

Phylogenetics

Course introduction

RL-V3 MPP

Rachel Warnock

16.04.24



About this course

Thursdays 14:00–16:00 CET at Henke Str.

8 weeks of lectures + 6 weeks for project work

Classes will consist of lectures and practical exercises

All lecture material available via the [course website](#)

Blue underlined text → [external links](#)

Course objectives

To learn the **application** of phylogenetics in paleobiology

- Tree building
- Substitution models
- Dating trees
- Clock models
- Tree models
- Diversification rates
- Morphological models
- Continuous trait evolution

Course evaluation

“**Phylogenetics**” is graded together with “**Introduction to Statistical Modelling**”
(course code: **RL-V3 MPP**)

Evaluation is based on a **written report** (info available on the [project page](#))

Class exercises will be in [R](#) or the Bayesian phylogenetic software [RevBayes](#)

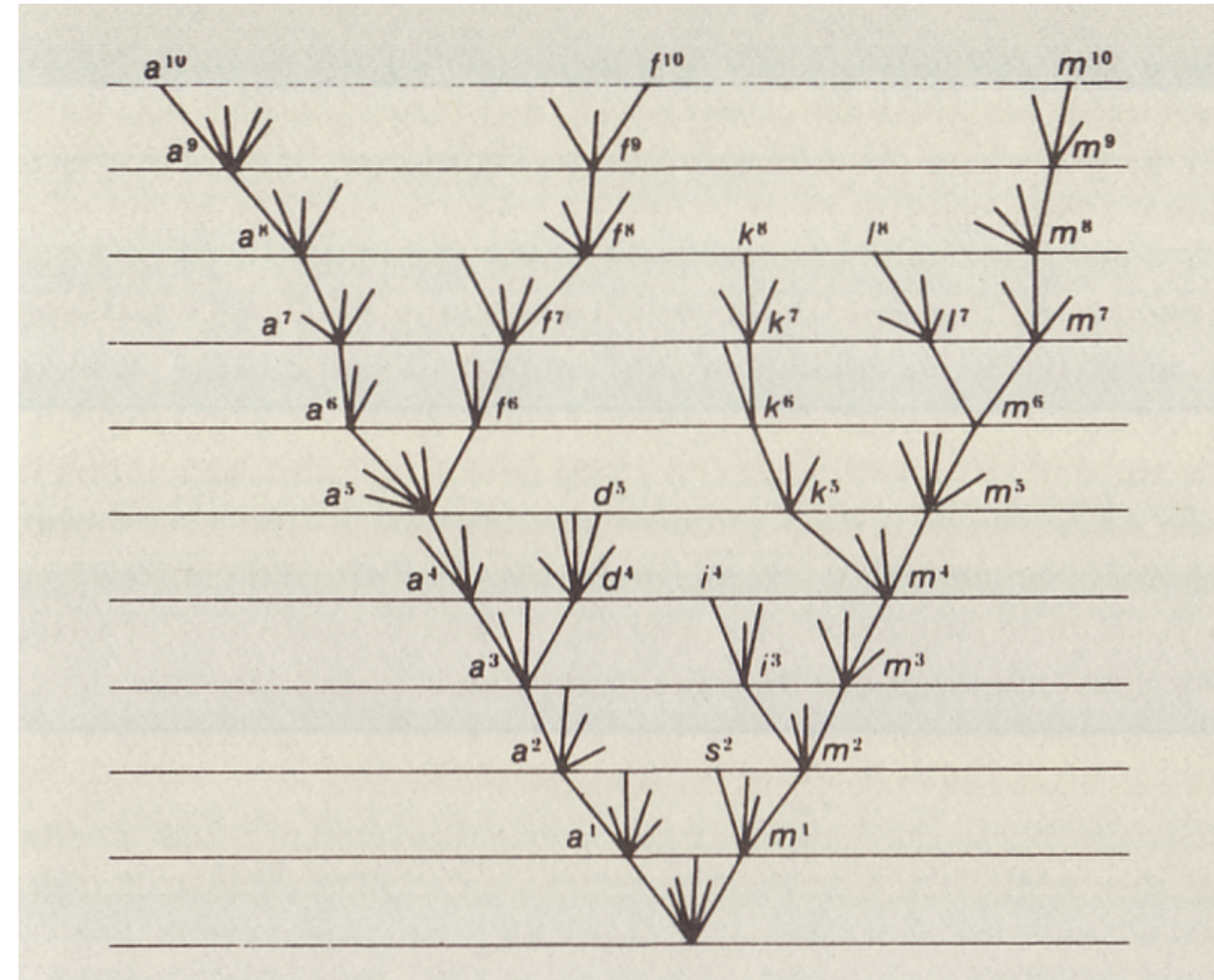
In addition, we have homework exercises that include videos and reading



Please ask questions!

Today's objectives

- Recap 'tree-thinking'
- Gain an understanding of the **parsimony** approach to tree-building and **statistical inconsistency**



Time tree from Darwin's *Origin of Species*

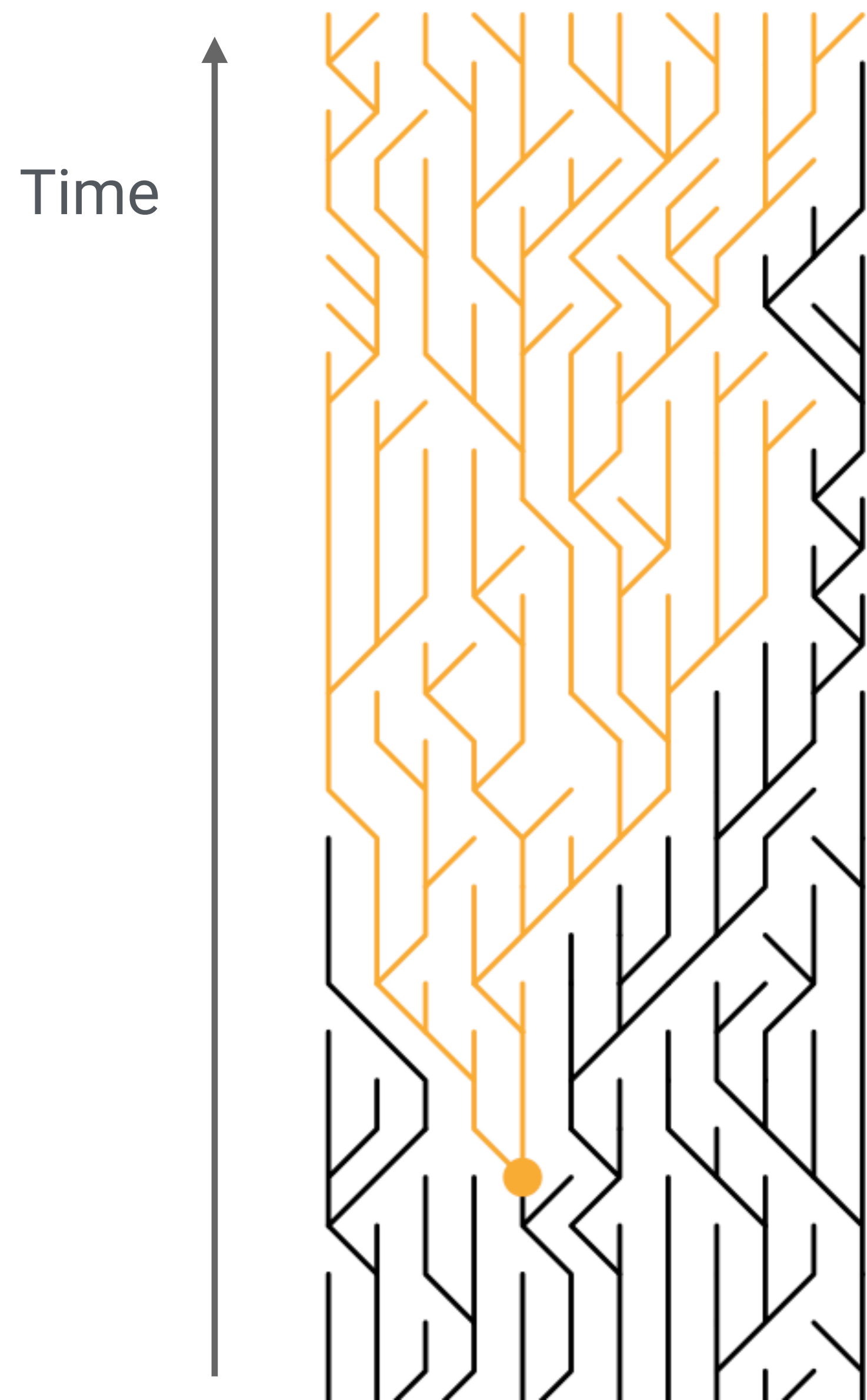
What is phylogenetics?

Phylogenetics

Phylogenetics aims to reconstruct the phylogeny of individual samples based on molecular or morphological character data

A phylogeny captures part of evolutionary history that is otherwise not directly observable

Phylodynamics aims to quantify the processes that gave rise to the tree, e.g., speciation, extinction

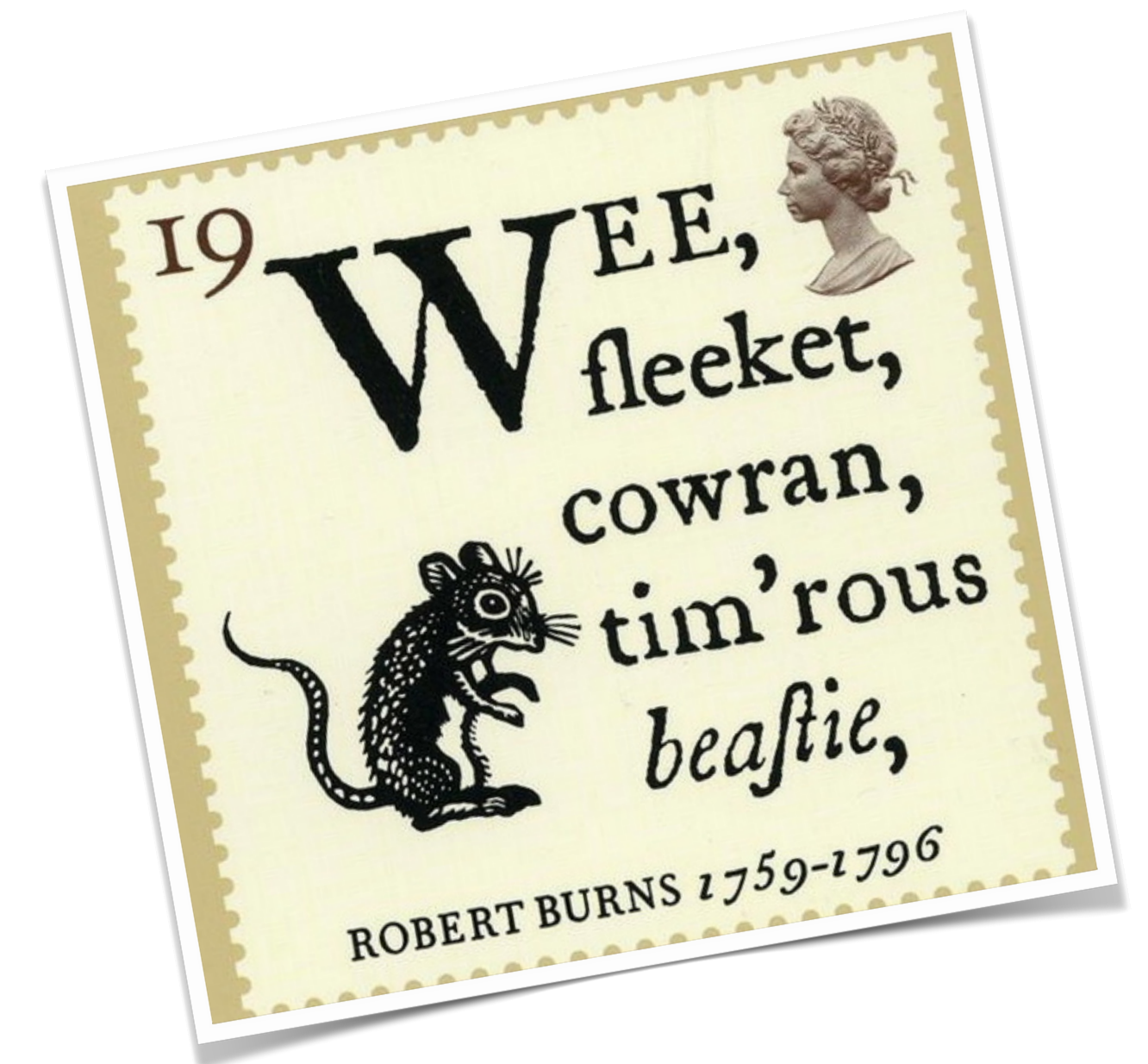


- populations
- species
- viruses
- cells
- languages

Data

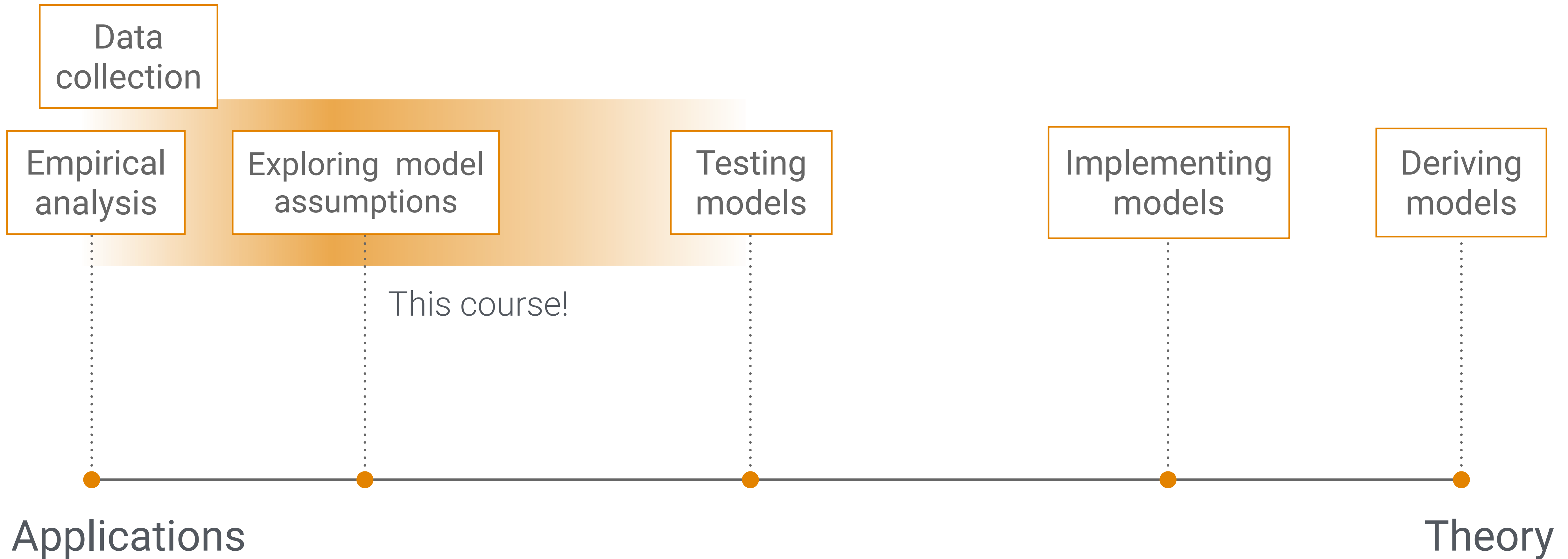
- DNA
- morphology
- words

In this course we mainly focus on trees that include **one representative per species**



[Scots poem](#) - also the [BEAST2](#) logo!

Research topics in phylogenetics

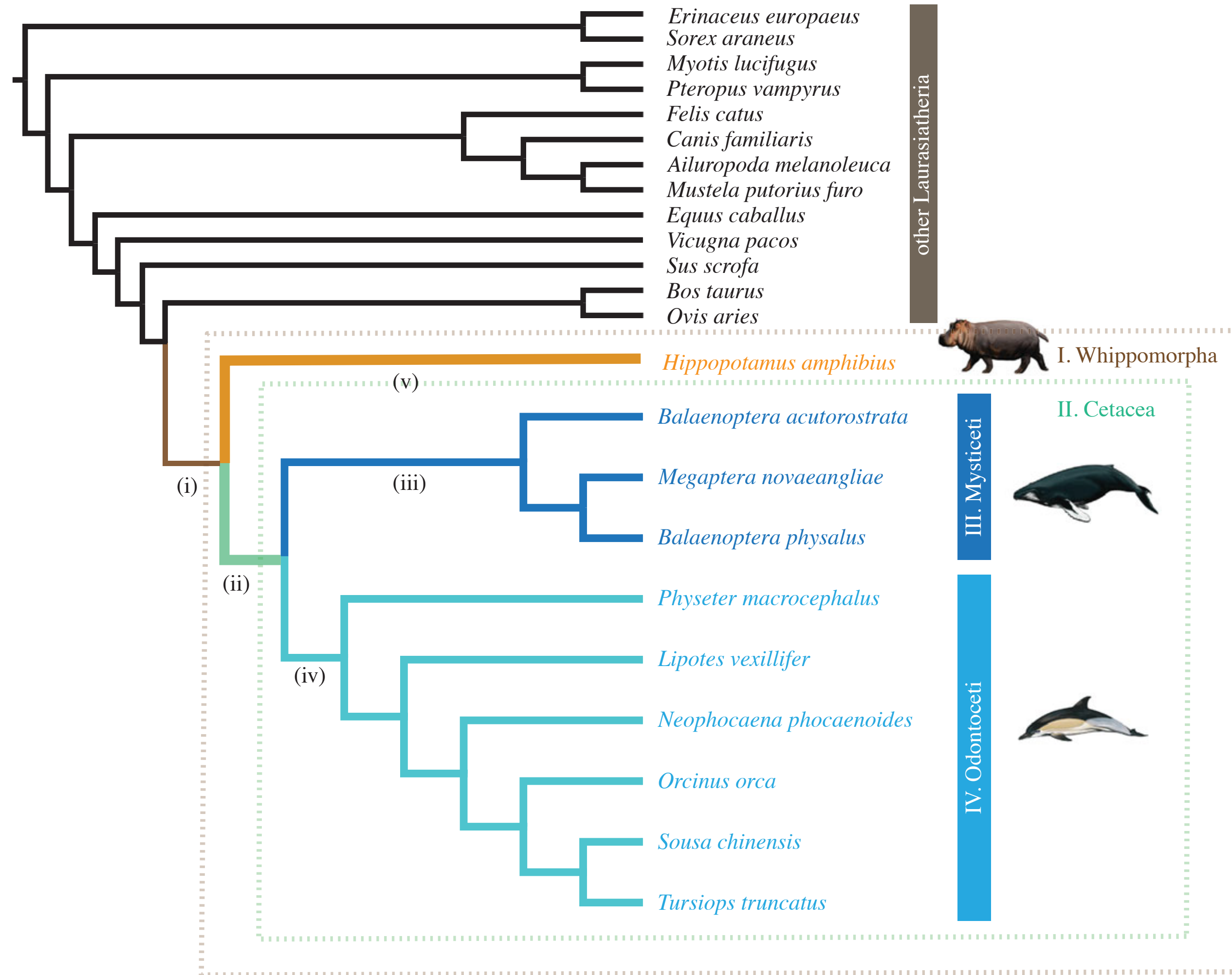


Nothing in biology makes sense except in the light of evolution – Theodosius Dobzhansky ([1973](#))

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

Trees in paleobiology

What can we learn from trees?

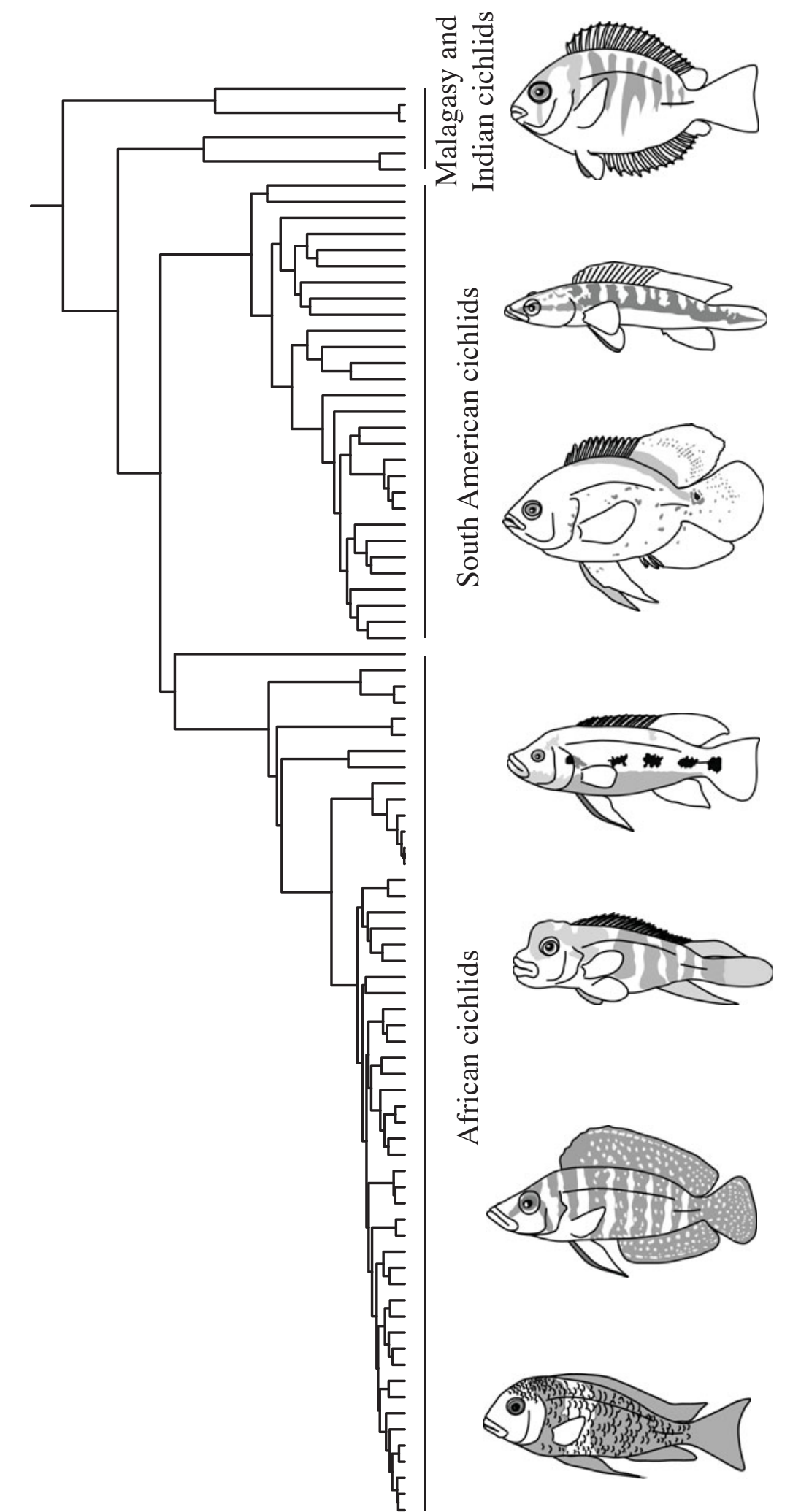


How are our favourite species related?

Does the phylogeny support the taxonomy?

What can we learn from trees?

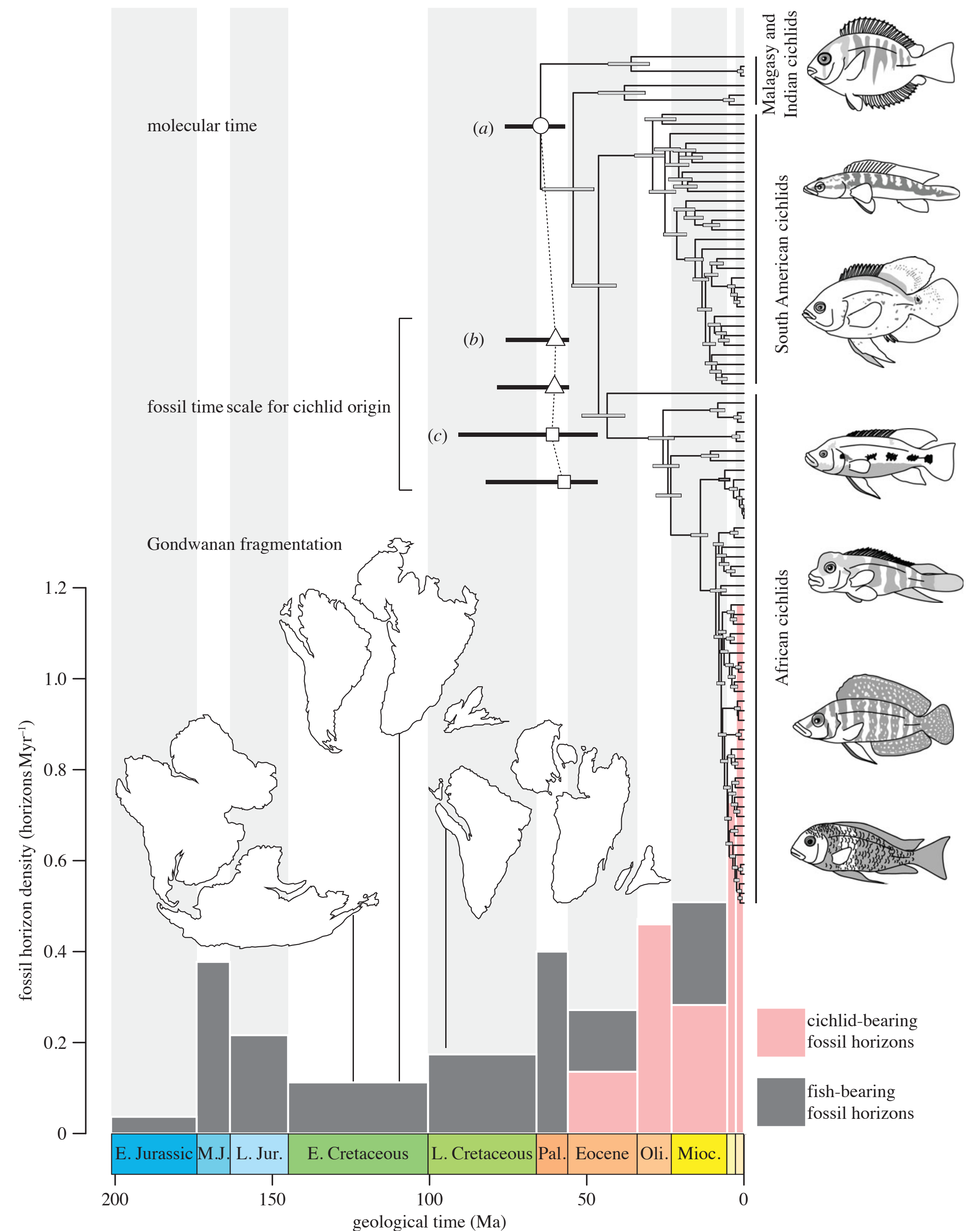
- Evolutionary relationships



What can we learn from trees?

- Evolutionary relationships
- Timing of diversification events
- Geological context
- Rates of phenotypic evolution
- Diversification rates

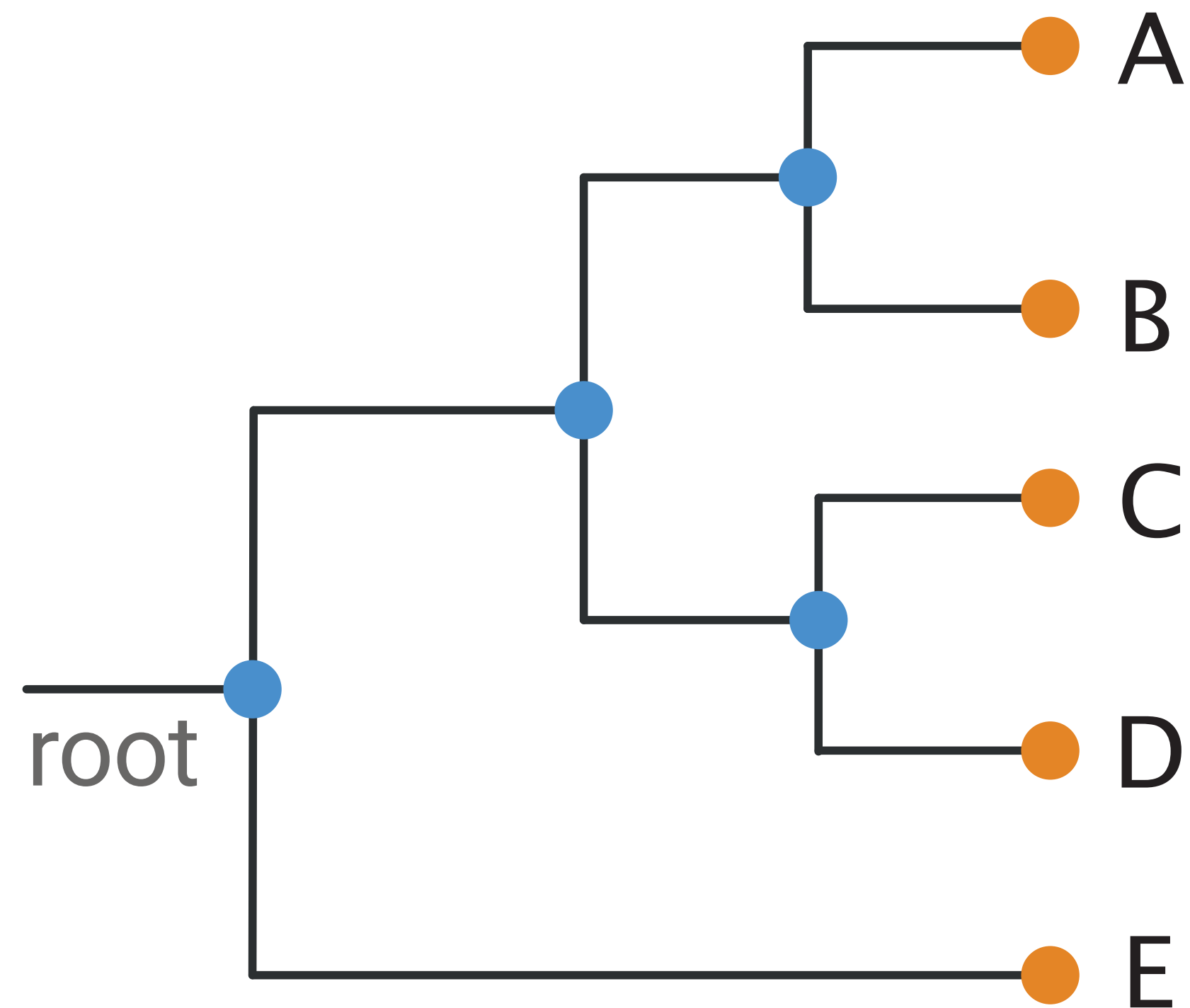
Image adapted from *Friedmann et al.* ([2013](#))



Where do we begin?

Some basic terms

MRCA = most recent common ancestor



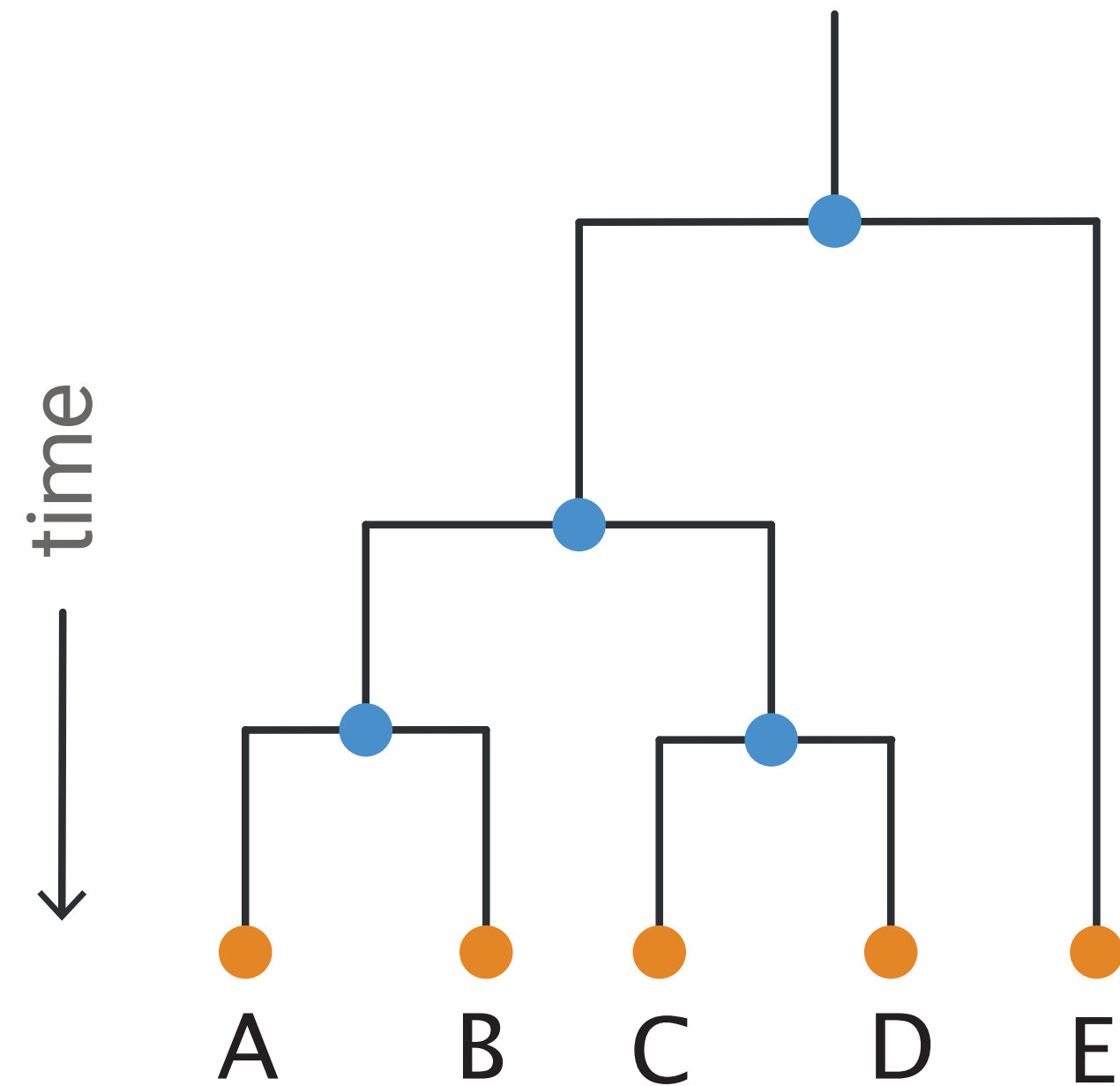
- internal nodes or MRCAs
- tips or leaves
- branches or edges

branch lengths = genetic distance or time

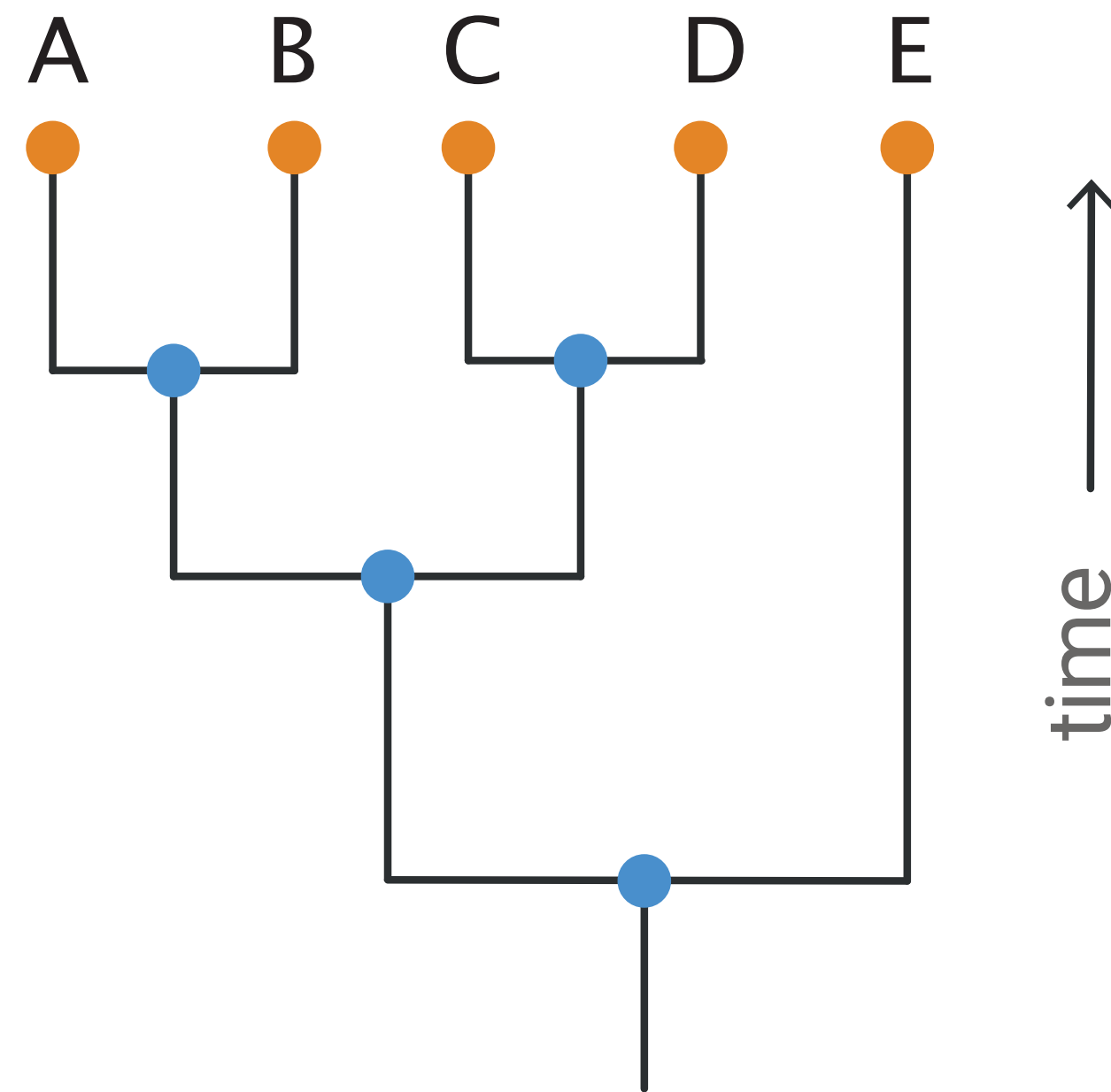
[How to read a phylogenetic tree](#)

The direction of time

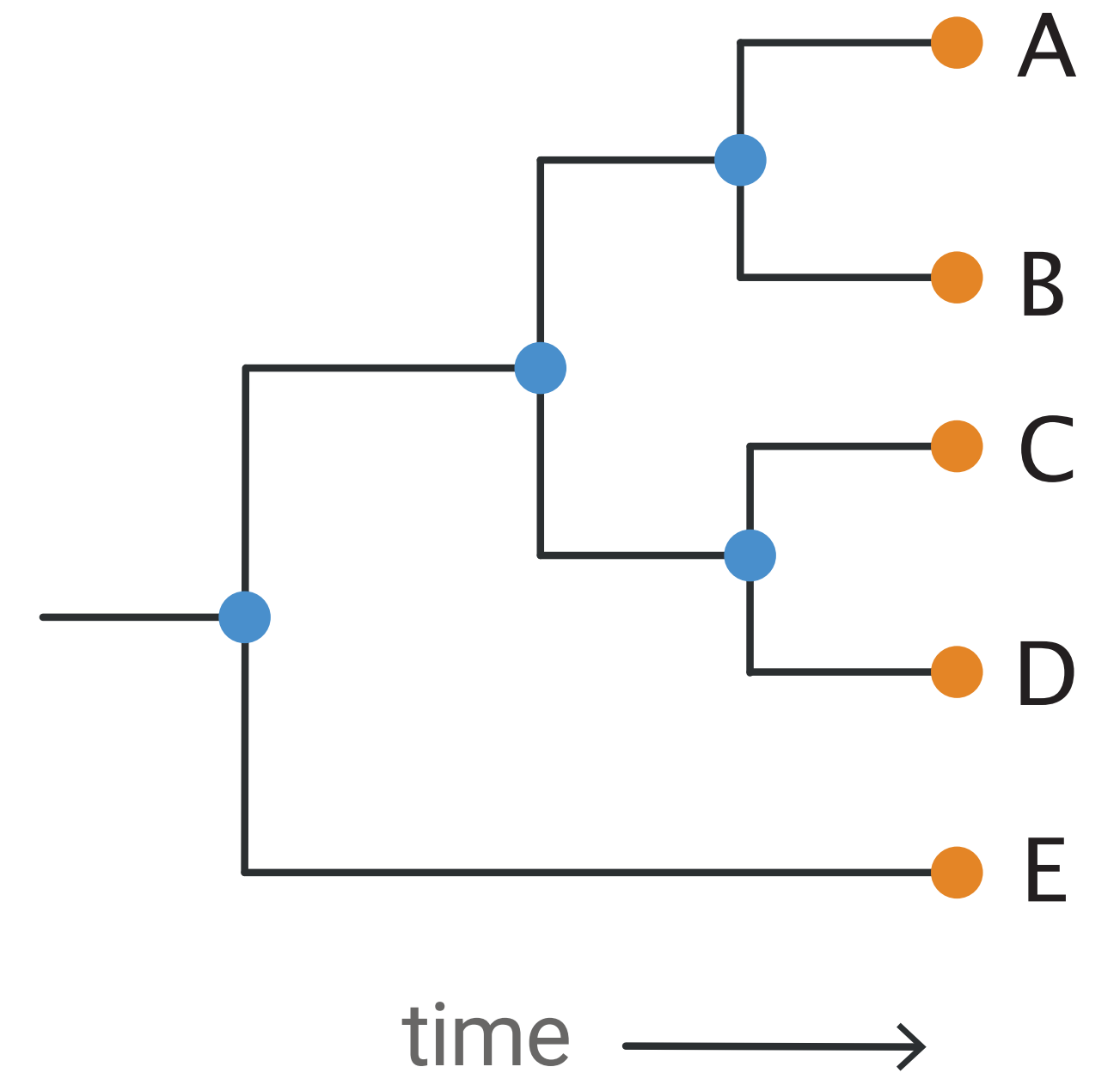
Tip look for the root!



Computer science, maths



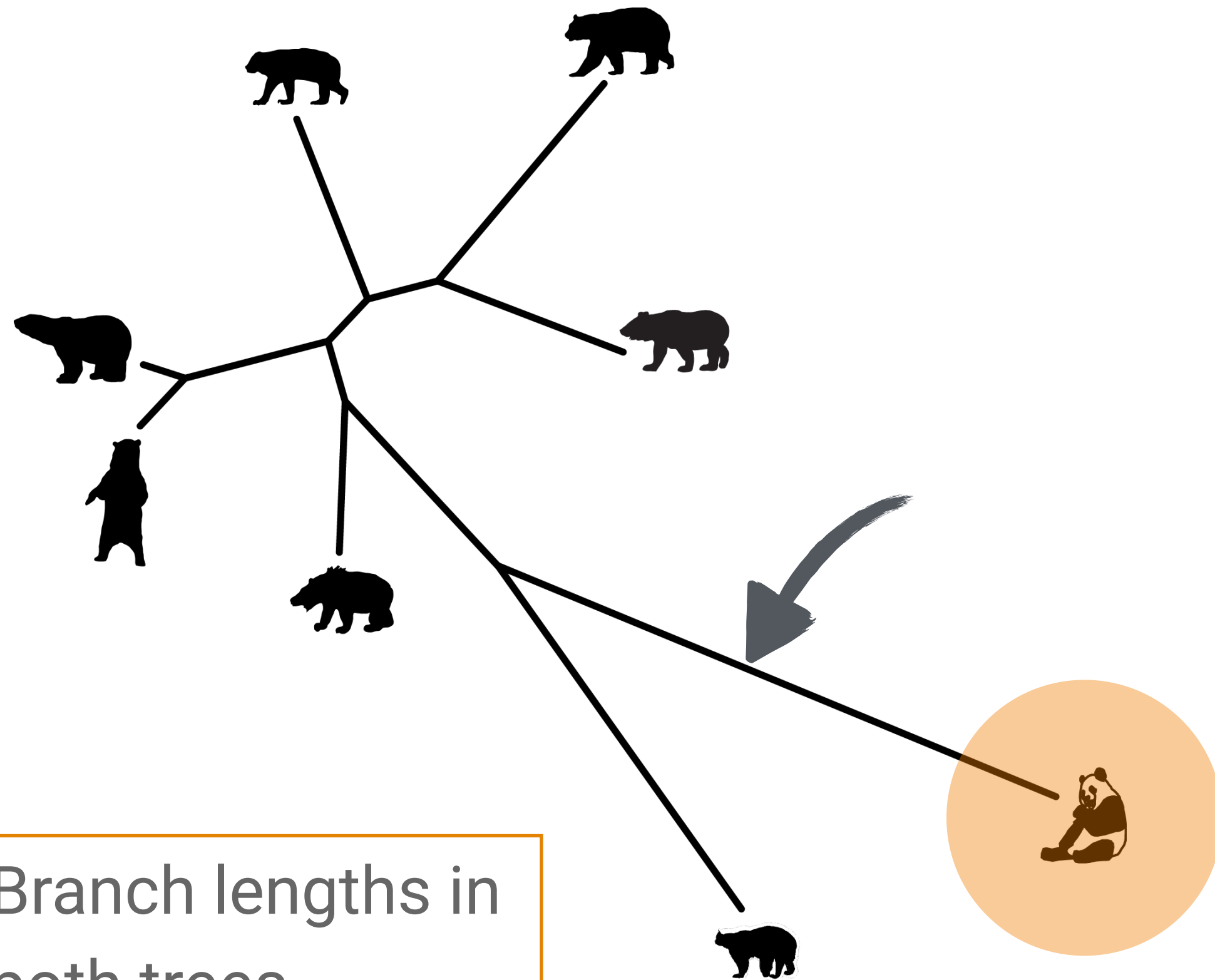
Geology



Evolutionary biology

Two types of trees

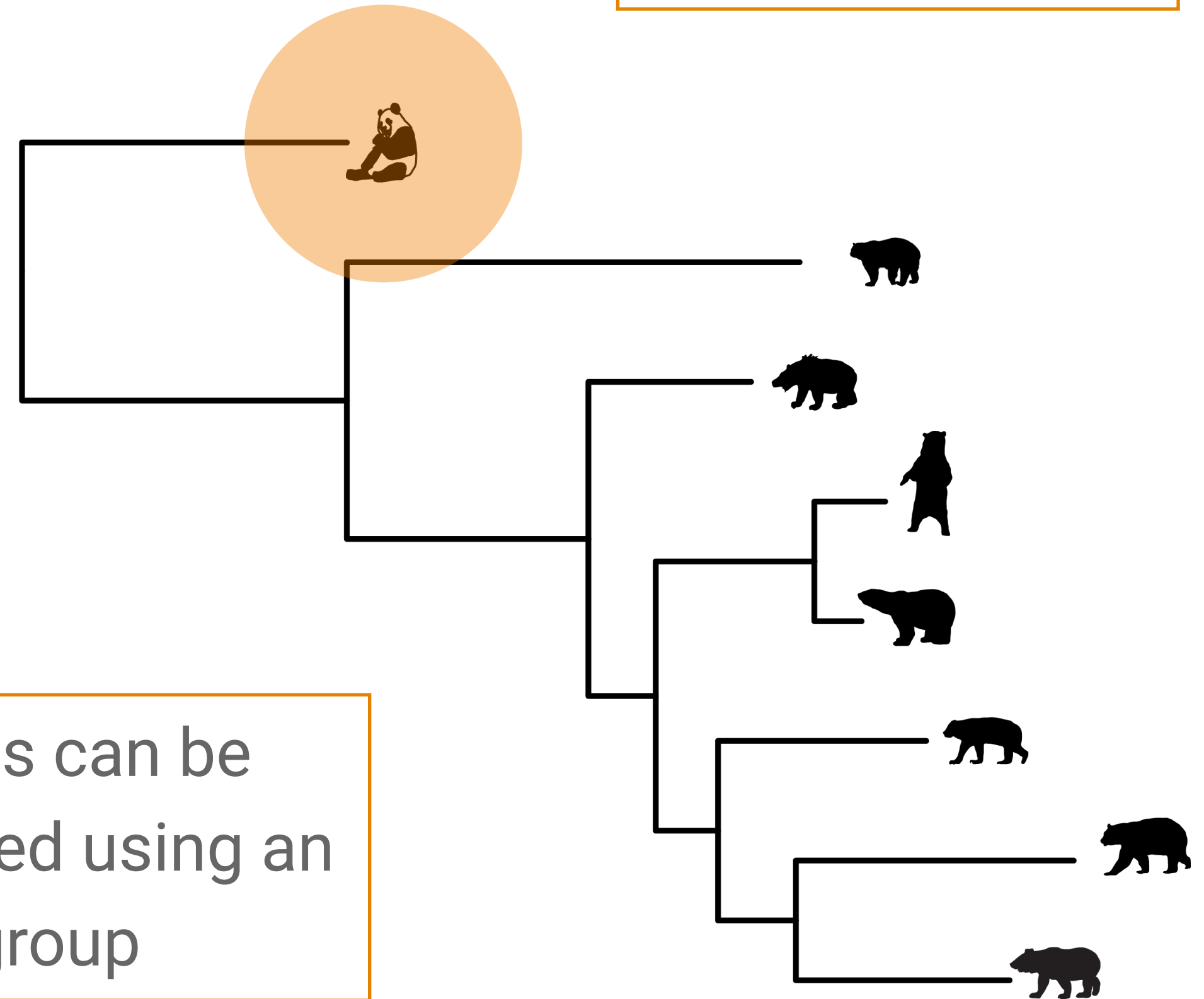
Unrooted vs. rooted trees



Branch lengths in both trees represents genetic distance

A rooted tree shows the direction of time

Trees can be rooted using an outgroup



Character	<u>Lamprey</u>	<u>Antelope</u>	<u>Bald eagle</u>	<u>Alligator</u>	<u>Sea bass</u>
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 **rooted** tree should be?
Write down your logic
- How many possible unrooted or rooted trees are there?

'0' and '1' represent **absence** or **presence**

There are a **huge** number of possible trees!

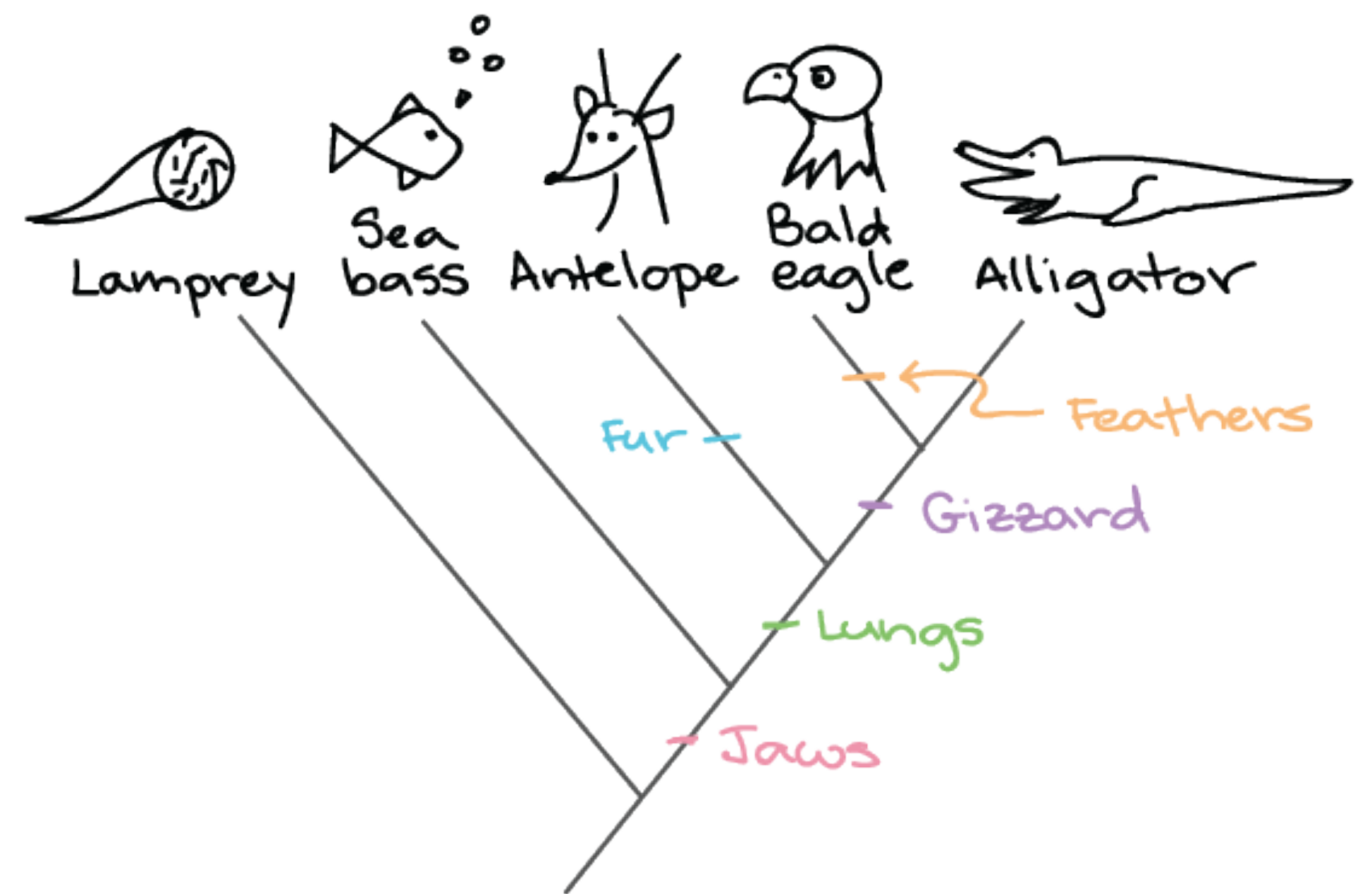
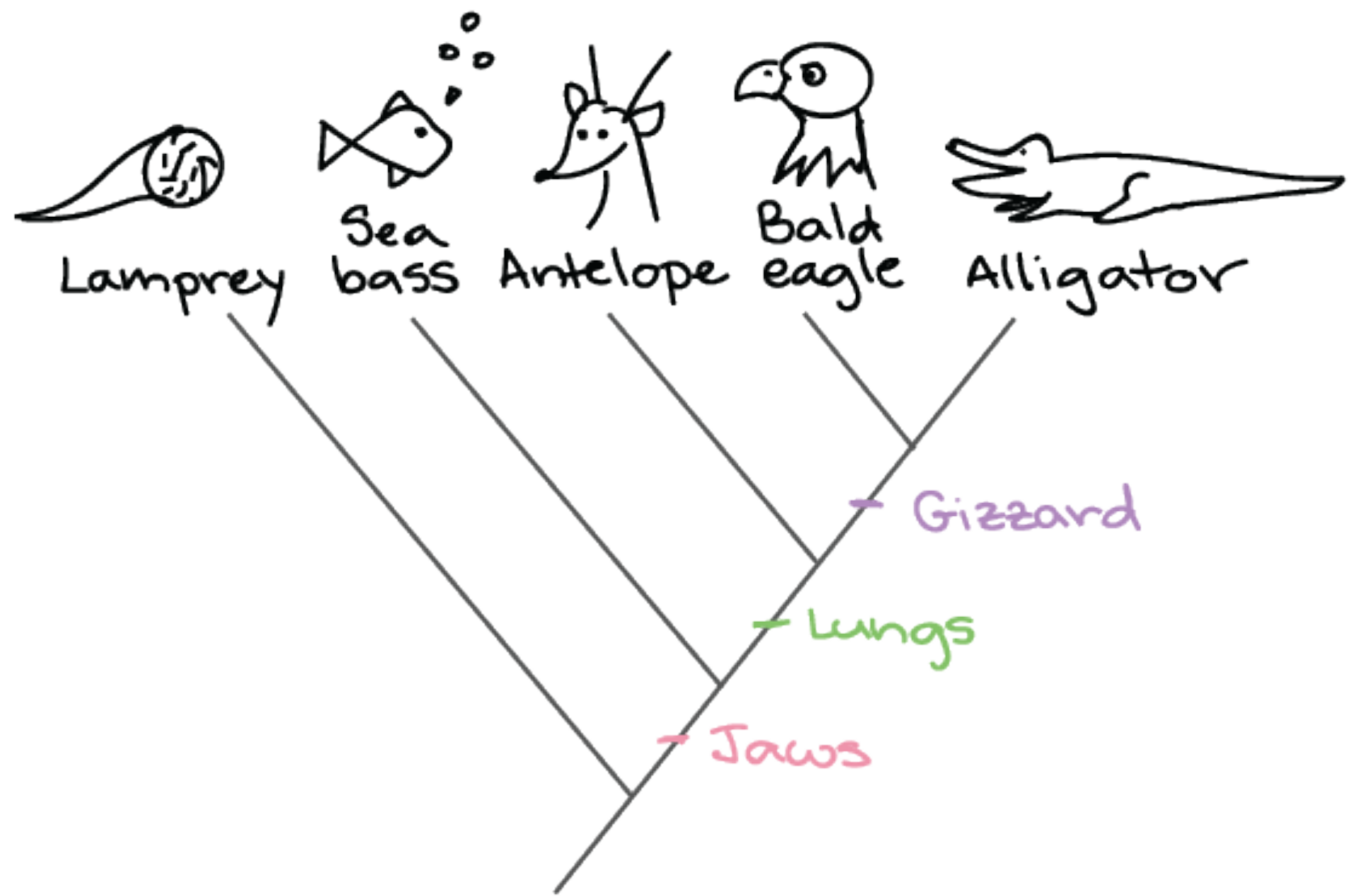
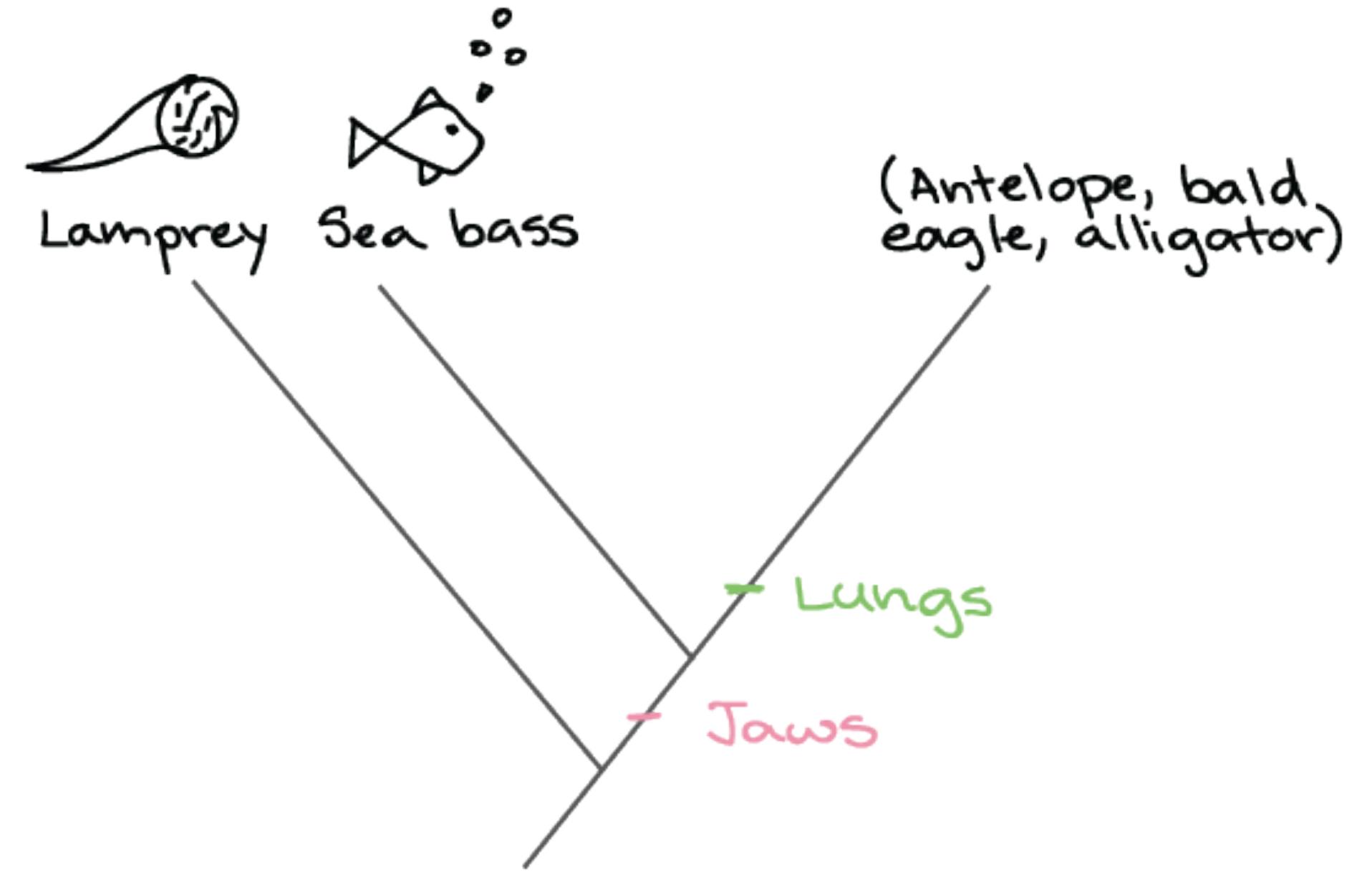
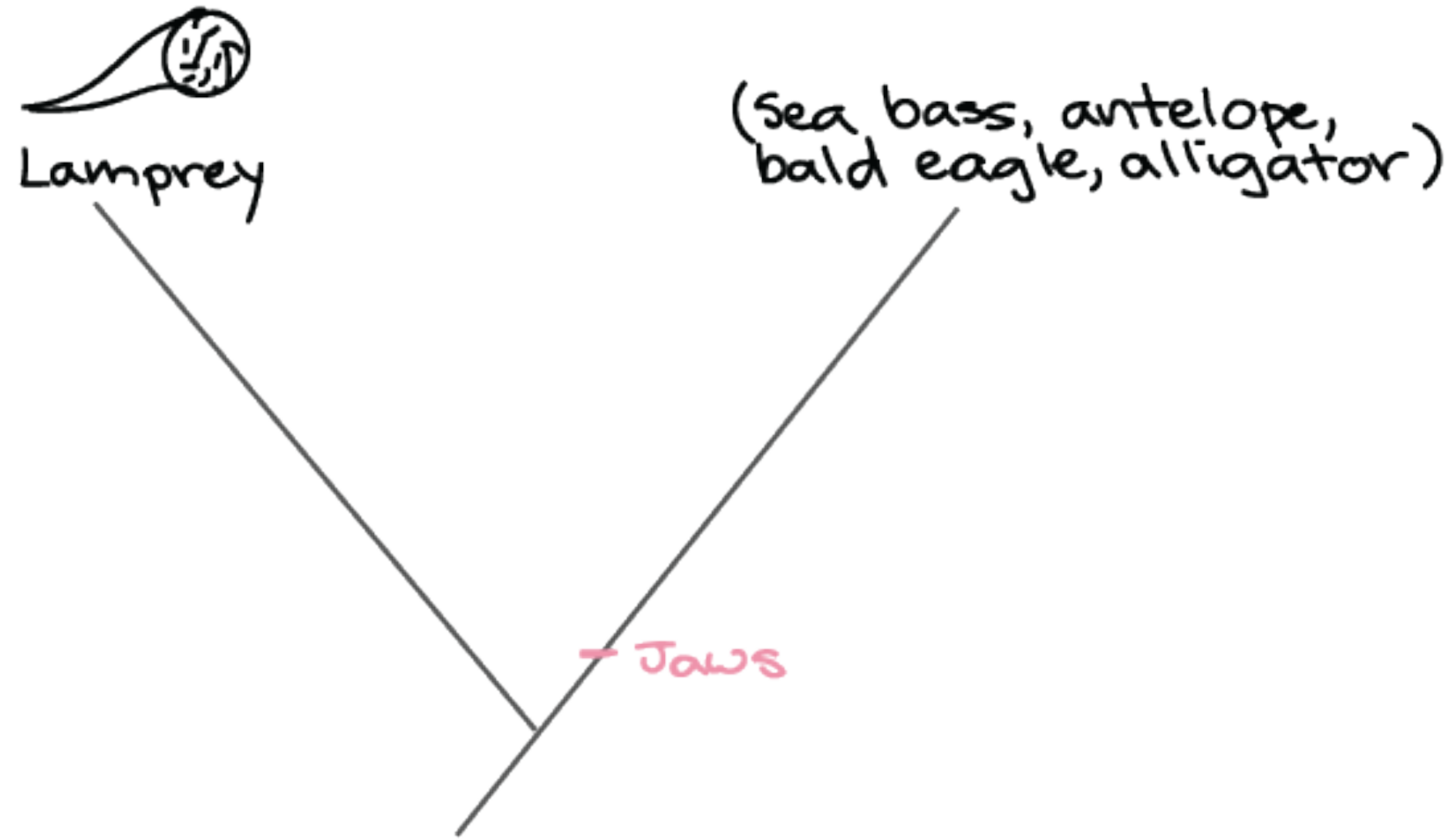
n tips	unrooted trees	rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

Number of **branches**

unrooted tree
 $2n-3$

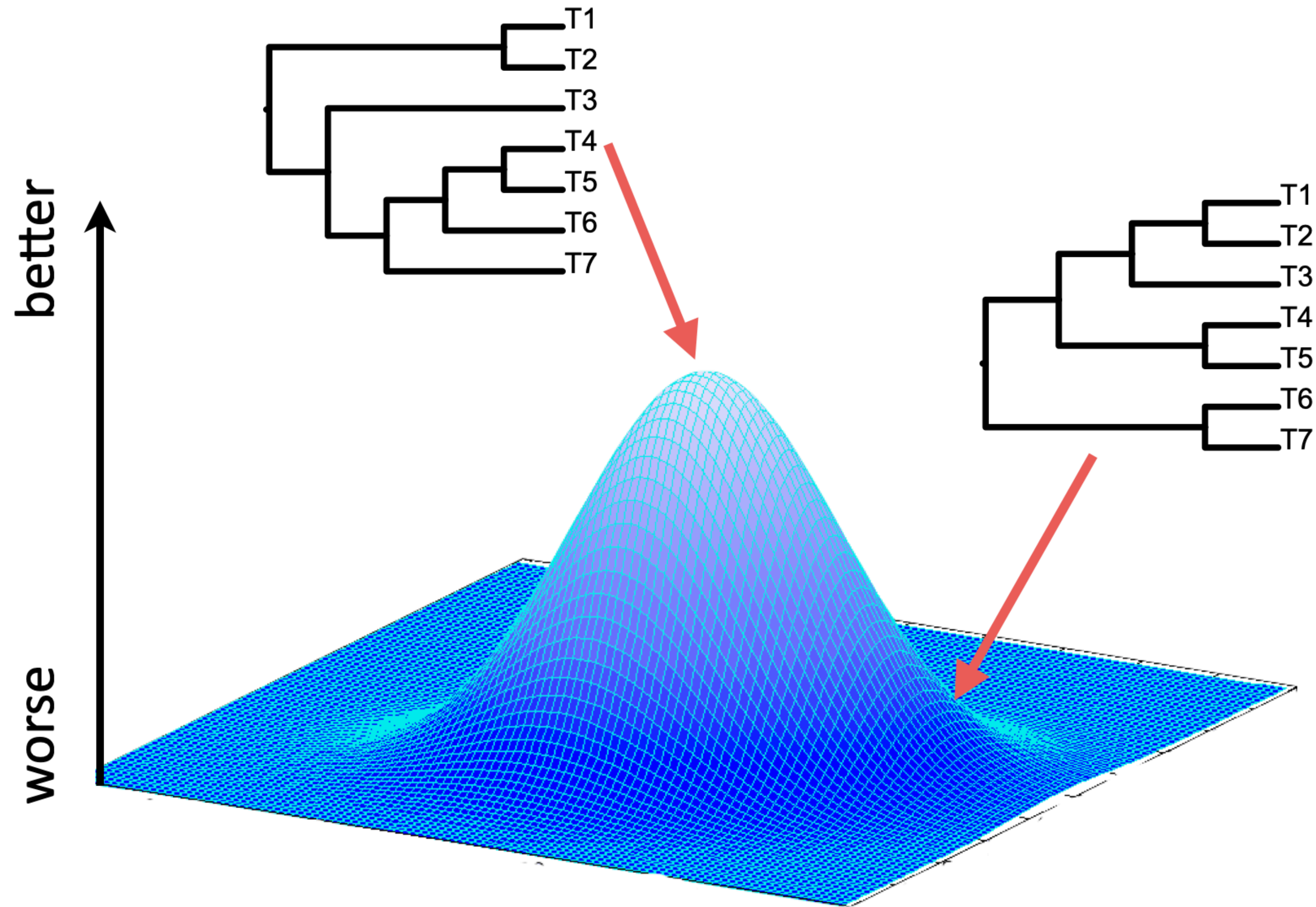
rooted tree
 $2n-2$

See wiki for more on [where these numbers come from](#)



- *Write down your logic*
 - Most people intuitively assume the tree with the **fewest** changes is correct
 - This approach to tree building is called **maximum parsimony**

How do we find the 'best' tree?



It depends how you measure 'best'

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

Both maximum likelihood and Bayesian inference are model-based approaches

Note these are not the only approaches to tree-building but they are the most widely used

Maximum parsimony

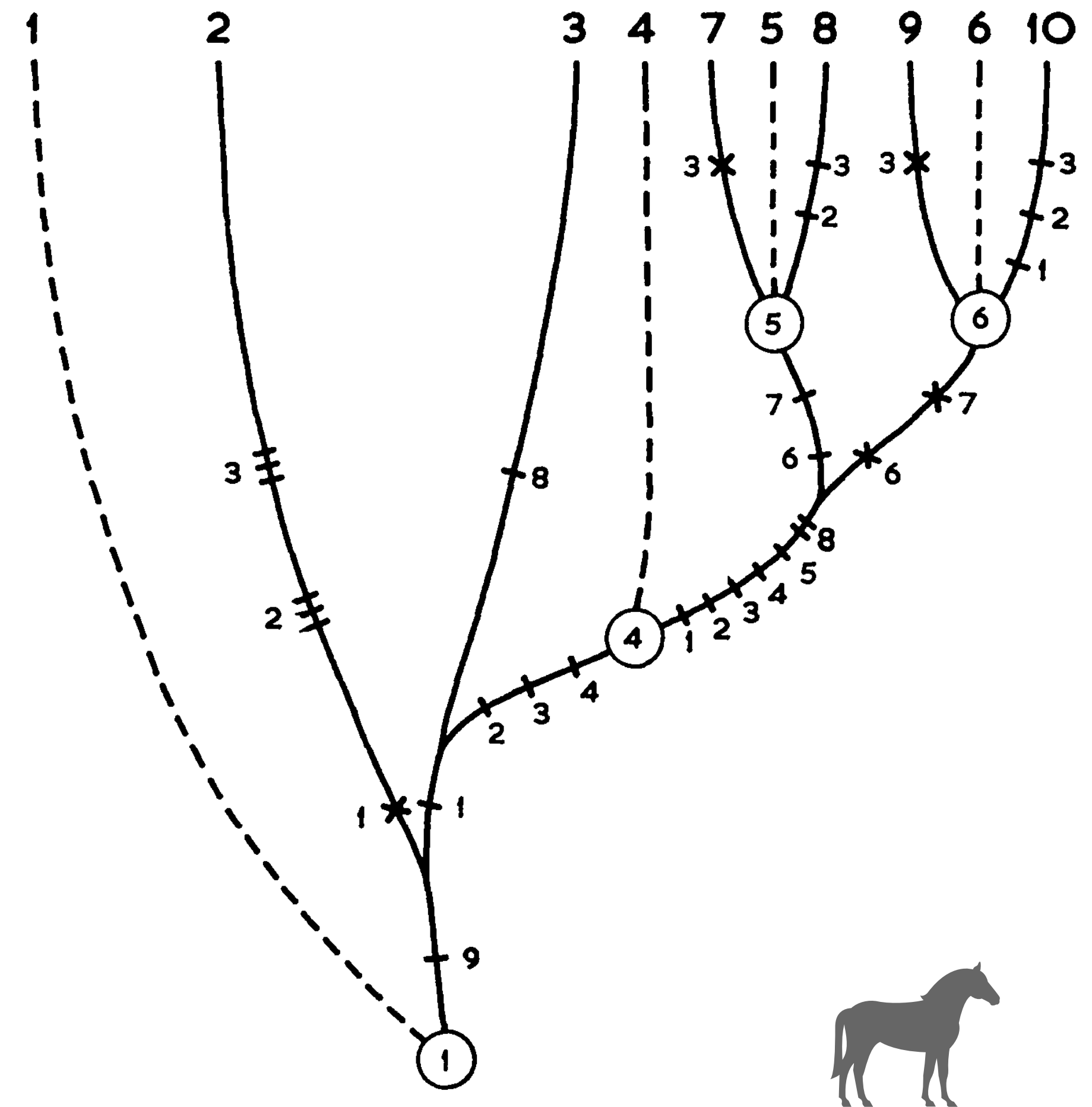
(also sometimes known as the minimum evolution method)

Maximum parsimony

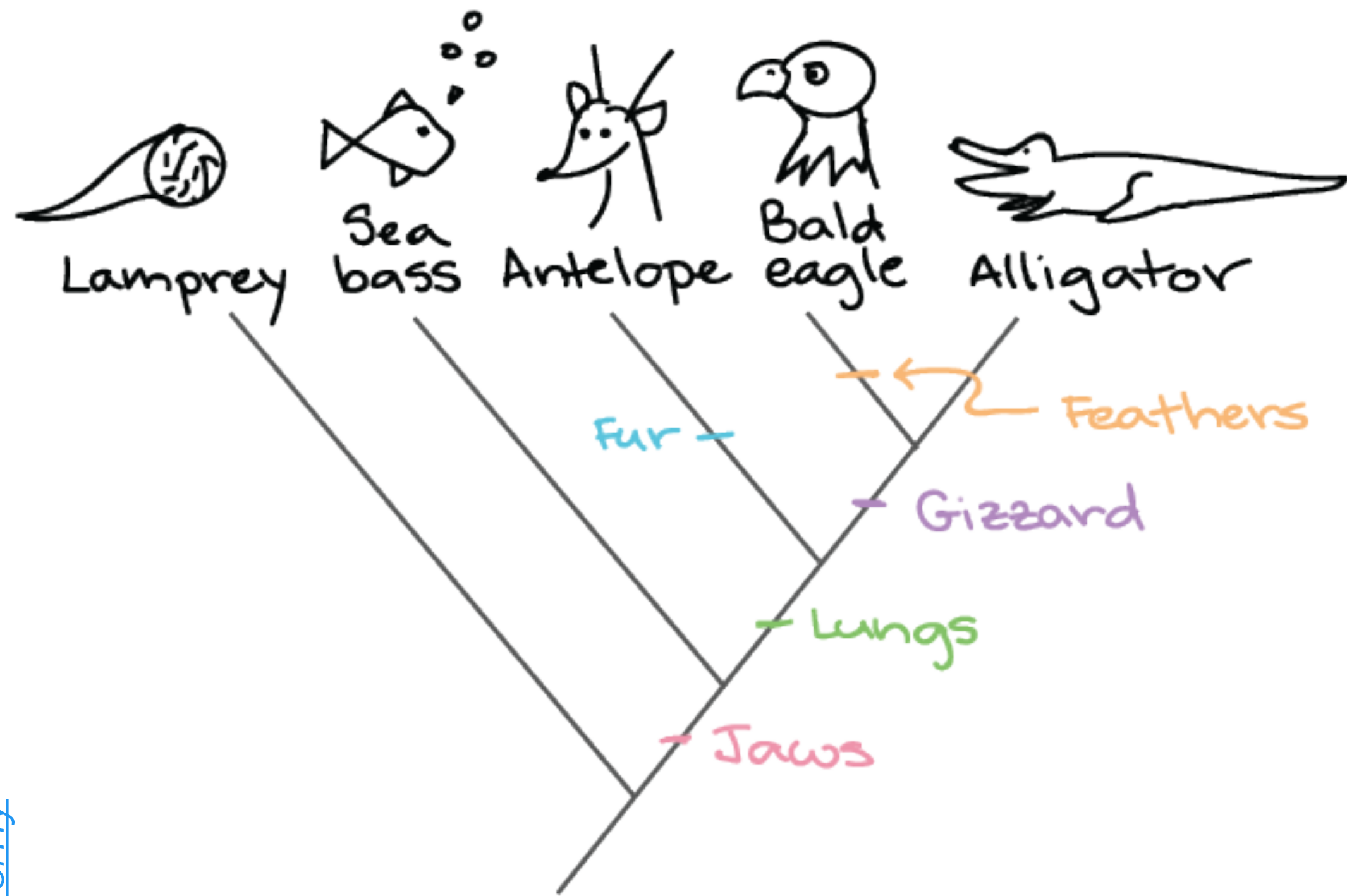
The **maximum parsimony** tree is the unrooted tree with the *lowest parsimony score*

The **parsimony score** of a tree is defined as the *minimum number of changes* required to explain the data summed across characters

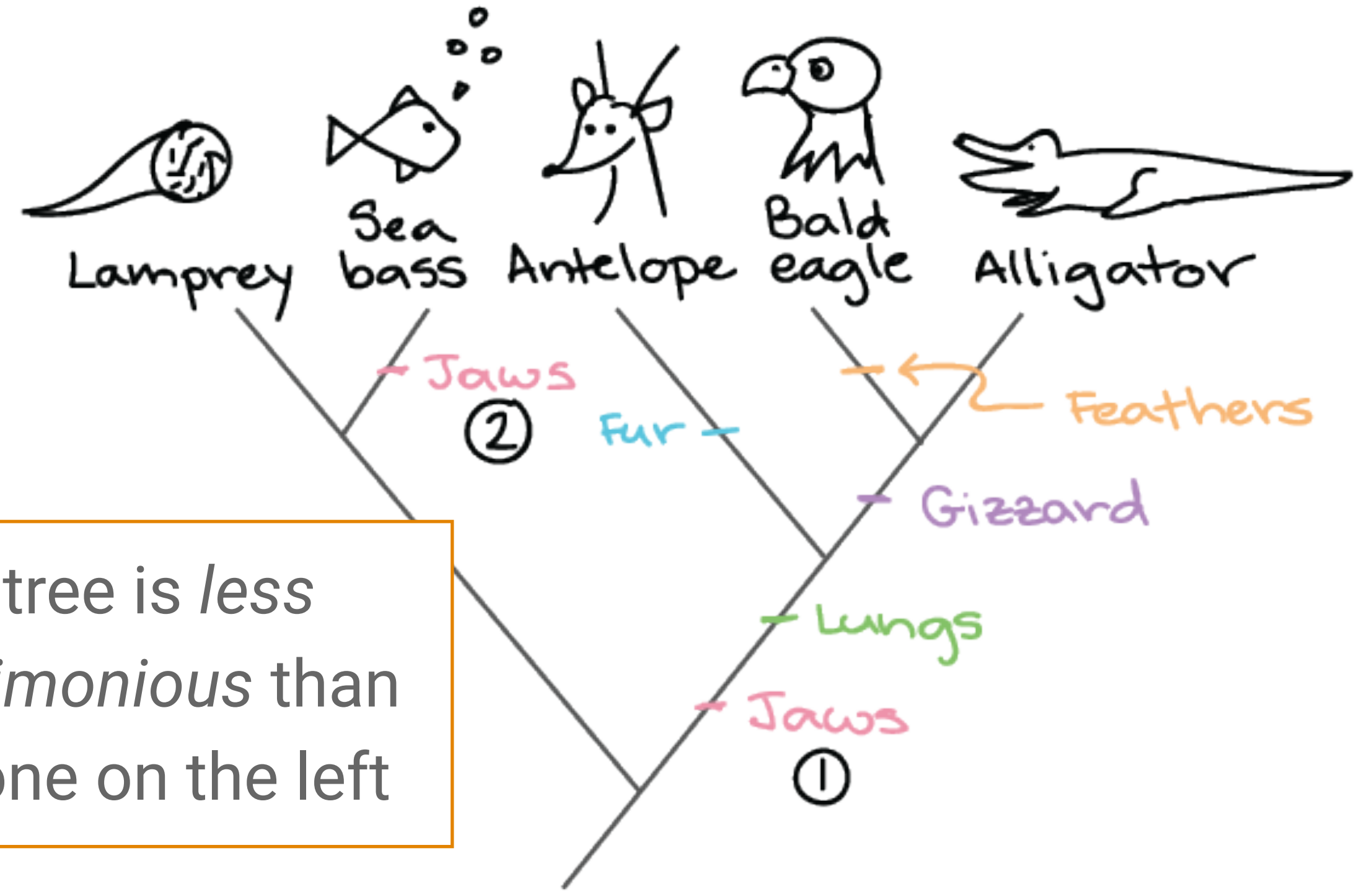
Maximum parsimony first described in *Edwards and Cavalli-Sforza* ([1964](#))



Phylogeny of fossil horses. It was (I think) the first tree constructed using parsimony and discrete morphological characters by *Camin and Sokal* ([1965](#)) — this study popularised the use of parsimony among systematists



This tree is *less parsimonious* than the one on the left



Maximum parsimony

Parsimony can not identify the location of the root, so we can use an outgroup to root the tree

There can be more than one tree with the same parsimony score

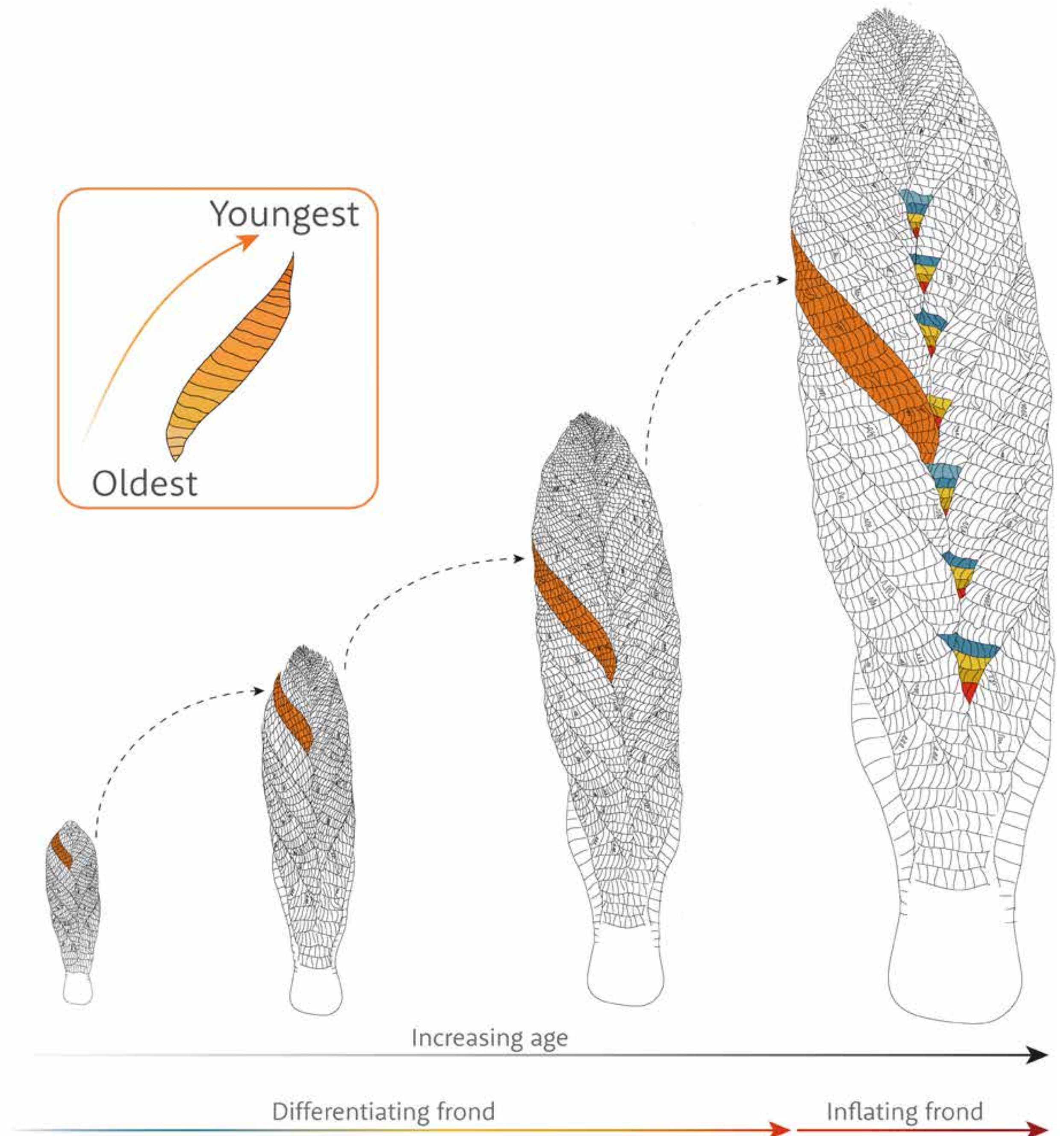
Parsimony does not make **explicit** assumptions about the evolutionary process

Maximum parsimony can only be used to estimate tree topology

Exercise

Tree building using
parsimony

Data from *Dunn (2016)*. See also *Dunn et al. (2021)*
Phylogenetic affinity of the enigmatic *Charnia*



#NEXUS

Initiates the nexus block

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS NTAX=10 NCHAR=117;

FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia	00?1??1?????????0???1?211201?????????0-?-?????????
Laccaria	110?0010000000000-000--00--00---00--0-----000-
Capsaspora	001011-00-----000-00---00-----0-----0-0-
Monosiga	001000-11-----000-00---00-----0-----0-0-
Sycon	00110011111111111110100110--0000000000----0---000-
Amphimedon	001100111111111111121100110--0000000000---0---000-
Trichoplax	001100111111100000-0000210--01111110000--00---000-
Mnemiopsis	001100111111100000-0011211200110010111121110011110
Nematostella	0011001111111000000112111001100101111111101111111
Hydra	001100111111100000-0011211100110010111111110111111

;

END;

```
#NEXUS
```

```
[Matrix modified from Dunn (2016)]
```

Comments go in square brackets

Great for keeping track of data sources

```
BEGIN DATA;
```

```
DIMENSIONS NTAX=10 NCHAR=117;
```

```
FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;
```

```
MATRIX
```

Charnia	00?1??1?????????0???1?211201?????????0-?-?????????
Laccaria	110?0010000000000-000--00--00---00--0-----000-
Capsaspora	001011-00-----000-00---00-----0-----0-0-
Monosiga	001000-11-----000-00---00-----0-----0-0-
Sycon	0011001111111111110100110--0000000000----0---000-
Amphimedon	00110011111111111121100110--0000000000---0---000-
Trichoplax	001100111111100000-0000210--01111110000--00---000-
Mnemiopsis	001100111111100000-0011211200110010111121110011110
Nematostella	0011001111111000000112111001100101111111101111111
Hydra	001100111111100000-0011211100110010111111110111111

```
;
```

```
END;
```



```
#NEXUS
```

```
[Matrix modified from Dunn (2016)]
```

```
BEGIN DATA;
```

Initiates the data block

```
DIMENSIONS  NTAX=10  NCHAR=117;
```

```
  FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;
```

```
MATRIX
```

```
Charnia      00?1??1?????????0???1?211201?????????0-?-?????????  
Laccaria     110?0010000000000-000--00--00---00--0-----000-  
Capsaspora  001011-00-----000-00---00-----0-----0-0-  
Monosiga     001000-11-----000-00---00-----0-----0-0-  
Sycon        00110011111111111110100110--0000000000---0---000-  
Amphimedon  001100111111111111121100110--0000000000---0---000-  
Trichoplax  001100111111100000-0000210--01111110000--00---000-  
Mnemiopsis  001100111111100000-0011211200110010111121110011110  
Nematostella 0011001111111000000011211100110010111111110111111  
Hydra        001100111111100000-0011211100110010111111110111111
```

```
;
```

```
END;
```

#NEXUS

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS NTAX=10 NCHAR=117;

FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia	00?1??1?????????0???1?211201?????????0-?-?????????
Laccaria	110?0010000000000-000--00--00---00--0-----000-
Capsaspora	001011-00-----000-00---00-----0-----0-0-
Monosiga	001000-11-----000-00---00-----0-----0-0-
Sycon	00110011111111111110100110--0000000000---0---000-
Amphimedon	001100111111111111121100110--0000000000---0---000-
Trichoplax	001100111111100000-0000210--01111110000--00---000-
Mnemiopsis	001100111111100000-0011211200110010111121110011110
Nematostella	00110011111110000000112111001100101111111101111111
Hydra	001100111111100000-00112111001100101111111101111111

;

END;

Number of taxa and characters in the matrix

```
#NEXUS
```

```
[Matrix modified from Dunn (2016)]
```

```
BEGIN DATA;
```

```
DIMENSIONS NTAX=10 NCHAR=117;
```

```
FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;
```

Details about the data

```
MATRIX
```

```
Charnia      00?1??1?????????0???1?211201?????????0-?-?????????  
Laccaria    110?0010000000000-000--00--00---00--0-----000-  
Capsaspora  001011-00-----000-00---00-----0-----0-0-  
Monosiga    001000-11-----000-00---00-----0-----0-0-  
Sycon       0011001111111111110100110--0000000000---0---000-  
Amphimedon  00110011111111111121100110--0000000000---0---000-  
Trichoplax  001100111111100000-0000210--01111110000--00---000-  
Mnemiopsis  001100111111100000-0011211200110010111121110011110  
Nematostella 001100111111100000011211100110010111111110111111  
Hydra       001100111111100000-0011211100110010111111110111111
```

```
;
```

```
END;
```

```
#NEXUS
```

```
[Matrix modified from Dunn (2016)]
```

```
BEGIN DATA;
```

```
DIMENSIONS NTAX=10 NCHAR=117;
```

```
FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;
```

MATRIX

```
Charnia      00?1??1?????????0???1?211201?????????0-?-?????????  
Laccaria    110?0010000000000-000--00--00---00--0-----000-  
Capsaspora  001011-00-----000-00---00-----0-----0-0-  
Monosiga    001000-11-----000-00---00-----0-----0-0-  
Sycon       00110011111111111110100110--0000000000---0---000-  
Amphimedon  001100111111111111121100110--0000000000---0---000-  
Trichoplax  001100111111100000-0000210--01111110000--00---000-  
Mnemiopsis  001100111111100000-0011211200110010111121110011110  
Nematostella 0011001111111000000011211100110010111111110111111  
Hydra       001100111111100000-0011211100110010111111110111111
```

```
;
```

```
END;
```

Initiates the data matrix

```
#NEXUS
```

```
[Matrix modified from Dunn (2016)]
```

```
BEGIN DATA;
```

```
DIMENSIONS NTAX=10 NCHAR=117;
```

```
FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;
```

```
MATRIX
```

```
Charnia      00?1??1????????0???1?211201????????0-?-?????????  
Laccaria    110?0010000000000-000--00--00---00--0-----000-  
Capsaspora  001011-00-----000-00---00-----0-----0-0-  
Monosiga    001000-11-----000-00---00-----0-----0-0-  
Sycon       00110011111111111110100110--0000000000----0---000-  
Amphimedon  001100111111111111121100110--0000000000---0---000-  
Trichoplax  001100111111100000-0000210--01111110000--00---000-  
Mnemiopsis  001100111111100000-0011211200110010111121110011110  
Nematostella 0011001111111000000011211100110010111111110111111  
Hydra       001100111111100000-0011211100110010111111110111111
```

```
;
```

```
END;
```

Phylogenetic data matrix

Taxon names left, characters right after tabs or spaces

```
#NEXUS
```

```
[Matrix modified from Dunn (2016)]
```

```
BEGIN DATA;
```

```
DIMENSIONS NTAX=10 NCHAR=117;
```

```
FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;
```

```
MATRIX
```

```
Charnia      00?1??1?????????0???1?211201?????????0-?-?????????  
Laccaria    110?0010000000000-000--00--00---00--0-----000-  
Capsaspora  001011-00-----000-00---00-----0-----0-0-  
Monosiga    001000-11-----000-00---00-----0-----0-0-  
Sycon       00110011111111111110100110--0000000000-0---0---000-  
Amphimedon  001100111111111111121100110--0000000000-0---0---000-  
Trichoplax  001100111111100000-0000210--01111110000--00---000-  
Mnemiopsis  001100111111100000-0011211200110010111121110011110  
Nematostella 0011001111111000000011211100110010111111110111111  
Hydra       001100111111100000-0011211100110010111111110111111
```



```
END;
```

Initiates the end of the data matrix

```
#NEXUS
```

```
[Matrix modified from Dunn (2016)]
```

```
BEGIN DATA;
```

```
DIMENSIONS NTAX=10 NCHAR=117;
```

```
FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;
```

```
MATRIX
```

```
Charnia      00?1??1?????????0???1?211201?????????0-?-?????????  
Laccaria    110?0010000000000-000--00--00---00--0-----000-  
Capsaspora  001011-00-----000-00---00-----0-----0-0-  
Monosiga    001000-11-----000-00---00-----0-----0-0-  
Sycon       0011001111111111110100110--0000000000---0---000-  
Amphimedon  00110011111111111121100110--0000000000---0---000-  
Trichoplax  001100111111100000-0000210--01111110000--00---000-  
Mnemiopsis  001100111111100000-0011211200110010111121110011110  
Nematostella 0011001111111000000011211100110010111111110111111  
Hydra       001100111111100000-0011211100110010111111110111111
```

```
;
```

```
END;
```

Initiates the end of the nexus block

Statistical inconsistency

The following slides are adapted from Tracy Heath (in turn adapted from Mark Holder)

Statistical consistency

Ideally, we want an inference method to return the correct answer if we provide enough data

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

It has been demonstrated that in some scenarios parsimony is statistically inconsistent. The issue is known as **long branch attraction**

Convergent evolution

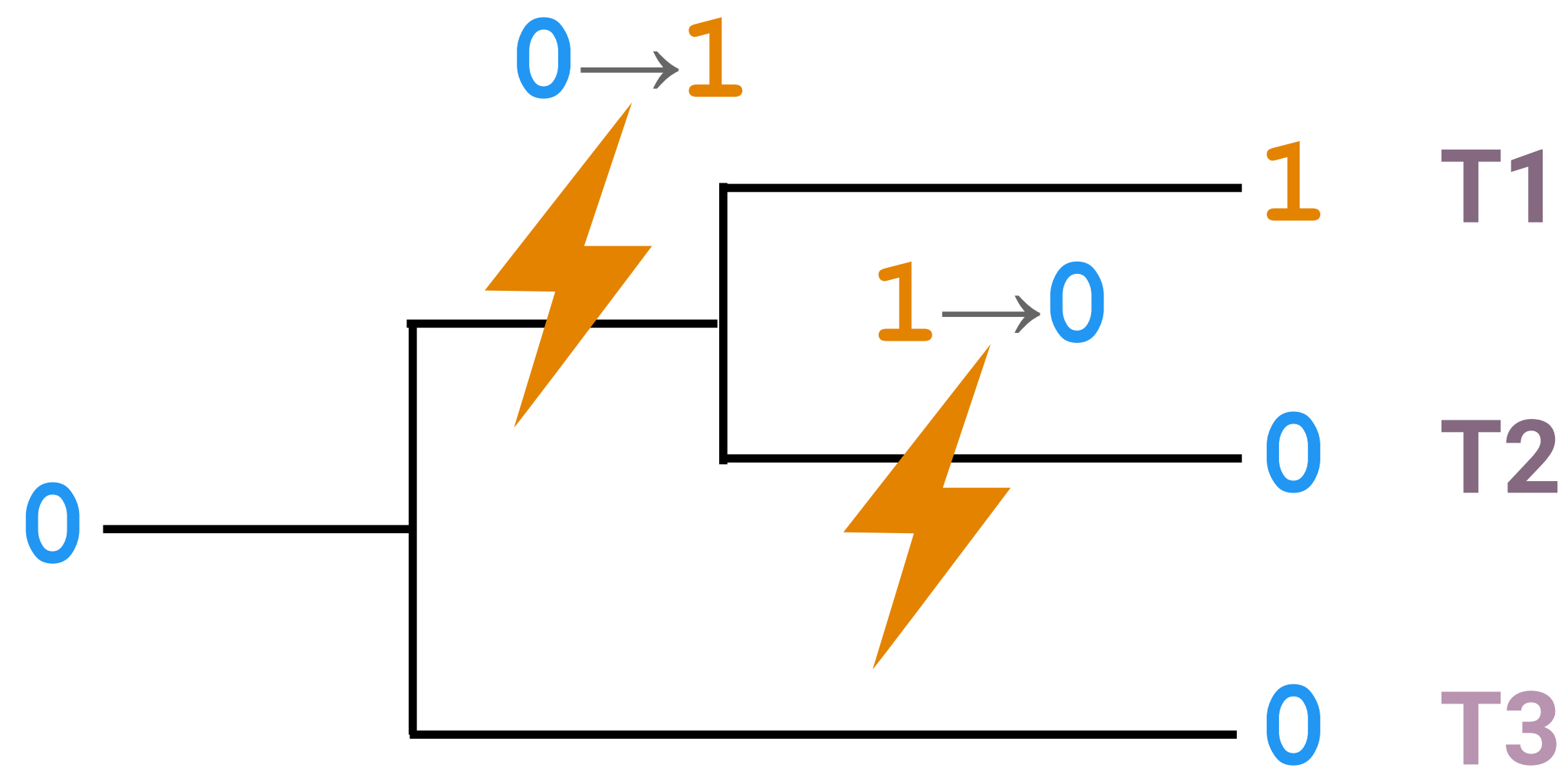
A trait that is found in two species, but not in their common ancestor is an example of **homoplasy**

We find widespread homoplasy in both morphological and molecular datasets



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

Convergence and parsimony



Hypothetical tree showing multiple transitions at the same character

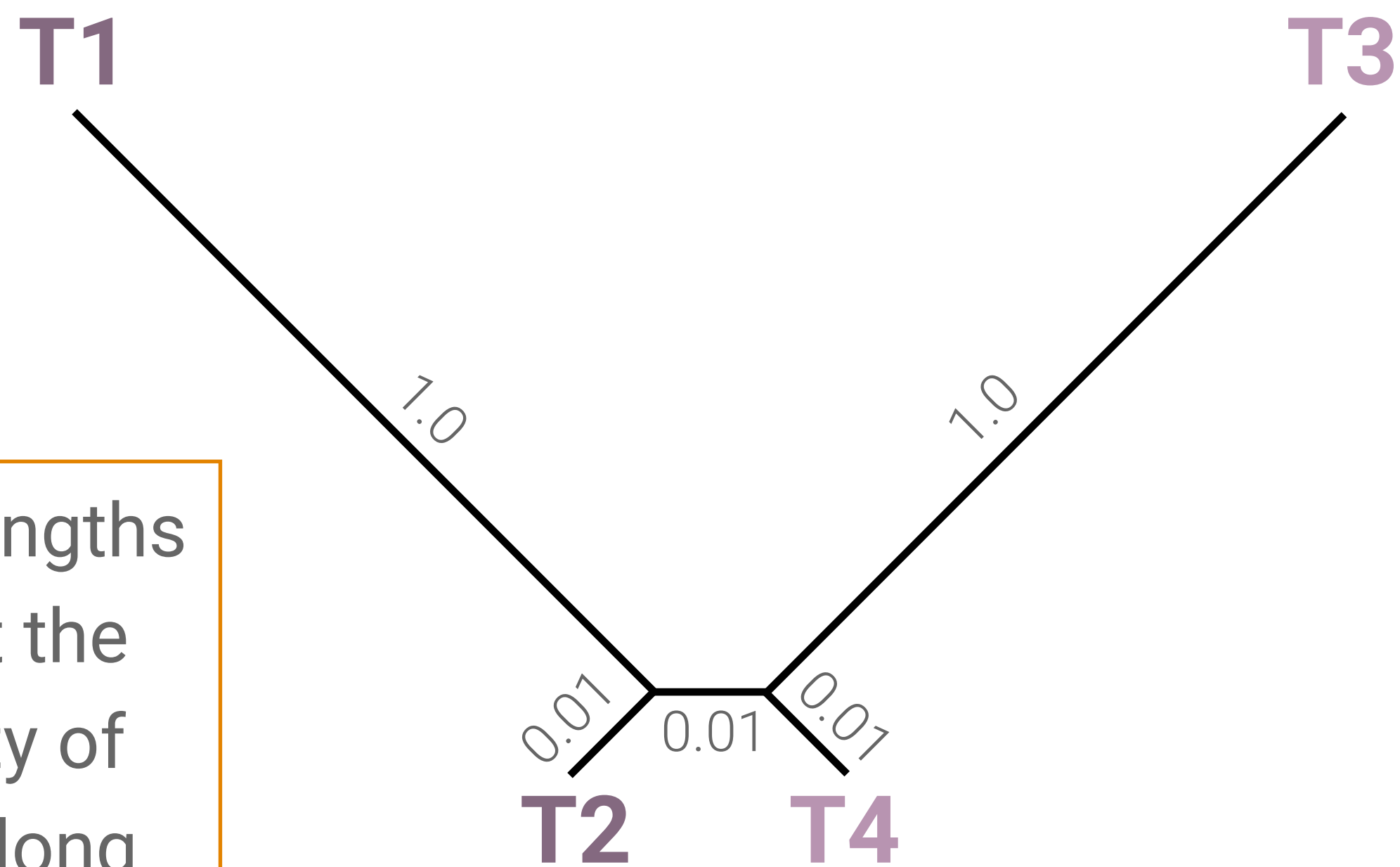
Parsimony will always favour the tree with the smallest number of changes

The method does not account for multiple transitions (or “hits”), e.g.,
 $0 \rightarrow 1 \rightarrow 0$

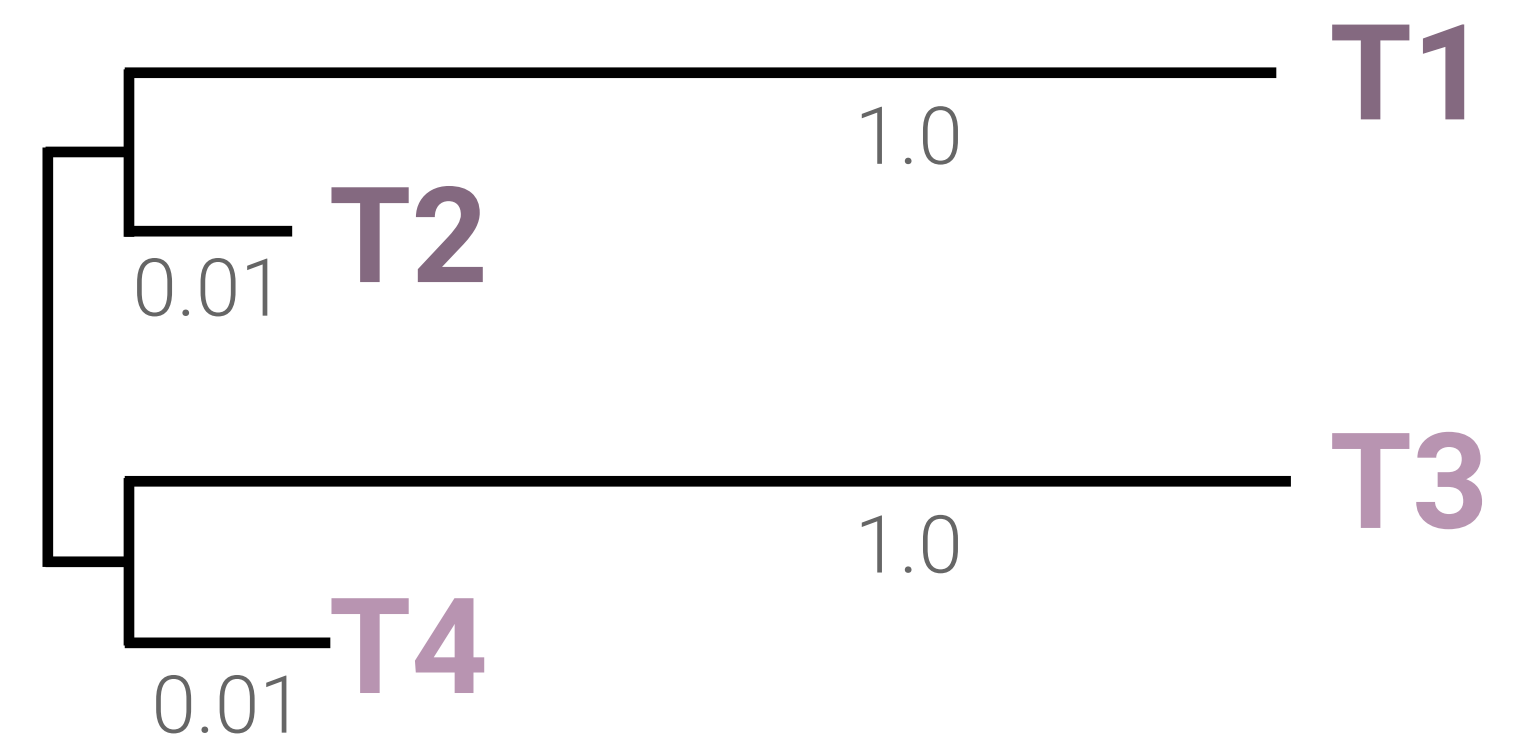
We can only invoke convergent evolution *ad hoc* after inference

Long branch attraction

Say we have the following tree, with 2 long and 2 short branches

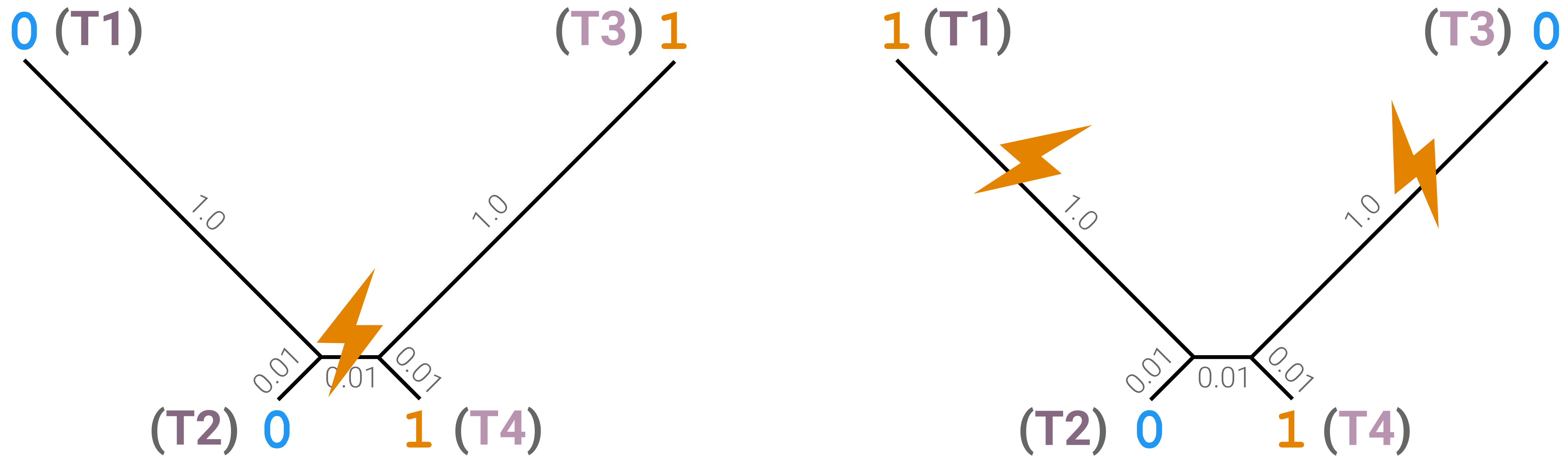


Branch lengths represent the probability of change along that branch



The rooted version of the same tree

Long branch attraction



Some time in the past

A transition somewhere along the branch that separates (T1, T2) and (T3, T4)

Subsequent transitions along the branches leading to T2 and T4

Long branch attraction

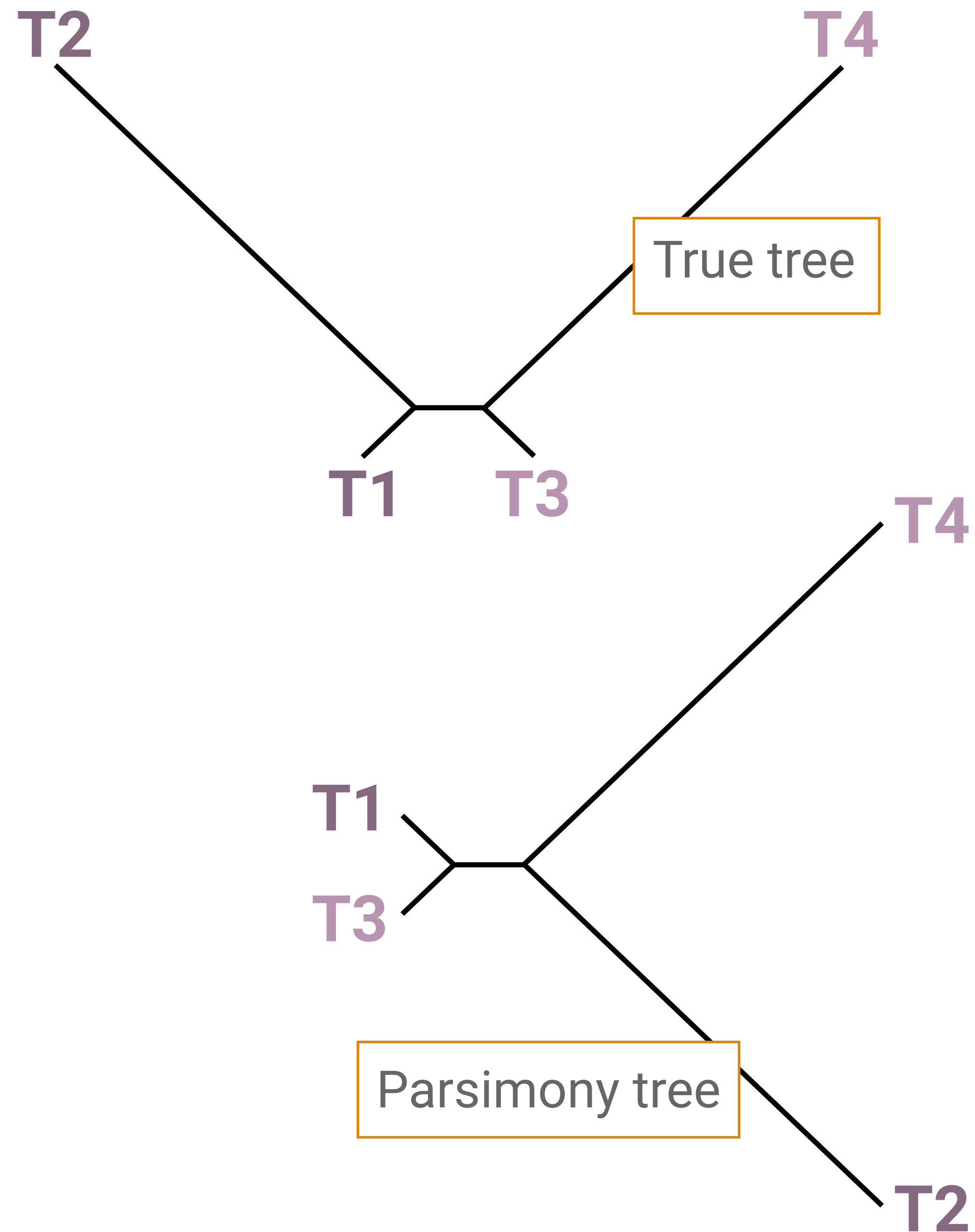
Parsimony is almost guaranteed to get this tree wrong

And more data makes this problem worse, meaning this approach is statistically inconsistent

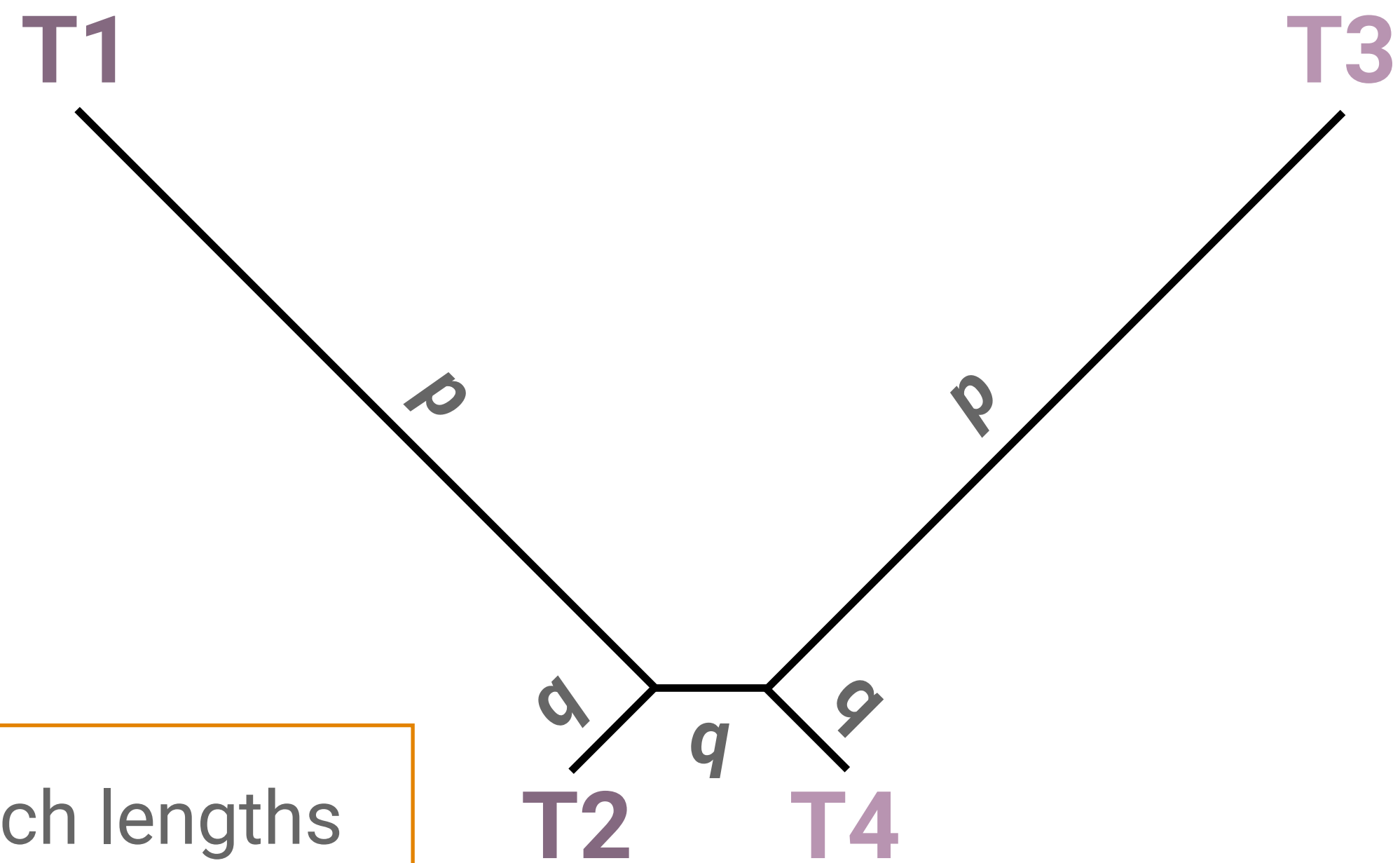
This issue has been thoroughly characterised for molecular data

Felsenstein ([1978](#))

Felsenstein (2004), *Inferring Phylogenies*

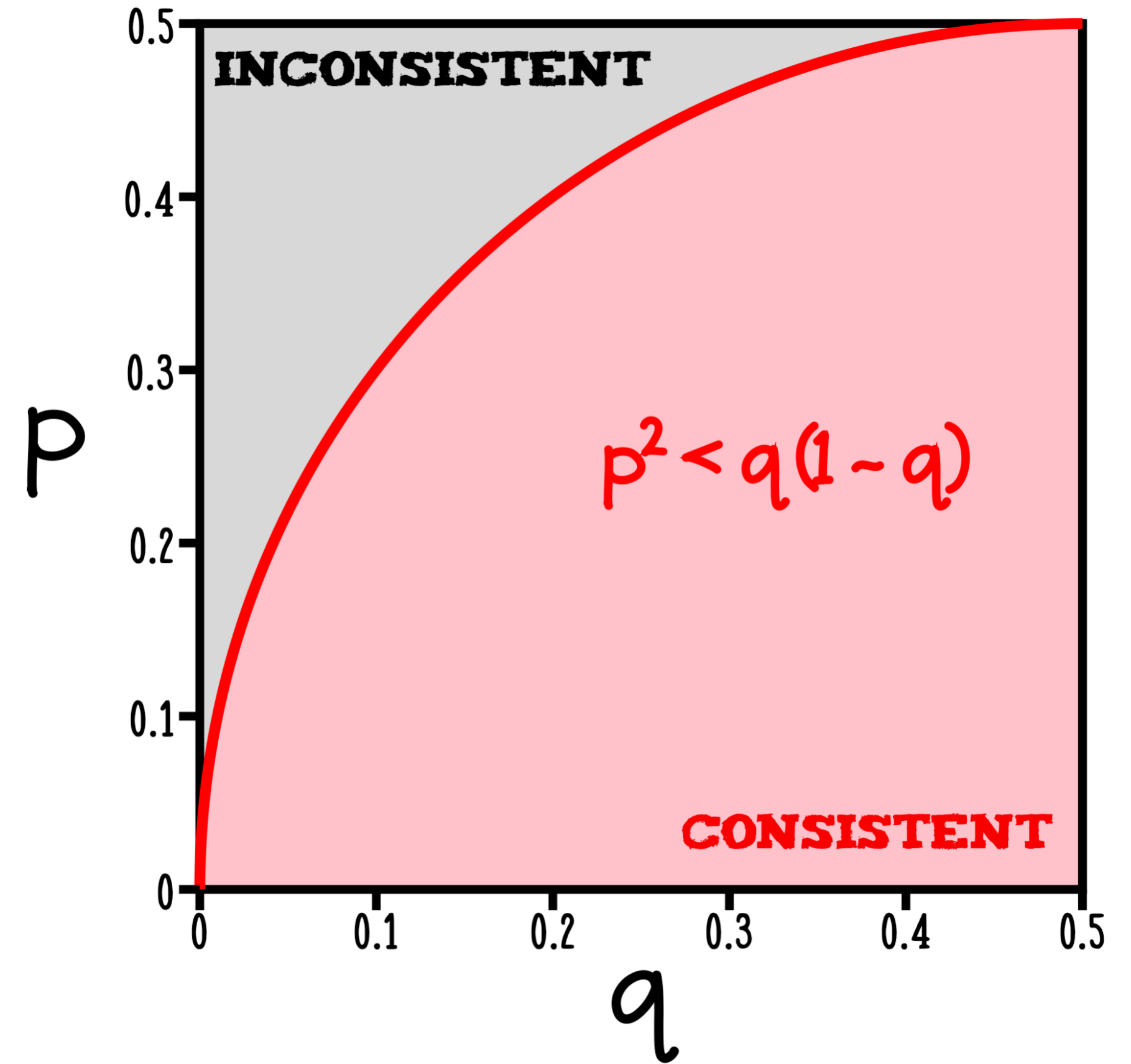


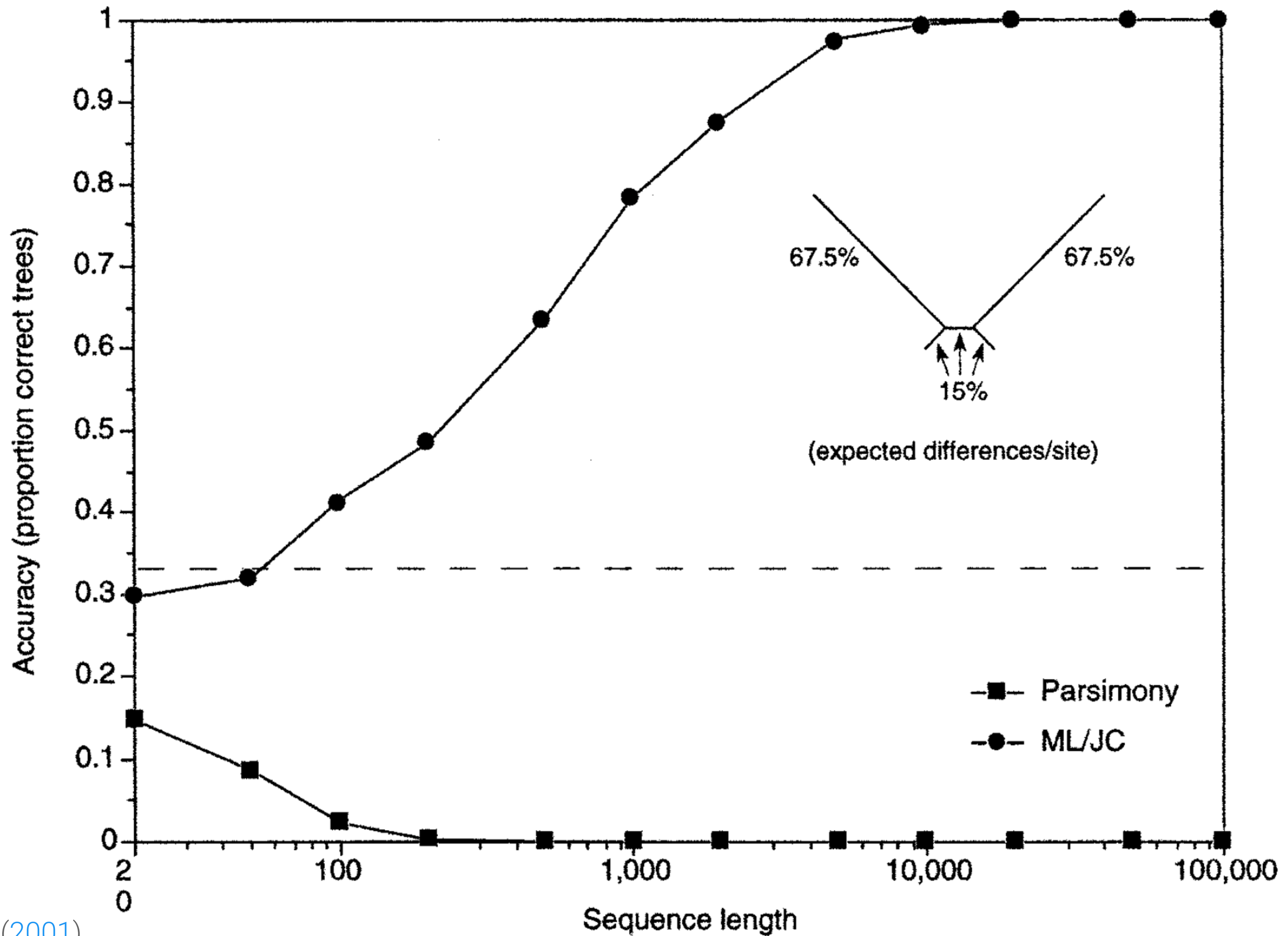
Long branch attraction



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

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



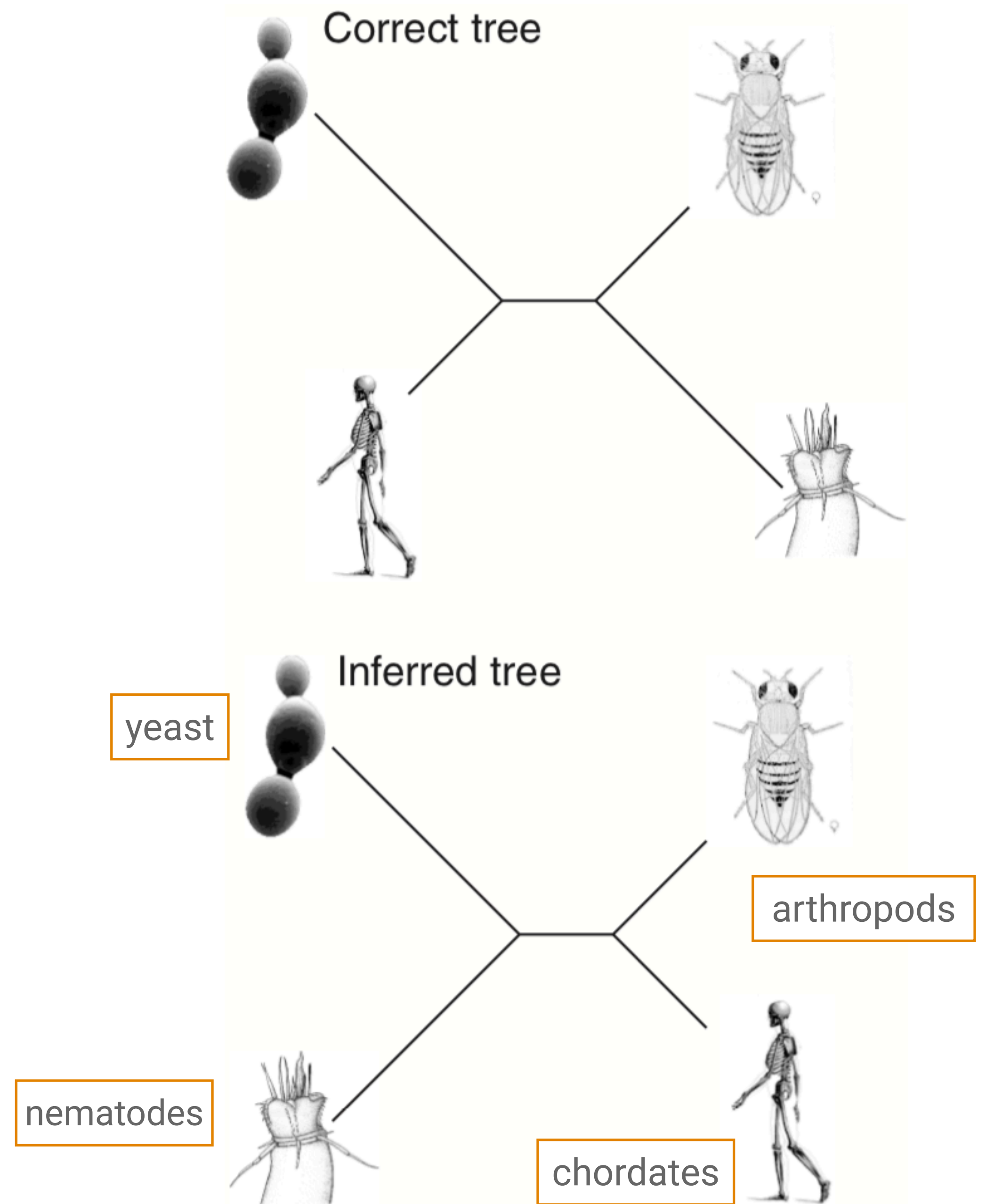


A classic case of LBA

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

- **Ecdysozoa***
(arthropods, nematodes), vertebrates
- **Coelomata**
(arthropods, vertebrates), nematodes

*widely accepted today, Image: *Telford et al.* ([2005](#))



Parsimony in paleobiology



Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data

April M. Wright*, David M. Hillis

Department of Integrative Biology, University of Texas at Austin, Austin, Texas, United States of America

Abstract

Despite the introduction of likelihood-based methods for estimating phylogenetic trees from phenotypic data, parsimony remains the most widely-used optimality criterion for building trees from discrete morphological data. However, it has been known for decades that there are regions of solution space in which parsimony is a poor estimator of tree topology. Numerous software implementations of likelihood-based models for the estimation of phylogeny from discrete morphological data exist, especially for the Mk model of discrete character evolution. Here we explore the efficacy of Bayesian estimation of phylogeny, using the Mk model, under conditions that are commonly encountered in paleontological studies. Using simulated data, we describe the relative performances of parsimony and the Mk model under a range of realistic conditions that include common scenarios of missing data and rate heterogeneity.

Citation: Wright AM, Hillis DM (2014) Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. PLoS ONE 9(10): e109210. doi:10.1371/journal.pone.0109210

Editor: Art F. Y. Poon, British Columbia Centre for Excellence in HIV/AIDS, Canada

Received: July 23, 2014; **Accepted:** September 3, 2014; **Published:** October 3, 2014

Copyright: © 2014 Wright, Hillis. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Data are available from Figshare using the DOI <http://dx.doi.org/10.6084/m9.figshare.1160541>.

Funding: AMW received financial support for material purchases related to this project from the National Science Foundation's Doctoral Dissertation Improvement Grant (number 201203267-001; http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5234). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: wright.aprilm@gmail.com

Introduction

For many decades, parsimony methods have been the most widely used approaches for estimation of phylogeny from discrete phenotypic data, despite the availability of likelihood-based methods for phylogenetic analysis. Maximum likelihood and Bayesian methods are commonly used in data sets combining molecules and morphology [1–5], but are used less frequently in morphology-only data sets [6]. As such, the efficacy of these methods under a range of conditions is not well-explored. In particular, the conditions that are investigated in most paleontological studies (many characters missing across sampled taxa, and rate heterogeneity among different sampled characters) lead some investigators to raise questions about the applicability of model-based approaches under these conditions [6–9].

At the present, the most widely implemented (in both pure likelihood and Bayesian contexts) model for estimating phylogenetic trees from discrete phenotypic data is the Mk model proposed by Lewis [10]. This model is a generalization of the 1969 Jukes-Cantor model of nucleotide sequence evolution [11]. The Mk model assumes a Markov process for character change, allowing for multiple character-state changes along a single branch. The probability of change in this model is symmetrical; in other words, the probability of changing from one state to another is the same as change in the reverse direction. This assumption can be relaxed in Bayesian implementations through the use of a hyperprior allowing variable change probabilities among states [12–14]. As many morphologists collect only variable

or parsimony-informative characters (i.e., characters that can be used to discriminate among different tree topologies under the parsimony criterion), the distribution of characters collected does not reflect the distribution of all observable characters. This sampling bias can lead to poor estimation of the rate of character evolution within a data set, as well as inflated estimates of character change along branches of the estimated tree. To counteract this bias, Lewis [10] introduced versions of the Mk model that correct for biases in character collection. These versions were subsequently shown to have the desirable quality of statistical consistency [15].

Sampled characters within data sets typically evolve under different rates, developmental processes, and modes of evolution [7,16,17]. Although heterogeneity in the underlying evolutionary processes can present challenges to the application of evolutionary models [18], a distribution of different evolutionary rates of characters can be helpful for resolving branches at different levels in the tree. Extremely labile characters, for example, are useful for resolving recently diverged lineages, whereas slowly evolving characters may be more useful for resolving deep divergences in the tree. Likelihood-based methods can benefit from this heterogeneity by accounting for different rates of character evolution and the amount of time available for change (based on the estimated branch lengths in the tree; [19]). In contrast, high levels of rate heterogeneity among characters can be more problematic for parsimony methods, especially if all character changes are weighted equally [20].

This was the first paper to show that the same LBA issues that affect molecular data probably also affect morphology

Wright and Hillis ([2014](#))

A slue of papers followed...



Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data

April M. Wright*, David M. Hillis

Department of Integrative Biology, University of Texas at Austin, Austin, Texas, United States of America

Abstract

Despite the introduction of likelihood-based methods for estimating phylogenetic trees from phenotypic data, parsimony remains the most widely-used optimality criterion for building trees from discrete morphological data. However, it has been known for decades that there are regions of solution space in which parsimony is a poor estimator of tree topology. Numerous software implementations of likelihood-based models for the estimation of phylogeny from discrete morphological data exist, especially for the Mk model of discrete character evolution. Here we explore the efficacy of Bayesian estimation of phylogeny, using the Mk model, under conditions that are commonly encountered in paleontological studies. Using simulated data, we describe the relative performances of parsimony and the Mk model under a range of realistic conditions that include common scenarios of missing data and rate heterogeneity.

Citation: Wright AM, Hillis DM (2014) Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. PLoS ONE 9(10): e109210. doi:10.1371/journal.pone.0109210

Editor: Art F. Y. Poon, British Columbia Centre for Excellence in HIV/AIDS, Canada

Received: July 23, 2014; **Accepted:** September 3, 2014; **Published:** October 3, 2014

Copyright: © 2014 Wright, Hillis. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Data are available from Figshare using the DOI <http://dx.doi.org/10.6084/m9.figshare.1160541>.

Funding: AMW received financial support for material purchases related to this project from the National Science Foundation's Doctoral Dissertation Improvement Grant (number 201203267-001); http://www.nsf.gov/funding/pgm_summ.jsp?ps1_id=5234. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: wright.aprilm@gmail.com

Introduction

For many decades, parsimony methods have been the most widely used approaches for estimation of phylogeny from discrete phenotypic data, despite the availability of likelihood-based methods for phylogenetic analysis. Maximum likelihood and Bayesian methods are commonly used in data sets combining molecules and morphology [1–5], but are used less frequently in morphology-only data sets [6]. As such, the efficacy of these methods under a range of conditions is not well-explored. In particular, the conditions that are investigated in most paleontological studies (many characters missing across sampled taxa, and rate heterogeneity among different sampled characters) lead some investigators to raise questions about the applicability of model-based approaches under these conditions [6–9].

At the present, the most widely implemented (in both pure likelihood and Bayesian contexts) model for estimating phylogenetic trees from discrete phenotypic data is the Mk model proposed by Lewis [10]. This model is a generalization of the 1969 Jukes-Cantor model of nucleotide sequence evolution [11]. The Mk model assumes a Markov process for character change, allowing for multiple character-state changes along a single branch. The probability of change in this model is symmetrical; in other words, the probability of changing from one state to another is the same as change in the reverse direction. This assumption can be relaxed in Bayesian implementations through the use of a hyperprior allowing variable change probabilities among states [12–14]. As many morphologists collect only variable

or parsimony-informative characters (i.e., characters that can be used to discriminate among different tree topologies under the parsimony criterion), the distribution of characters collected does not reflect the distribution of all observable characters. This sampling bias can lead to poor estimation of the rate of character evolution within a data set, as well as inflated estimates of character change along branches of the estimated tree. To counteract this bias, Lewis [10] introduced versions of the Mk model that correct for biases in character collection. These versions were subsequently shown to have the desirable quality of statistical consistency [15].

Sampled characters within data sets typically evolve under different rates, developmental processes, and modes of evolution [7,16,17]. Although heterogeneity in the underlying evolutionary processes can present challenges to the application of evolutionary models [18], a distribution of different evolutionary rates of characters can be helpful for resolving branches at different levels in the tree. Extremely labile characters, for example, are useful for resolving recently diverged lineages, whereas slowly evolving characters may be more useful for resolving deep divergences in the tree. Likelihood-based methods can benefit from this heterogeneity by accounting for different rates of character evolution and the amount of time available for change (based on the estimated branch lengths in the tree; [19]). In contrast, high levels of rate heterogeneity among characters can be more problematic for parsimony methods, especially if all character changes are weighted equally [20].

Syst. Biol. 69(5):897–912, 2020.
© The Author(s) 2020. Published by Oxford University Press on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syaa012
Advance Access publication February 19, 2020

Morphological Phylogenetics Evaluated Using Novel Evolutionary Simulations

JOSEPH N. KEATING^{1,2}, ROBERT S. SANSOM^{1*}, MARK D. SUTTON³, CHRISTOPHER G. KNIGHT¹
AND RUSSELL J. GARWOOD^{1,4,*}

¹Department of Earth and Environmental Sciences, University of Manchester, Williamson Building, Oxford Road, Manchester M13 9PL, UK; ²School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK; ³Department of Earth Science and Engineering, South Kensington Campus, Imperial College London, London SW7 2AZ, UK; and ⁴Earth Sciences Department, Natural History Museum, Cromwell Rd, South Kensington, London SW7 5BD, UK

*Correspondence to be sent to: Department of Earth and Environmental Sciences, University of Manchester, Manchester M13 9PL, UK; E-mail: russell.garwood@manchester.ac.uk; robert.sansom@manchester.ac.uk

Received 15 July 2019; reviews returned 31 January 2020; accepted 7 February 2020
Associate Editor: Jeanne Serb

Abstract.—Evolutionary inferences require reliable phylogenies. Morphological data have traditionally been analyzed using maximum parsimony, but recent simulation studies have suggested that Bayesian analyses yield more accurate trees. This debate is ongoing, in part, because of ambiguity over modes of morphological evolution and a lack of appropriate models. Here, we investigate phylogenetic methods using two novel simulation models—one in which morphological characters evolve stochastically along lineages and another in which individuals undergo selection. Both models generate character data and lineage splitting simultaneously; the resulting trees are an emergent property, rather than a fixed parameter. Standard consensus methods for Bayesian searches (Mk) yield fewer incorrect nodes and quartets than the standard consensus trees recovered using equal weighting and implied weighting parsimony searches. Distances between the pool of derived trees (most parsimonious or posterior distribution) and the true trees—measured using Robinson-Foulds (RF), subtree prune and regraft (SPR), and tree bisection reconnection (TBR) metrics—demonstrate that this is related to the search strategy and Morphological coherence, which has previously not been considered in this context, proves to be a more important factor for phylogenetic accuracy than homoplasy. Selection-based models exhibit relatively lower homoplasy, lower morphological coherence, and higher inaccuracy in inferred trees. Selection is a dominant driver of morphological evolution, but we demonstrate that it has a confounding effect on numerous character properties which are fundamental to phylogenetic inference. We suggest that the current debate should move beyond considerations of parsimony versus Bayesian, toward identifying modes of morphological evolution and using these to build models for probabilistic search methods. [Bayesian; evolution; morphology; parsimony; phylogenetics; selection; simulation.]

Phylogenetic trees provide a vital framework for evolutionary inferences. Consequently, the accuracy of phylogenetic estimates built using empirical characters underpins our understanding of evolutionary history. Morphology was fundamental to the conception and development of phylogenetic methods (Hennig 1950, 1965). However, in the genomic age, sequence data have replaced morphology as both the dominant source of phylogenetic information for estimating tree topology and the basis of numerical phylogenetic method development (Lee and Palci 2015; Lartillot et al. 2016). Molecular characters are more numerous than morphological ones and their evolution can be modeled based on empirical observations (Kimura 1980; Felsenstein 1981; Hasegawa et al. 1985). Nevertheless, morphology still plays a fundamental role. It is the only form of data by which we can incorporate fossils, and thus a deep-time perspective, in phylogenies. Fossil taxa allow the calibration of molecular clocks (Donoghue and Yang 2016); offer an independent test of evolutionary developmental hypotheses (Raff 2007); break long branches, and thus clarify otherwise intractable relationships (Donoghue et al. 1989; Wiens and Soltis 2005; Legg et al. 2013); and provide the only means of understanding diversity and evolution in deep time (Raupe and Sepkoski 1982). For these reasons, integrating fossils in phylogenies is necessary in order

to derive accurate phylogenies and reconstructions of character and clade evolution. Morphological data have conventionally been analyzed using maximum parsimony (Kitching et al. 1998) in which trees that necessitate the fewest character changes are considered optimal. Characters are either treated as weighted equally or rescaled in relation to their homoplasy, for example, using implied weighting (IW; Goloboff 2013). Likelihood-based models have also been used to analyze morphological data, primarily through Bayesian analysis using the Mk model of character evolution (Lewis 2001). The Mk model is a *k*-parameter model, where *k* is the number of possible unordered states for a discrete morphological character (e.g., in an M2 model, characters could have *k* = 2 states). The model assumes that character state changes follow a Markov process, and thus the likelihood of changing from one state to another is determined only by the current state. The basic Mk model assumes that all state changes are equally likely and occur at the same rate, although these assumptions are not always true (Lewis 2001). Some characters might be gained or lost much faster or slower than others; as such for asymmetrical evolutionary rates. For instance, the symmetrical (SYM) and all-rates-different (ARD) models (Paradis et al. 2004) are two extensions of the Mk model that can relax this assumption.

ton 2004,

comparisons their fit morphological data have conventionally been analyzed using maximum parsimony (Kitching et al. 1998) in which trees that necessitate the fewest character changes are considered optimal. Characters are either treated as weighted equally or rescaled in relation to their homoplasy, for example, using implied weighting (IW; Goloboff 2013). Likelihood-based models have also been used to analyze morphological data, primarily through Bayesian analysis using the Mk model of character evolution (Lewis 2001). The Mk model is a *k*-parameter model, where *k* is the number of possible unordered states for a discrete morphological character (e.g., in an M2 model, characters could have *k* = 2 states). The model assumes that character state changes follow a Markov process, and thus the likelihood of changing from one state to another is determined only by the current state. The basic Mk model assumes that all state changes are equally likely and occur at the same rate, although these assumptions are not always true (Lewis 2001). Some characters might be gained or lost much faster or slower than others; as such for asymmetrical evolutionary rates. For instance, the symmetrical (SYM) and all-rates-different (ARD) models (Paradis et al. 2004) are two extensions of the Mk model that can relax this assumption.

time, the data sets, in some morphological

Downloaded from <https://academic.oup.com/sysbio/article/69/5/897/598275> by Erlangen Nurnberg University user on 13 April 2024

ers

de^{1,3}

lata

ndez^{1,2},
oghue¹

stics

cla.12205

tic

rees. s of nes. nies the rices rating ony gical sult data

rating ak up n the tion efore, over been cated uracy (odel) a [4]. tions mor- x and marac- ation, matic gions; [11], issing / con- ected

of life's or the s have re less re we shape lysis that all trees, is the lution nplied model tion of n from in rea- uence species small para- ructed

of life on ssils has lar phy- ther [1]. iversity ind cali- without become genetic simony id prob- neity in

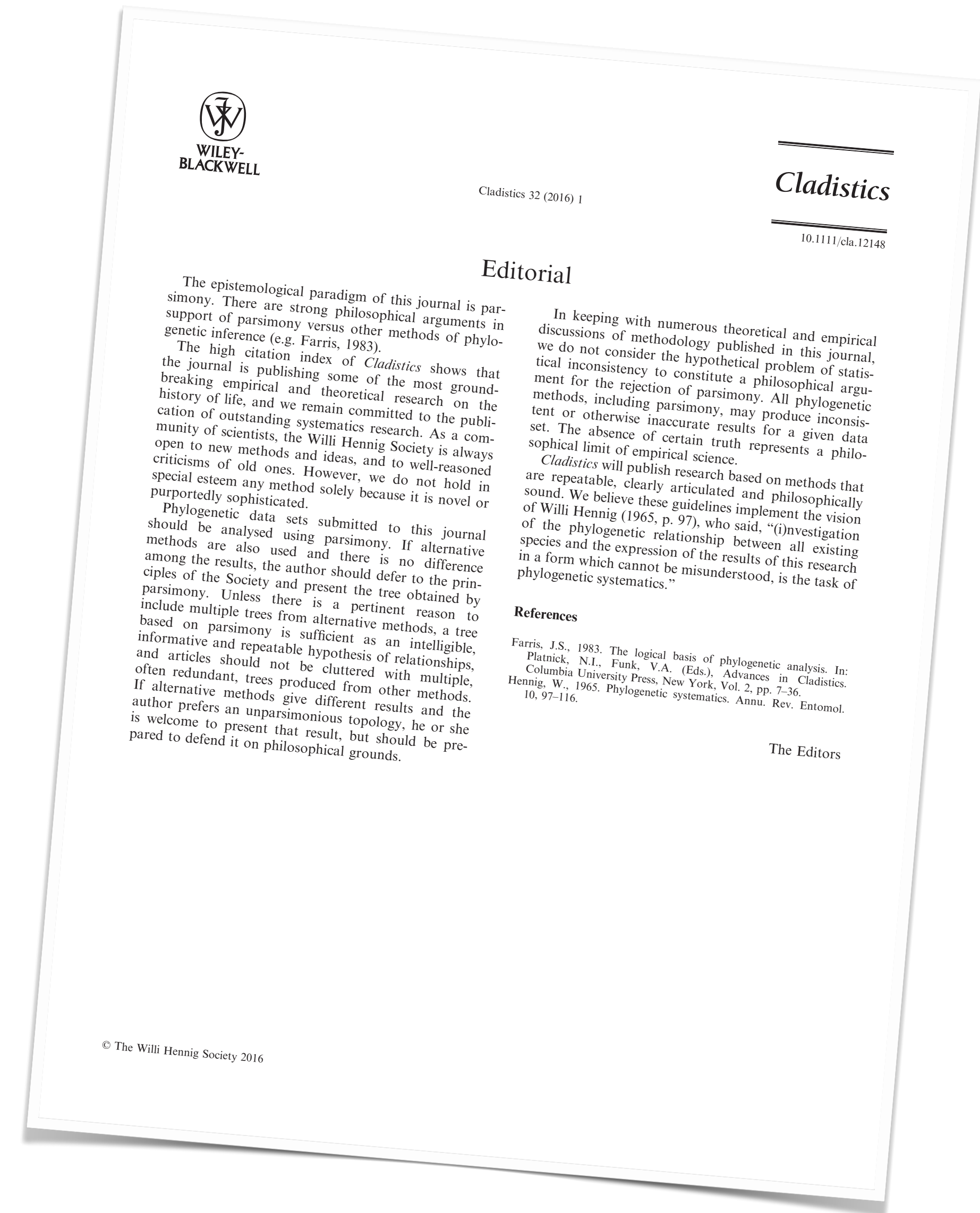
influence premise the best ructions. mselves, nsidering l model- quences, morpho- modelled rt along volution rsimony

equal y et al. zes bet- nplied methods. the pre- instead ological p. 447) did no l-based equal). Con- simony ighting. ers, by to the Puttick O'Reilly

In 2016 the journal *Cladistics* published an editorial stating

“If alternative methods give different results and the author prefers an unparsimonious topology, he or she is welcome to present that result, but should be prepared to defend it on philosophical grounds”

The debacle was written up in [Wired](#) Magazine
See also #ParsimonyGate: [The Perspective of a Reformed ‘Hardcore’ Cladist](#) by Prosanta Chakrabarty



Final notes on parsimony

The greatest advantage of parsimony is its beautiful simplicity (Yang, 2014)

Computationally fast, often produces sensible results and still serves practical purposes

Some argue that parsimony is assumption free. Others argue parsimony does make assumptions, even if we don't know what they are, referred to as **implicit** assumptions

We are often interested in more than just the topology

Model-based approaches on the other hand make **explicit** assumptions about evolutionary processes

They are also flexible and have many more applications, e.g., rate estimation, phylogenetic dating

We will therefore turn our focus to model-based inference

Homework



Re-familiarise yourself with [how to read a phylogenetic tree and rooting](#)



[A Brief History of Computational Phylogenetics](#) *Joe Felsenstein*