Universidade Federal do Paraná

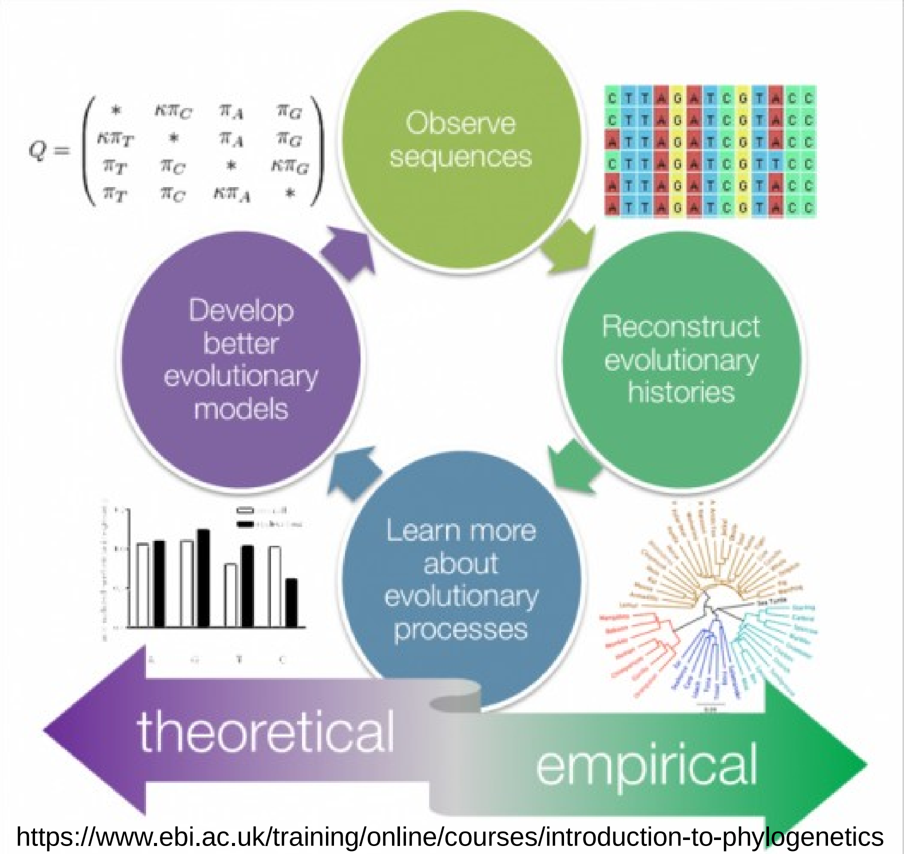
Phylogenetics
An Introduction

Prof. Leonardo M. Cruz
Programa de Pós-Graduação em Ciências – Bioquímica

What is phylogenetics?

- Study of evolutionary relationships among biological entities
  - species
  - individuals
  - genes

The major elements of phylogenetics



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

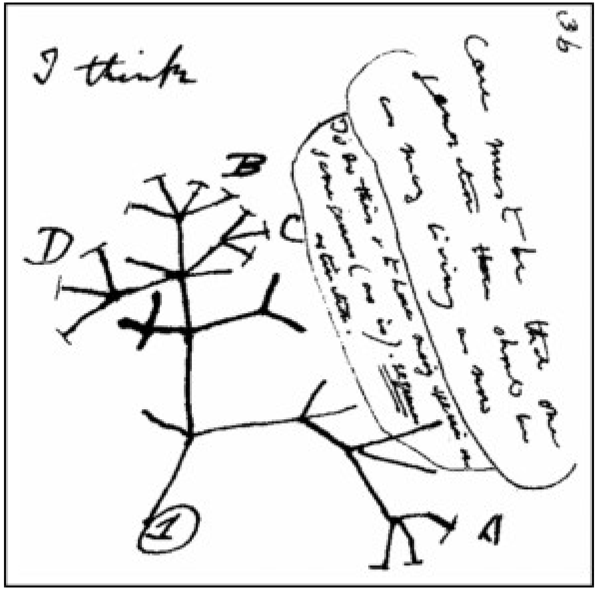
Why use molecular data?

- Today, almost all evolutionary relationships are inferred from molecular sequence data
- This is because:
  - DNA is the inherited material
  - We can now easily, quickly, inexpensively and reliably sequence genetic material
  - Sequences are highly specific and are often information rich

https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Tree thinking

- One the first sketches of a phylogenetic tree was made by Charles Darwin



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

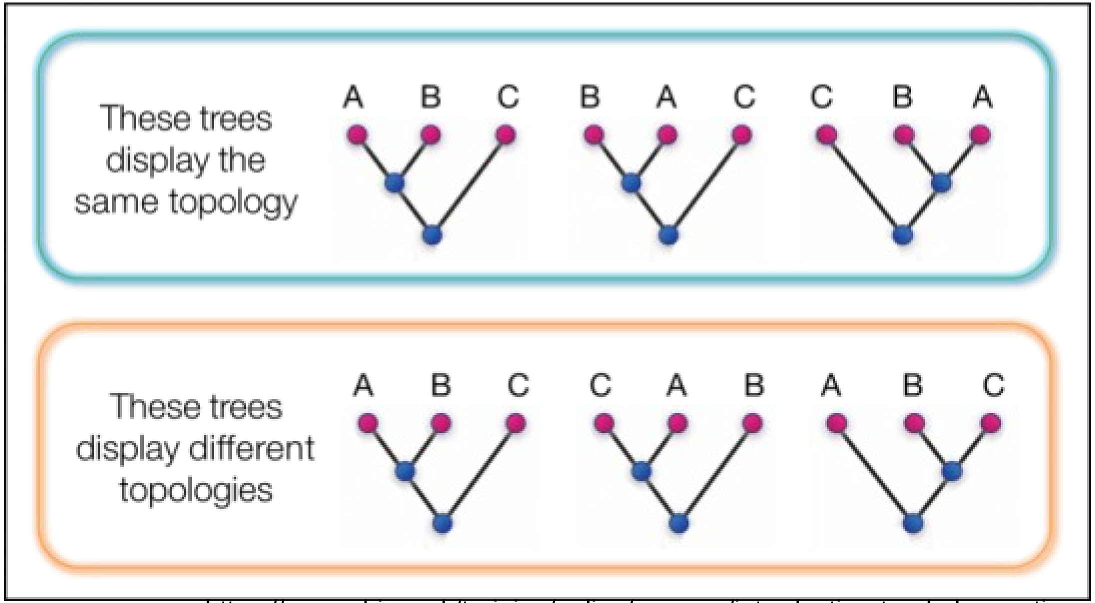
Aspects of phylogenies

- Topology – inferred from phylogeny
  - Branches – their length can be inferred from phylogeny
  - Nodes
    - Tips
    - Internal nodes
    - Root
- Confidence

https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Topology

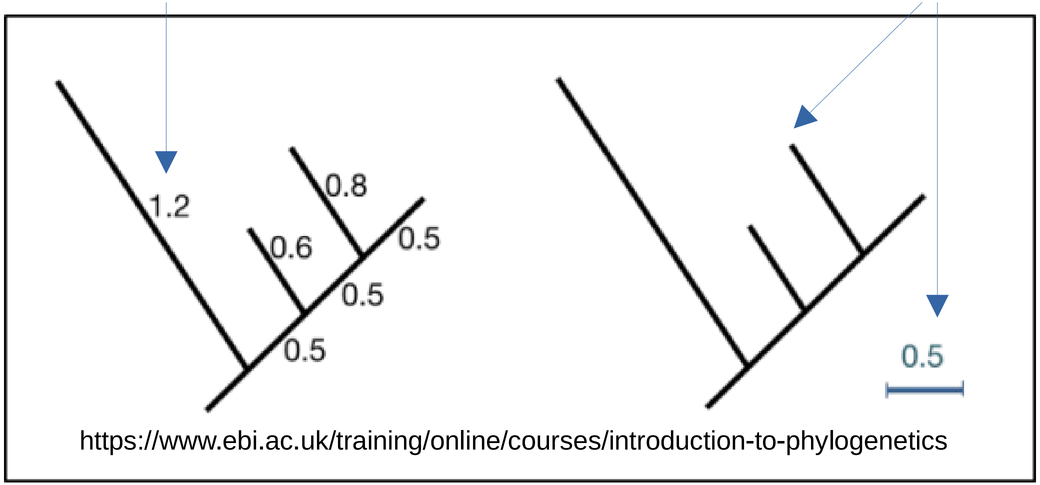
- Is the branching structure of the tree
- It is of particular biological significance because
  - indicates patterns of relatedness among taxa
  - trees with the same topology have the same biological interpretation



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Branches

- Show the path of transmission of genetic information
- **Branch lengths** indicate genetic change
  - the longer the branch, the more genetic change (or divergence) has occurred
  - estimates the average number of nucleotide or protein substitutions per site



Branch lengths shown

Branch lengths scaled

https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

How do we estimate genetic change?

- A naïve method
  - align pairs of sequences
  - count up the number of differences
  - divide by the sequence length
- Assumes that...
  - Every substitution was observed
    - Does not account multiple substitutions
  - Every substitution is equally likely
- Evolutionary models correct for these unrealistic situations

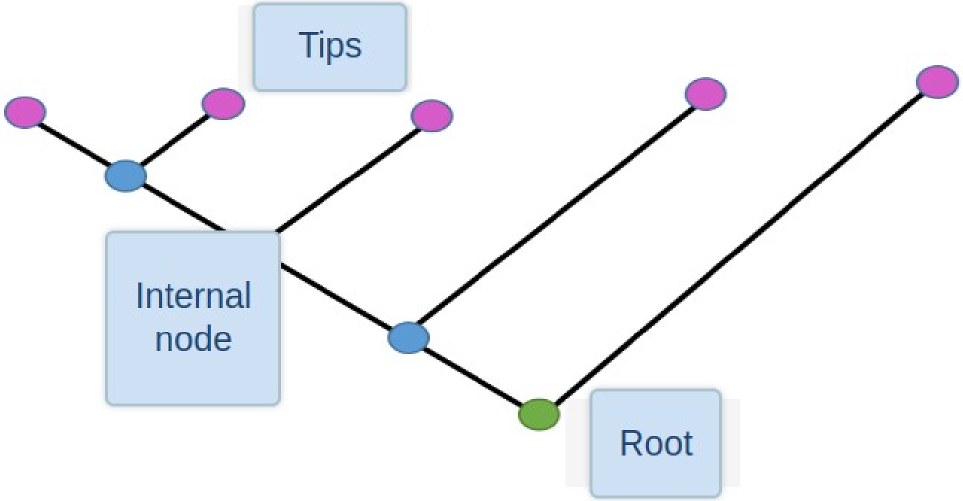
Human **ATGTTGACTC**
Mouse **ATGCTGACTC**

1 difference
10 sites
1/10 = 0.1 substitutions per site (*p* distance)

https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Nodes

- For the particular group of organisms being studied
  - Tips = currently sequences/organisms
  - Internal node = ancestral sequences/organisms
  - Root = common ancestral sequence/organism

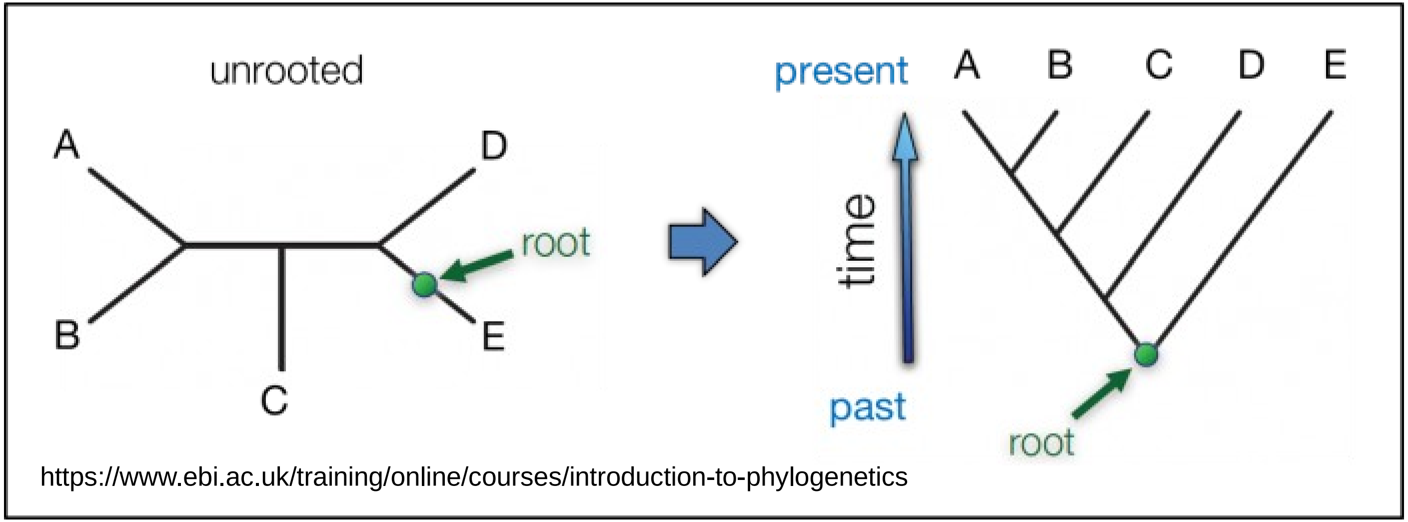


https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Internal nodes and root are hypothetical (inferred by phylogenetic methods)

Root

- Is the **most recent common ancestor** (**MRCA**) of all of the *taxa* in the tree
    - the oldest part of the tree
    - tells us the direction of evolution
- Most of phylogenetic methods do not estimate the position of the root
    - in part because this increases the number of possible trees



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Rooting a tree

- There are two main approaches that we can use to root a tree
  - **Outgroup rooting**
  - **Midpoint rooting**

https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Outgroup rooting

- Include one or more sequences that we know are more distantly related to our sequences of interest than they are to one another
- The root estimate is then simply the point at which our outgroup(s) join the rest of our tree of interest
- The best possible outgroups are those most closely related to our sequences of interest
    - If outgroups are too distantly related then they can be unreliable as they may be difficult to align or become saturated with substitutions

https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Midpoint rooting

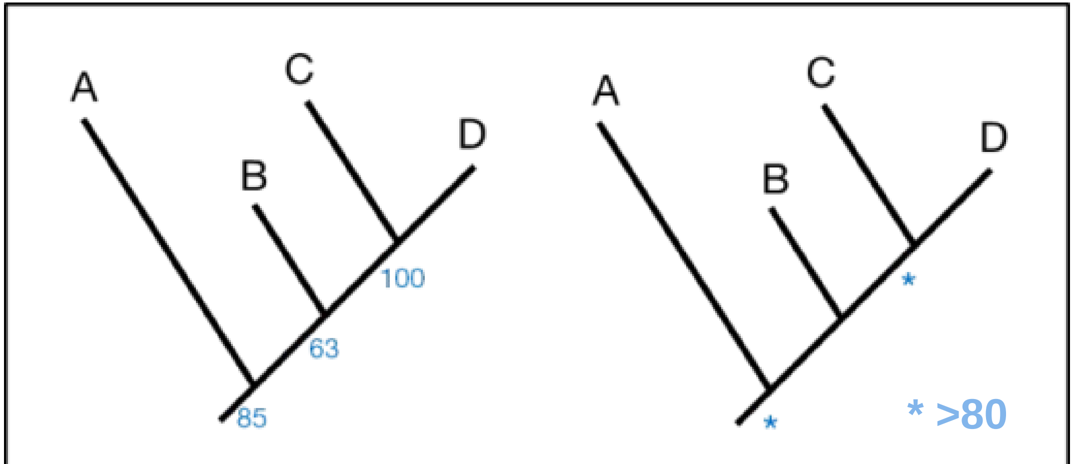
- Assume that all sequences evolve at the same rate
  - this assumption does not hold for many biological datasets!
- The root is positioned at the midpoint between the two longest branches

https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Approaches for estimating confidence

- Several approaches to estimate confidence in tree topology...
  - Bootstraps
  - Likelihood
  - Bayesian
- Confidence estimates refer to the internal branches

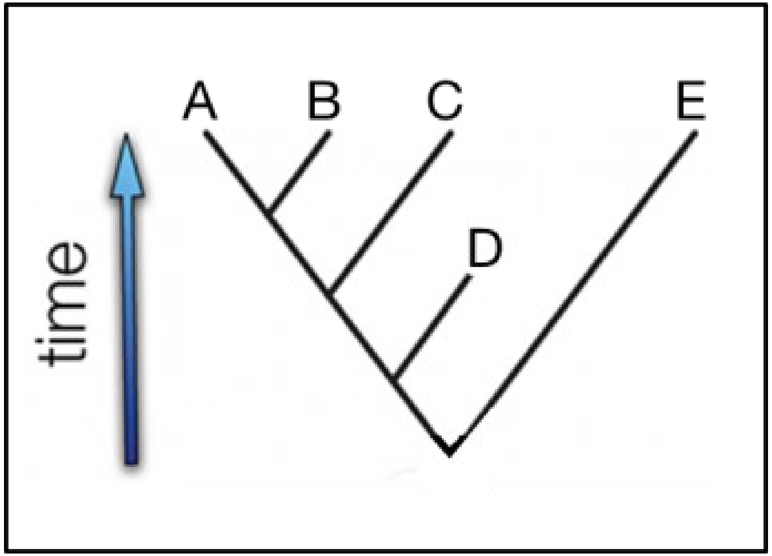
typical representations of
confidence on part of a phylogeny



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Relating distance, rate and time

- Evolutionary rate and time are confounded
  - The genetic change is calculated by a combination of the rate of substitution and the time that has elapsed
  - **Genetic change (substitutions/site) =** **Evolutionary rate (substitutions/site/year) x Divergence time (years)**



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

without any other prior information about the tree, we do not know whether branch **D** represents an extinct or frozen lineage that was sampled a long time ago; a lineage with a relatively slow evolutionary rate; or some combination of both

Evolutionary rate and time are confounded

- Genetic change/distance – **substitutions/site**
  - Calculated by a combination of…
    - Rate substitution (**substitutions/site/year**)
    - Time that has elapsed (**years**)
    - **GENETIC CHANGE =**
      EVOLUTIONARY RATE X DIVERGENCE TIME

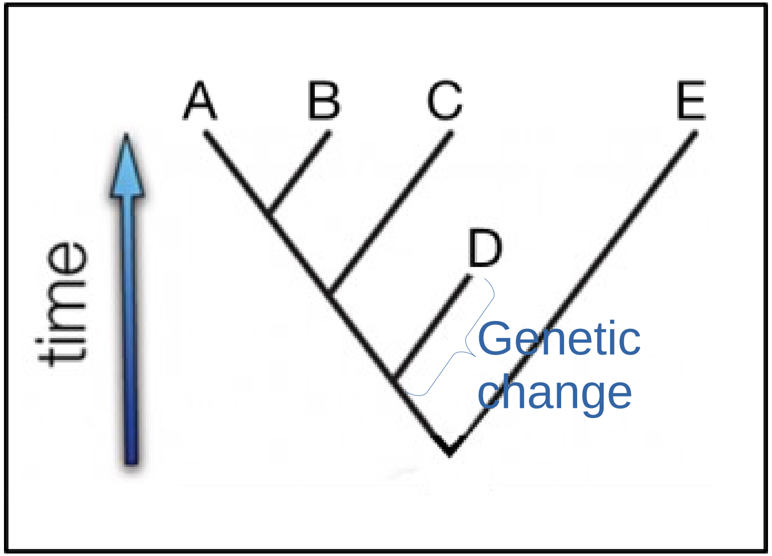
Genetic change → the branch lengths estimated in a phylogenetic tree

Relating distance, rate and time

Genetic change	**=**	**Evolutionary rate**	**X**	**Divergence time**
(substitutions/site)		(substitutions/site/year)		(years)

Not possible to know whether branch **D** represents

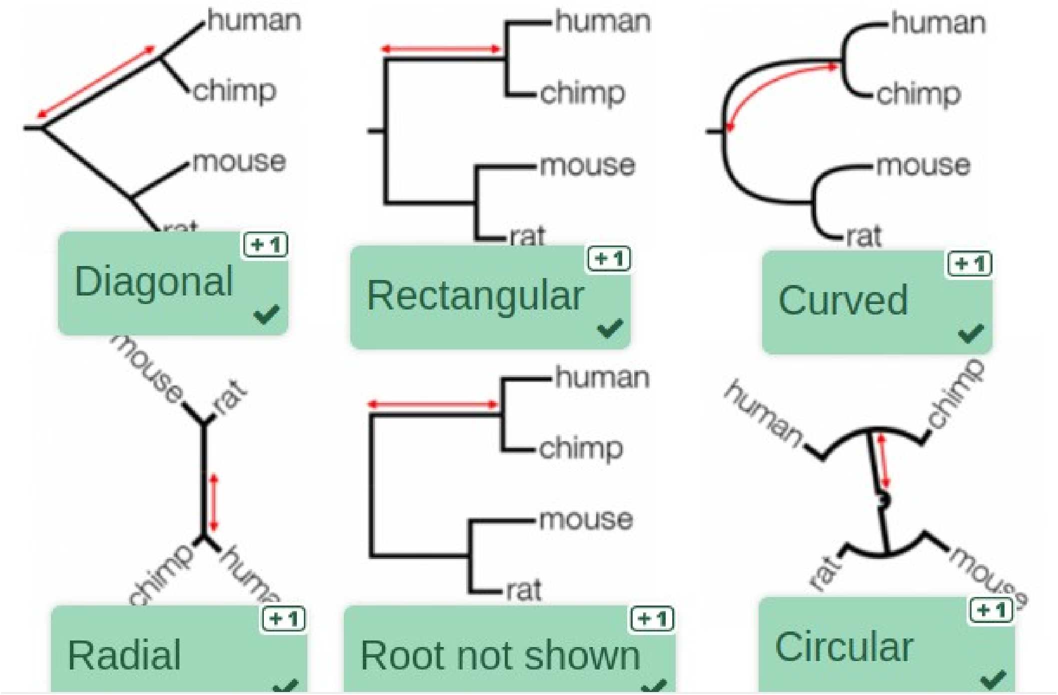
→ an extinct lineage
→ a lineage with a relatively slow evolutionary rate
→ or some combination of both



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Alternative representation of phylogenies

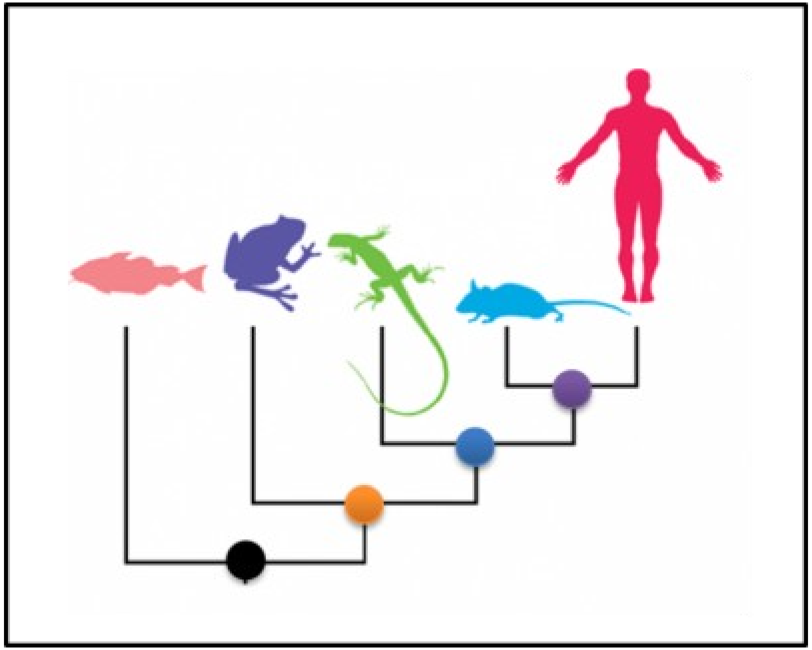
- The same topology can be drawn in lots of different ways
- Red lines indicate the same branch in each representation
- Trees can be rotated on the page and still depict the same tree
- These trees are not drawn to the same scale.



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Interpreting patterns of relatedness

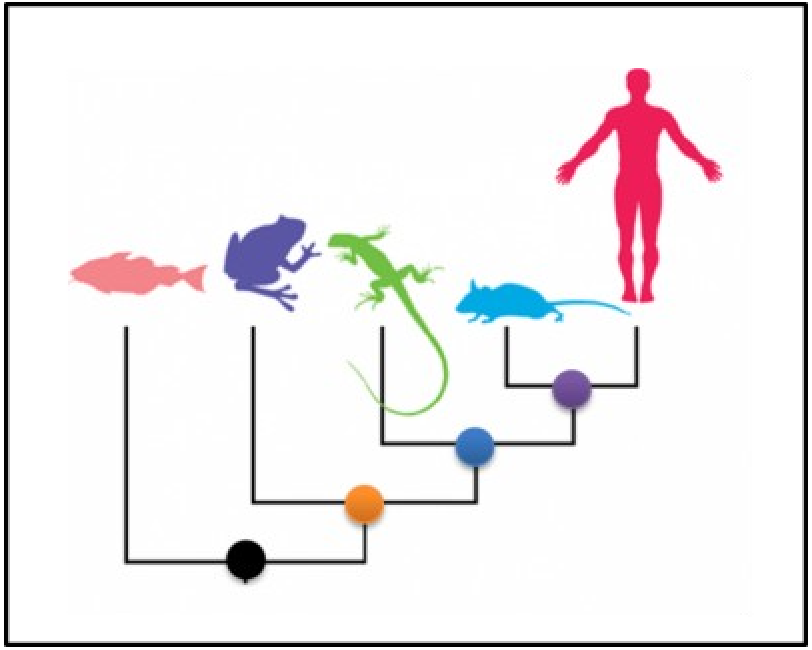
- The key to interpreting patterns of relatedness in an evolutionary context is to trace back to the point in the tree where taxa share their most recent common ancestor (MRCA)



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Interpreting patterns of relatedness

- From figure tthe following statements can be made
  - Humans are closer to mice than to lizards
    - Because humans and mice share a common ancestor more recently (purple spot)
  - Frogs are closer to lizards then to fish
    - Because frogs and lizards share a common ancestor more recently (orange spot)
  - Fish are equally related to mice and frogs
    - Mice and frogs share the same common ancestor (black spot) with fish
    - So neither species is more closely related to fish
- Remember
  - These interpretations rely on our tree being rooted because the root is needed to define the direction of evolution
  - If we rotate the branches to change the topology of the tree then the tree still has the same biological meaning and evolutionary relationships



https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Further reading

- Phylogenetics – an Introduction
  - Online Tutorial – EMBL-EBI Training
  - https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics
- How to read a phylogenetic tree
  - https://artic.network/how-to-read-a-tree.html
- Understanding Evolution
  - UC Museum of Paleontology – University of California, Berkeley
  - https://evolution.berkeley.edu/teach-evolution
  - Misconceptions about evolution (https://evolution.berkeley.edu/teach-evolution/misconceptions-about-evolution)

Further reading

- Baum DA, Smith SD, Donovan SS. Evolution. The tree-thinking challenge. Science (New York, N.Y.). 2005 Nov;310(5750):979-980. DOI: 10.1126/science.1117727. PMID: 16284166.
- Baldauf SL. Phylogeny for the faint of heart: a tutorial. Trends in Genetics : TIG. 2003 Jun;19(6):345-351. DOI: 10.1016/s0168-9525(03)00112-4. PMID: 12801728.
- Chapter 1. Molecular Evolution and phylogeny: what, when, why and how? Kolekar, P; Kale, M; Kulkarni-Kale, U. In: Lopes, HS; Cruz, LM (ed.) Computational biology and applied bioinformatics. InTech, 2011
  - https://www.intechopen.com/books/249
- Chapter 5. Protein Function Prediction. Cruz LM et al. Kaufmann, M.; Klinger, C.; Savelsbergh, A. (eds.) Functional Genomics - Methods and Protocols. Third Edition - Methods in Molecular Biology 1654 - Springer protocols. Humana Press.
- Barry G. Hall, Building Phylogenetic Trees from Molecular Data with MEGA, Molecular Biology and Evolution, Volume 30, Issue 5, May 2013, Pages 1229–1235
  - https://doi.org/10.1093/molbev/mst012