Phylogenetics

Introduction to tree building (and German snacks)

Rachel Warnock, Laura Mulvey
rachel.warnock@fau.de, laura.l.mulvey@fau.de

September 5, 2022

Analytical Paleobiology Workshop, Erlangen 2022

About this course

Mon 5–Wed 7 Sept 09:00–17:00 CET at Henke Str.

A mix of lectures and exercises

All material available via the Course website

Course objectives

To develop a working knowledge of models used for phylogenetics in palaeobiology.

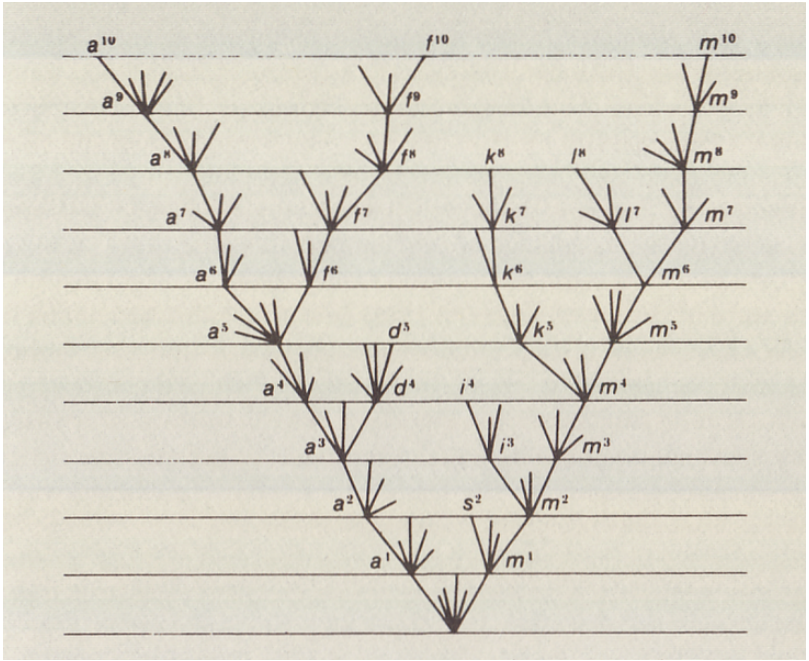
- Tree building
- Substitution models (DNA)
- Dating trees
- Clock models
- Tree models
- Substitution models (morphology)
- Model selection
- Model adequacy

We will apply these models in a Bayesian phylogenetic framework using the software RevBayes.

Phylogenetics is full of jargon, so don't hesitate to ask for clarification / ask questions!

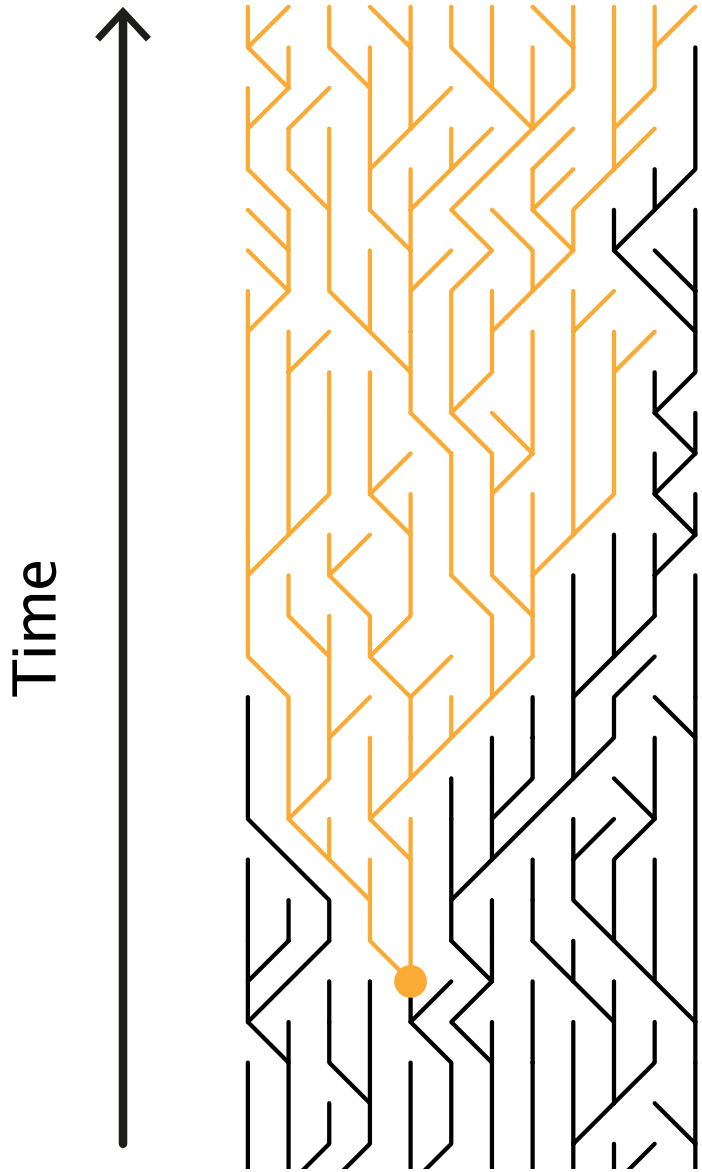
Part 1 objectives

- Begin “tree-thinking”
- Gain an understanding of the parsimony approach to tree-building and statistical inconsistency



From Darwin’s *Origin of Species*

What is phylogenetics?



- populations
- species
- viruses
- cells
- languages

Data

- DNA
- morphology
- words

What is phylogenetics?

At the broadest level – the discipline that lets us study the relationships between individuals.

We can apply the same principles to any case where we have hierarchical (ancestor & descendant) relationships.

The data is anything that can tell us about the relationships between individuals.

What is phylogenetics?

TO

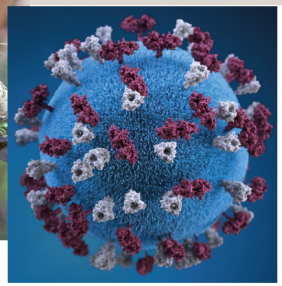
A MOUSE,

On turning her up in her Neſt, with the Plough,

November, 1785.

WEE, ſleeket, cowran, tim'rous *beaſtie*,
O, what a panic's in thy breaſtie!
Thou need na ſtart awa ſae haſty,
Wi' bickering brattle!
I wad be laith to rin an' chaſe thee,
Wi' murd'ring *pattle!*

I'm truly ſorry Man's dominion
Has broken Nature's ſocial union,
An' juſtifies that ill opinion,
Which makes thee ſtartle,
At me, thy poor, earth-born companion,
An' *fellow-mortal!*



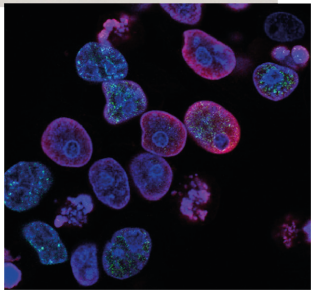
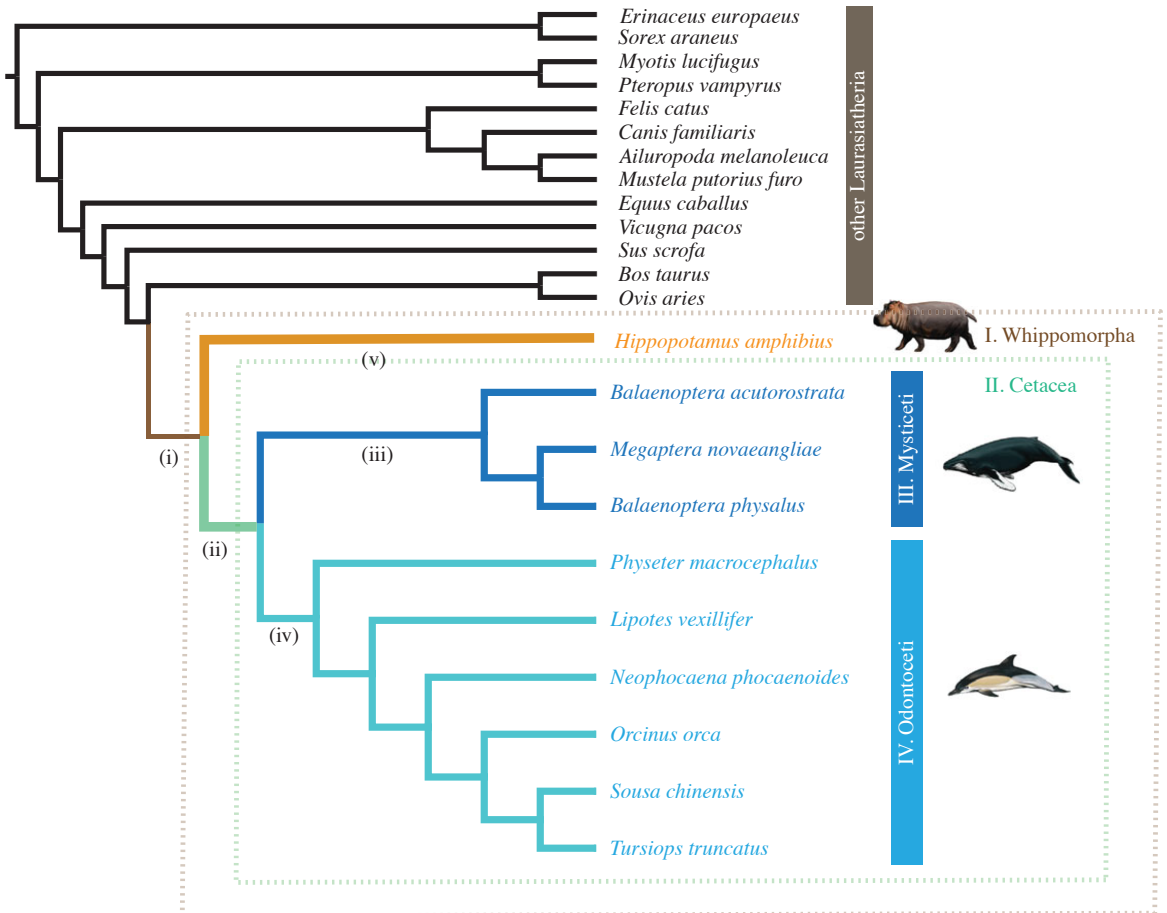


Image source upsplash.org

What can we learn from trees?

How are our favourite species related?
How does phylogeny reflect taxonomy?



Tsagkogeorga et al. (2015) *Royal Soc Open Science*

What can we learn from trees?

Evolutionary relationships

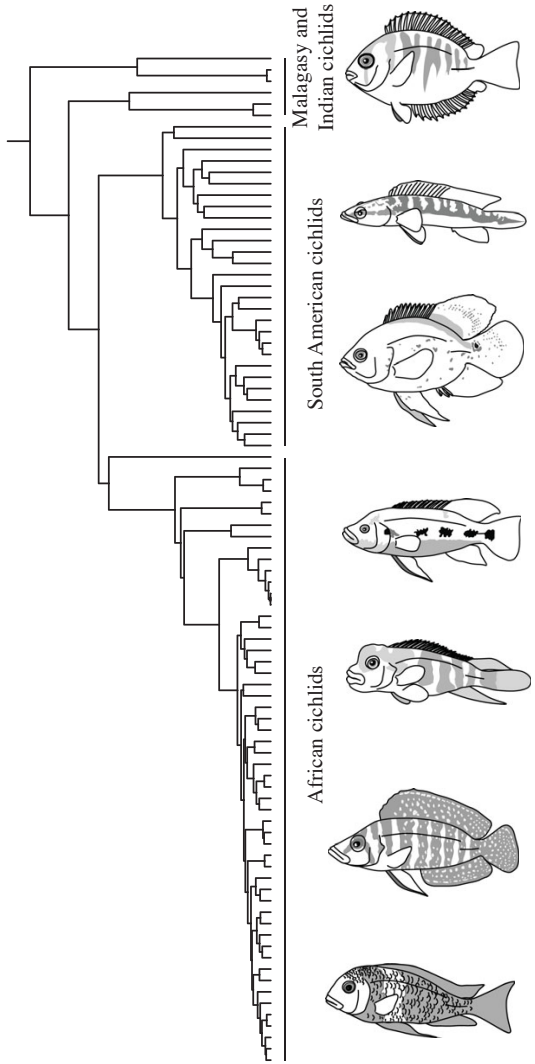


Image adapted from Friedmann et al. 2013. PRSB

What can we learn from trees?

Evolutionary relationships

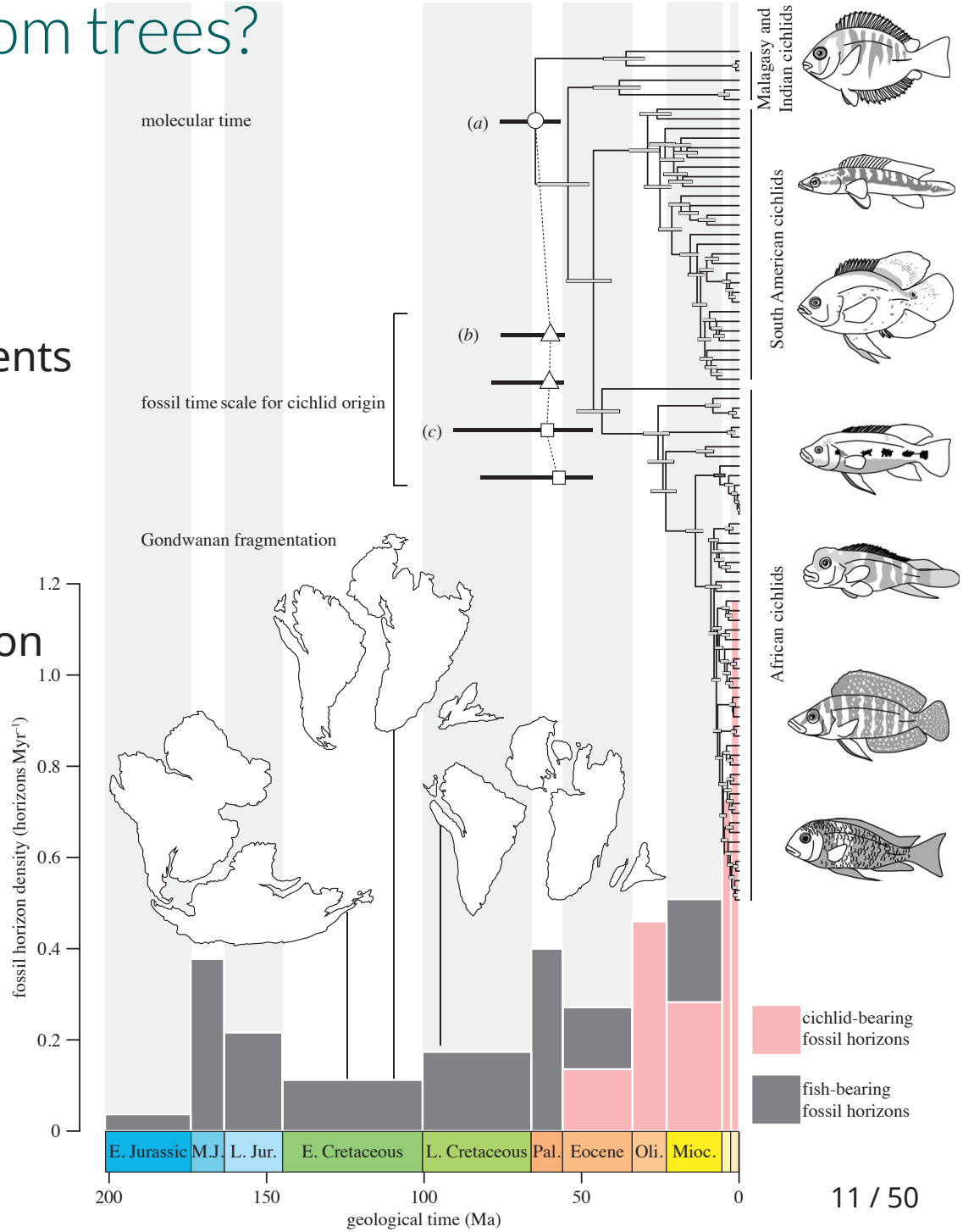
Timing of diversification events

Geological context

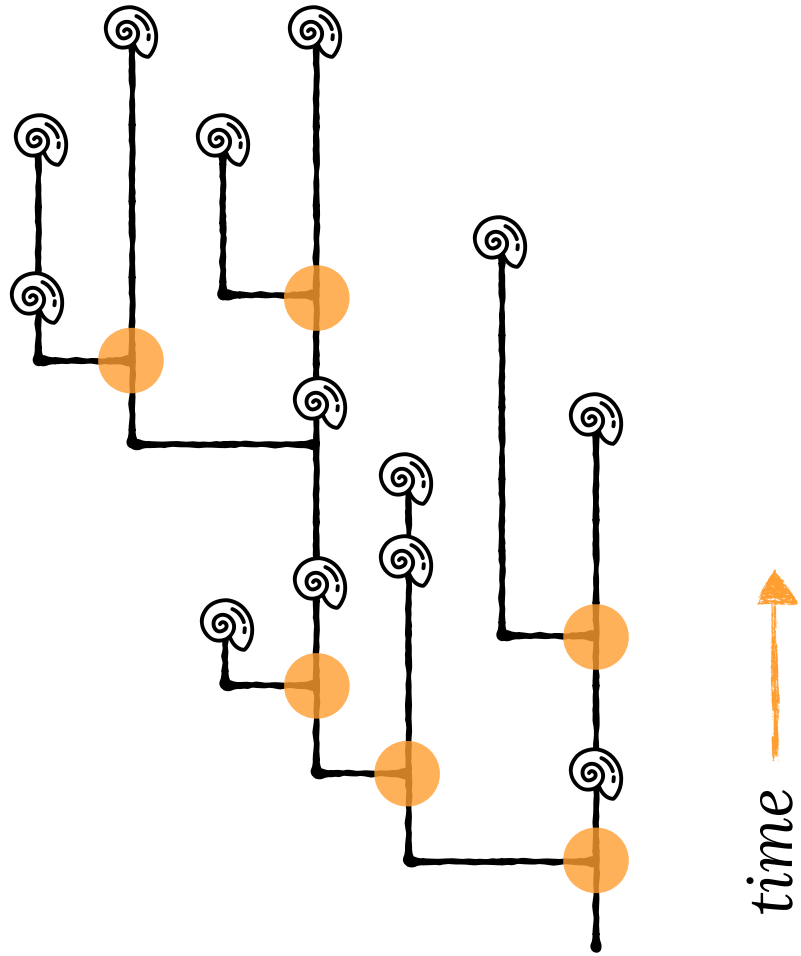
Rates of phenotypic evolution

Diversification rates

...



A phylogenetic tree captures part of evolutionary history that is otherwise not directly observable.

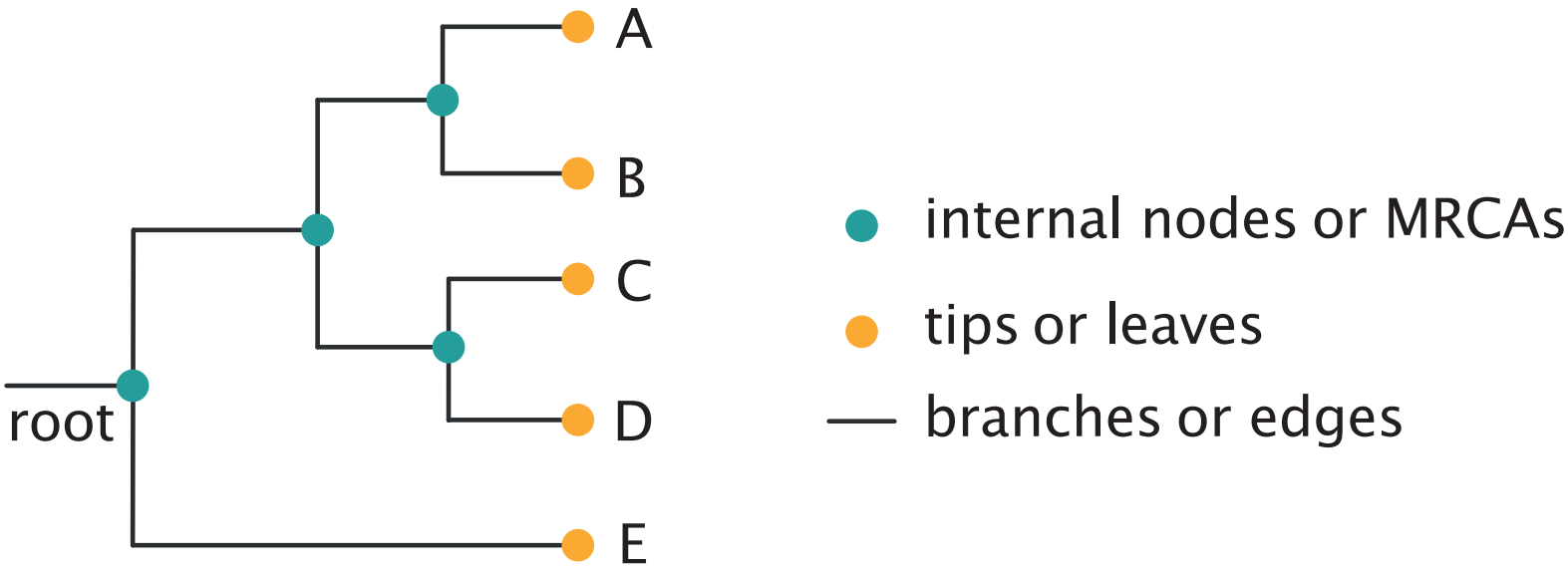


Nothing in biology makes sense except in the light of evolution — Theodosius Dobzhansky (1973)

Nothing in evolution makes sense except when seen in the light of phylogeny — Jay M. Savage (1997)

Where do we begin?

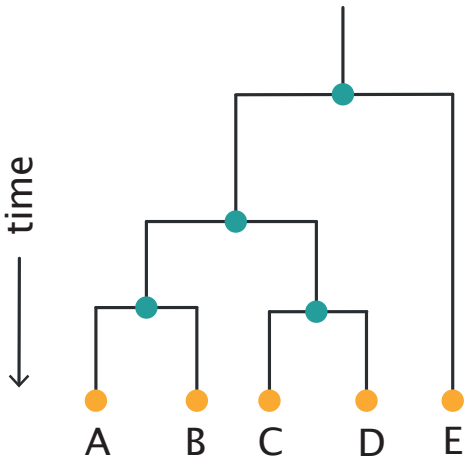
Some basic terms



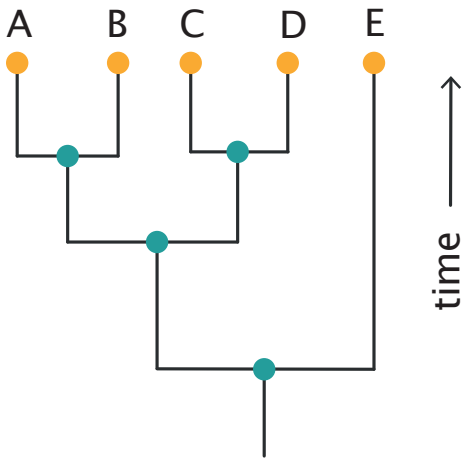
branch lengths =genetic distance OR time

Note: genetic distance = rate x time

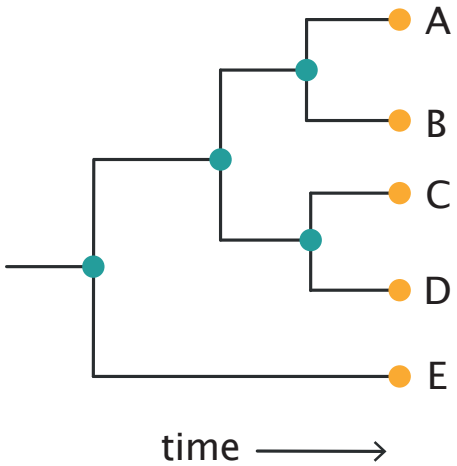
The direction of time



Computer science, maths



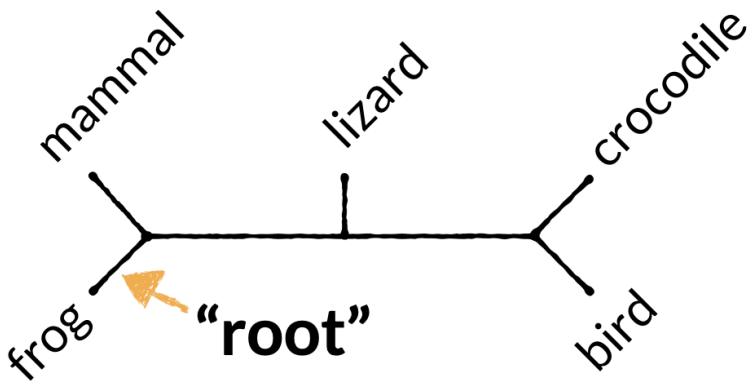
Geology

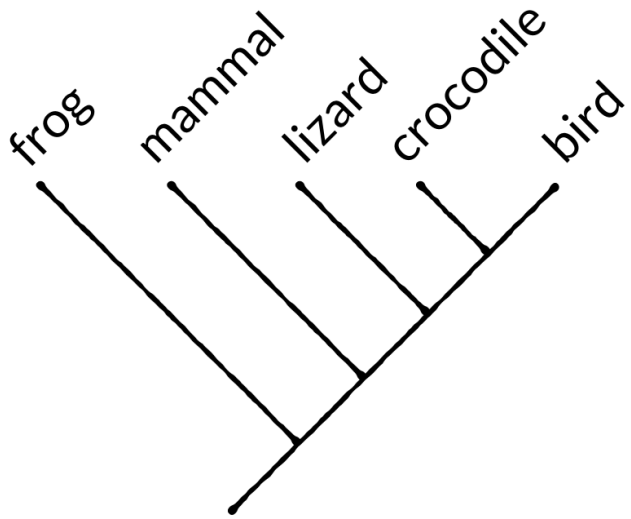


Evolutionary biology

Tip: look for the root!

Rooted versus unrooted trees

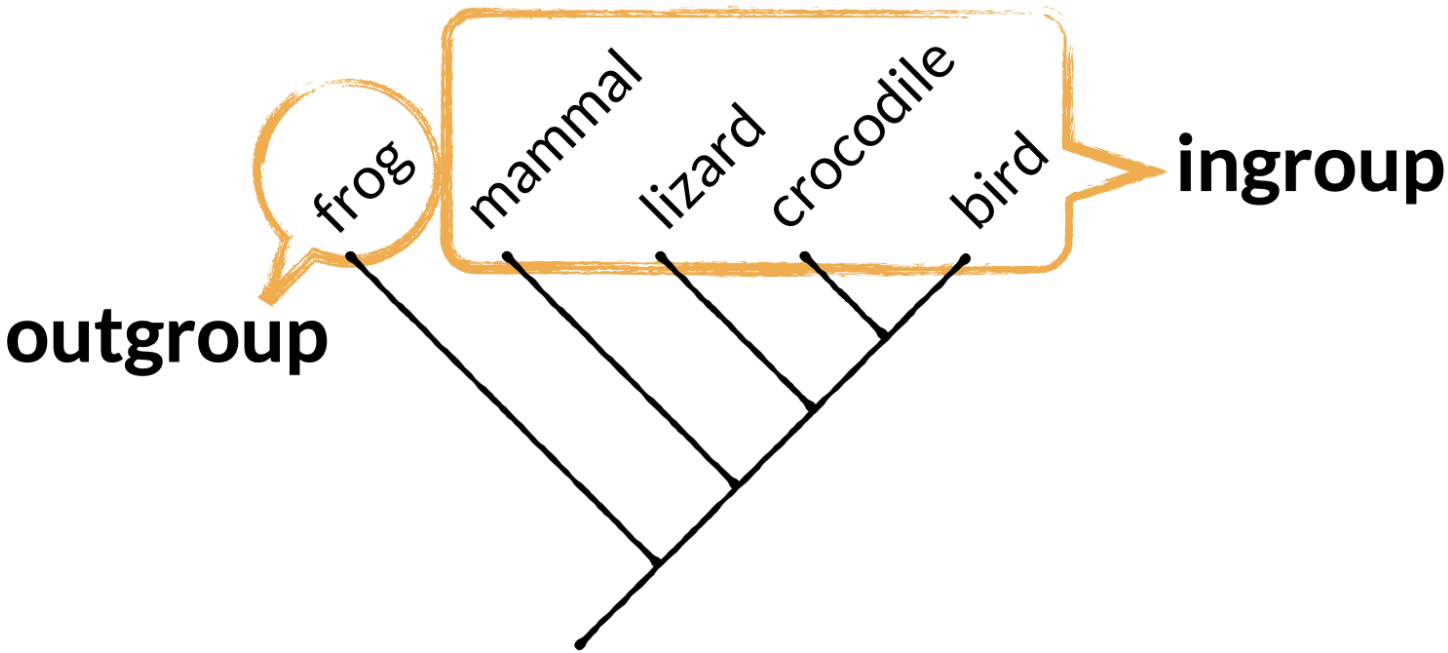




Phylogenies are unrooted by default, because phylogenetic characters don't contain information about the direction of time.

Image adapted from Phil Donoghue

Rooted versus unrooted trees



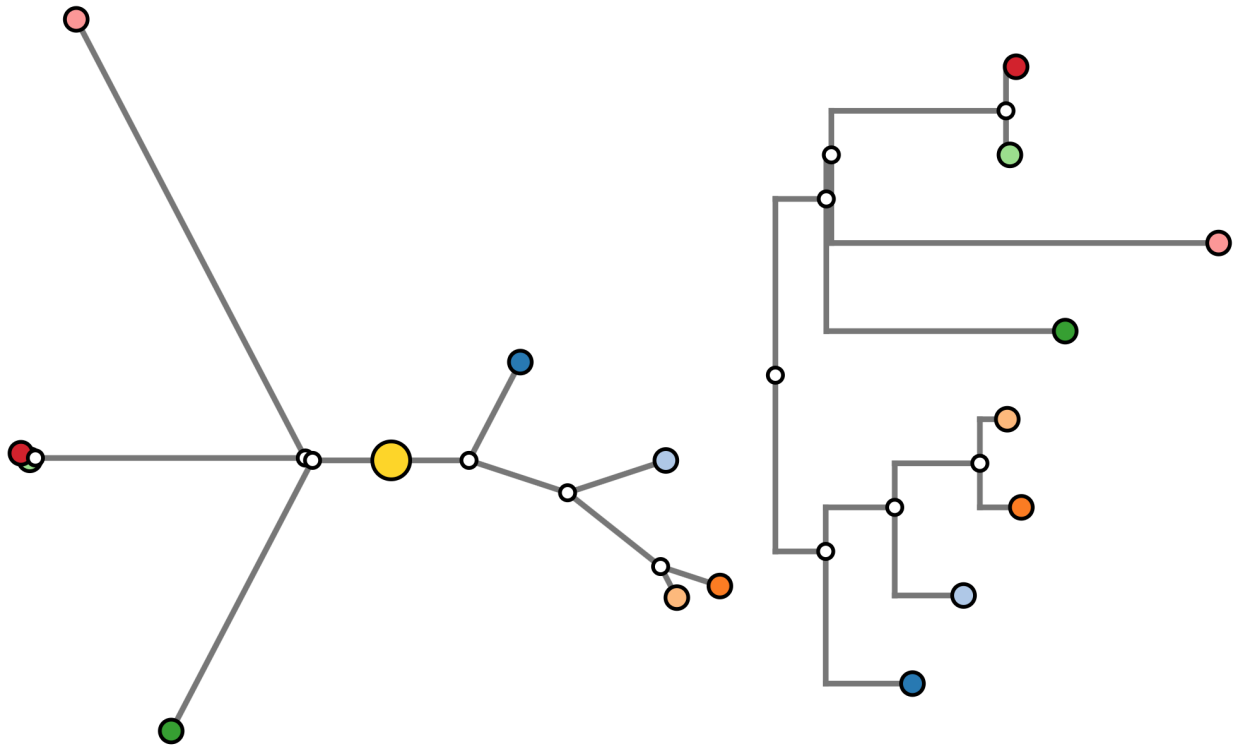
We have to find a way of breaking one of the branches in two, where the break represents the oldest point in our tree.

The most common approach is to use an outgroup – a taxon that we know is more distantly related than any of the taxon within the ingroup.

Image adapted from Phil Donoghue

Rooted versus unrooted trees

By default phylogenies are not rooted.



We need an **outgroup** OR a model that includes **time**.

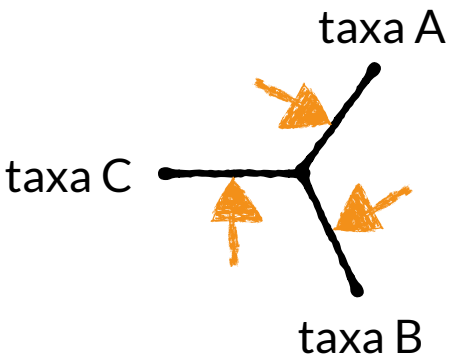
Use Art Poon's online tool to explore this further. Click here to learn more about reading trees.

Rooted versus unrooted trees

How many possible trees are there for 3 species?

Rooted versus unrooted trees

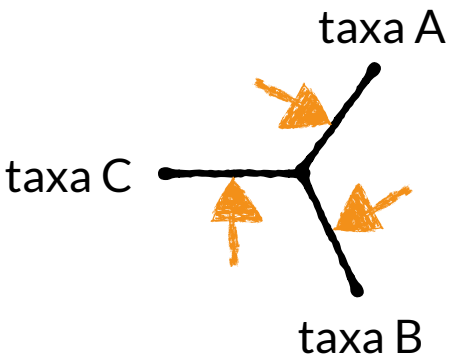
How many possible trees are there for 3 species?

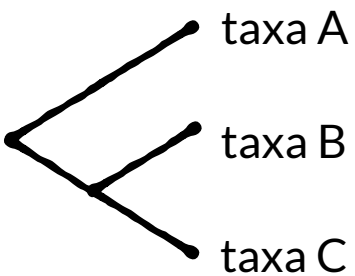


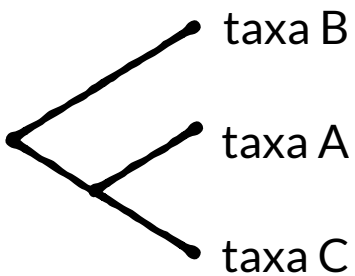
unrooted = 1

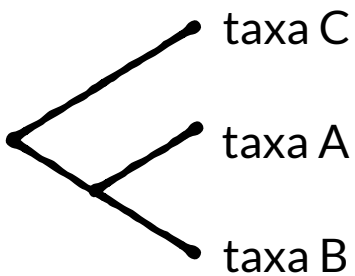
Rooted versus unrooted trees

How many possible trees are there for 3 species?







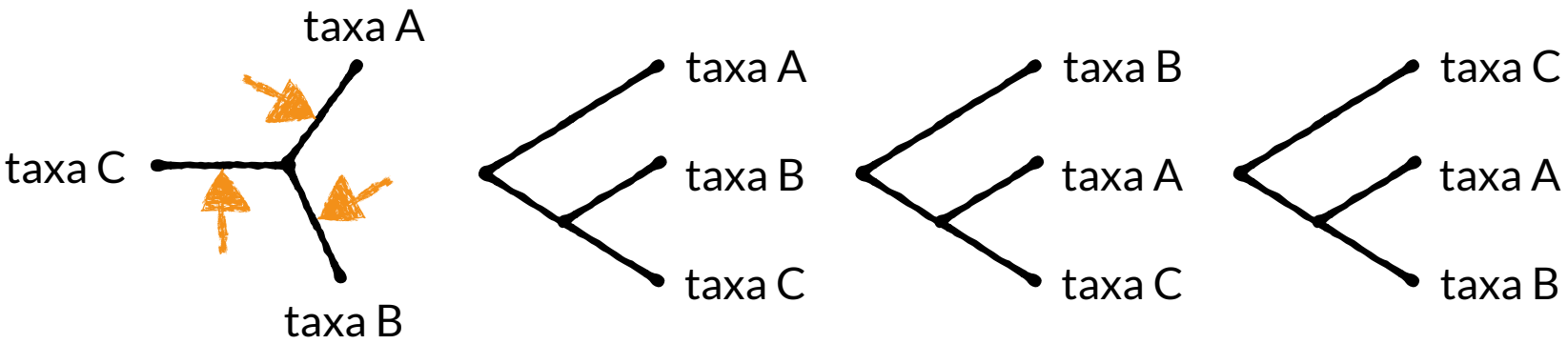


unrooted = 1

rooted = 3

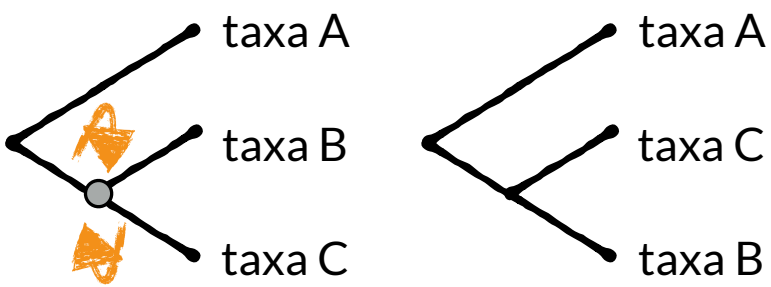
Rooted versus unrooted trees

How many possible trees are there for 3 species?



unrooted = 1

rooted = 3



Note these 2 trees are the same! B and C are more closely related.

Group exercise

Let's build a tree $\rightarrow$ form groups of 5 and come to the front to collect the exercise materials!

We're going to try building a rooted tree of snacks. Go to the Course website for more info.

Break

Exercise

Character	taxa A	taxa B	taxa C	taxa D	taxa E
Lungs	0	1	1	1	0
Jaws	0	1	1	1	1
Feathers	0	0	1	0	0
Gizzard	0	0	1	1	0
Fur	0	1	0	0	0

- How many possible unrooted or rooted trees are there?
- What do you think the correct rooted tree should be?
- Write down your logic.

Exercise

Character	taxa A	taxa B	taxa C	taxa D	taxa E
Lungs	0	1	1	1	0
Jaws	0	1	1	1	1
Feathers	0	0	1	0	0
Gizzard	0	0	1	1	0
Fur	0	1	0	0	0

- How many possible trees are there?

There are a huge number of possible trees!

# species	# unrooted trees	# rooted trees
3	1	3
4	3	15
5	**15**	**105**
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

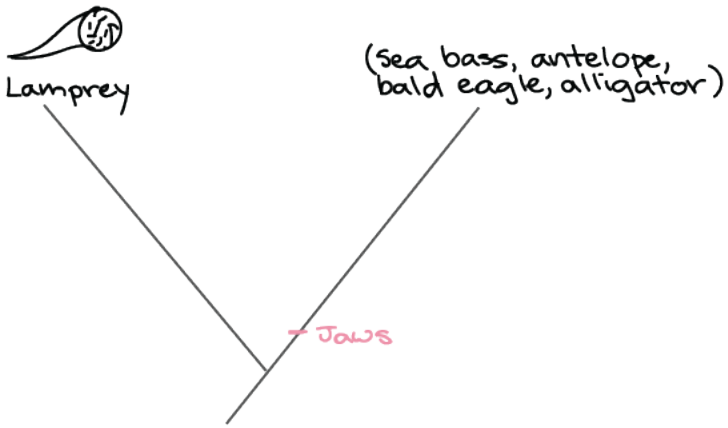
Exercise

Character	taxa A	taxa B	taxa C	taxa D	taxa E
Lungs	0	1	1	1	0
Jaws	0	1	1	1	1
Feathers	0	0	1	0	0
Gizzard	0	0	1	1	0
Fur	0	1	0	0	0

- What do you think the correct tree should be?

Exercise

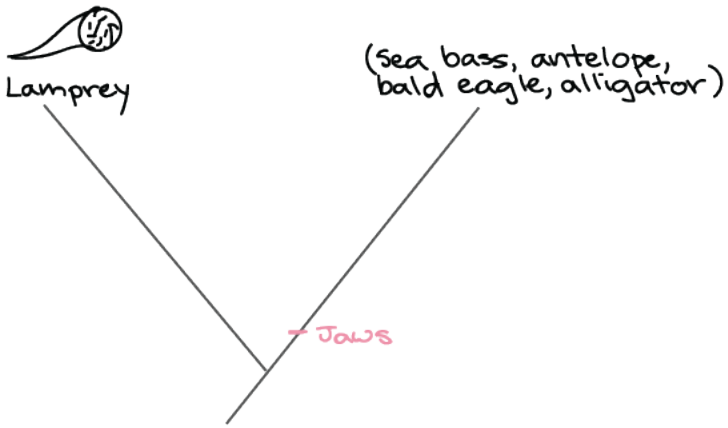
- What do you think the correct tree should be?

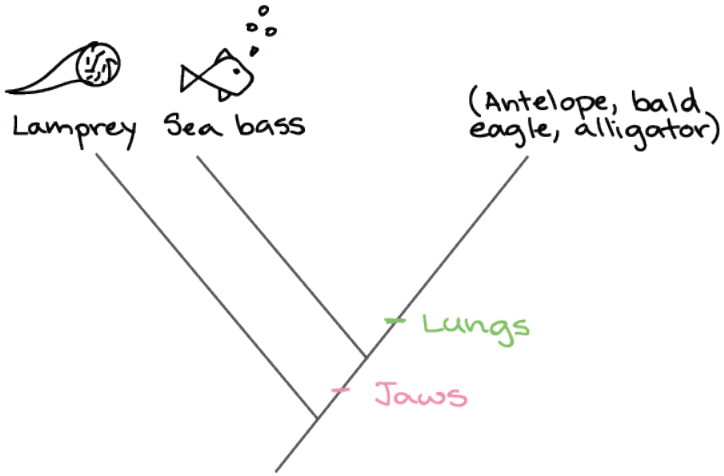


A = Lamprey, B = Antelope, C = Bald eagle, D = Alligator, E = Sea bass

Exercise

- What do you think the correct tree should be?

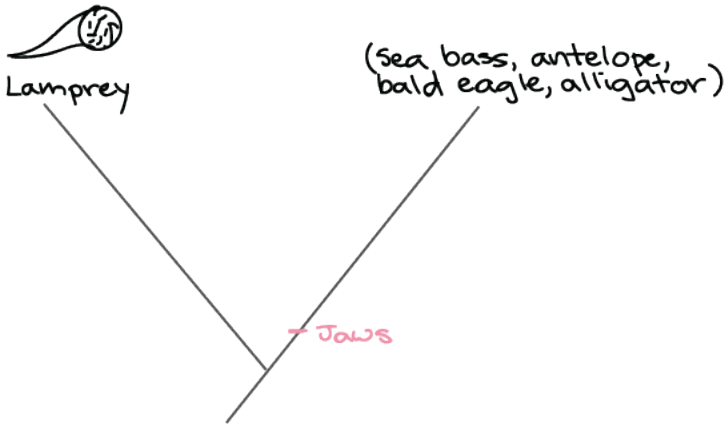


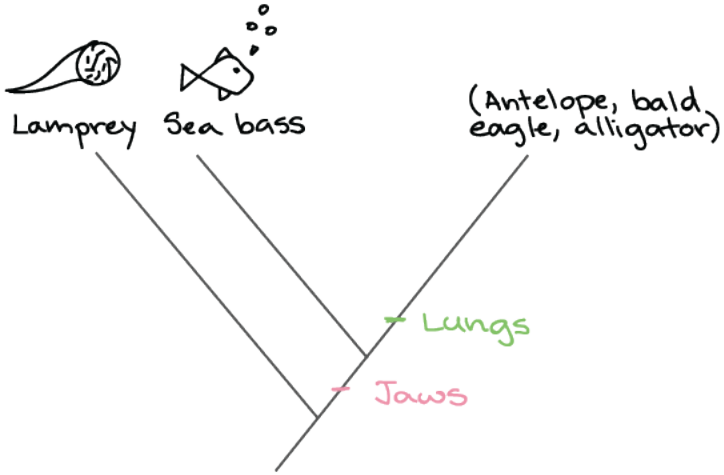


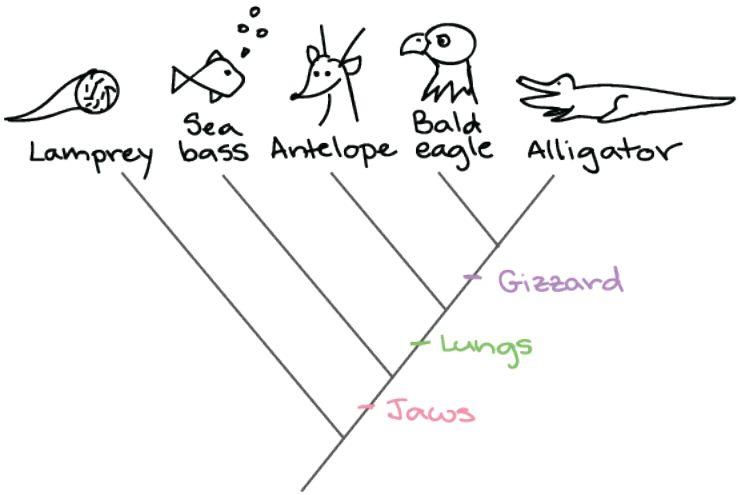
A = Lamprey, B = Antelope, C = Bald eagle, D = Alligator, E = Sea bass

Exercise

- What do you think the correct tree should be?



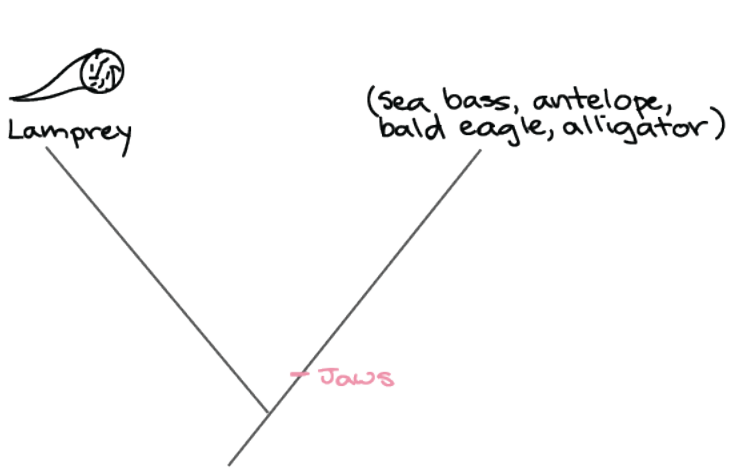


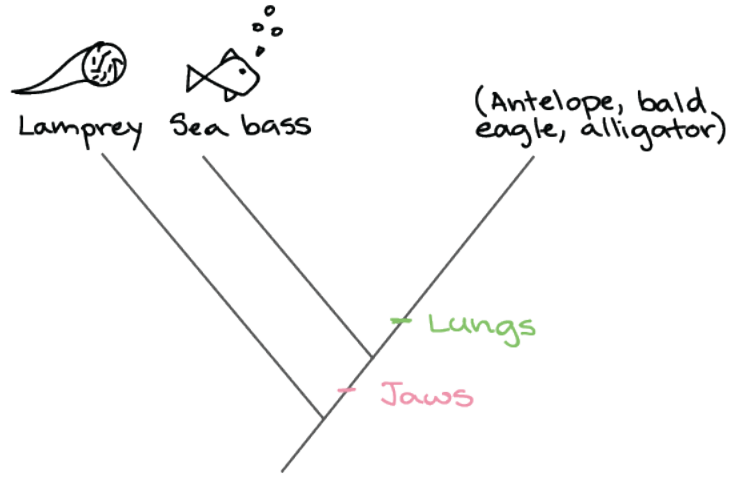


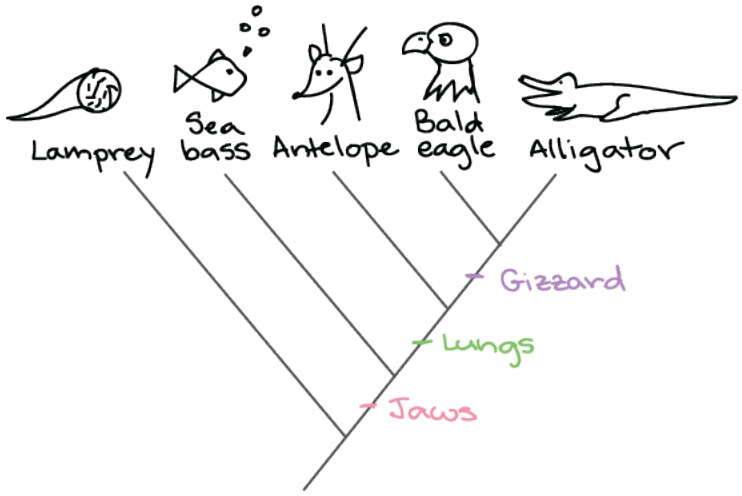
A = Lamprey, B = Antelope, C = Bald eagle, D = Alligator, E = Sea bass

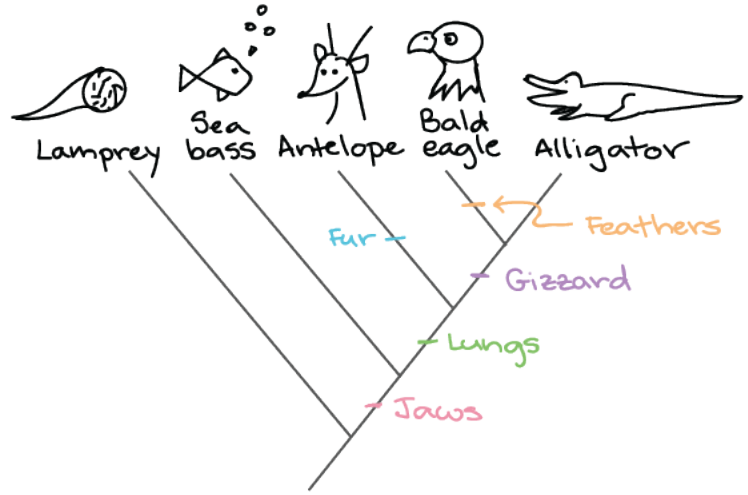
Exercise

- What do you think the correct tree should be?





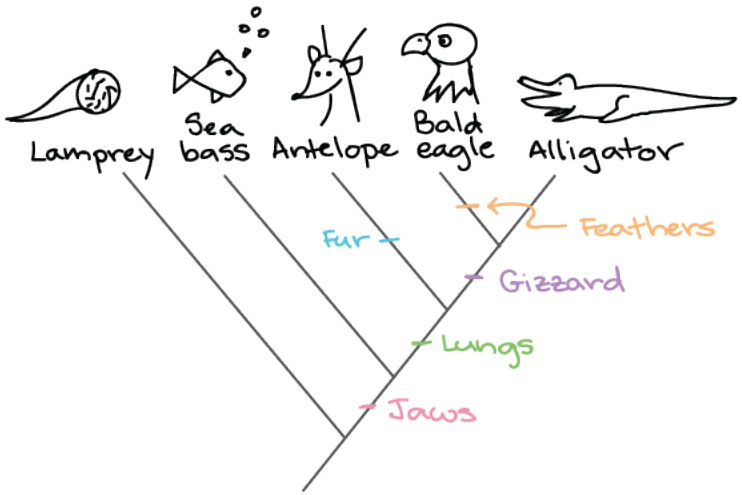


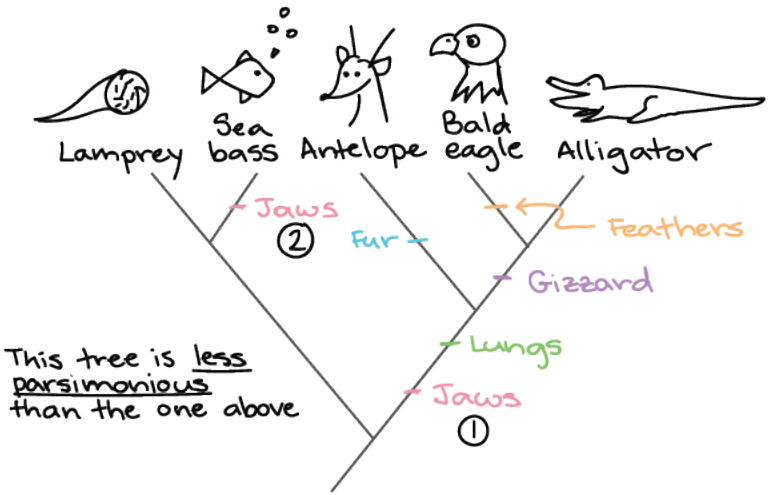


A = Lamprey, B = Antelope, C = Bald eagle, D = Alligator, E = Sea bass

Exercise

- What do you think the correct tree should be?





A = Lamprey, B = Antelope, C = Bald eagle, D = Alligator, E = Sea bass

Source Khan Academy

Exercise

- Write down your logic.

Exercise

- Write down your logic.

  $\rightarrow$ Most people intuitively assume the tree with the *fewest* changes is correct.

Exercise

- Write down your logic.

  $\rightarrow$ Most people intuitively assume the tree with the *fewest* changes is correct.

  $\rightarrow$ This approach to tree building is called parsimony.

How do we find the "best" tree?

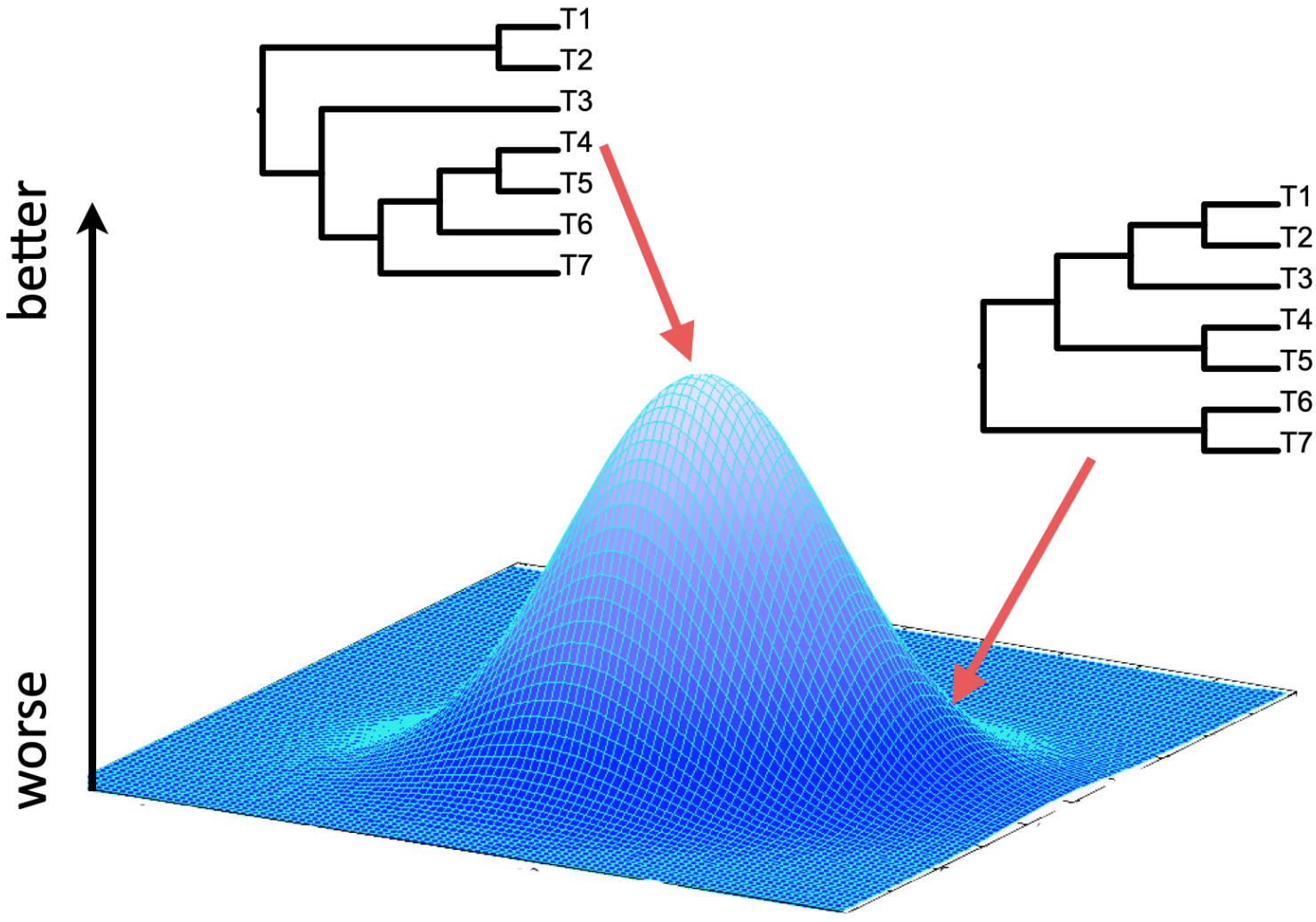


Image source: Tracy Heath

It depends how you measure "best"

Method	Criterion (tree score)
Maximum parsimony	Minimum number of changes
Maximum likelihood	Log likelihood score, optimised over branch lengths and model parameters
Bayesian	Posterior probability, integrating over branch lengths and model parameters

Both maximum likelihood and Bayesian inference are model-based approaches.

Parsimony

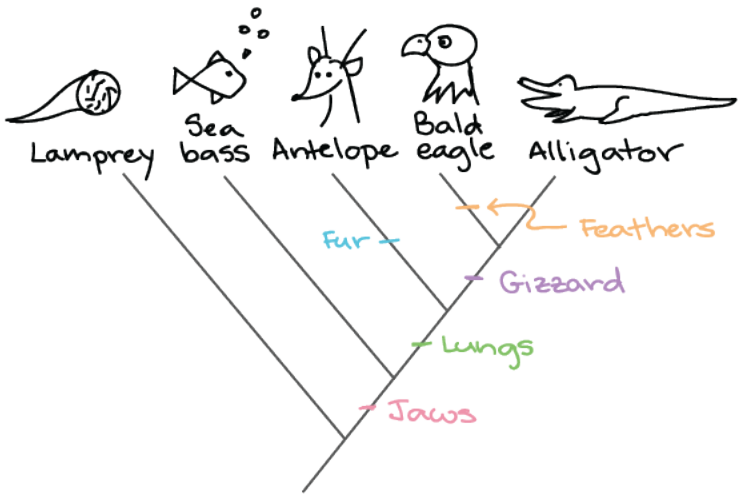
Maximum parsimony selects the tree (or trees) that require the fewest number of changes.

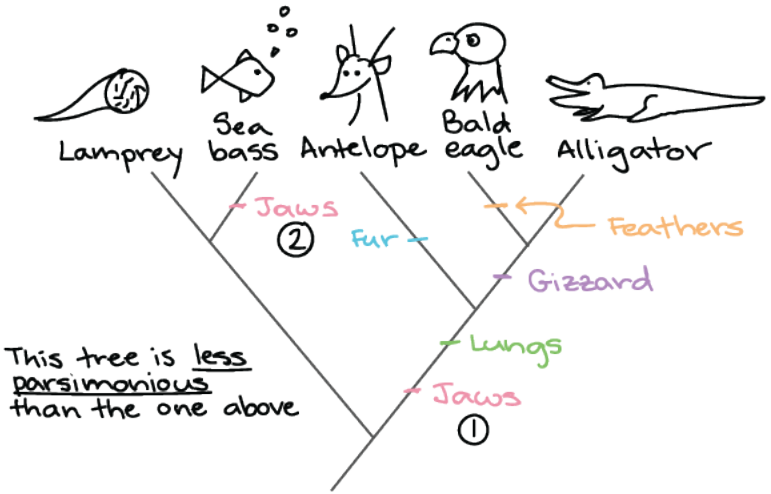
Given two trees, the one minimising the parsimony score (i.e., the minimum number of changes) is the better one.

Parsimony

Maximum parsimony selects the tree (or trees) that require the fewest number of changes.

Given two trees, the one minimising the parsimony score (i.e., the minimum number of changes) is the better one.

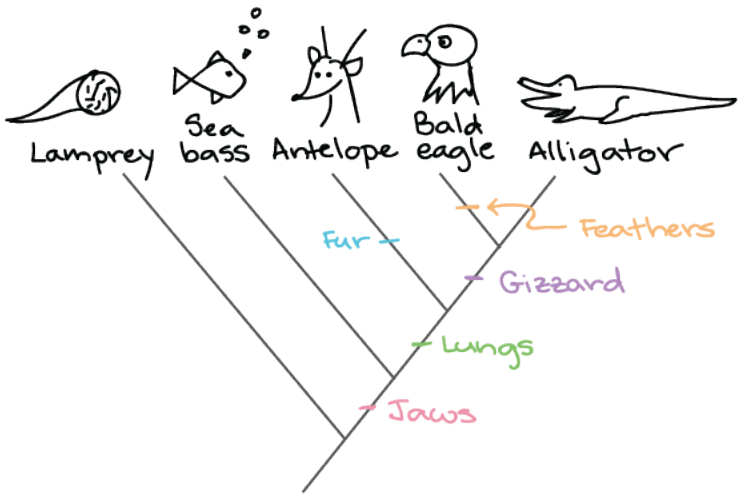


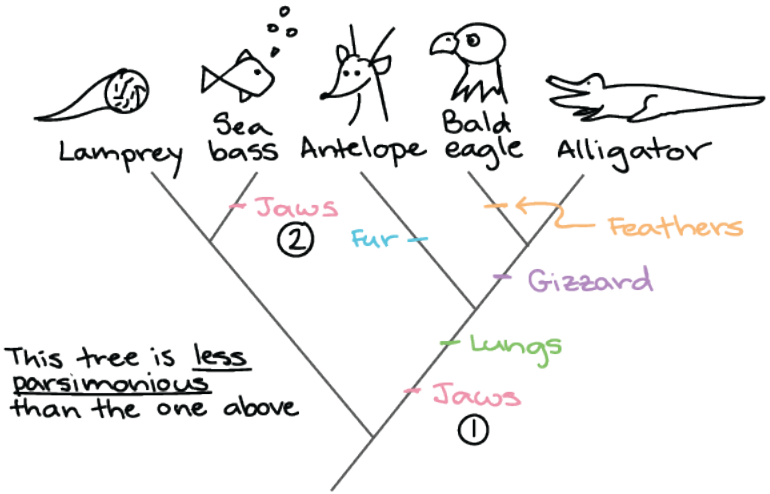


Parsimony

Maximum parsimony selects the tree (or trees) that require the fewest number of changes.

Given two trees, the one minimising the parsimony score (i.e., the minimum number of changes) is the better one.





Branch lengths = *number of observed changes or steps*

Parsimony

It is based on the **parsimony principle**: assume simpler explanations are better than complex ones.

Parsimony does not make **explicit** assumptions about the evolutionary process that generated the observed data.
$\rightarrow$ However, the method makes **implicit** assumptions.

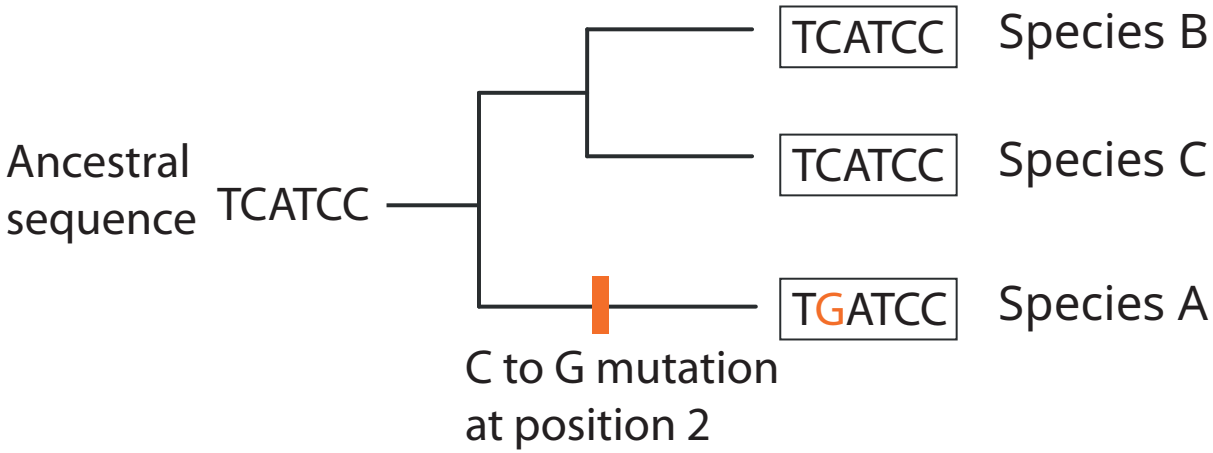
Convergence or homoplasy

Homoplasy: a trait that is found in two species, but not in their common ancestor.

The bluebird, Pterosaur (extinct) and fruit bat: 3 different vertebrates independently lightened bones and transformed hands into wings.

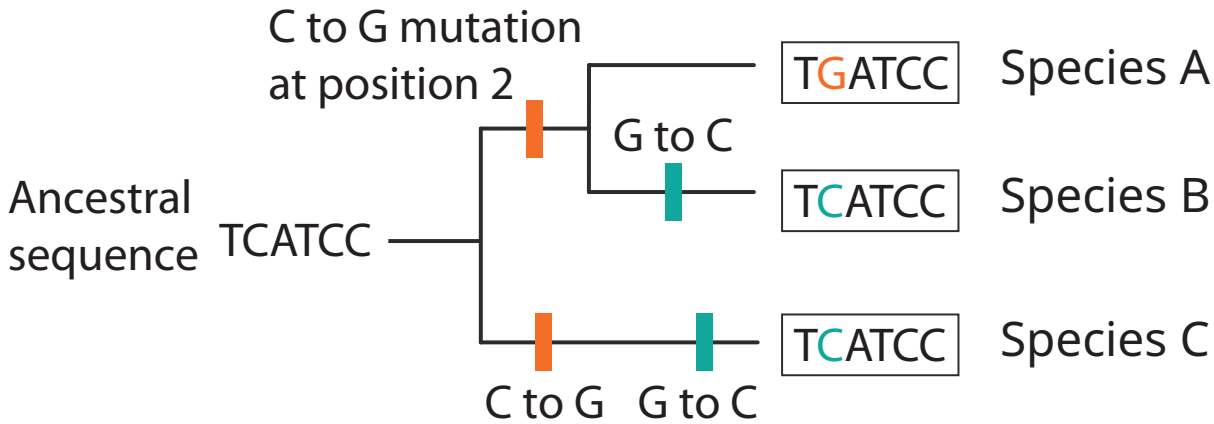
Image source: Convergent Evolution: an introduction

Molecular convergence



If we assume the simplest solution is correct, this could mislead our inference if the underlying process is more complex.

Molecular convergence



If we assume the simplest solution is correct, this could mislead our inference if the underlying process is more complex.

Parsimony

When we build a tree using parsimony and observe convergence, **ad hoc** explanations (e.g., convergence, reversals) are required to explain the patterns.

In the case of birds, pterosaurs and bats, we know based on other anatomical features that these taxa are distantly related, but convergence can also interfere with our ability to recover the correct tree. In fact, this is very common.

Parsimony

Parsimony has been demonstrated to be statistically inconsistent.

An estimator is consistent if it is guaranteed to get the correct answer with an infinite amount of data.

Felsenstein (1978) demonstrated that in some situations, parsimony is inconsistent, i.e., it will recover the wrong tree, even with an infinite amount of data.

Long branch attraction

If you have long branches (due to higher rates of evolution), the probability of misleading parsimony due to convergence is much higher.

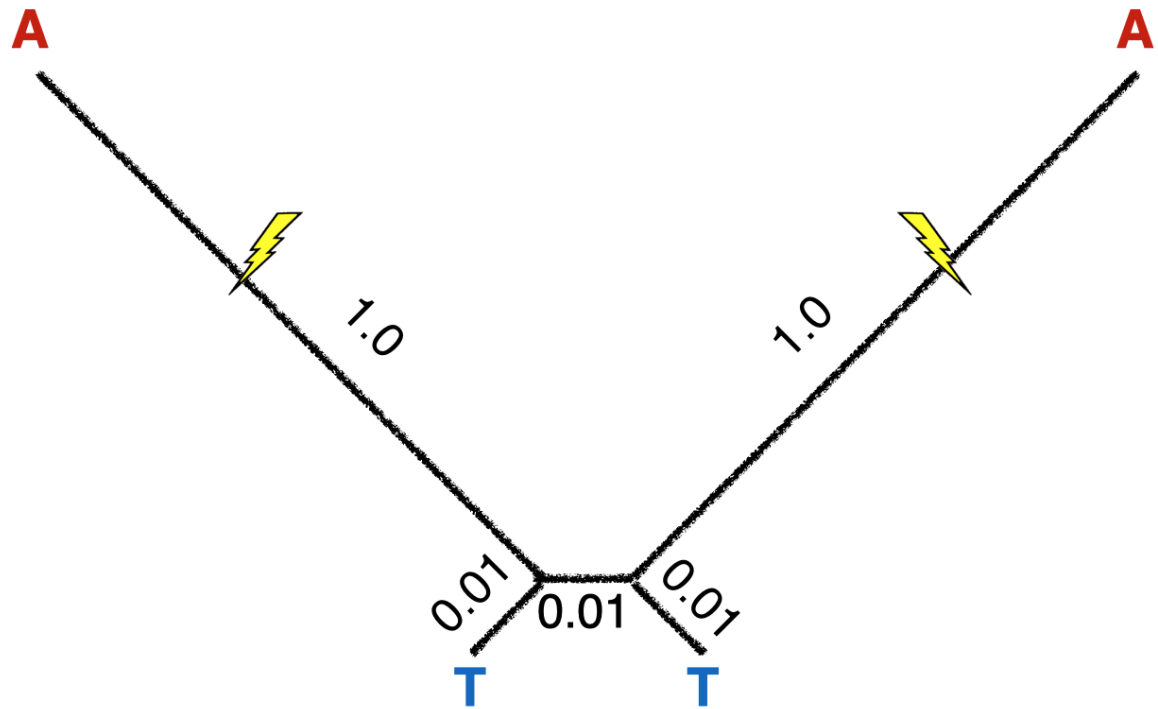


Image source: Tracy Heath

Long branch attraction

Parsimony is almost guaranteed to get the tree below wrong. It will incorrectly place two long branches (T1,T3) together as sister lineages.

More data will make the problem worse, making this approach statistically inconsistent.

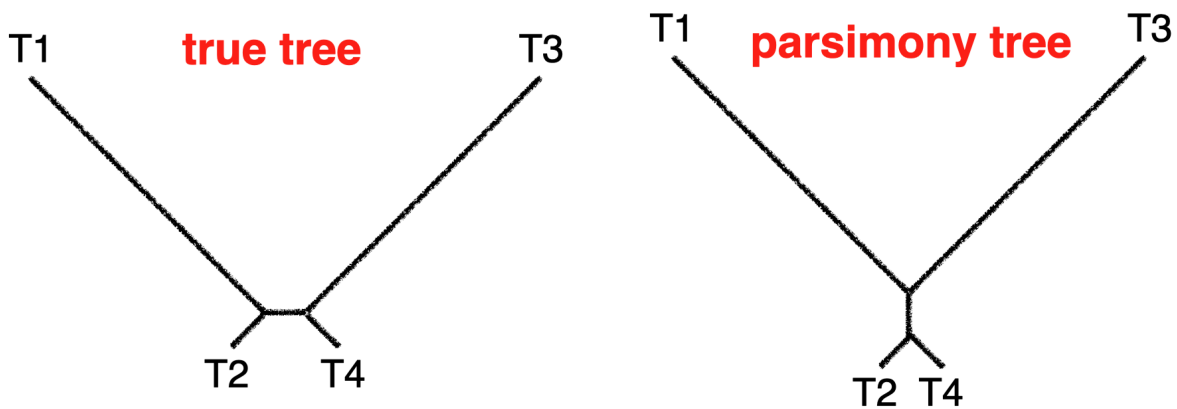
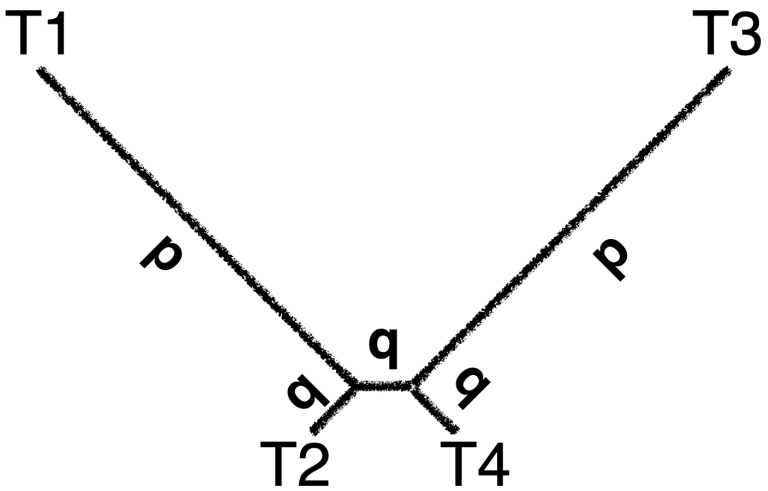
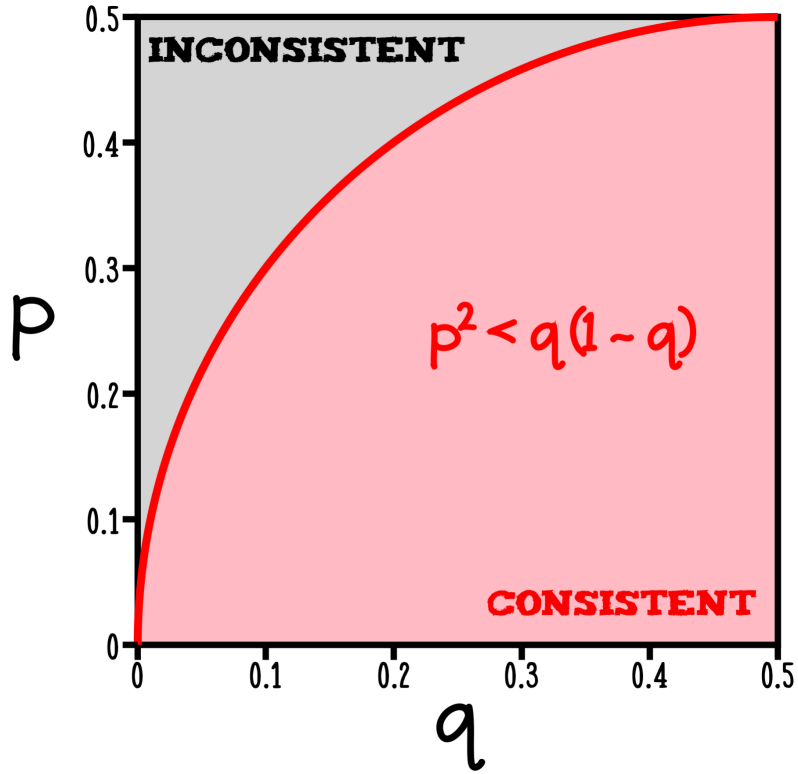


Image source: Tracy Heath

Long branch attraction

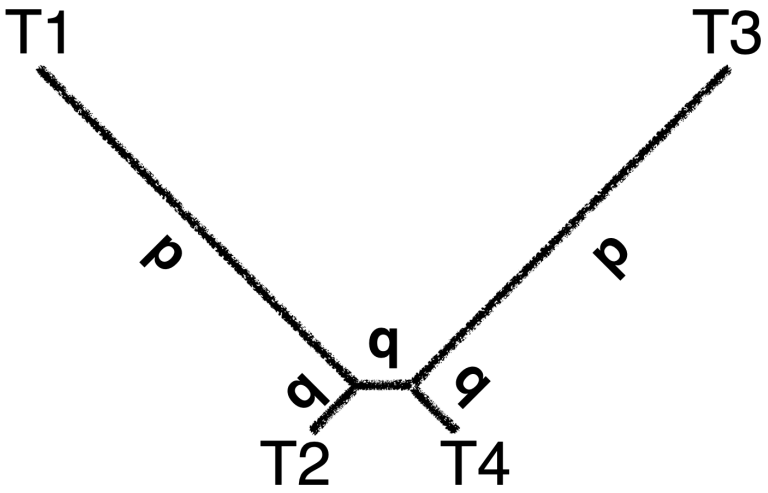


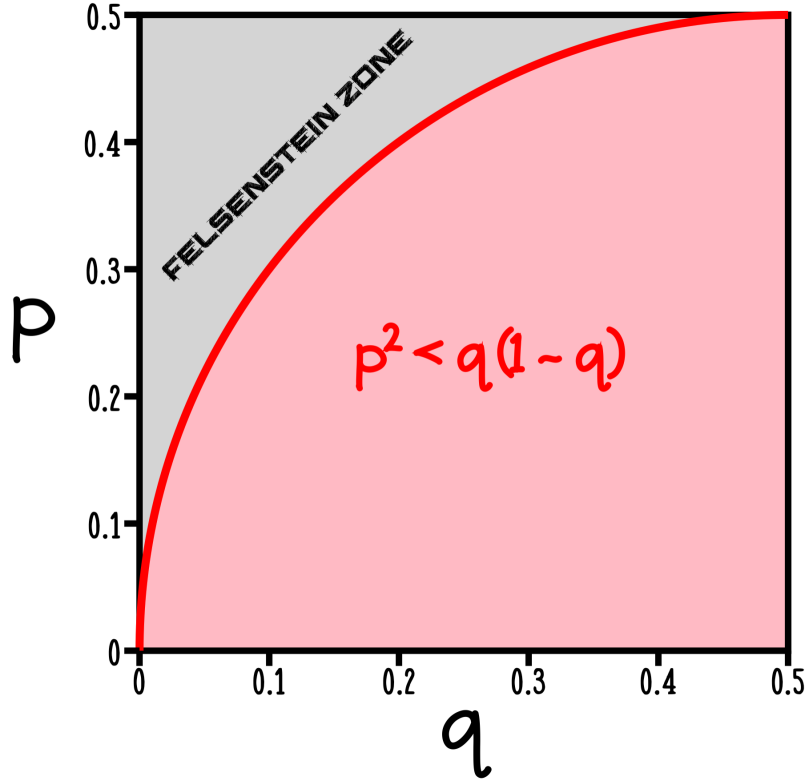
Here, the branch lengths represent probability (p, q) of change along that branch.



Felsenstein, *Inferring Phylogenies*, (2004), Image source: Tracy Heath

Long branch attraction



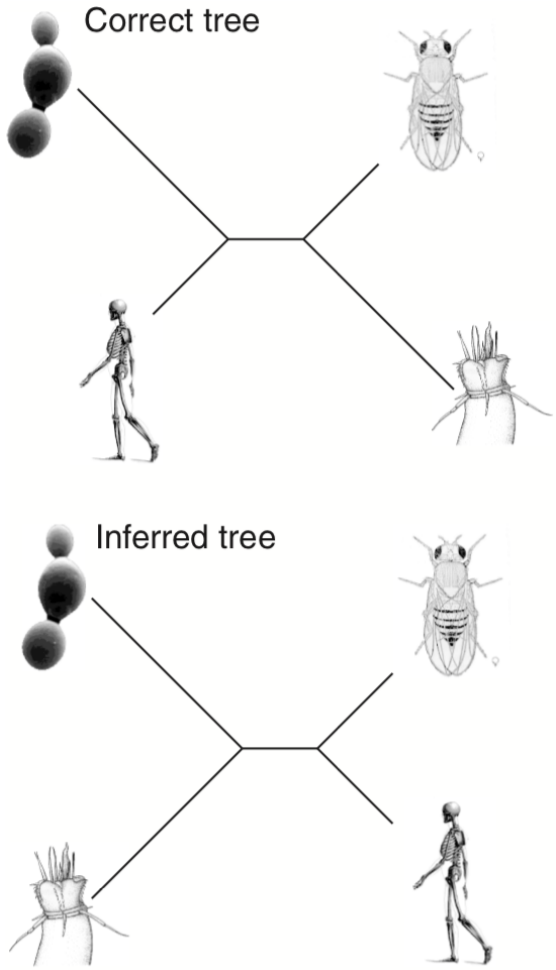


The area shaded in grey, is the area of parameter space where you are practically guaranteed to recover the wrong tree, with increasing certainty the more data you have.

Felsenstein, *Inferring Phylogenies*, (2004), Image source: Tracy Heath

A classic case of long branch attraction

The relationship between nematodes, arthropods and chordates was misunderstood for a long time.



- outgroup = yeast
- Ecdysozoa
  arthropods + nematodes,
  ex. vertebrates*
- Coelomata
  arthropods + vertebrates,
  ex. nematodes

*widely accepted today, Image: Telford et al. (2005) *Curent Biology*

A classic case of long branch attraction

The branch leading nematodes is long, reflecting high rates of evolution along this lineage, relative to other animals.

Because the outgroup used to root the tree inevitably has a long branch, it can incorrectly 'attract' long branching species, such as nematodes, towards the base of the tree.

Important note: this issue can affect all tree building methods! And all types of data (e.g., DNA, morphology).

Things that help: (sometimes) high quality data, increased taxon sampling inc. shorter branching outgroups, models that more reliably capture the variation in evolutionary rates.

Felsenstein (1978) *Systematic Zoology*, Telford et al. (2005) *Curent Biology*

Parsimony: advantages and disadvantages

The greatest advantage of parsimony is its beautiful simplicity

Computationally fast

Often produces sensible results

Parsimony: advantages and disadvantages

The greatest advantage of parsimony is its beautiful simplicity

Computationally fast

Often produces sensible results

Some argue that parsimony is assumption free

Others argue parsimony *does* make assumptions, even if we don't know what they are

Yang (2014) Molecular Evolution: A Statistical Approach

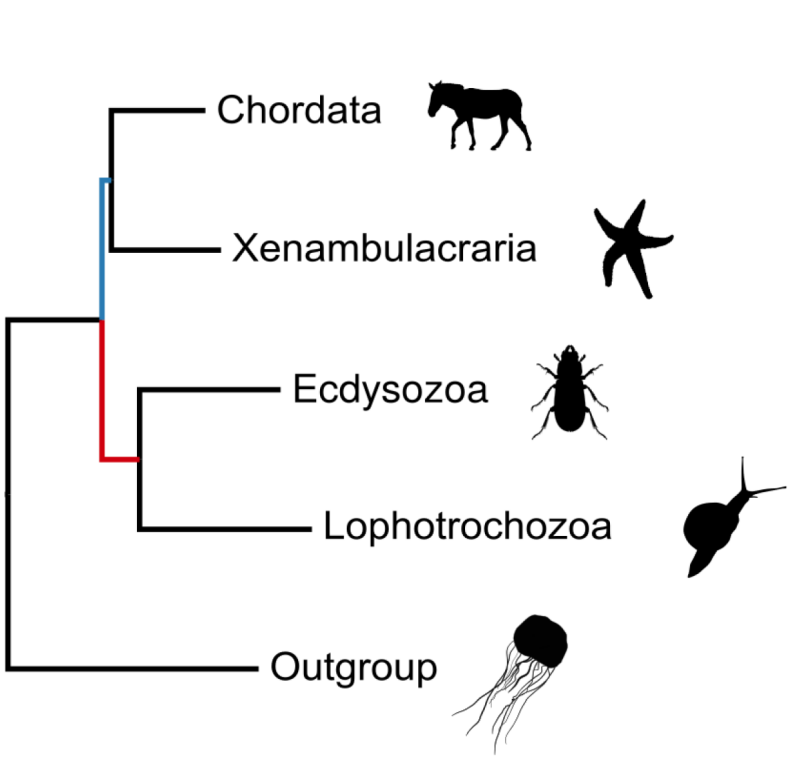
Parsimony vs. model-based approaches

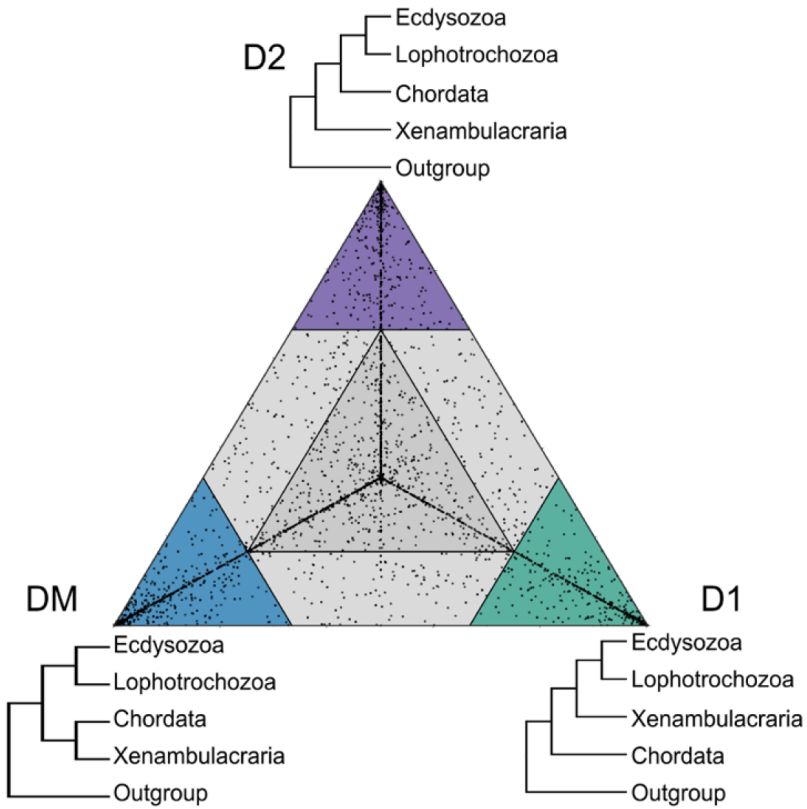
Model-based approaches assume an explicit model of molecular or morphological evolution.

If evolutionary distance is relatively small, model based approaches and parsimony will often recover the same tree.

As distance increases, the amount of homoplasy (i.e., convergent or parallel changes) also increases, parsimony is more likely to recover the wrong tree.

Important note: short internal branches pose a huge challenge for any approach to tree building





Kapli et al. (2021) *Science Advances* – support for deuterostomes (chordates + echinoderms) varies across datasets and analyses under different models, probably caused by the extremely short (blue) branch associated with this group.

Take homes

Parsimony is simple and intuitive but makes **implicit** assumptions about the evolutionary process.

Next, we'll explore model-based approaches – these are more flexible and make **explicit** assumptions → it's very important to try understanding what these are!

Suggested listening / reading

Check out this fascinating interview with Joseph Felsenstein by Mary Kuhner. Joe played an important role in establishing the field of statistical phylogenetics.

Kapli et al. (2021) Lack of support for Deuterostomia prompts reinterpretation of the first Bilateria. *Science Advances*.

Quick demo – tree building using parsimony in R.

End of Part 1