Day 1

Pathogen phylogenetics: From sequences to trees

Workshop

14.01.2020 – 17.01.2020

Organizational Matters & Introduction

- What is going to happen and with whom? -

Organizational Matters

Theoretical sessions: 09:00 – 12:00

Practical sessions: 13:00 – 16:00

Workshop material will be online available at:
https://github.com/Bioinformatics-Core-Facility-Jena/Phylogenetics-Workshop-2020

For Wifi connect to network: `eduroam`
username: `tagung11@uni-jena.de`
password: `Mo8zhn1nov`

Who Is Who?

Manja Marz

Emanuel Barth

Konstantin Riege

Denise Kühnert

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

fli Leibniz Institute on Aging – Fritz Lipmann Institute

Max-Planck-Institut für Menschheitsgeschichte
Max Planck Institute for the Science of Human History

BiCS Bioinformatics Core Facility

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

Who Is Who?

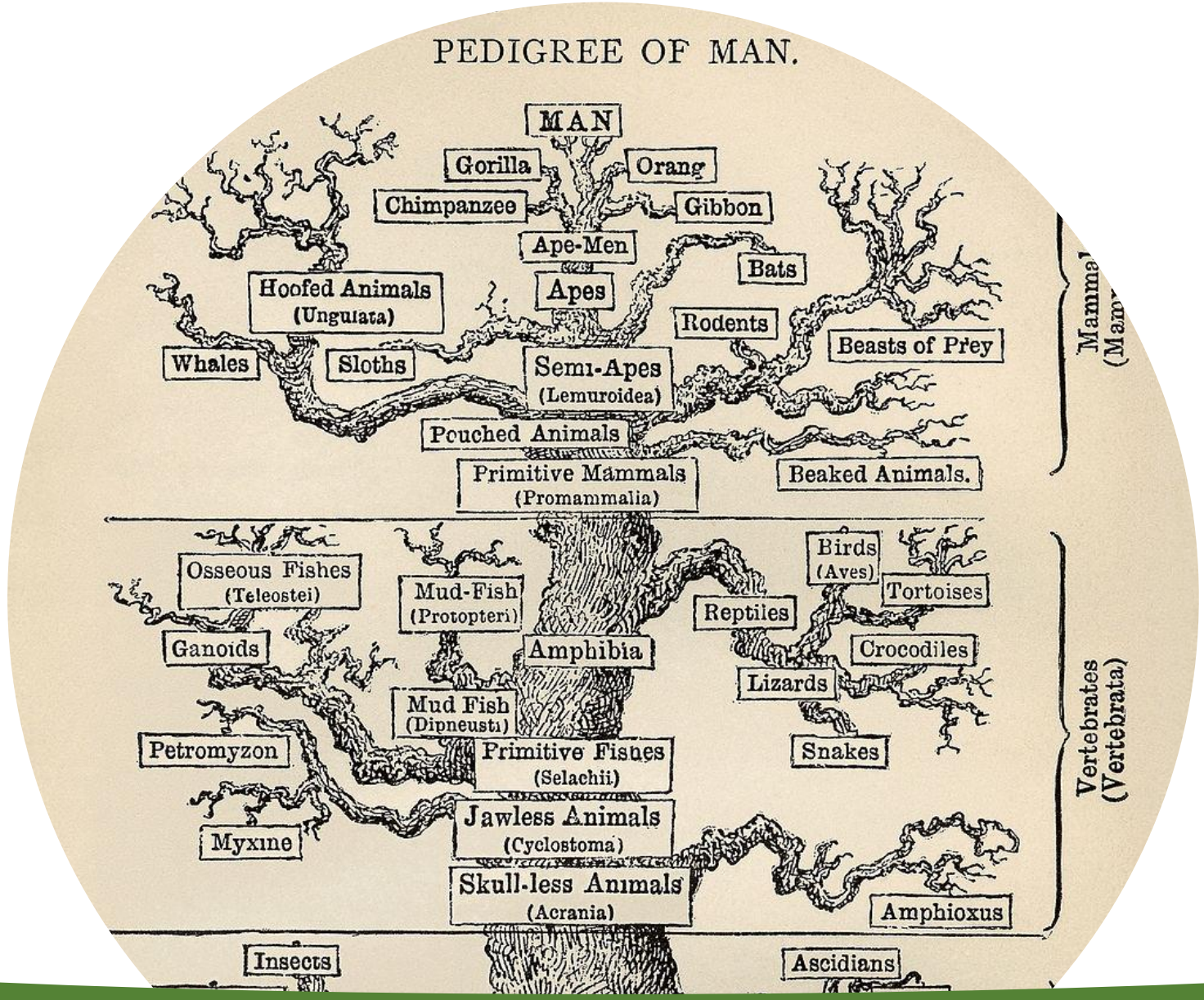
- Shortly introduce yourself!
- What do you expect from/what do you want to learn in this workshop?
- Do you have already a (phylogenetic) problem or data at hand that you would like to understand?

Workshop Content

- Basic phylogenetics
- Sequence alignments
- Standard methods of algorithmic phylogenetics
- Bayesian evolutionary analysis
- Phylodynamics

Objectives Of The Day

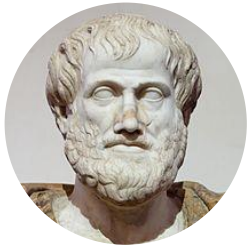
- You are familiar with (or have recalled) the most important basics and terms of phylogenetics.
- You are aware of the existing problems and limitations of phylogenetic analysis and phylogenetic trees.
- You know what an optimal sequence alignment is, why it is absolutely essential for phylogenetics and why it is nevertheless almost impossible to ever get one.



The (Abridged) History Of Phylogenetics

- A Brief Journey From Aristoteles to Walter Fitch. -

The (Abridged) History Of Phylogenetics



Aristoteles
„Historia animalium“

~400 BC

Charles Darwin
„On the Origin of Species“

1859

Walter M. Fitch
„Construction of Phylogenetic Trees“

1967

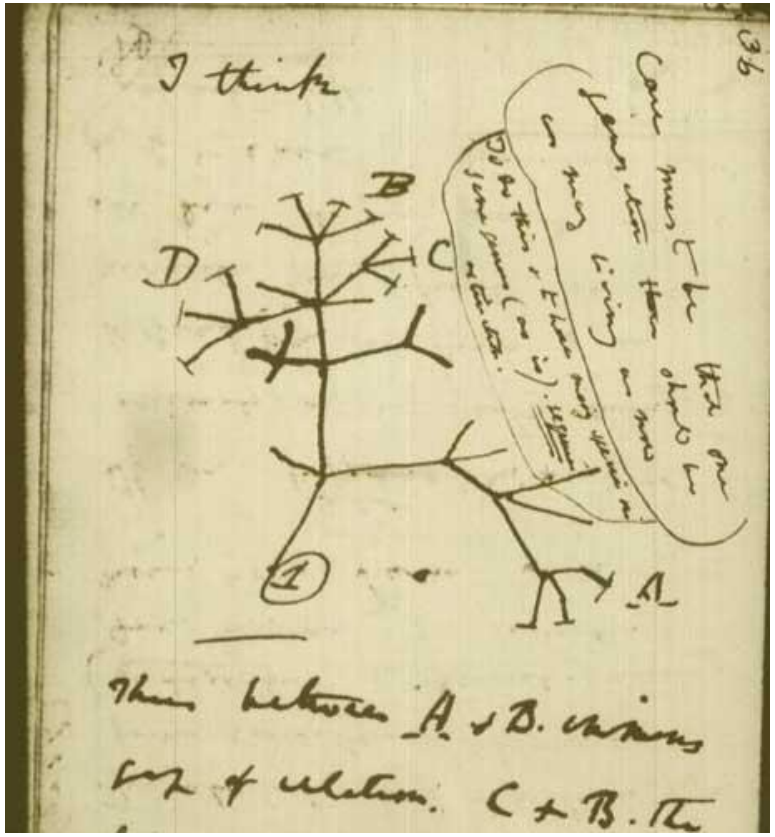
1753

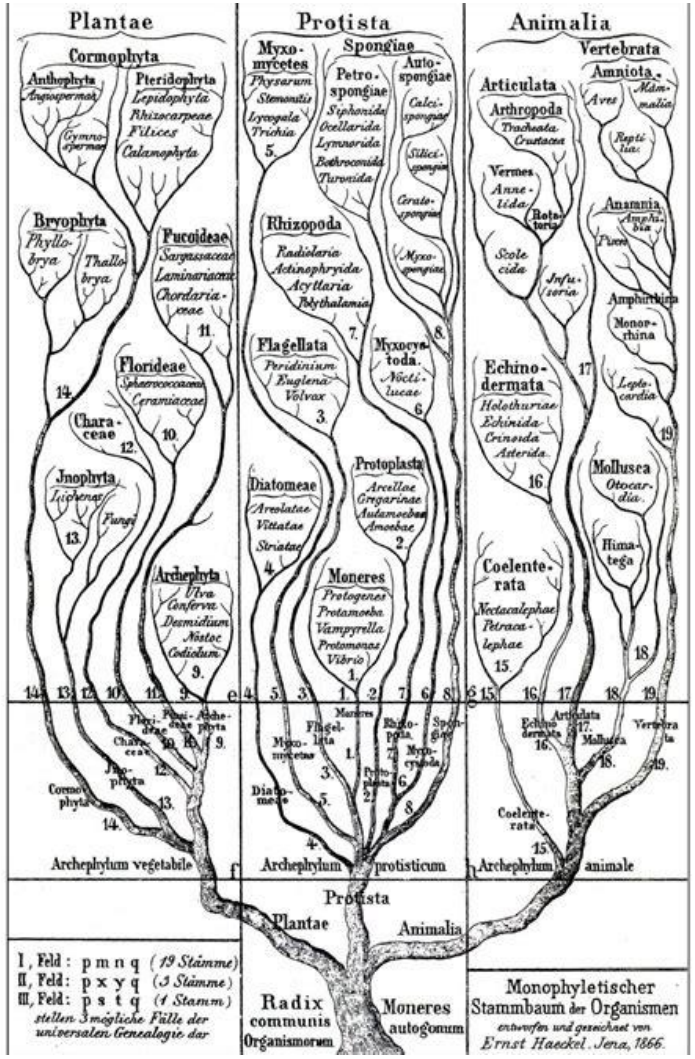
Carl von Linné
„Systema Naturae“

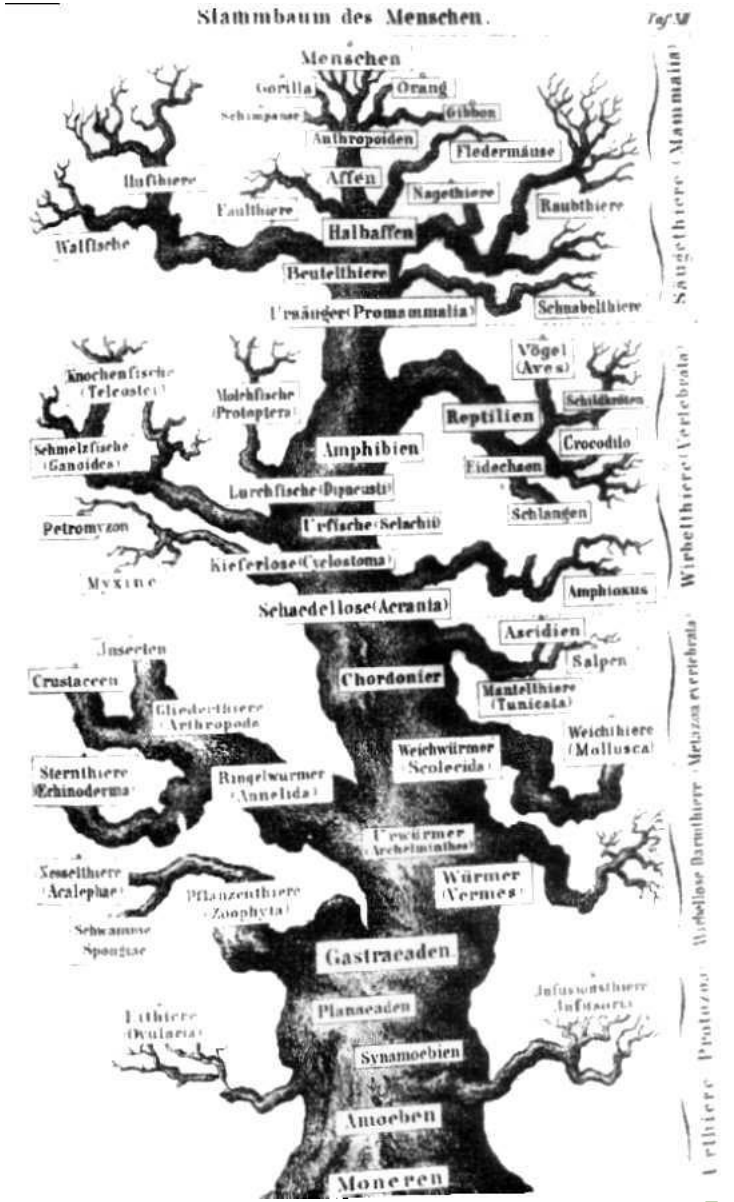
1874

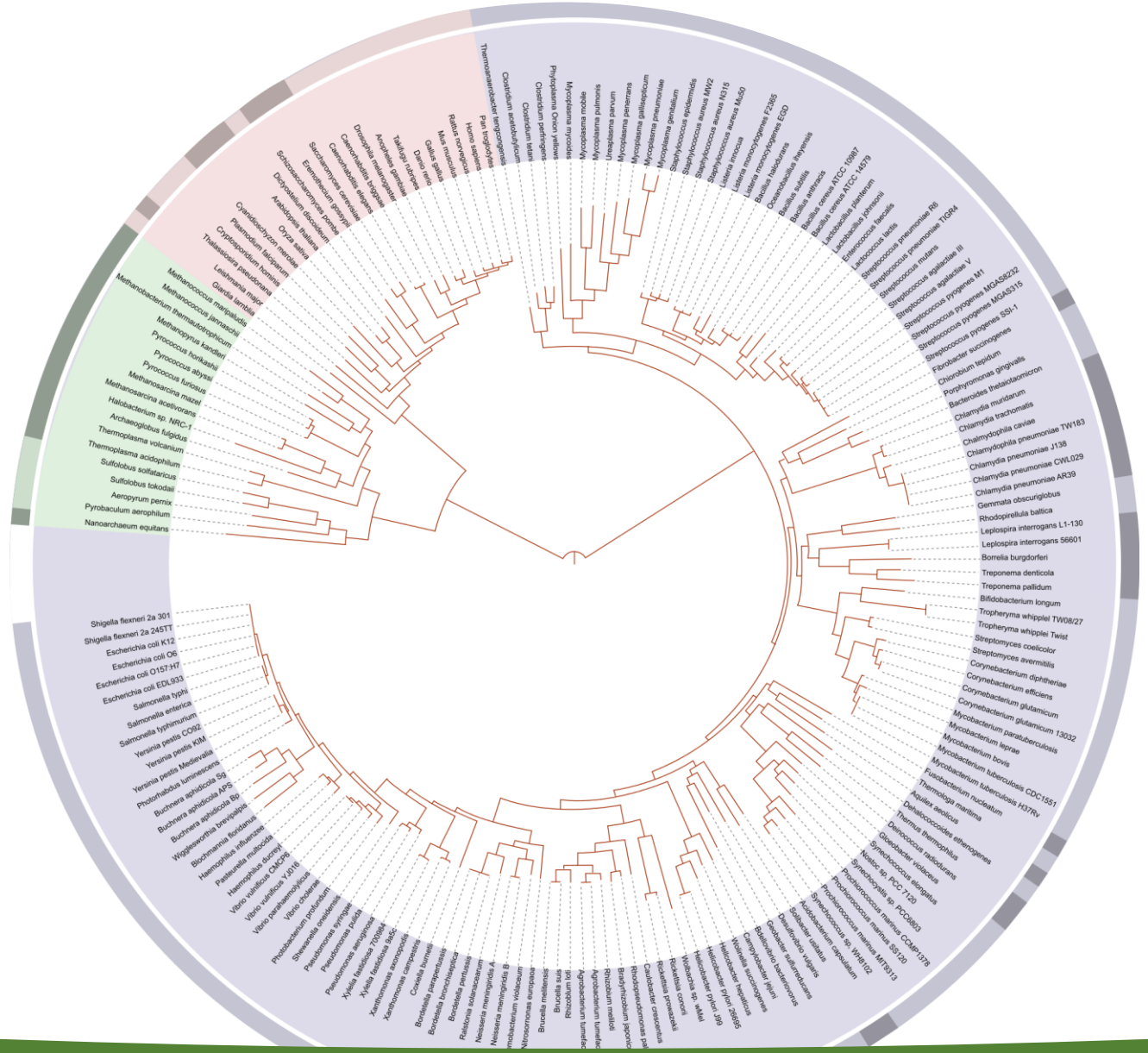
Ernst Häckel
„Anthropogenie“

The (Abridged) History Of Phylogenetics





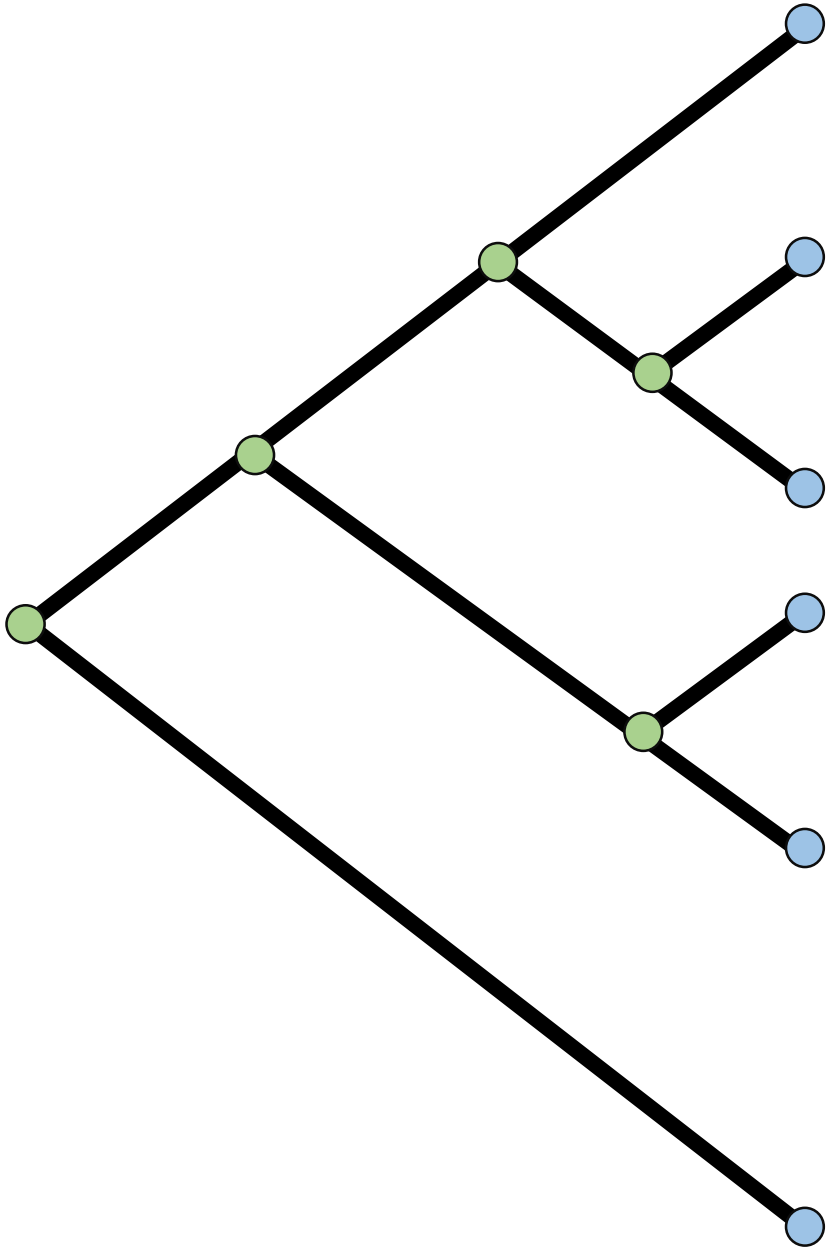


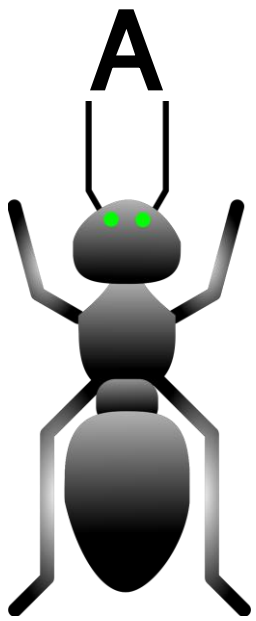


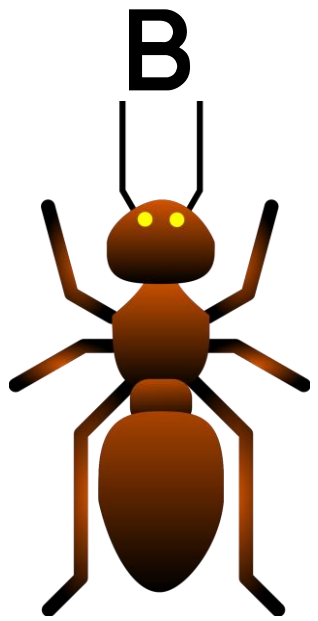
Phylogenetics

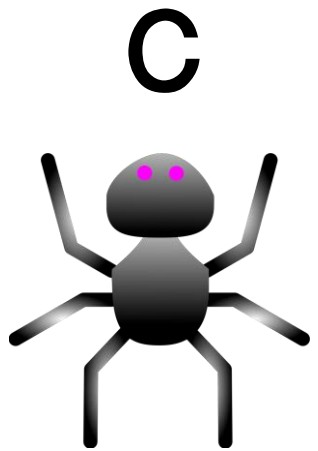
- "Nothing in biology makes sense except in the light of evolution." (T. Dobzhansky) -

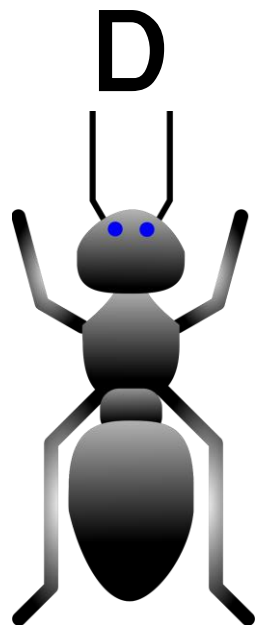
Phylogenetics: An Introductory Example

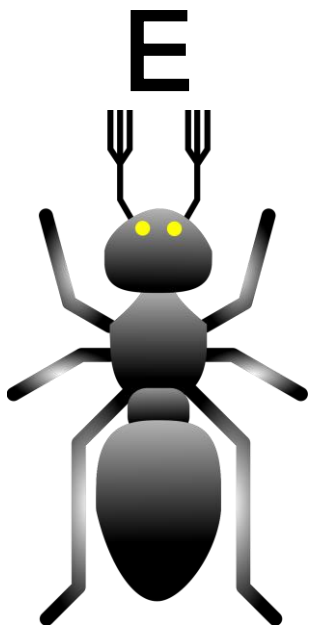


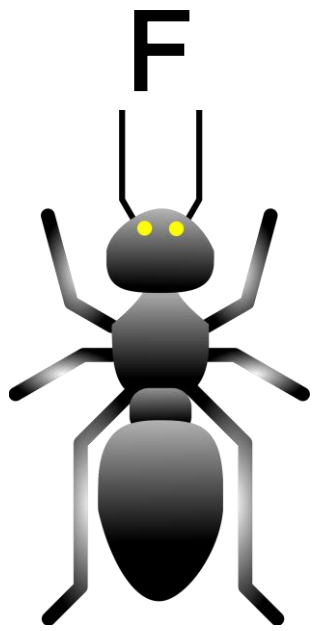




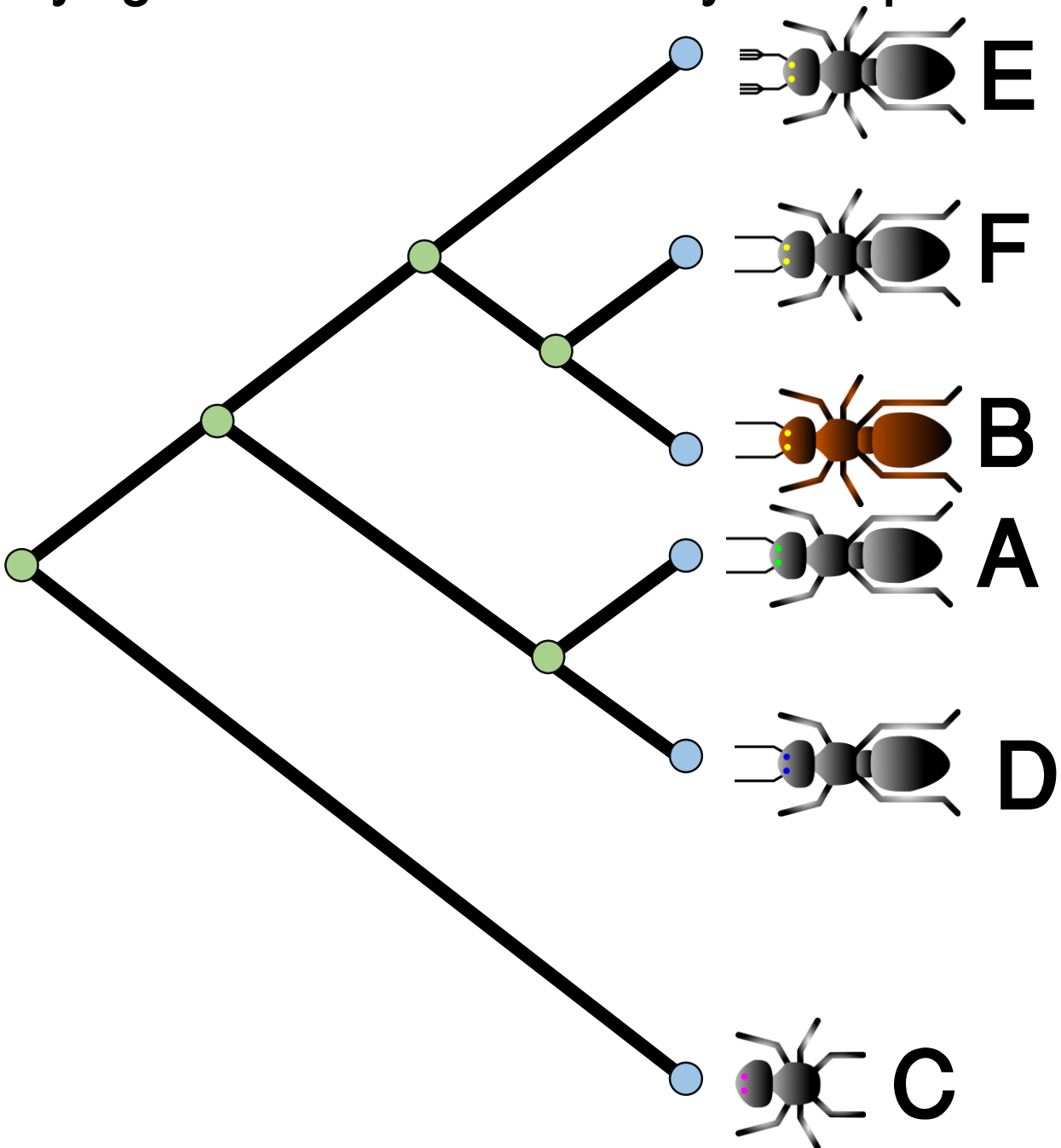




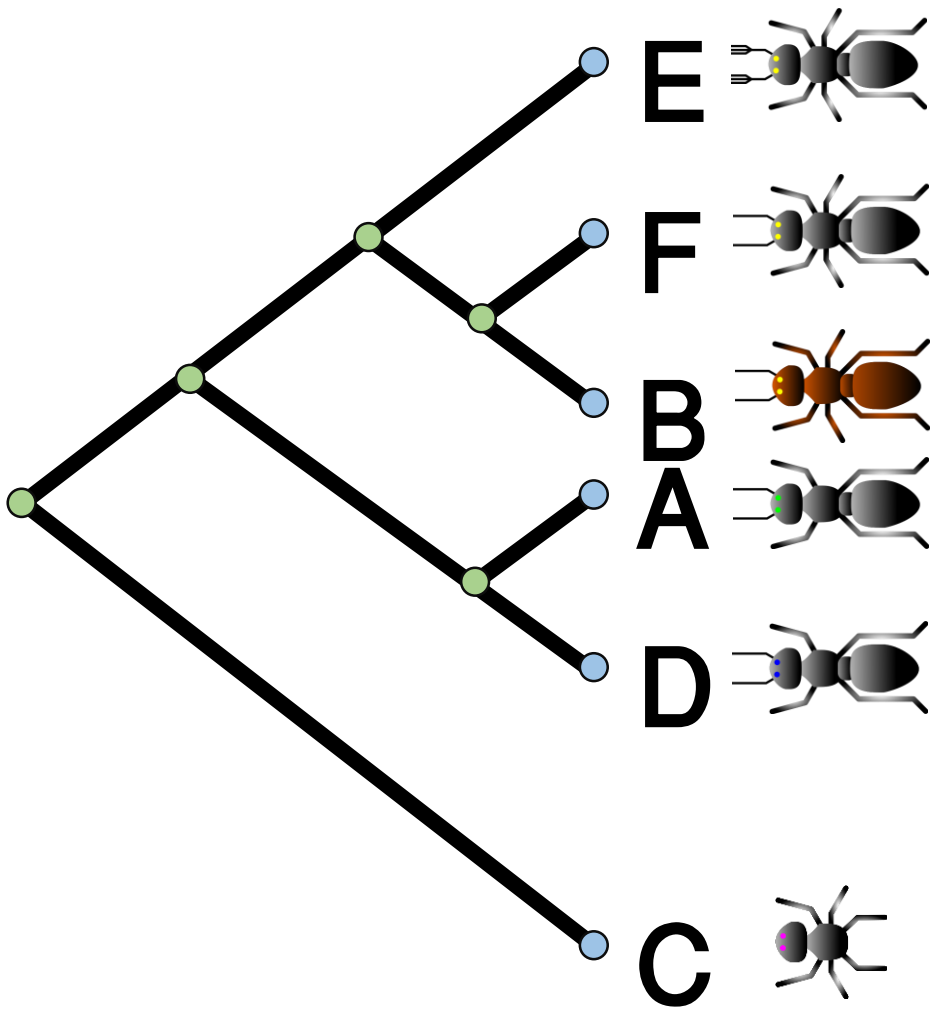




Phylogenetics: An Introductory Example



Phylogenetics: An Introductory Example



feature matrix

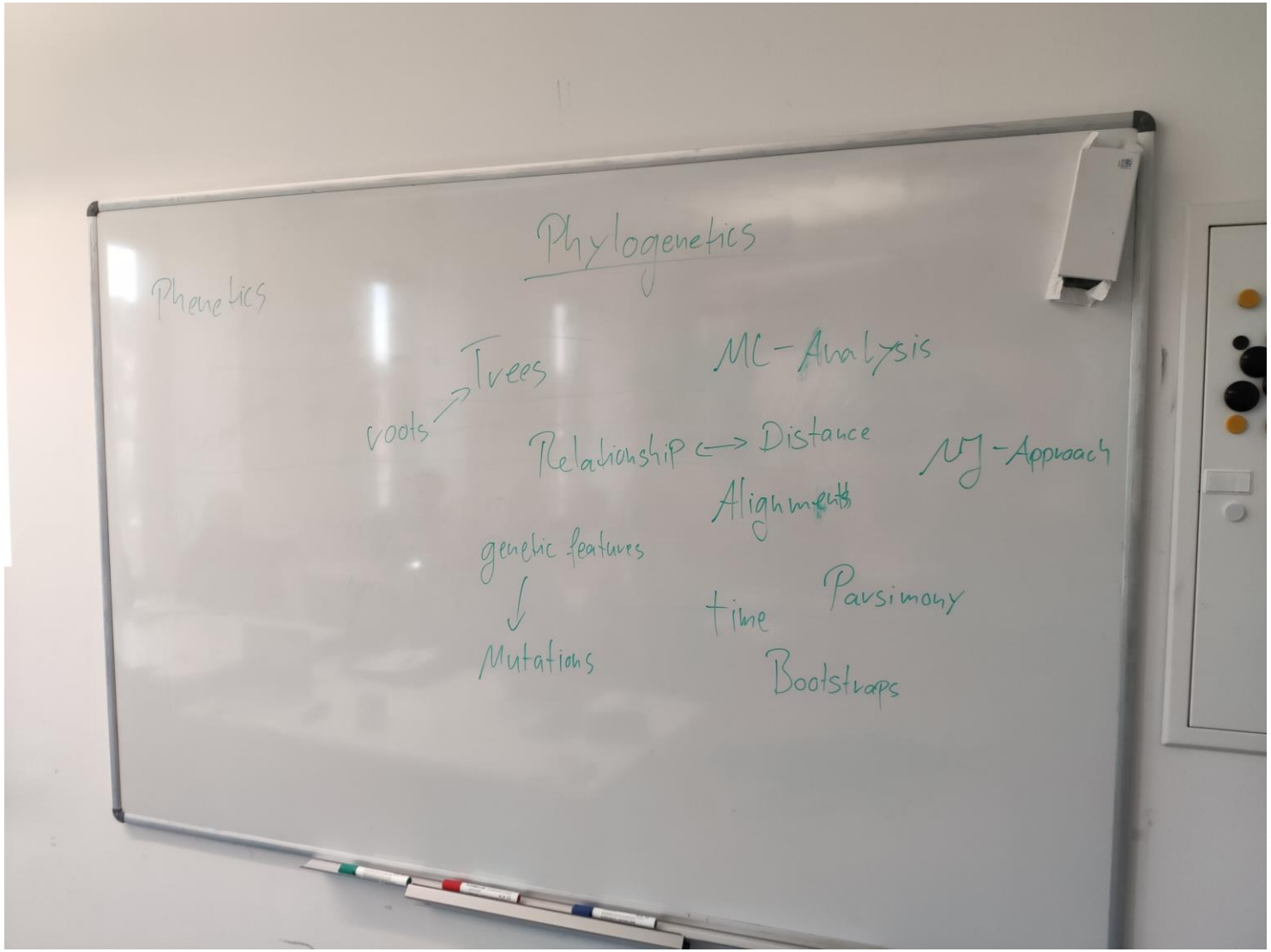
Species	No. of legs	No. of antennae	Color	Eyes	Segments
A	4	2			3
B	6	2			3
C	6	0			1
D	4	2			3
E	6	6			3
F	6	2			3

distance matrix

	A	B	C	D	E	F
A	0	3	4	1	3	2
B		0	4	3	2	1
C			0	4	3	3
D				0	3	1
E					0	1
F						0

Phylogenetics: Definition

- What do you understand by or what do you associate with the term phylogeny?



Phylogenetics: Definition

- What do you understand by or what do you associate with the term phylogeny?

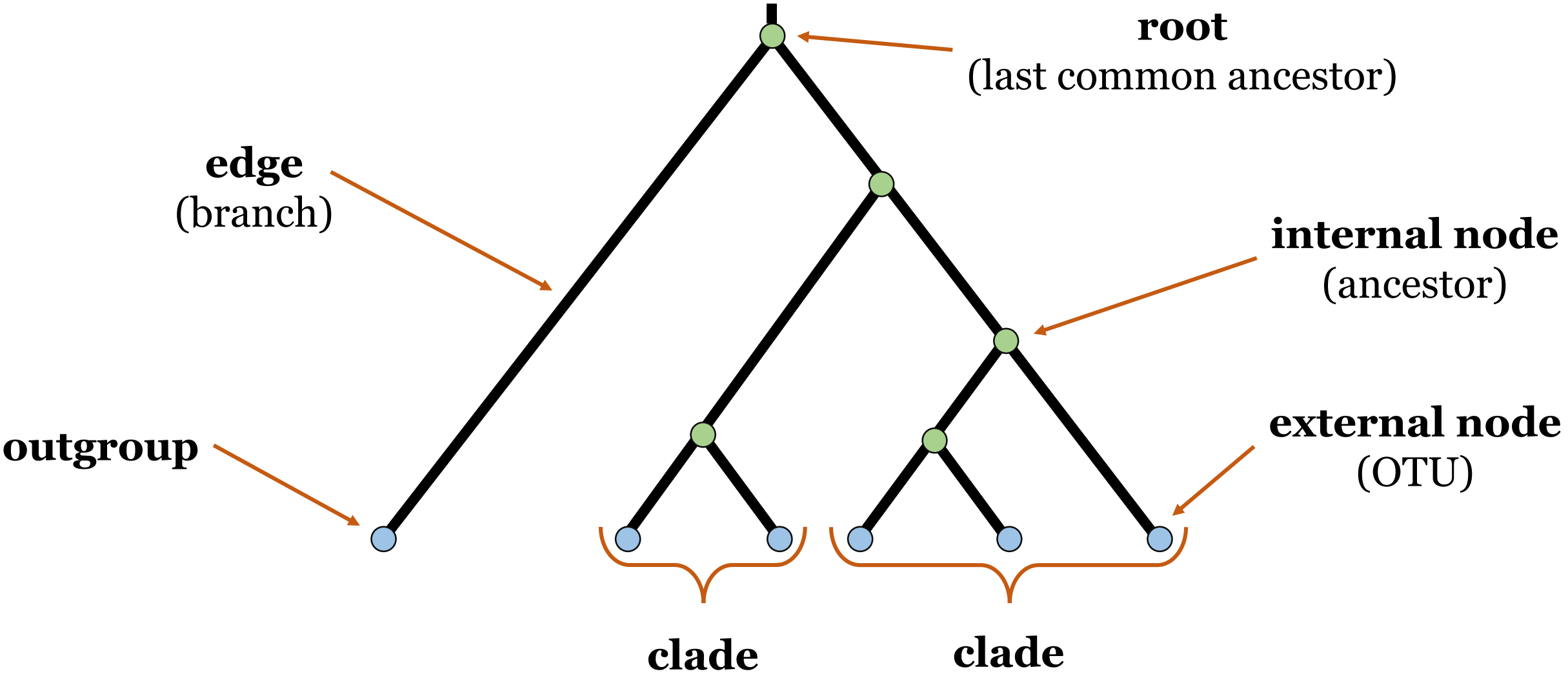
- ~ is the study of the evolutionary history and relationships among biological species or other entities (e.g. individuals, species, groups of organisms, populations).

- These relationships are discovered through phylogenetic inference methods that evaluate observed heritable similarities and differences in their physical or genetic characteristics (such as DNA sequences or morphology) under a model of evolution of these characteristics.

- The result of these analyses is a phylogenetic tree — a diagrammatic hypothesis about the history of the evolutionary relationships of a group of organisms.

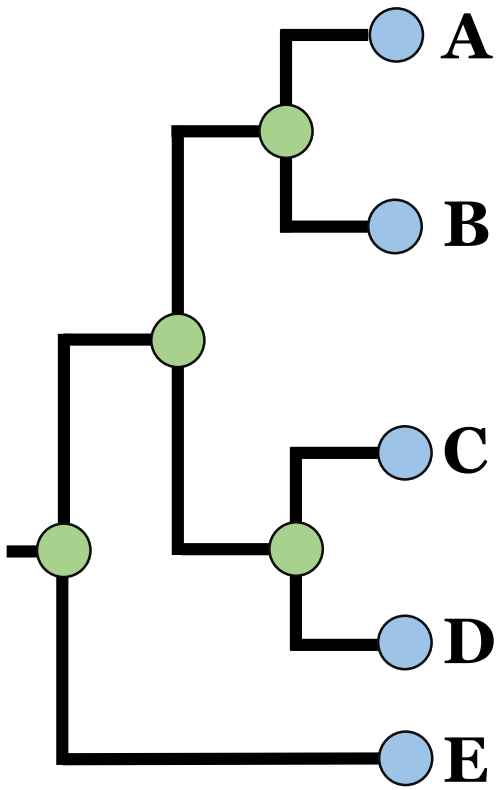
- The goal of phylogenetics is to reconstruct the tree of live.

Phylogenetics: Basic Terms



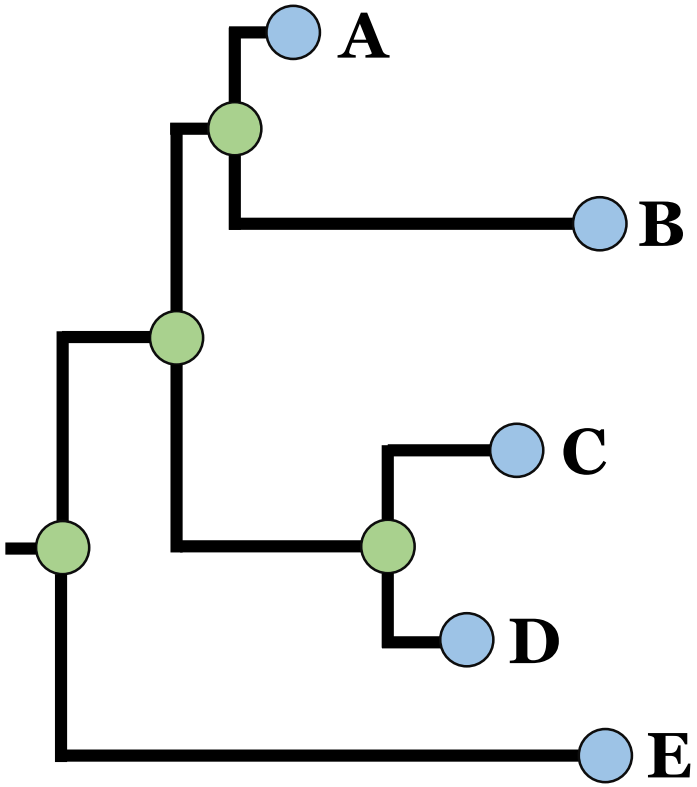
Phylogenetics: Different Trees, Different Meanings

Cladogram



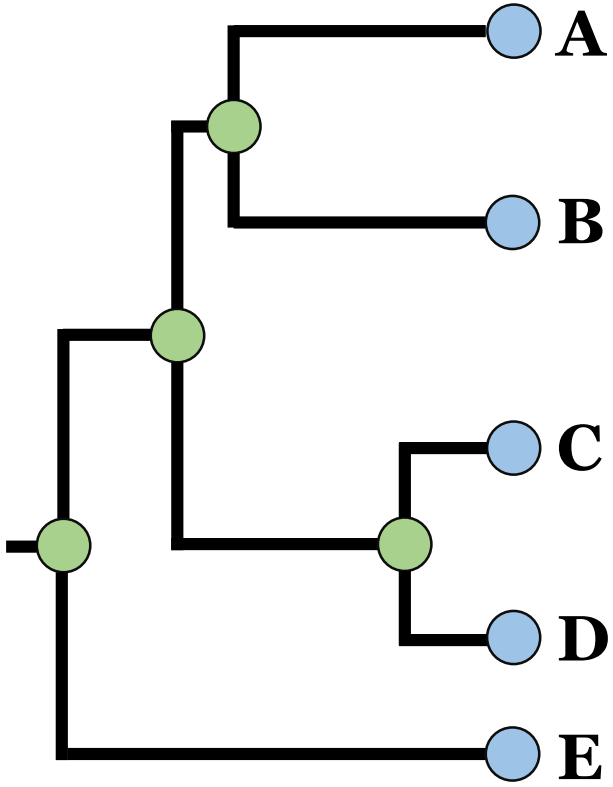
no meaning

Phylogram



change of characteristics

Chronogram

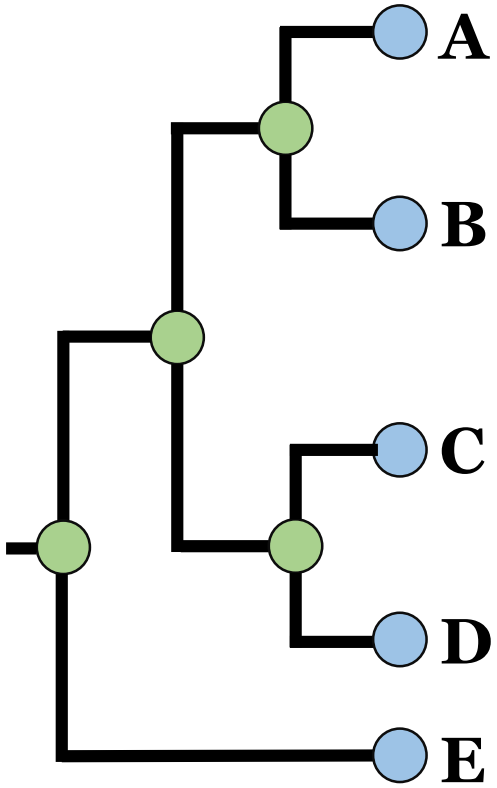


course of time

(molecular clock hypothesis → workshop day 3)

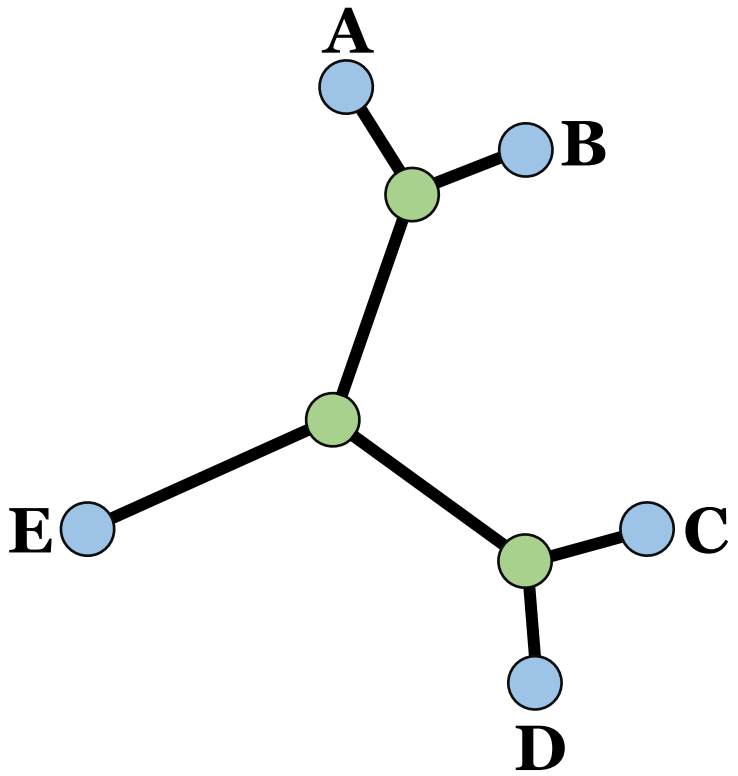
Phylogenetics: Different Trees, Different Meanings

rooted tree



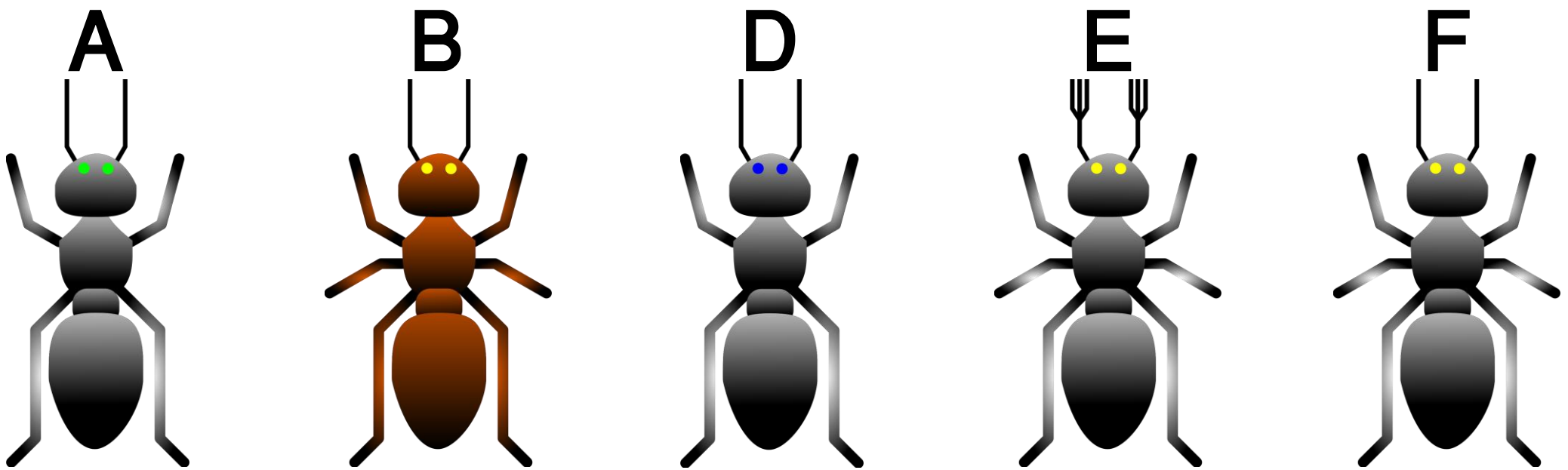
specifies a direction of ancestry

unrooted tree

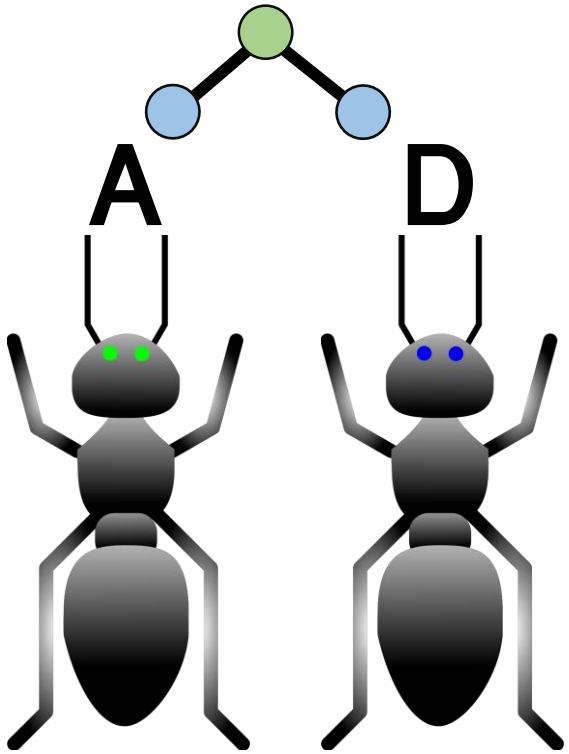


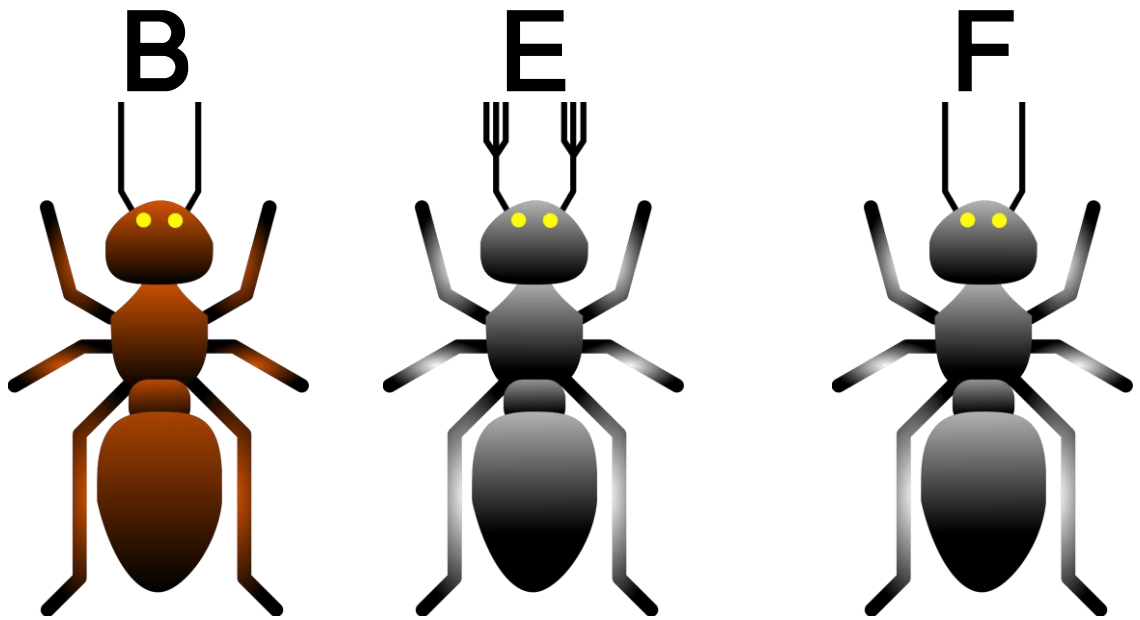
only illustrates relationship between taxa

Phylogenetics: Different Trees, Different Meanings

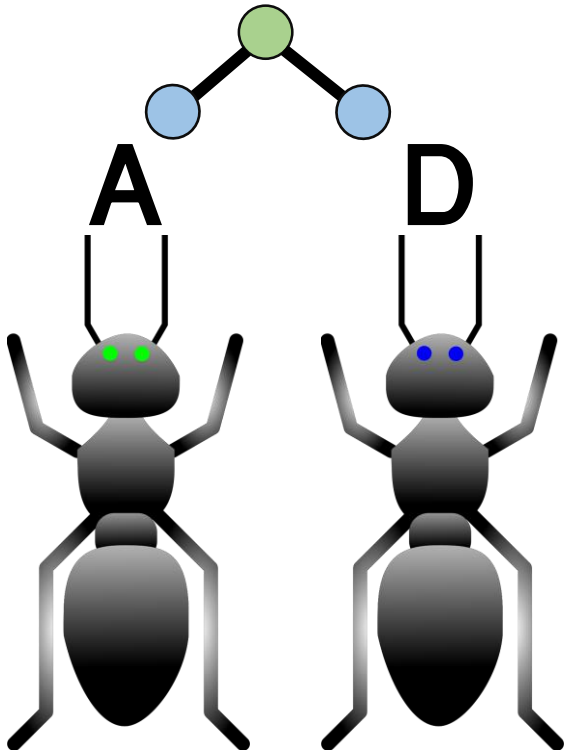


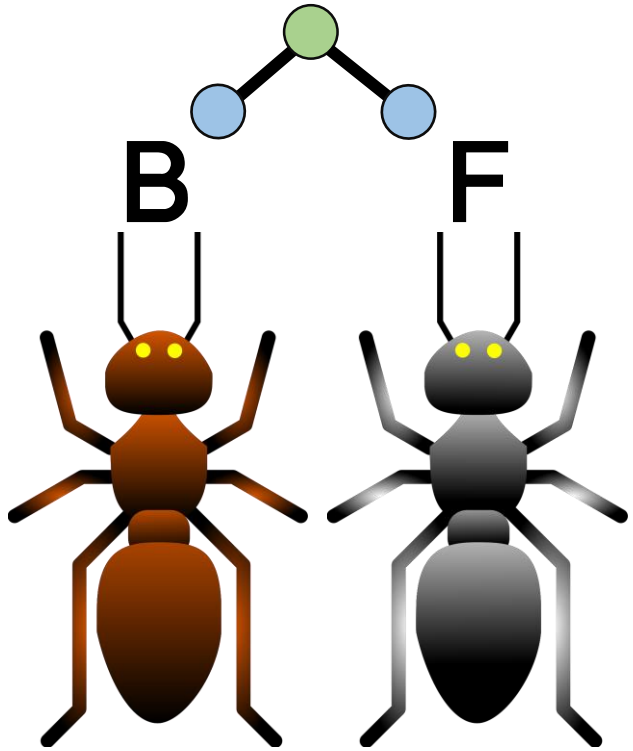
Phylogenetics: Different Trees, Different Meanings

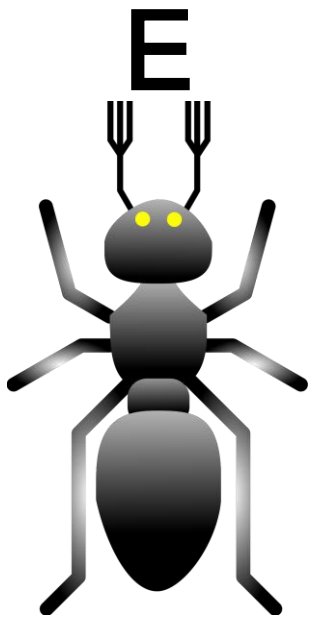




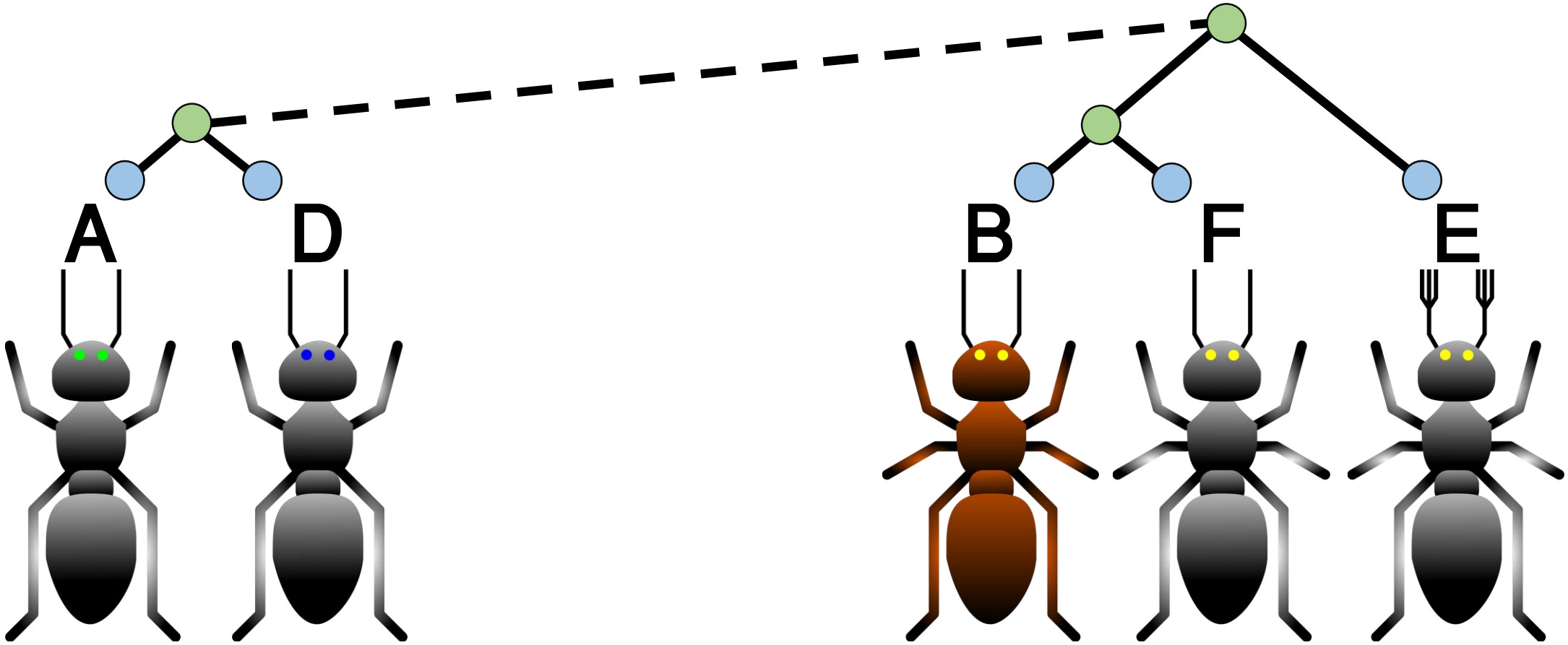
Phylogenetics: Different Trees, Different Meanings



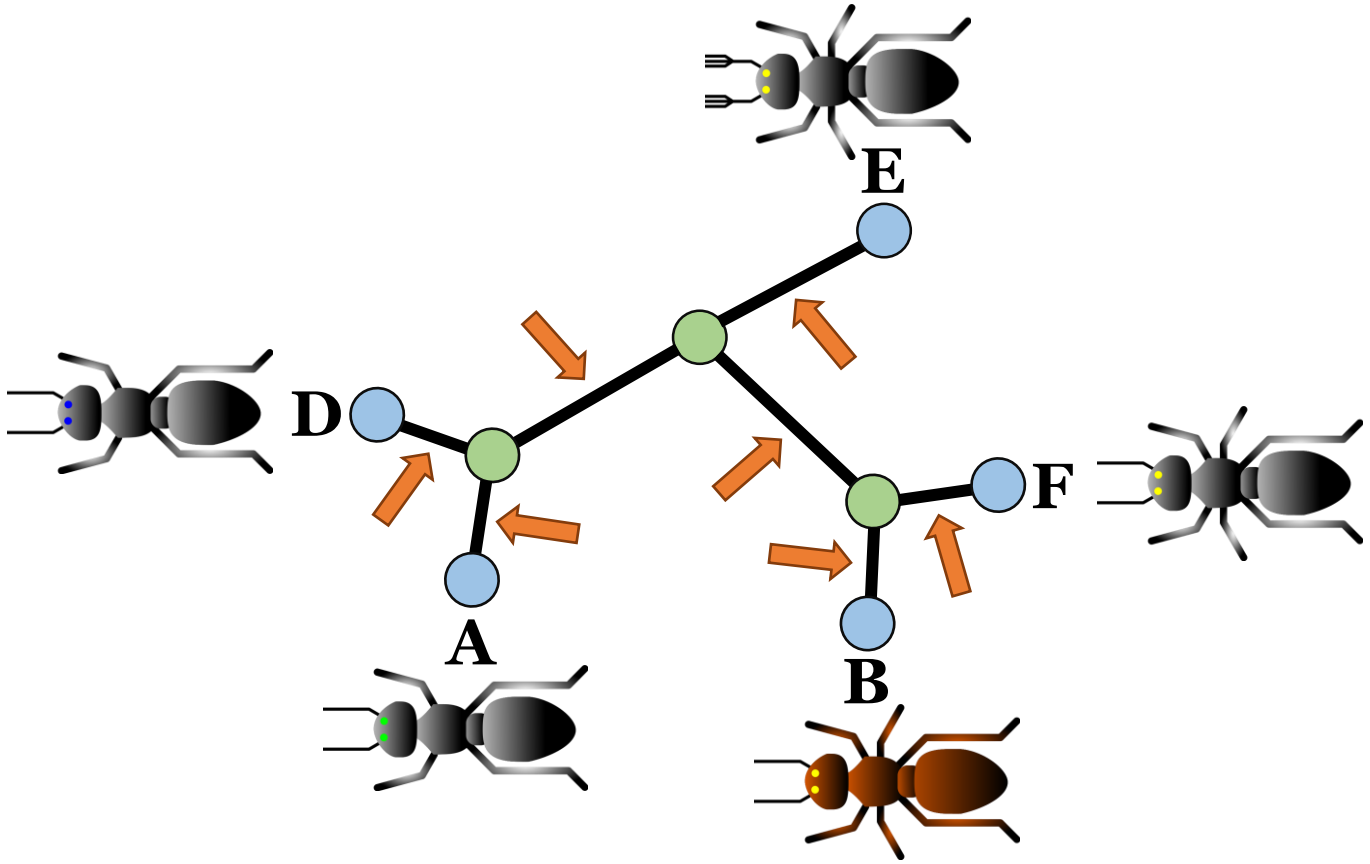




Phylogenetics: Different Trees, Different Meanings

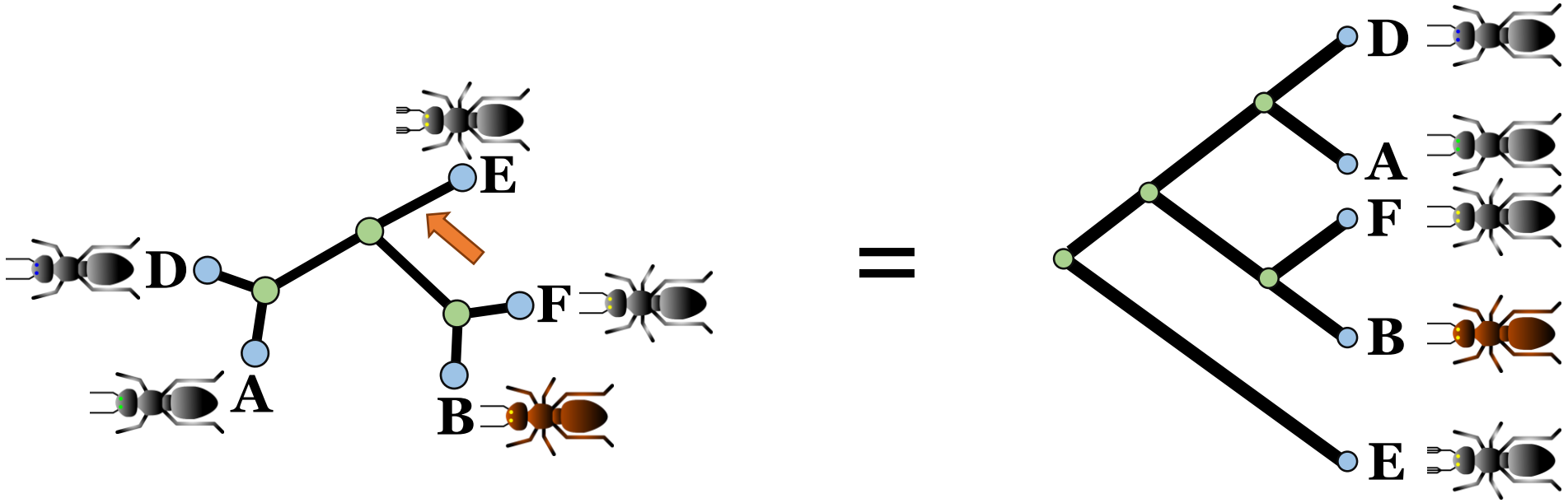


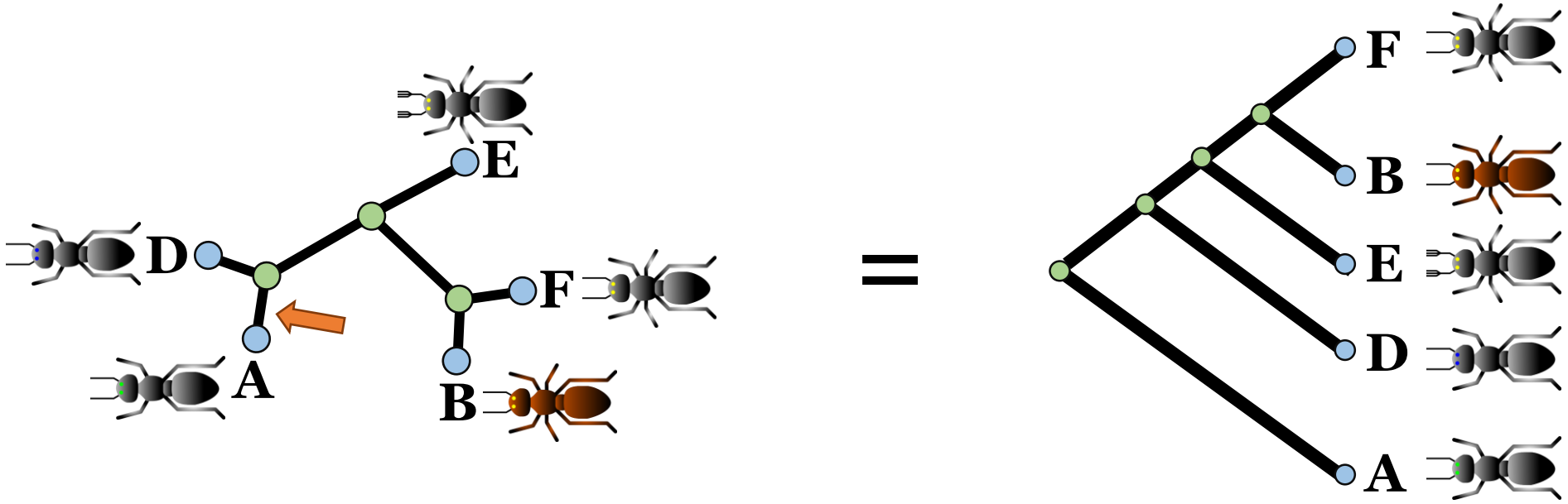
Phylogenetics: Different Trees, Different Meanings



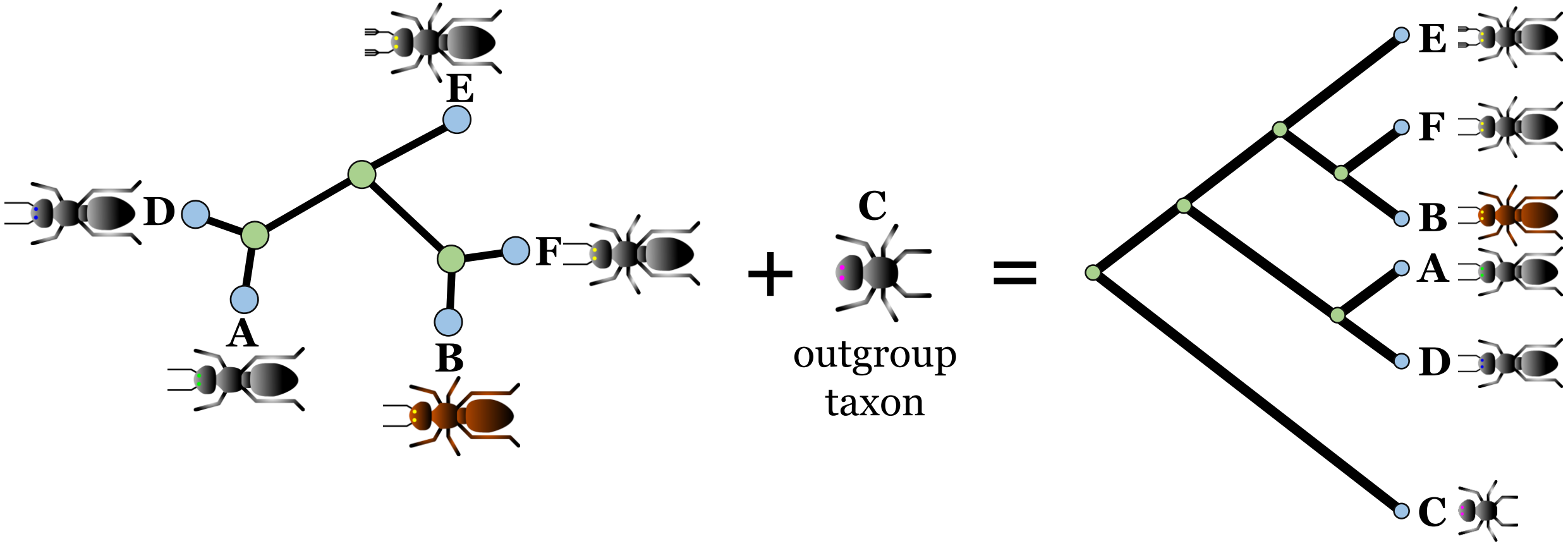
What did the last common ancestor look like?
Where to insert the root?

Phylogenetics: Different Trees, Different Meanings



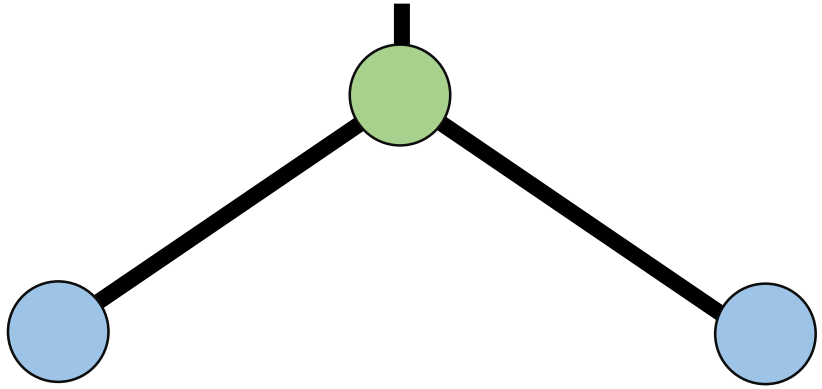


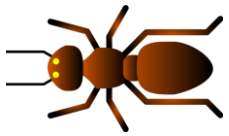
Phylogenetics: Different Trees, Different Meanings



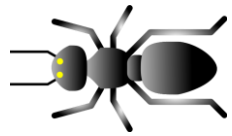
An outgroup is a taxon which is closely related to the group studied, but which is known to be clearly outside the clade of the studied taxa.

Phylogenetics: What Is Our Data?





phenotypic traits



Gene A | Gene B | Gene C | Gene D

gene groups

Gene C | Gene B | Gene A | Gene D

Gene A

genes

Gene A

MLEVNAMEYPPAHDIRGIS*

protein sequences

MYPPAHDIRGISLEVNAME*

ATGACGTCGATCTAG

nucleotide sequences

ATGAGTCCGTCATAG

Phylogenetics: When to use which type of data?

- In modern molecular phylogenetics, sets of conserved sequences are often used for generating phylogenetic trees, as it can be assumed that organisms with similar sequences are closely related.

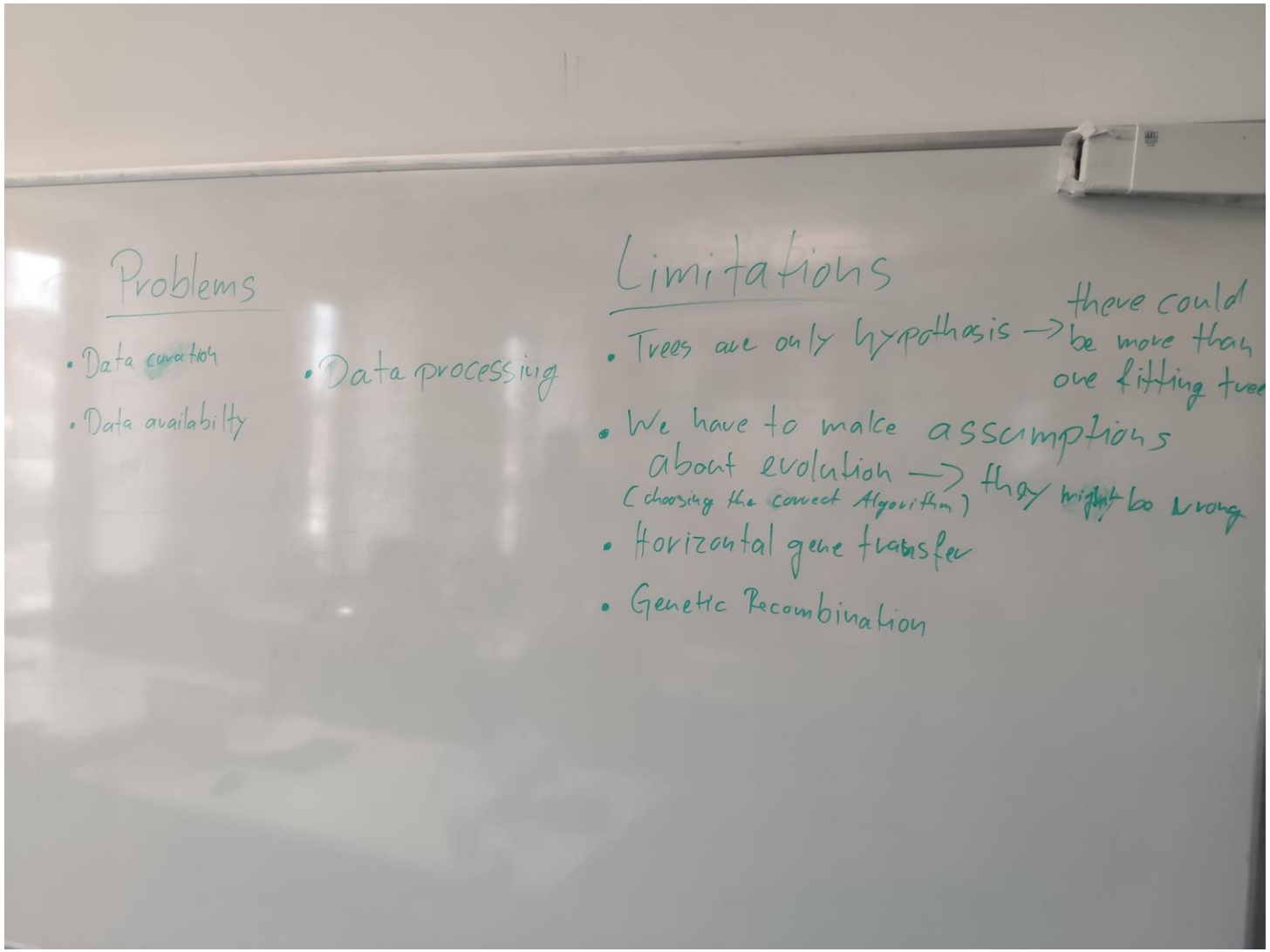
- The choice of sequences may vary depending on the taxonomic scope of the study:
  - Highly conserved genes like the 16S RNA (or other rRNA sequences) are used to reconstruct deep phylogenetic relationships.
  - Housekeeping genes of a specific clade, can be used to study species relationships.

Phylogenetics: When to use which type of data?

- rRNA and tRNA sequences
- presence of specific genes in the respective genome
- genomic rearrangements
- presence of specific introns
- amino acid sequence of proteins
- nucleotide sequence of protein-coding genes
- secondary structures of non-coding genes
- intron sequences
- intergenic sequences
- single mutations (e.g. SNPs)

Phylogenetics: Limitations And Problems

- What limitations and problems still exist in modern molecular phylogenetics today?



Phylogenetics: Limitations And Problems

- What limitations and problems still exist in modern molecular phylogenetics today?

- Phylogenetic trees are only hypothesis and don't have to be true.
  (→ gene trees versus species trees)

Problems:

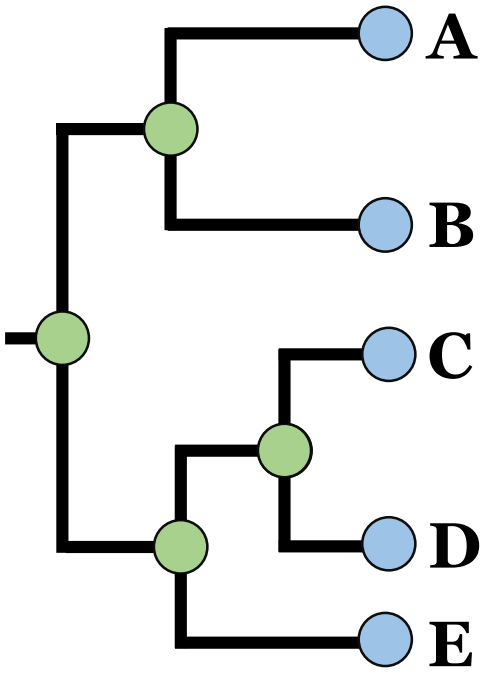
- Noisy data
- Missing data
- Depending on the taxonomic scope, wrong characteristic chosen for phylogenetic analysis

Limitations:

- Hybridization events
- Horizontal/lateral gene transfer
- Genetic recombination
- Convergent evolution

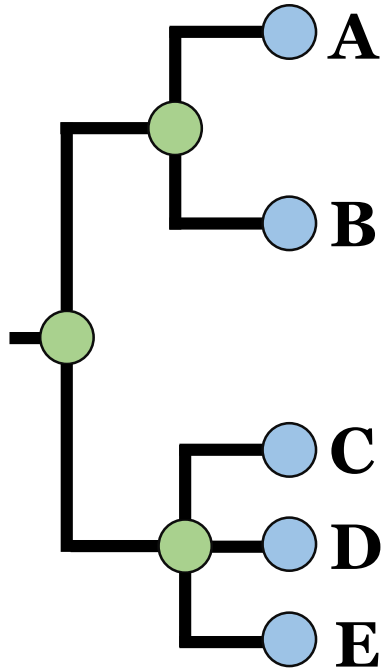
Phylogenetics: Dichotomy Versus Polytomy

dichotomic tree



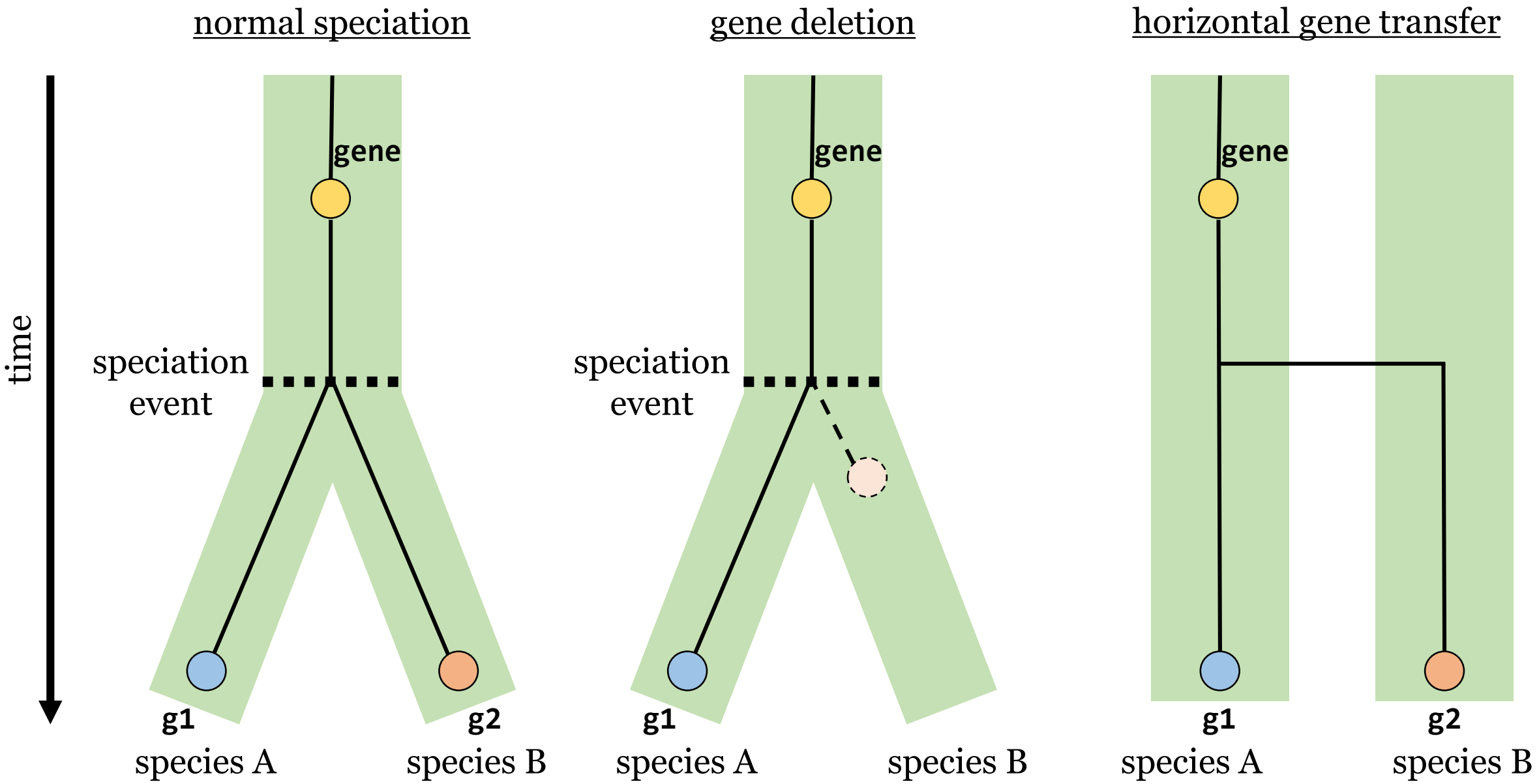
fully resolved
binary tree

polytomic tree

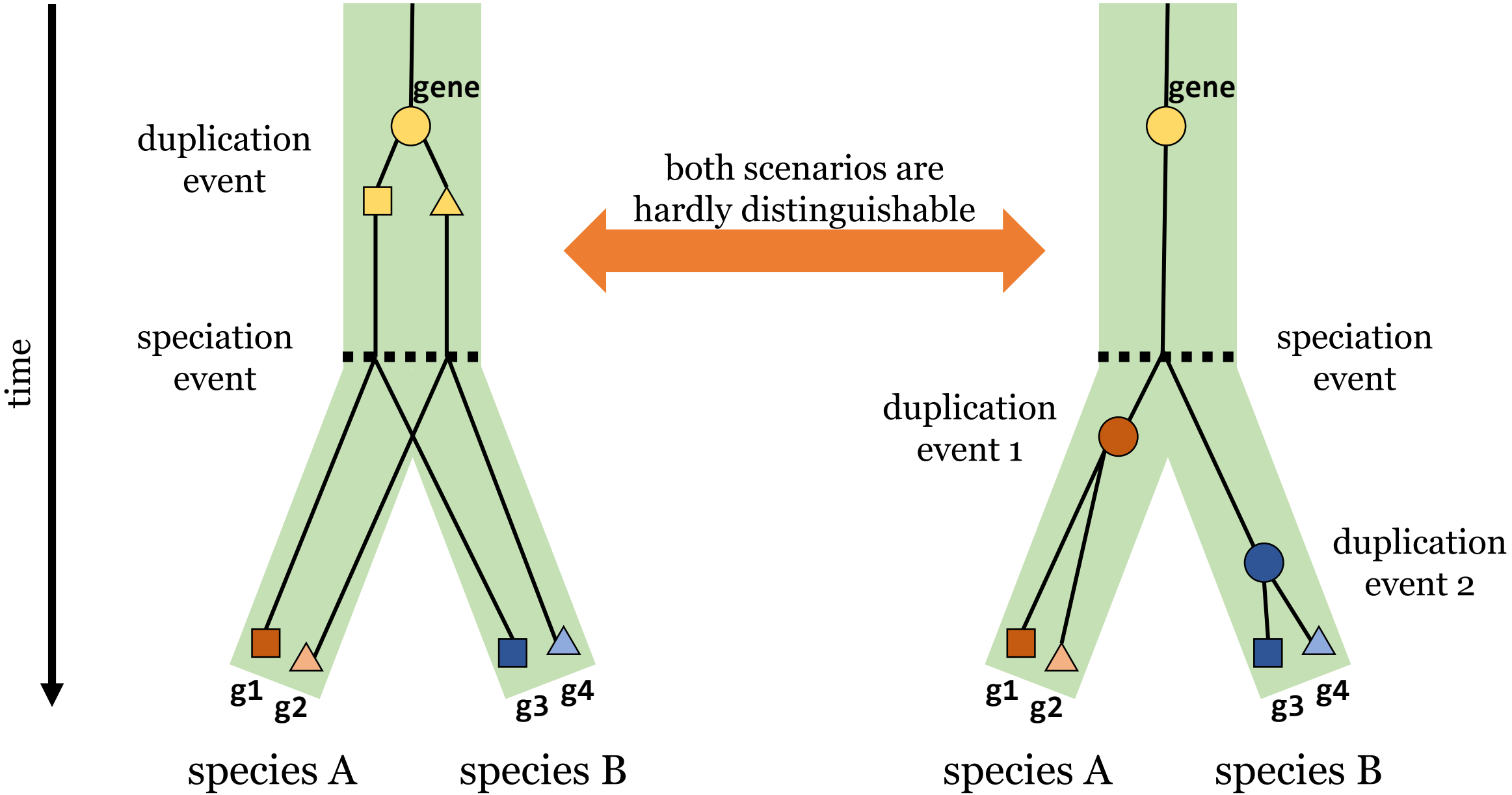


partially resolved
non-binary tree

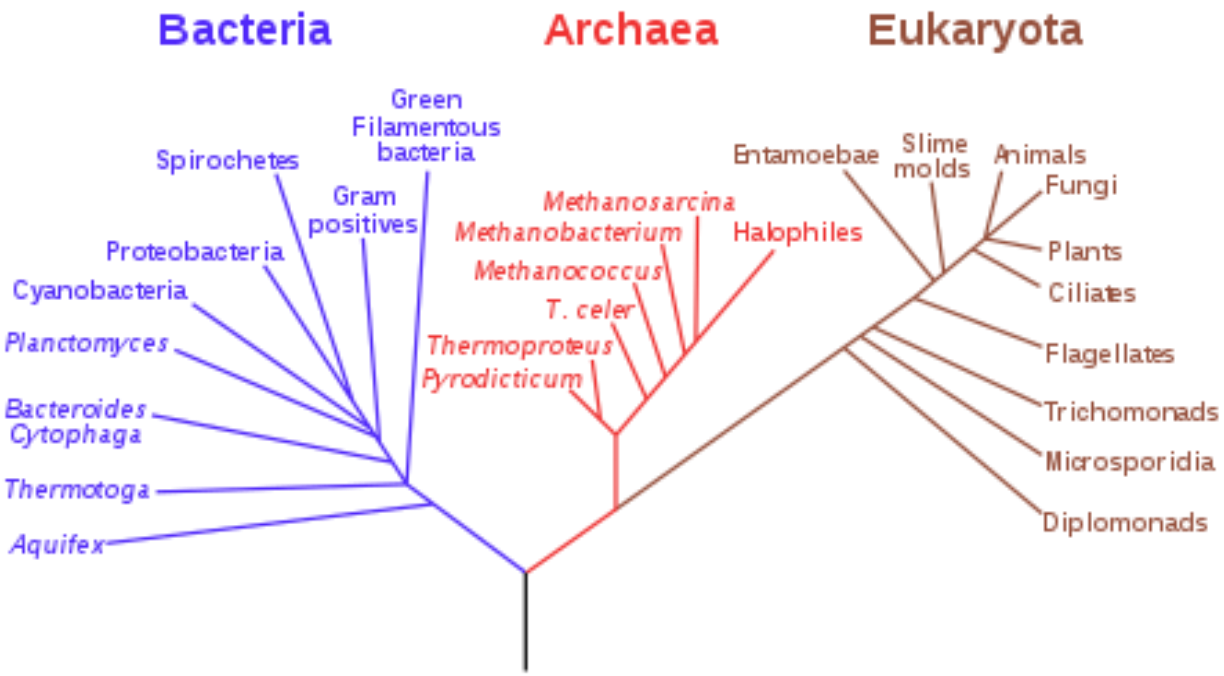
Phylogenetics: Gene Trees Versus Species Trees

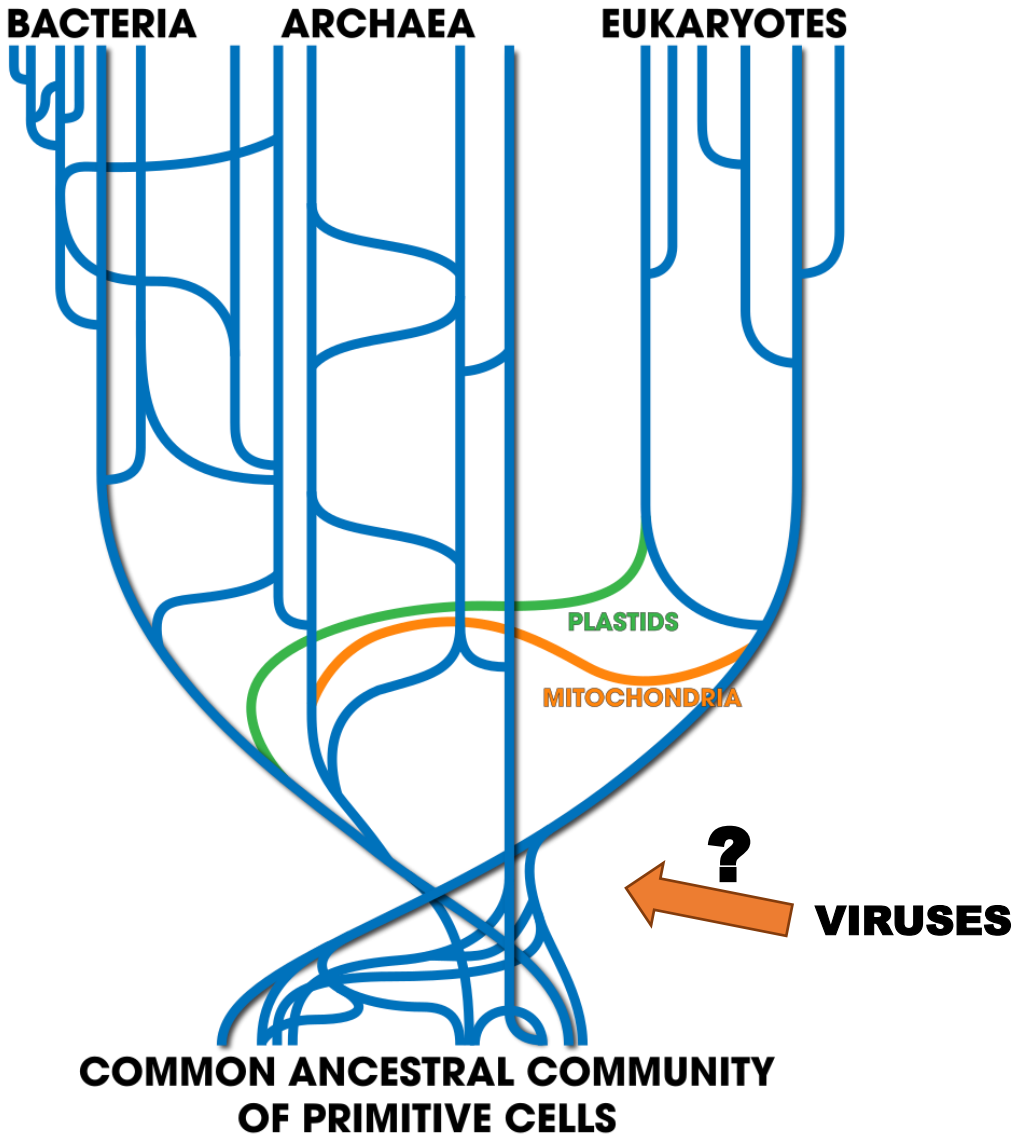


Phylogenetics: Gene Trees Versus Species Trees



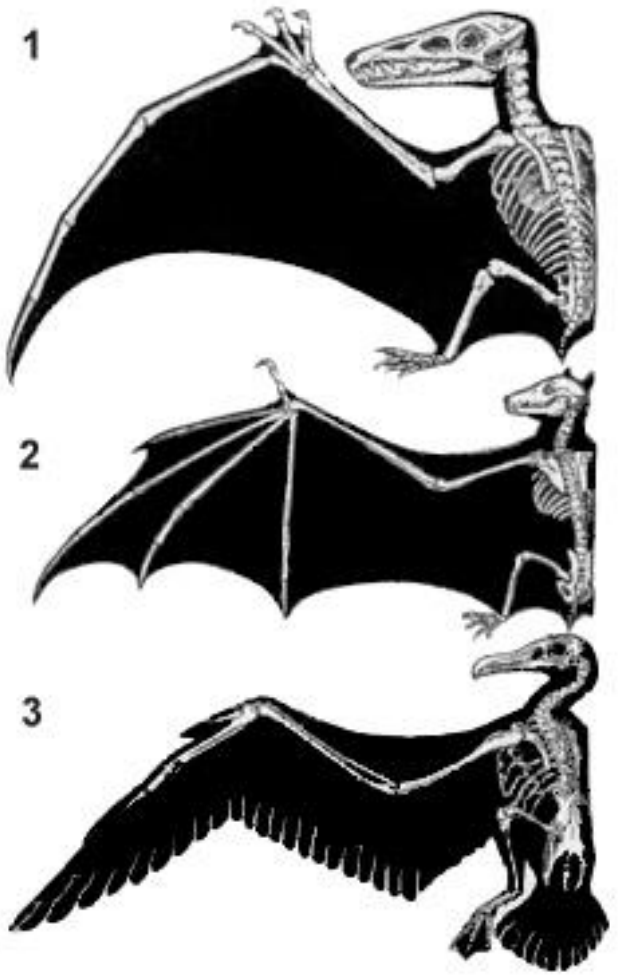
Phylogenetics: Horizontal Gene Transfer

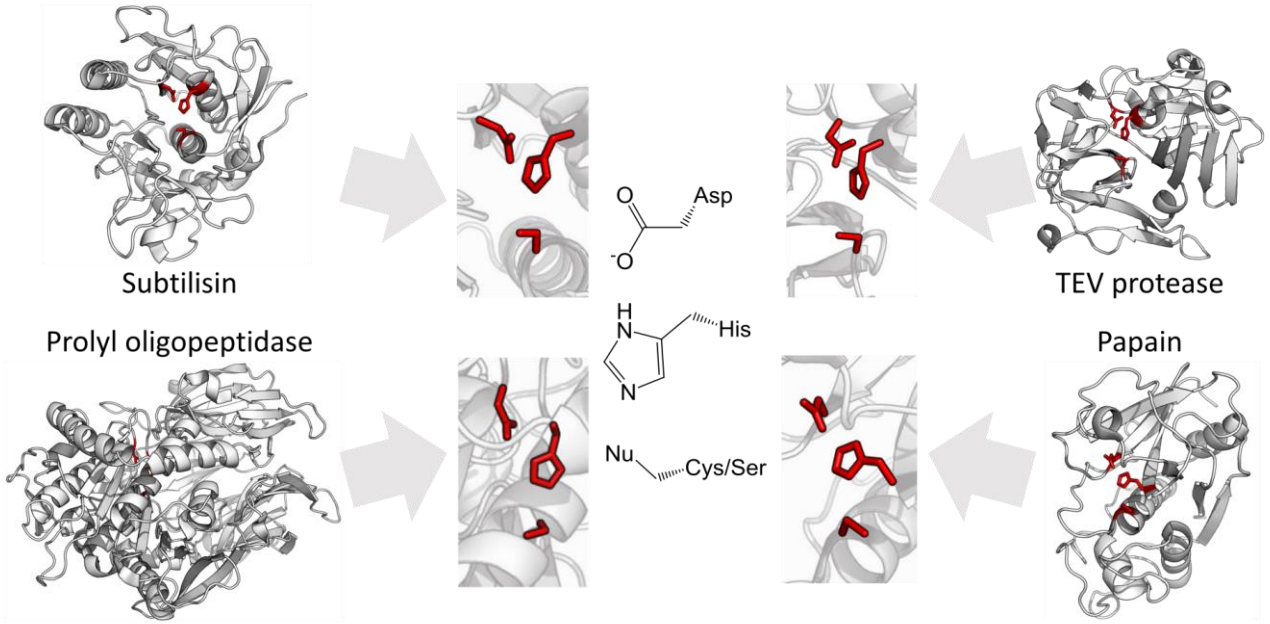




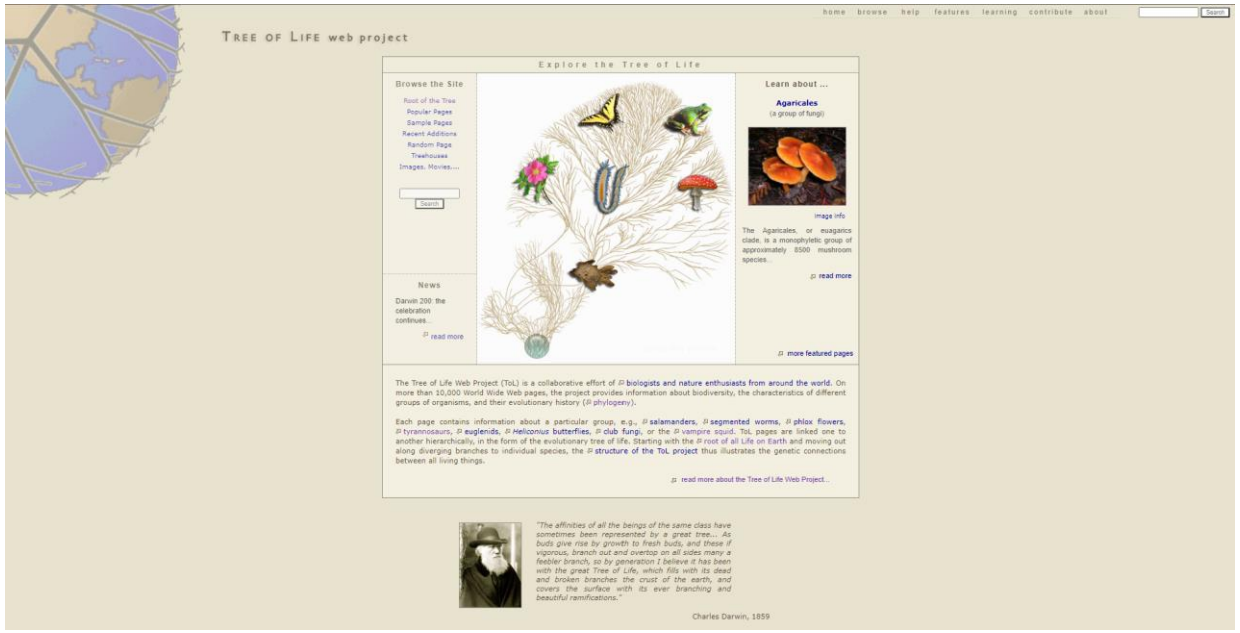
Smets, B., Barkay, T. Horizontal gene transfer: perspectives at a crossroads of scientific disciplines. *Nat Rev Microbiol* **3,** 675–678 (2005) doi:10.1038/nrmicro1253

Phylogenetics: Convergent Evolution

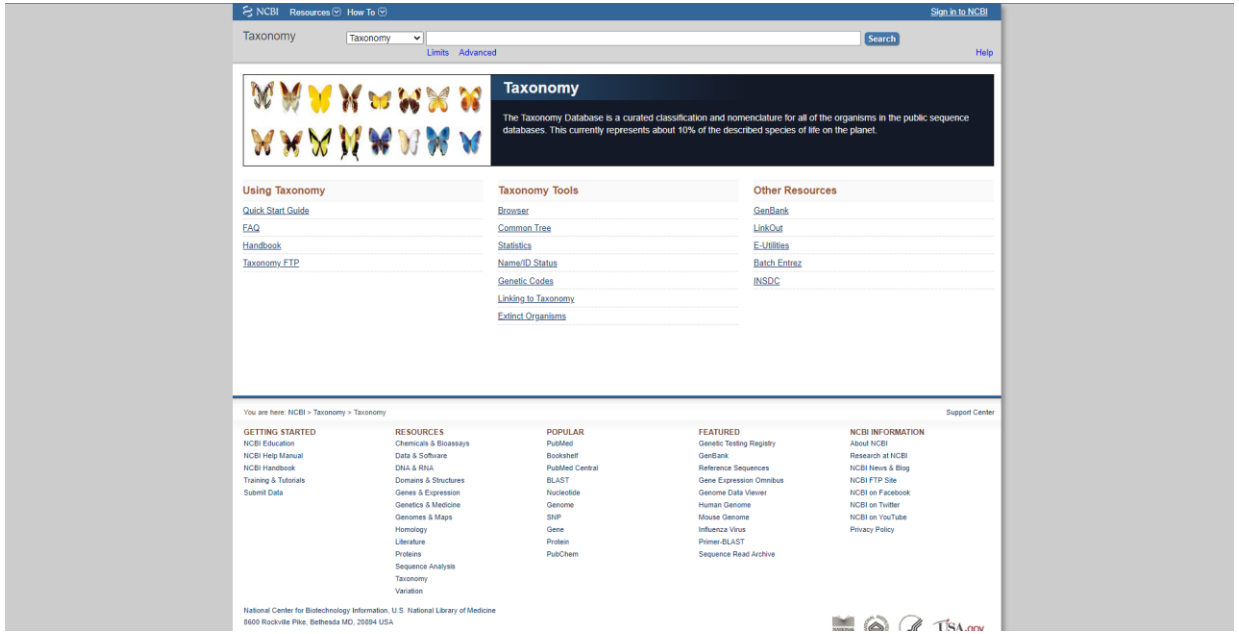




Useful resources



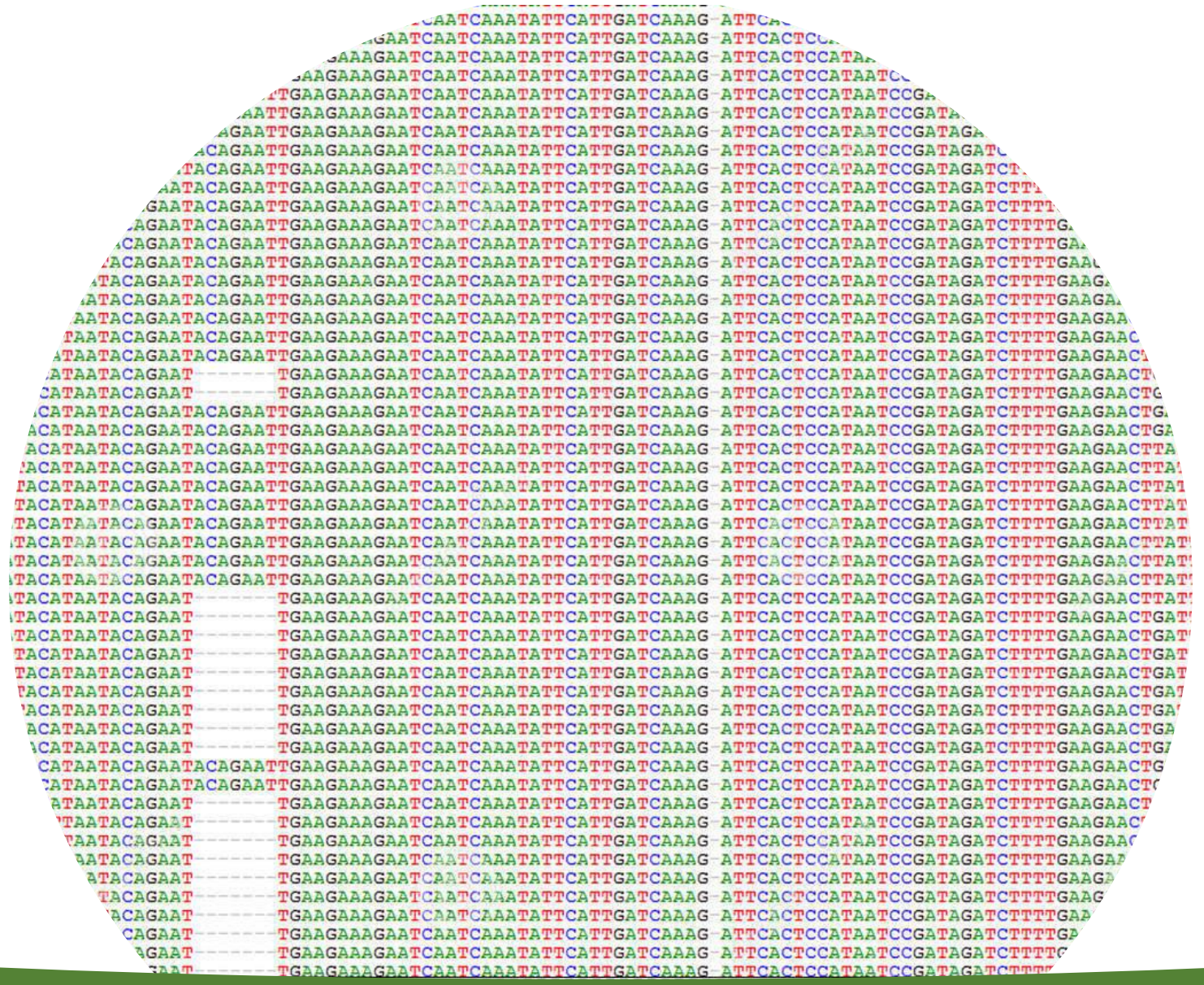
https://www.tolweb.org



https://www.ncbi.nlm.nih.gov/taxonomy

Short Summary

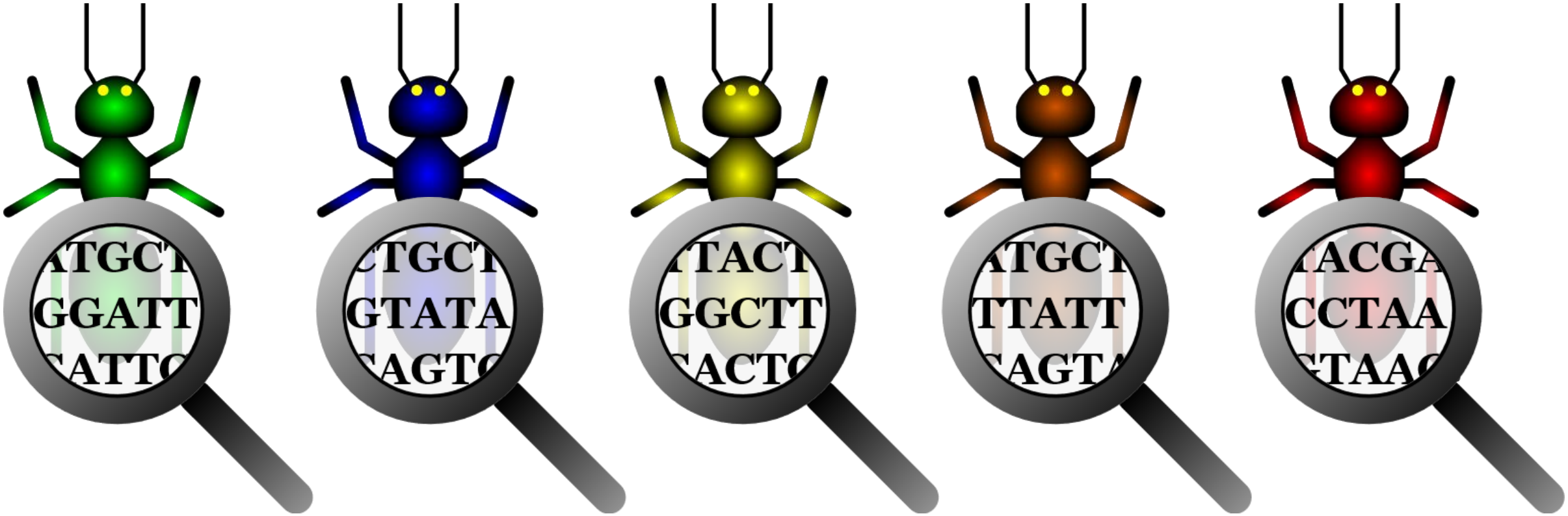
- Phylogenetics tries to model the evolution of different taxa (molecules, species, groups, populations) by evaluating heritable characteristics.
- Different kinds of phylogenetic trees exist, but they are always only hypothesis about evolution.
- Different types of data are useful for different phylogenetic scopes.
- There are still many problems and limitations for modern phylogenetic analysis.
- Gene trees and species trees can disagree with each other and can still be true on their own.



Sequence Alignments

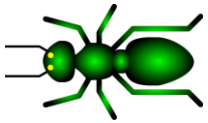
- The cornerstone of algorithmic phylogenetics. -

Sequence Alignments: The All-Purpose Tool Of Molecular Biology



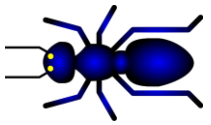
Sequence Alignments: Definition

- A pairwise sequence alignment is an arrangement of two DNA, RNA, or protein sequences, that reflects their similarity to each other (possibly due to an evolutionary relationship).



sequence 1

ATCAGCCTAATCGCATCATGC



sequence 2

TTGCAGCAATCGCGTCATGCC

sequence 1 AT-CAGCCTAATCGCATCATGC-

sequence 2 TTGCAGC--AATCGCGTCATGCC

Mismatch Gap Match

Sequence Alignments: Another Introductory Example

- An optimal sequence alignment maximizes the total number of matches, while minimizing the number of mismatches and gaps.

sequence 1 = GCATTTA

sequence 2 = ATTAGG

```
(GCATT-TA)
(--ATTAGG)
```

```
(GCATTT-A)
(--ATTAGG)
```

```
(GCATTTA-)
(--ATTAGG)
```

```
(GCATTTA--)
(--ATT-AGG)
```

```
(GCATTTA--)
(--AT-TAGG)
```

```
(GCATTTA--)
(--A-TTAGG)
```

Sequence Alignments: Optimal Alignments And Scoring Functions

- Scoring functions are needed to calculate the quality of a sequence alignment.
- Scoring functions are either cost functions or similarity functions.
- An alignment of sequences with the highest quality (i.e. the lowest cost score or the highest similarity score) is called an optimal alignment of these sequences.
- Different scoring functions can produce different alignments.

$$\delta = \begin{cases} 0, & a = b \\ 1, & a \neq b \end{cases} \qquad \forall a, b \in \{A, C, G, T, -\}$$

Sequence Alignments: Optimal Alignments And Scoring Functions

- How can we calculate the optimal pairwise alignment of two sequences?

- Naive approach:
  1. Produce all possible alignments of the given sequences.
  2. Calculate the quality of each alignment using a scoring function.
  3. Select the alignment(s) with the highest quality.

sequence 1 = T sequence 2 = G → $\binom{T}{G}$ $\binom{T-}{-G}$ $\binom{-T}{G-}$

sequence 1 = TT sequence 2 = TG → 13 possible alignments

Sequence Alignments: Optimal Alignments And Scoring Functions

- How can we calculate the optimal pairwise alignment of two sequences?

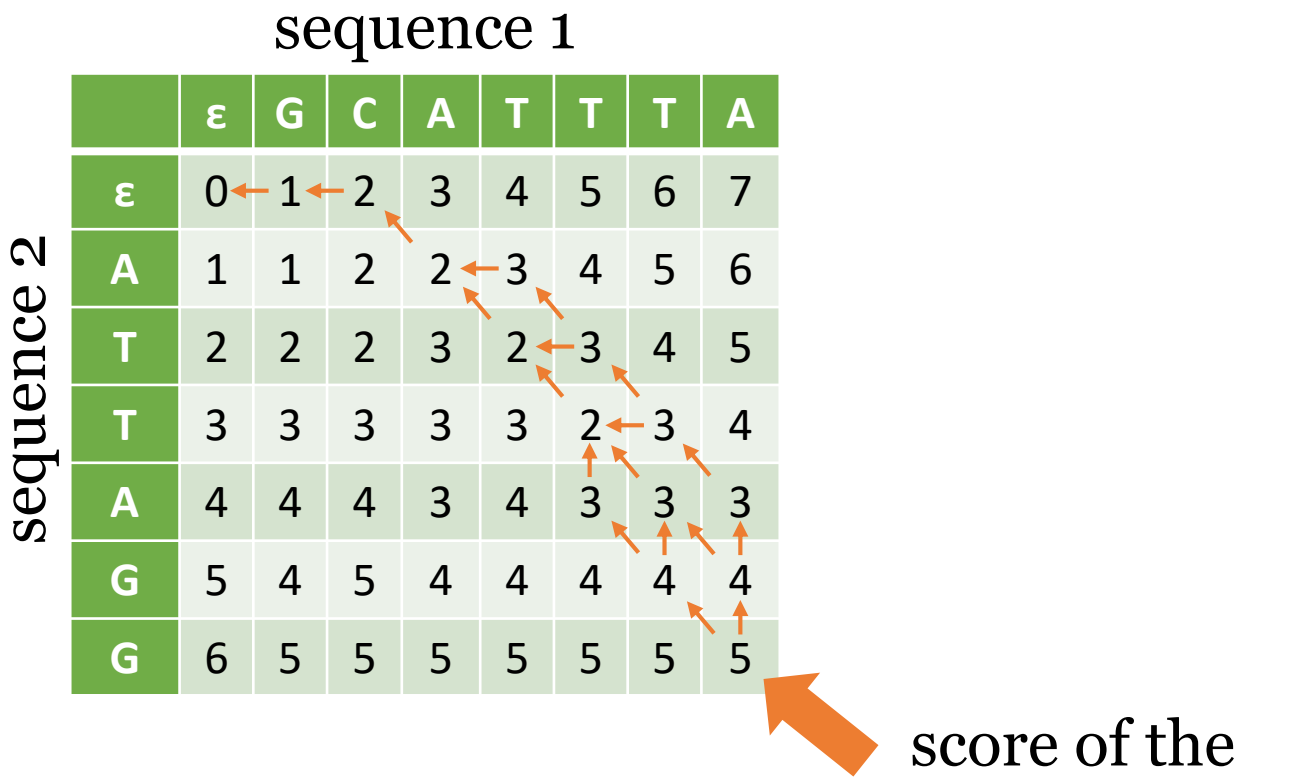
- Naive approach:
  1. Produce all possible alignments of the given sequences.
  2. Calculate the quality of each alignment using a scoring function.
  3. Select the alignment(s) with the highest quality.

	0	1	2	3	4	. . .	10	. . .	20
0	1	1	1	1	1	…	1	…	1
1	1	3	5	7	9	…	…	…	…
2	1	5	13	25	41	…	…	…	…
3	1	7	25	63	129	…	…	…	…
4	1	9	41	129	321	…	…	…	…
. . .	…	…	…	…	…	…	…	…	…
10	1	…	…	…	…	…	8,097,453	…	…
. . .	…	…	…	…	…	…	…	…	…
20	1	…	…	…	…	…	…	…	260,543,813,797,441

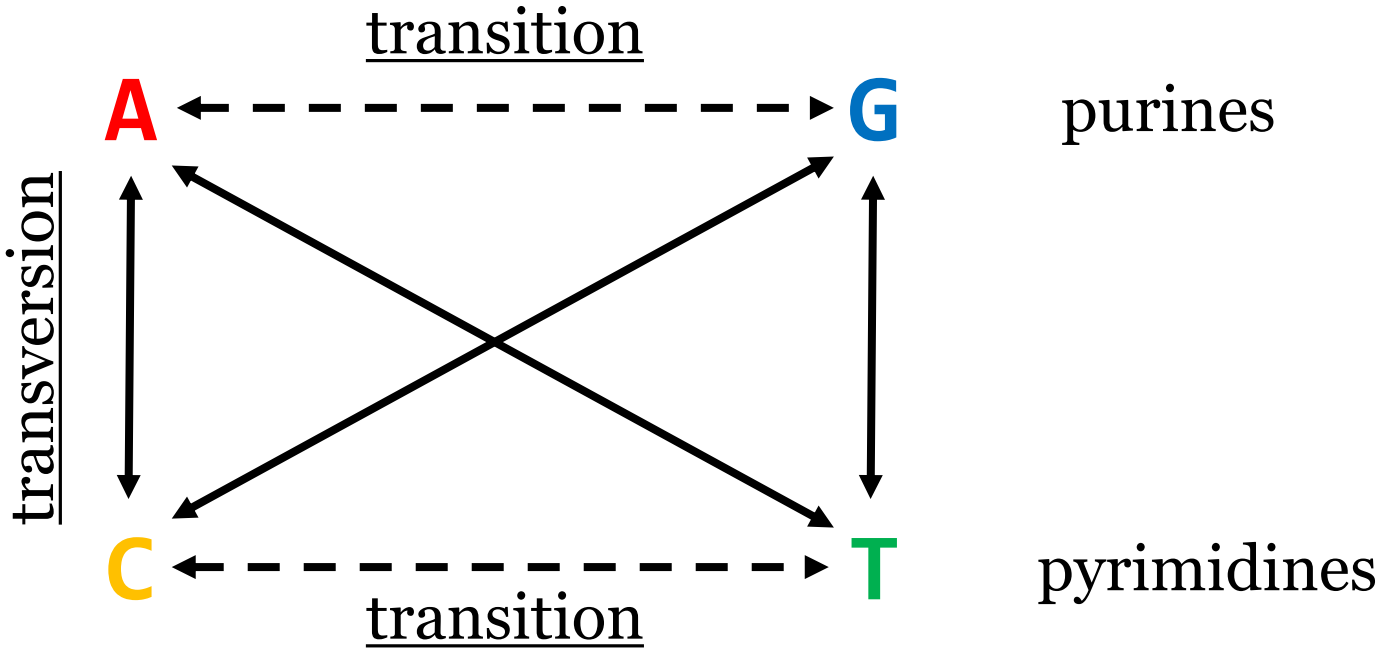
Sequence Alignments: Optimal Alignments And Scoring Functions

- How can we calculate the optimal pairwise alignment of two sequences?

- Dynamic programming approach: Find the optimal solution of a problem by first finding the optimal solutions of smaller subproblems.



Sequence Alignments: Optimal Alignments And Scoring Functions



- For biochemical reasons, transitions occur more frequently than transients, although statistically they should be less frequent.

Sequence Alignments: Optimal Alignments And Scoring Functions

- We can adjust our cost function to be aware of transitions and transversions.

standard cost function

$$\delta = \begin{cases} 0, & a = b \\ 1, & a \neq b \end{cases}$$

adjusted cost function

$$\delta = \begin{cases} 0, & a = b \\ 1, & a \neq b \text{ and } a, b \in \{\text{A}, \text{G}\} \\ 1, & a \neq b \text{ and } a, b \in \{\text{C}, \text{T}\} \\ 2, \text{else} \end{cases}$$

	ε	A	T	C	A	C	A	C	T	T	A
ε	0	1	2	3	4	5	6	7	8	9	10
A	1	0	1	2	3	4	5	6	7	8	9
G	2	1	1	2	3	4	5	6	7	8	9
T	3	2	1	2	3	4	5	6	6	7	8
G	4	3	2	2	3	4	5	6	7	7	8
C	5	4	3	2	3	3	4	5	6	7	8
A	6	5	4	3	2	3	3	4	5	6	7
C	7	6	5	4	3	2	3	3	4	5	6
A	8	7	6	5	4	3	2	3	4	5	5
C	9	8	7	6	5	4	3	2	3	4	5
A	10	9	8	7	6	5	4	3	3	4	4

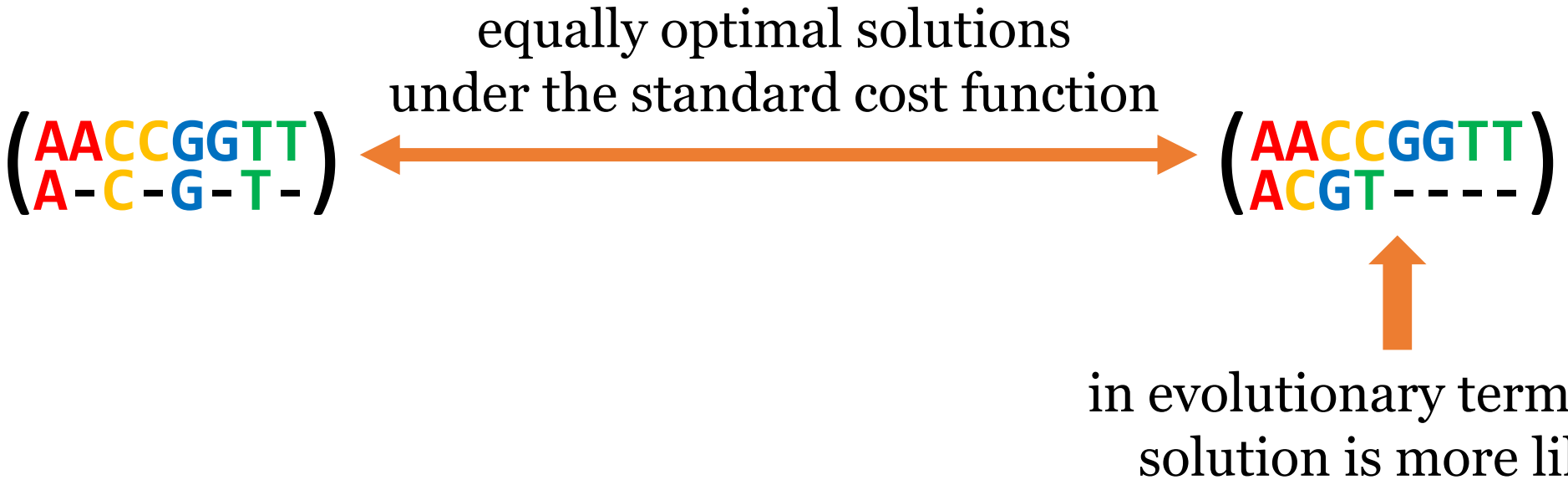
	ε	A	T	C	A	C	A	C	T	T	A
ε	0	2	4	6	8	10	12	14	16	18	20
A	2	0	2	4	6	8	10	12	14	16	18
G	4	2	2	4	5	7	9	11	13	15	17
T	6	4	2	3	5	6	8	10	11	13	15
G	8	6	4	4	4	6	7	9	11	13	14
C	10	8	6	4	6	4	6	7	9	11	13
A	12	10	8	6	4	6	4	6	8	10	11
C	14	12	10	8	6	4	6	4	6	8	10
A	16	14	12	10	8	6	4	6	6	8	8
C	18	16	14	12	10	8	6	4	6	7	9
A	20	18	16	14	12	10	8	6	6	8	7

$$\begin{pmatrix} \text{A-T-CACACTTA} \\ \text{AGTGCACAC--A} \end{pmatrix}$$

$$\begin{pmatrix} \text{ATCACACTTA} \\ \text{AGTGCACACA} \end{pmatrix}$$

Sequence Alignments: Optimal Alignments And Scoring Functions

- Gaps represent insertions and deletions that occurred during the evolution of related sequences.

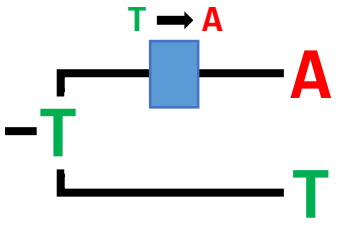


- There are several gap penalty functions that take this fact into account and produce alignments with a few large gaps instead of many small ones.

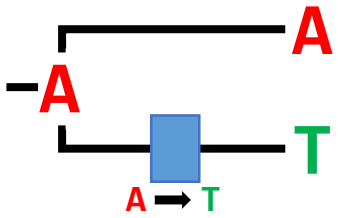
Sequence Alignments: Optimal Alignments And Scoring Functions

substitution result: different nucleotides

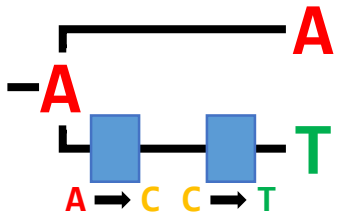
single substitution



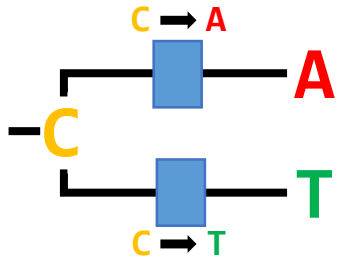
single substitution



multiple substitution

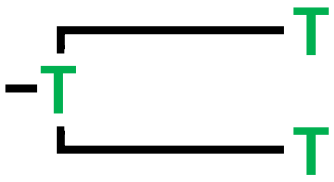


simultaneous substitution

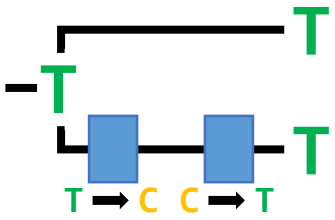


substitution result: equal nucleotides

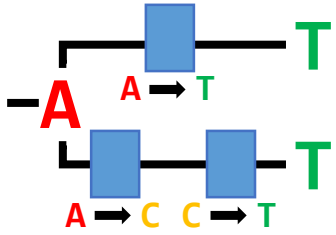
no substitution



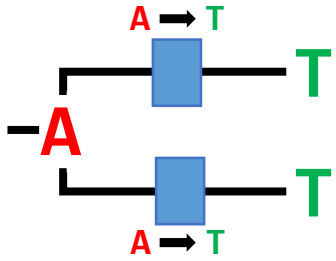
back substitution



convergent substitution



parallel substitution

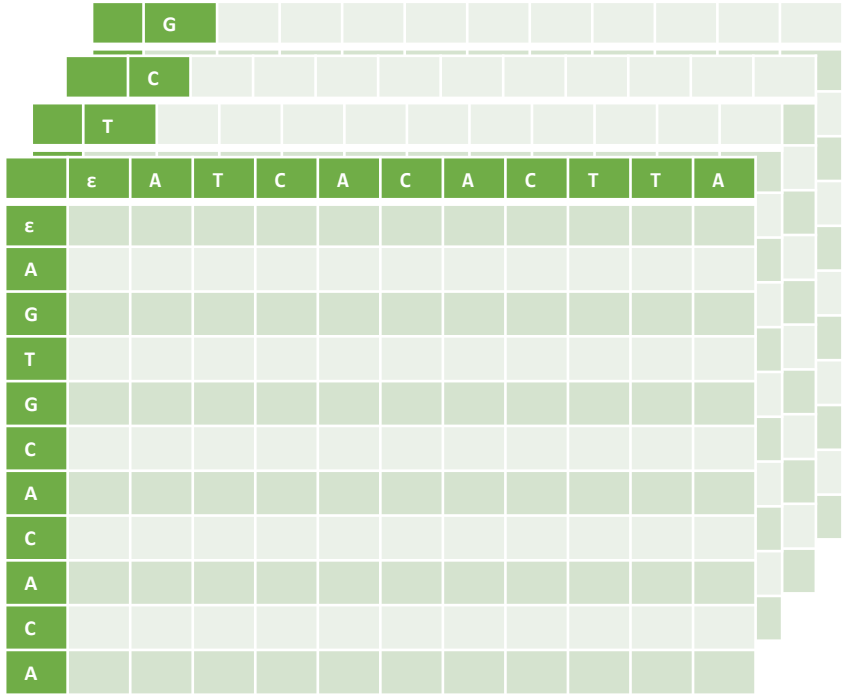


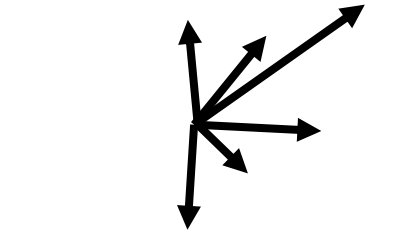
- There are different possible scenarios that can explain observable and unobservable substitutions. (→ more on this on workshop day 2)

Sequence Alignments: Complexity

- Calculating the optimal pairwise alignment of two sequences still takes relatively long computational time.

- Calculating the optimal multiple alignment of ***k*** sequences is even worse. (→ this problem is NP-hard, meaning it can never be solved efficiently)



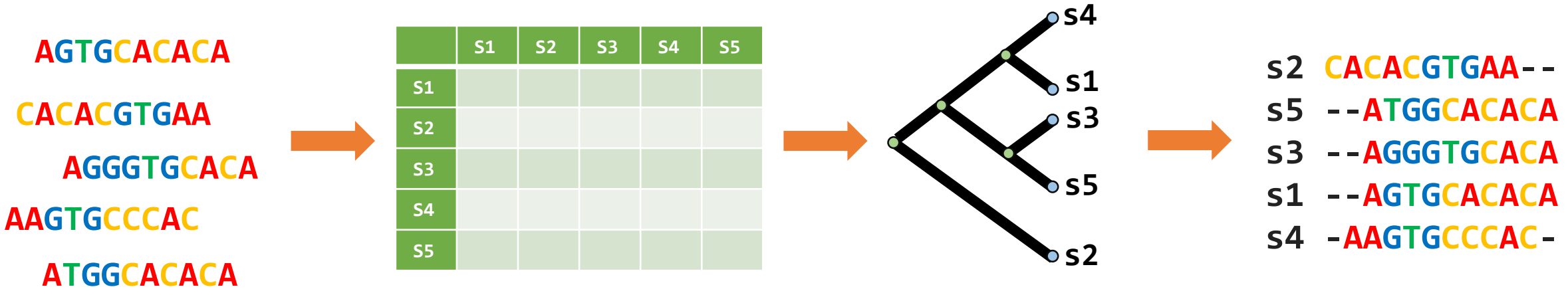


k-dimensions

Sequence Alignments: Complexity

- Instead of using an exact but slow algorithm, we have to use heuristics that may not find the optimal solution but are much faster.

- Many modern multiple sequence alignment heuristics are based on progressive construction methods:
  1. Calculate the pairwise similarities of all sequences.
  2. Based on these similarities, align those sequences first that are most similar to each other.
  3. Progress until all sequences are included into the alignment.



Sequence Alignments: Tools

- Many different alignment tools exist, which all follow their own approach, producing different alignments for the same input sequences.

- Blast
- ClustalΩ
- MAFFT
- MUSCLE
- …

Short Summary

- Sequence alignments are the basis of molecular phylogenetics, because sequences (hopefully) reflect evolutionary relationships most accurately.
- An optimal sequence alignment maximizes the total number of matches, while minimizing the number of mismatches and gaps.
- Different scoring functions exist to better model evolutionary processes.
- Calculating useful optimal alignments is impossible → we use heuristics which may not find the best alignment.

What we have learned today

- You are familiar with (or have recalled) the most important basics and terms of phylogenetics.
- You are aware of the existing problems and limitations of phylogenetic analysis and phylogenetic trees.
- You know what an optimal sequence alignment is, why it is absolutely essential for phylogenetics and why it is nevertheless almost impossible to ever get one.

What we will learn tomorrow

- Algorithmic phylogenetics

- Four main approaches:
  - Parsimony analysis
  - Distance matrix approaches
  - Maximum Likelihood analysis
  - Bayesian inference