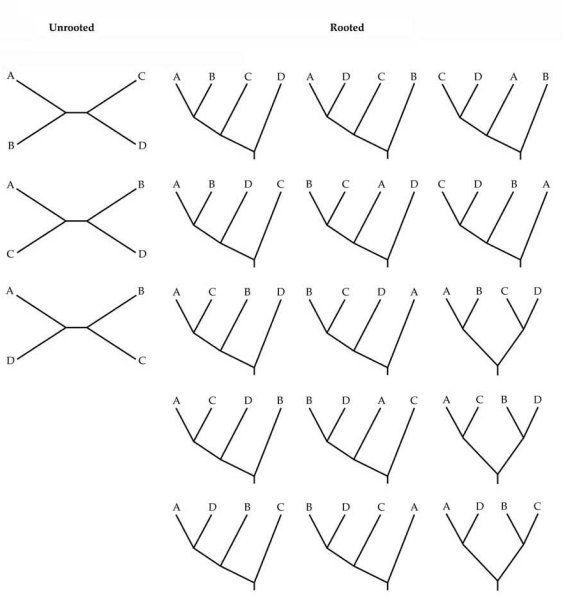
3 OTUs

Operational Taxonomic Units -- **operational** definition used to classify groups of closely related individuals



1 unrooted tree = 3 rooted trees

4 OTUs



3 unrooted trees = 15 rooted trees

The number of possible bifurcating rooted trees (N_R) for $n \geq 2$ OTUs

$$N_{\mathrm{R}} = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

The number of possible bifurcating unrooted trees (N_U) for $n \geq 3$ OTUs

$$N_{\mathrm{U}} = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Number of OTUs	Number of possible rooted tree
2	1
3	3
4	15
5	105
6	954
7	10,395
8	135,135
9	2,027,025
10	34,459,425
15	213,458,046,676,875
20	**8,200,794,532,637,891,559,375**

Evolution is an historical process.

Only one historical narrative is true.

From 8,200,794,532,637,891,559,375 possibilities, 1 possibility is true and 8,200,794,532,637,891,559,374 are false.

Truth is one, falsehoods are many.

How do we know which of the 8,200,794,532,637,891,559,375 trees is true?

We don't, we infer by using decision criteria.

True and inferred trees

The sequence of speciation events that has led to the formation of a group of OTUs is historically unique. A tree representing the true evolutionary history is called the **true tree**.

A tree that is obtained by using a certain set of data and a certain method of tree reconstruction is called an **inferred tree**.

An **inferred tree** may or may NOT be the **true tree**.

Constructing phylogeny

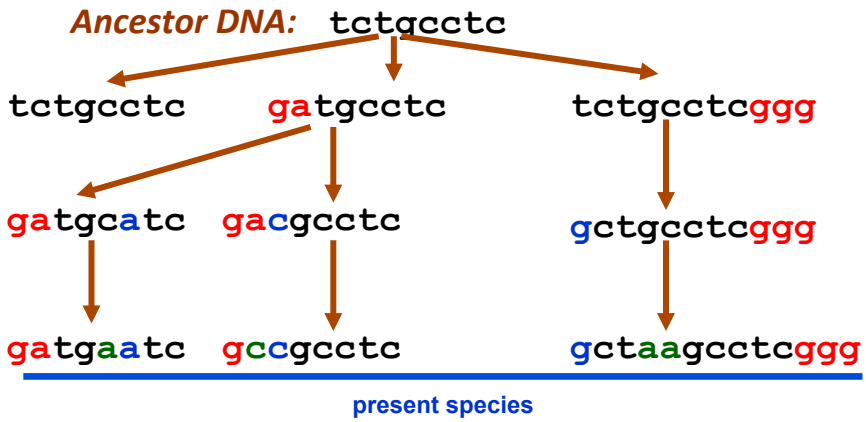
- Phylogenetic inference -- the process by which the branching pattern of evolutionary relationship (phylogeny) is estimated.

- A phylogenetic tree is a hypothesis; it is subject to reevaluation upon the discovery of new evidence.

Molecular data **vs.** Morphology / Physiology

• Strictly heritable entities	• Can be influenced by environmental factors
• Data is unambiguous	• Ambiguous modifiers: "reduced", "slightly elongated", "somewhat flattened"
• Regular & predictable evolution	• Unpredictable evolution
• Quantitative analyses	• Qualitative argumentation
• Ease of homology assesment	• Homology difficult to assess
• Relationship of distantly related organisms can be inferred	• Only close relationships can be confidently inferred
• Abundant and easily generated with PCR and sequencing	• Problems when working with micro-organisms and where visible morphology is lacking

The molecular basis of evolution



Distance and Character

A tree can be based on

1. **quantitative measures** like the **distance** or **similarity** between species, or

2. based on **qualitative aspects** like **common characters**.

Types of data used in phylogenetic inference:

Characters

```
Species A    ATGGCTATTCTTATAGTACG
Species B    ATCGCTAGTCTTATATTACA
Species C    TTCACTAGACCTGTGGTCCA
Species D    TTGACCAGACCTGTGGTCCG
Species E    TTGACCAGTTCTCTAGTTCG
```

A character provides information about an individual OTU.

Distances

	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	----

A distance represents a quantitative statement concerning the dissimilarity between two OTUs.

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

Most molecular data yield **character states** that are subsequently **converted** into **distances**.

CHARACTERS ⟶ DISTANCE

DISTANCE ⟶/ CHARACTERS

Distance matrices

- There are many ways of building phylogenetic trees, one family of methods uses a **distance matrix** as a starting point.
- A distance matrix is a table that indicates pairwise dissimilarity, for instance...

	Cat	Dog	Rat	Cow
Cat	0	2	4	7
Dog	2	0	5	6
Rat	4	5	0	3
Cow	7	6	3	0

	A	B	C	D
B	400	-	-	-
C	300	300	-	-
D	250	150	250	-
E	250	250	500	200

The 4-Point Condition

- Distances that fit exactly on a tree can be characterised by a condition on any quartet i, j, k, l (i.e. it must hold true for any 4 taxa).
- We write $d(x,y)$ for the distance between x and y.

Given 4 taxa $i, j, k, l,$ of the 3 sums

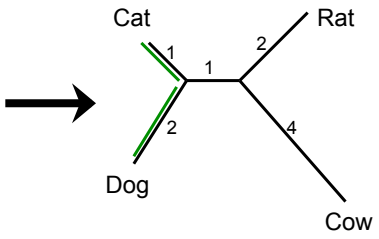
- $d(i,j) + d(k,l)$
- $d(i,k) + d(j,l)$
- $d(i,l) + d(j,k)$

The largest two are equal.

- Distances with this property are called **additive,** because the weights on the paths along the tree **add up** to the values in the distance matrix.

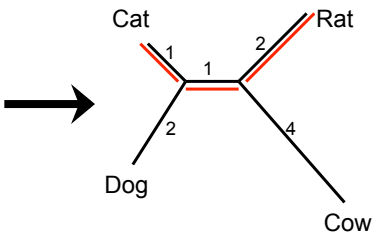
Perfectly "tree-like" distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



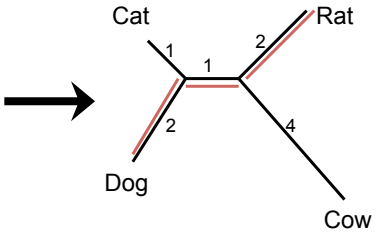
Perfectly "tree-like" distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



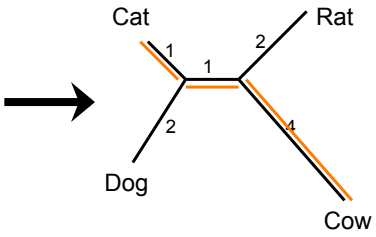
Perfectly "tree-like" distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



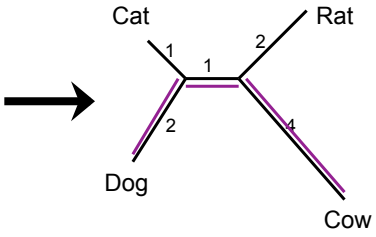
Perfectly "tree-like" distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



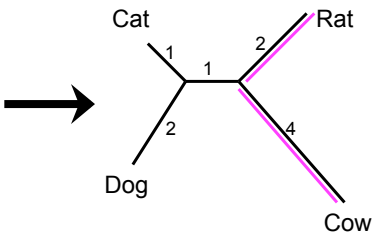
Perfectly "tree-like" distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



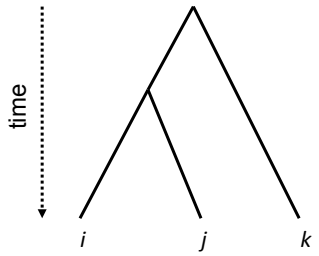
Perfectly "tree-like" distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



Clock-like distances

- An even stricter condition on distances is that they fit on a clock-like tree.
- Distances with this property are called **ultrametric**.



$d(i,k) = d(j,k) > d(i,j)$

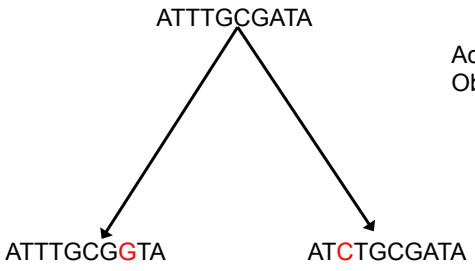
Where do we get distances from?

- Distances can be derived from Multiple Sequence Alignments (MSAs).
- The most basic distance is just a count of the number of sites which differ between two sequences divided by the sequence length. These are sometimes known as **p-distances**.

Cat	ATTTGCGGTA
Dog	ATCTGCGATA
Rat	ATTGCCGTTT
Cow	TTCGCTGTTT

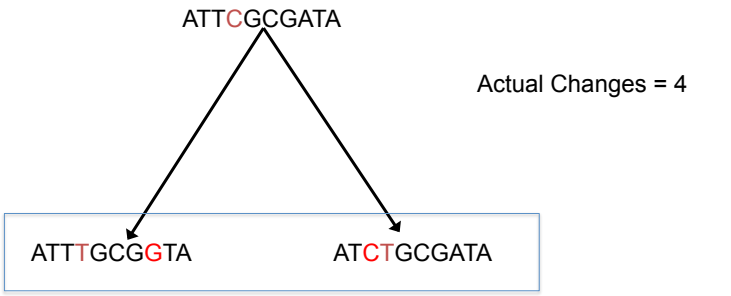
	Cat	Dog	Rat	Cow
Cat	0	0.2	0.4	0.7
Dog	0.2	0	0.5	0.6
Rat	0.4	0.5	0	0.3
Cow	0.7	0.6	0.3	0

Observed distances usually underestimate the true number of changes



Actual Changes = 2
Observed Changes = 2

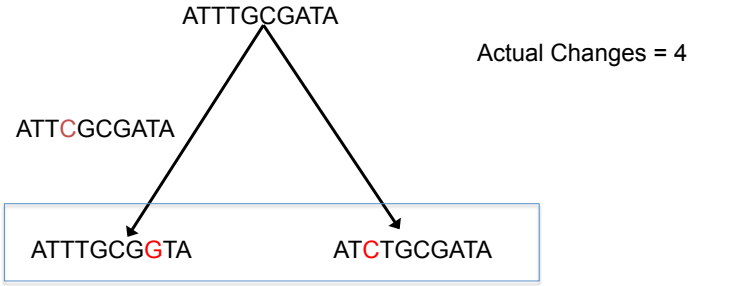
- Parallel changes
- Reversals
- Superimposed changes



Actual Changes = 4

Observed Changes = 2

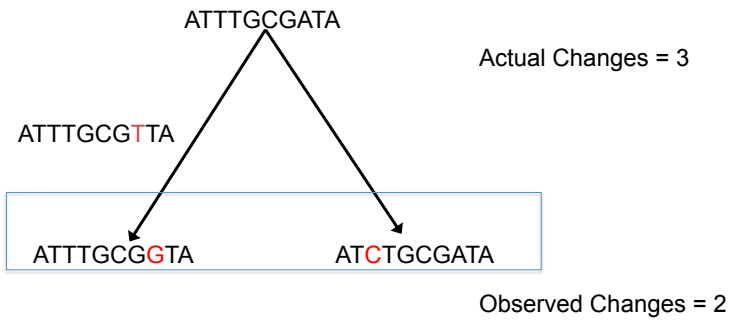
- Parallel changes
- Reversals
- Superimposed changes



Actual Changes = 4

Observed Changes = 2

- Parallel changes
- Reversals
- Superimposed changes



Correcting for hidden changes

- Given a statistical model of how point mutations occur it is possible to estimate the true genetic distance from the observed distance.