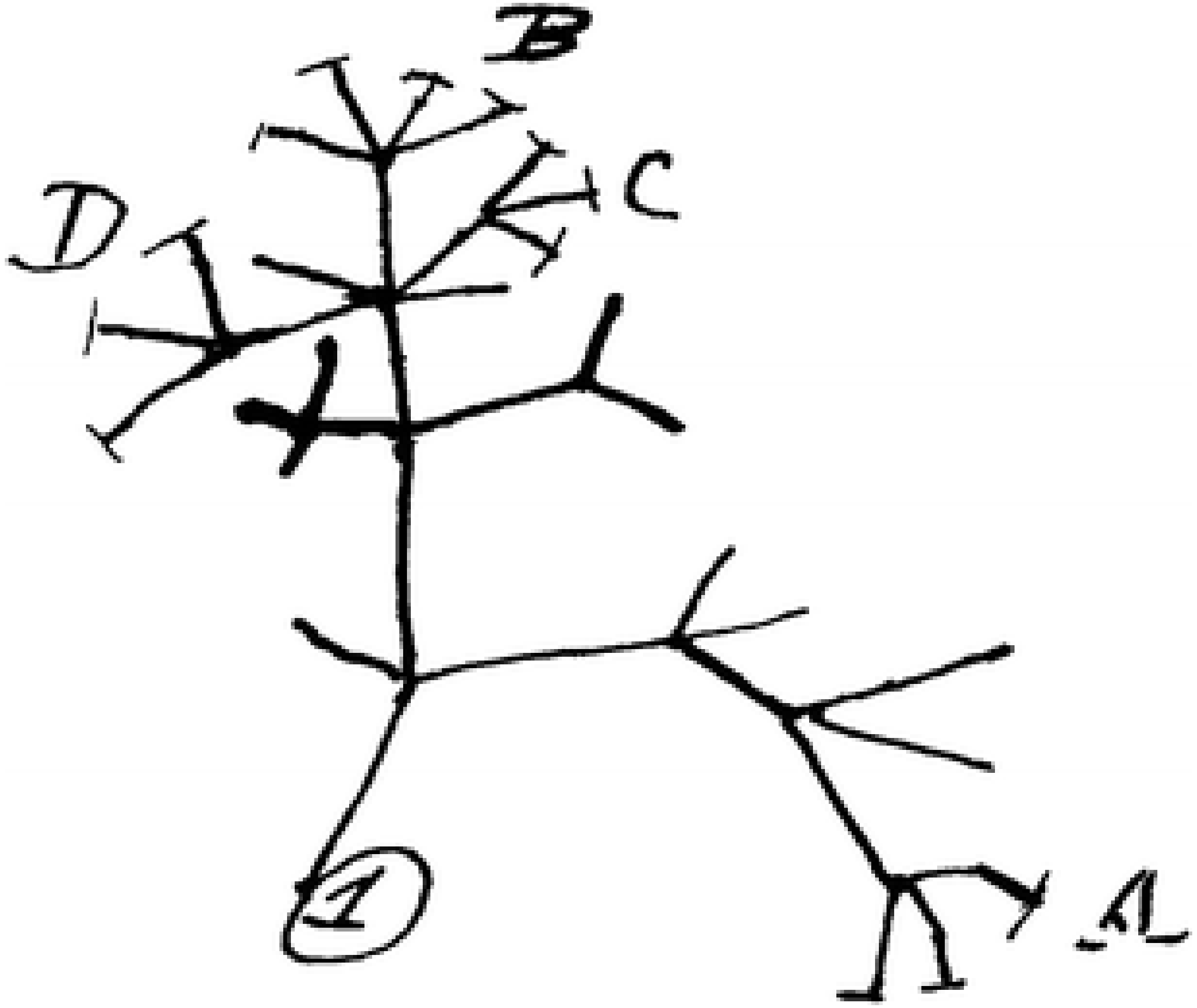
DTU

Dr Thom Booth – 25/10/2024

Practical Phylogenetics - 29907

I think



COURSE OUTLINE:

DAY ONE:

Session 1: Introduction

Session 2: From Sequences to Trees

Session 3: Interpreting Trees

Free Study – Until 4 pm

DAY TWO:

Session 4: Advanced Phylogenetic Techniques

Session 5: Presentation Session

Free Study – Until 4 pm

COURSEWORK

To pass this course you must complete a **3 – 5 minute presentation** on the phylogenetics of a gene/protein/organism of interest.

The presentation must cover:

1. Breif background of your sequence of interest
2. The evolutionary hypothesis you wanted to test
3. The methods you used to find the homologues, make the alignment and build the tree
4. A description of the quality of the resulting tree
5. Whether the tree supported or confirmed your hypothesis

Feedback Reminder

Introductions

Tell us:

- Your name (and write it down on the paper in front of you!),
- Your lab,
- A one sentence summary of your research,
- Your experience with phylogenetics so far and what (if any) tools are you using already,
- The you want to do this course and,
- The most important thing you want to learn from this course.

Practical Phylogenetics

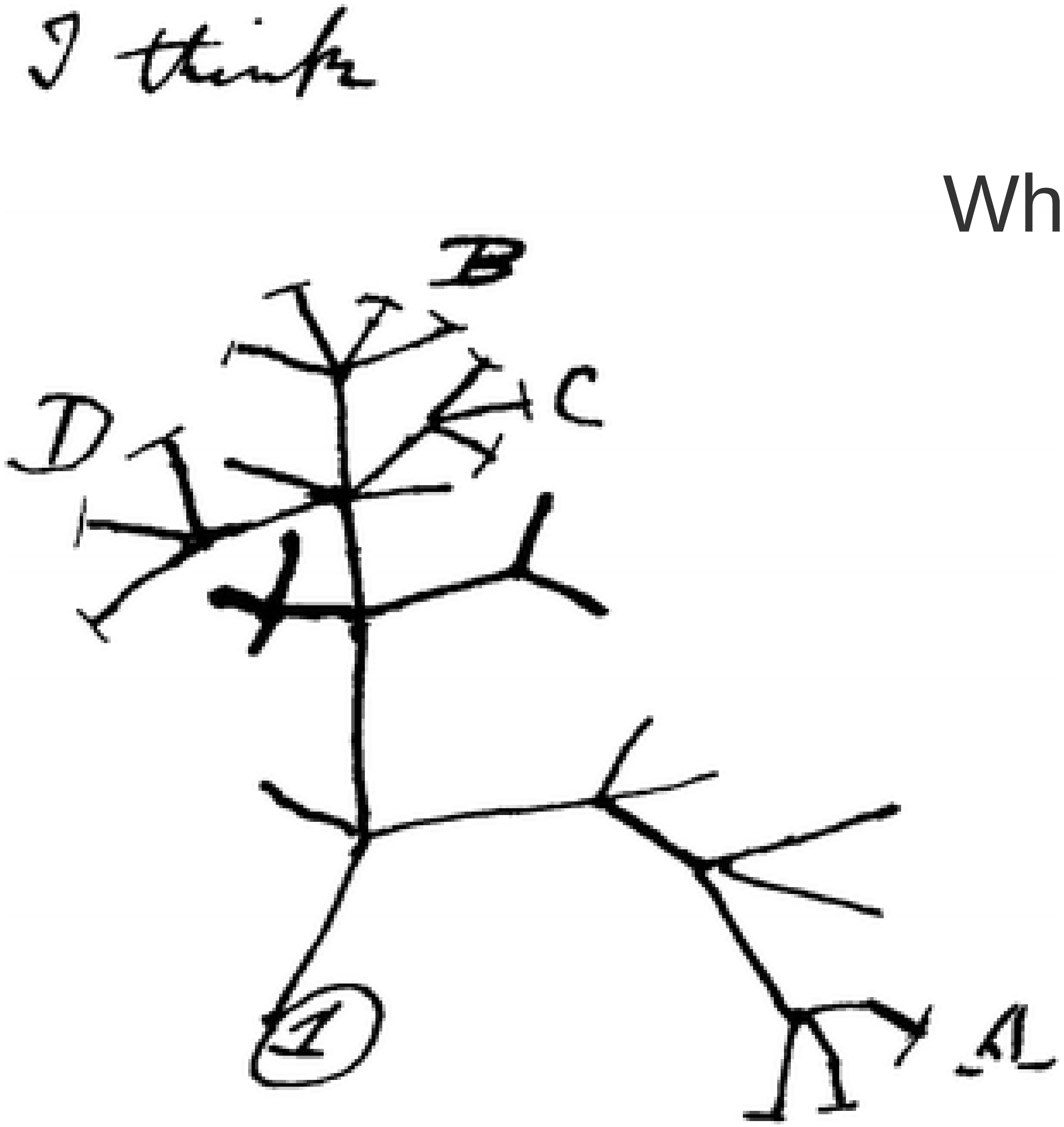
Session 1: Introduction

Session 1: Learning Objectives

- Explain the different types of data that can be used to infer a tree. Understand the advantages of using DNA or protein sequences.

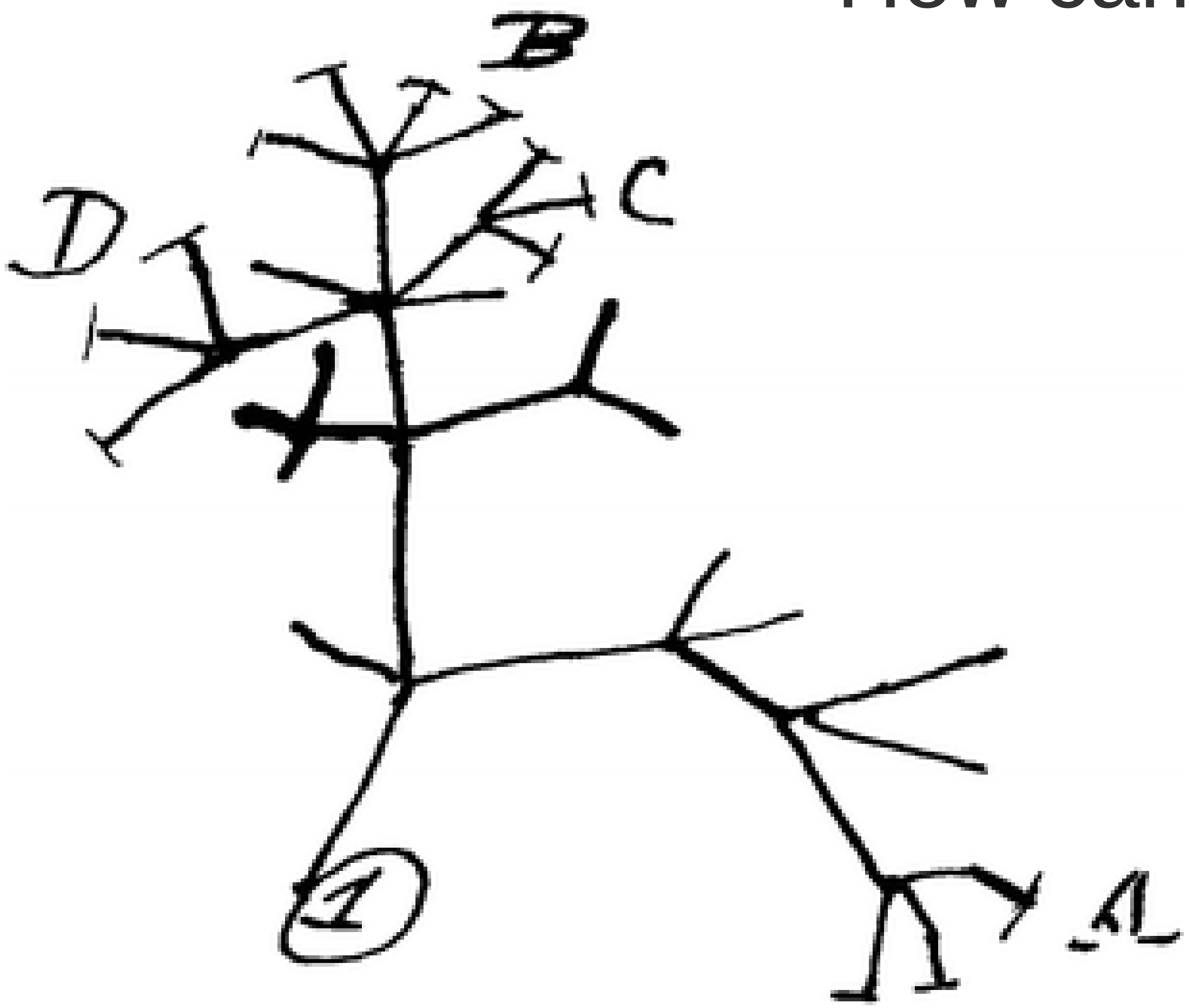
- Create a simple distance matrix and draw simple trees by hand.

What is a phylogeny?

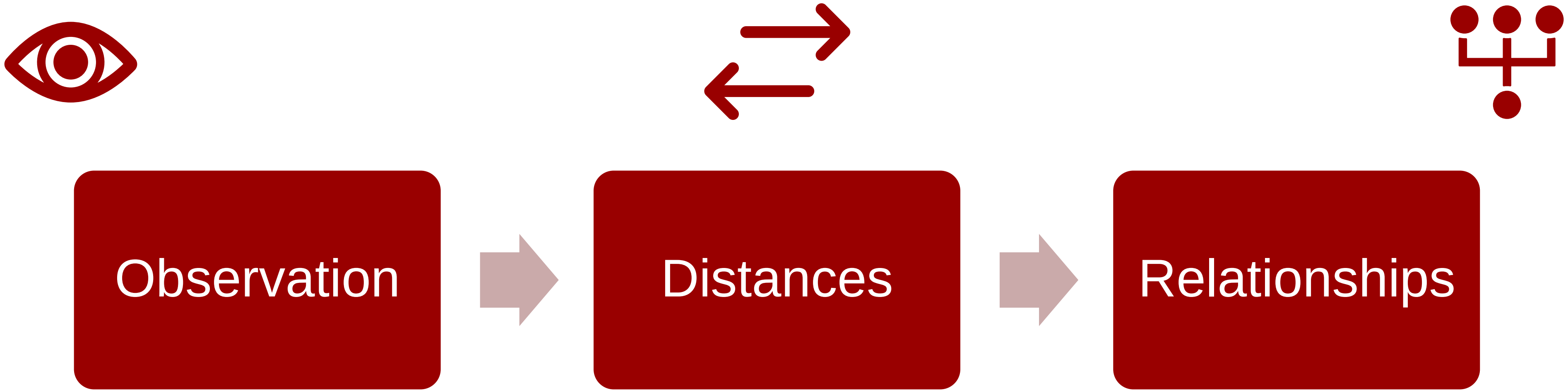


I think

How can we ***infer*** relationships?



How can we ***infer*** relationships?



Observation

Distances

Relationships

Let’s play 20-questions…

Let's play 5-questions…

- You have to guess the animal I am thinking.
- You can ask me 5 questions.
- The questions must be binary (i.e. yes/no).
- We are going to play 10 times in a row.
- You must use the same 5 questions.
- You must win every time.

- **Discuss with your partner and write down 5 questions.**

Let’s play 5-questions…

- Is it warm blooded?
- Does it have a backbone?
- Does it have wings?
- Does it have four legs?
- Does it have a tail?

Let’s play 5-questions…

Fruit Fly

Drosophilla melanogaster

Zebra Fish

Danio rerio

Chicken

Gallus gallus

Human (allegedly…)

Homo sapiens

Stan

Tyranosaurus rex

House Mouse

Mus musculus

Let’s play 5-questions…

	Fly	Fish	Chicken	Human	T.rex	Mouse
Is it warm blooded?	0	0	1	1	1	1
Does it have a backbone?	0	1	1	1	1	1
Does it have wings?	1	0	1	0	0	0
Does it have four legs?	0	0	0	0	0	1
Does it have a tail?	1	1	1	0	1	1

Let’s play 5-questions…

	Fly	Fish	Chicken	Human	T.rex	Mouse
Is it warm blooded?	0	0	1	1	1	1
Does it have a backbone?	**0**	**1**	**1**	**1**	**1**	**1**
Does it have wings?	1	0	1	0	0	0
Does it have four legs?	**0**	**0**	**0**	**0**	**0**	**1**
Does it have a tail?	**1**	**1**	**1**	**0**	**1**	**1**

Uninformative characterisitcs!

Let’s play 5-questions…

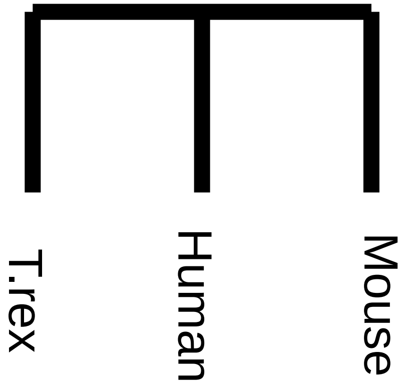
	Fly	Fish	Chicken	Human	T.rex	Mouse
Is it warm blooded?	0	0	1	1	1	1
Is it bigger than a loaf of bread?	0	0	1	1	1	0
Does it have wings?	1	0	1	0	0	0
Does it have two legs?	0	0	1	1	1	1
Does it have hair?	0	0	0	1	0	1

Let’s play 5-questions…

	Fly	Fish	Chicken	Human	T.rex	Mouse
Fly	**0**		-	-	-	-
Fish	1	**0**	-	-	-	-
Chicken	3	4	**0**	-	-	-
Human	5	4	2	**0**	-	-
T.rex	4	3	2	1	**0**	-
Mouse	4	3	3	1	2	**0**

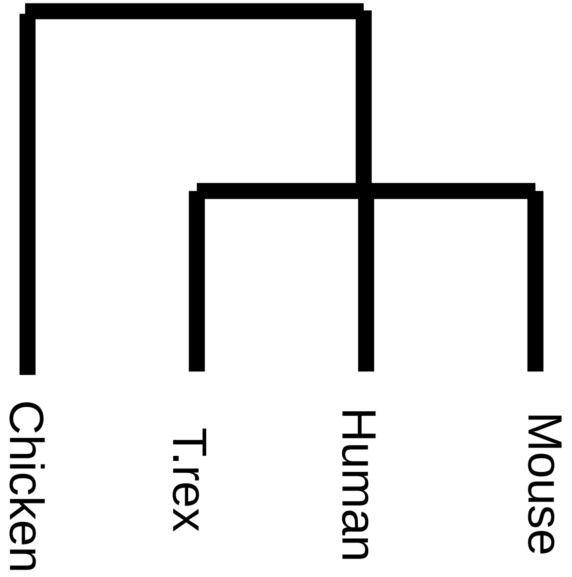
Let’s play 5-questions…

	Fly	Fish	Chicken	Human	T.rex	Mouse
Fly	**0**	1	-	-	-	-
Fish	1	**0**	-	-	-	-
Chicken	3	4	**0**	-	-	-
Human	**5**	**4**	**2**	**0**	-	-
T.rex	**4**	**3**	**2**	**1**	**0**	-
Mouse	**4**	**3**	**3**	**1**	**2**	**0**



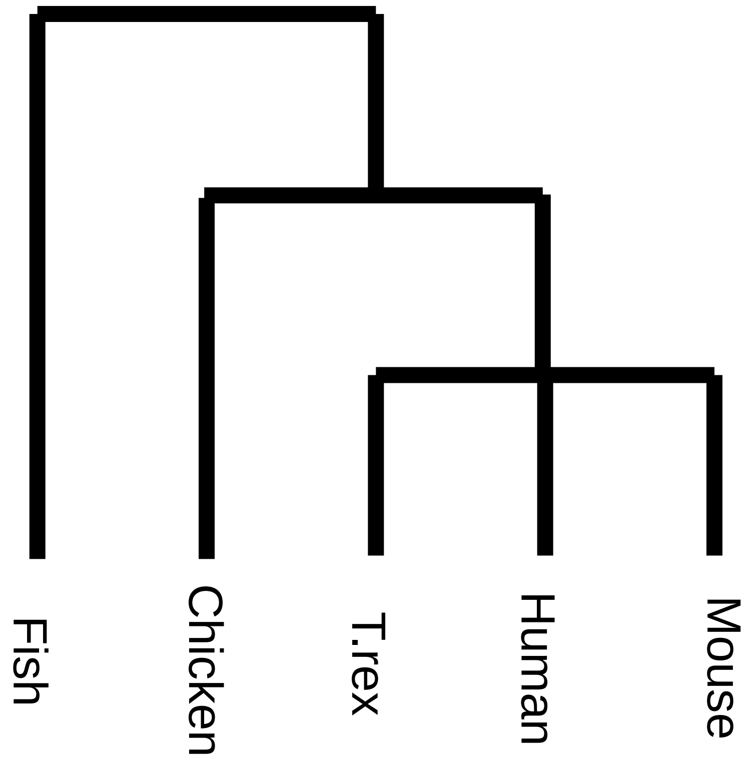
Let's play 5-questions…

	Fly	Fish	Chicken	Human	T.rex	Mouse
Fly	0	1	-	-	-	-
Fish	1	0	-	-	-	-
Chicken	3	4	0	-	-	-
Human	5	4	2	0	-	-
T.rex	4	3	2	1	0	-
Mouse	4	3	3	1	2	0
	13	11	7			



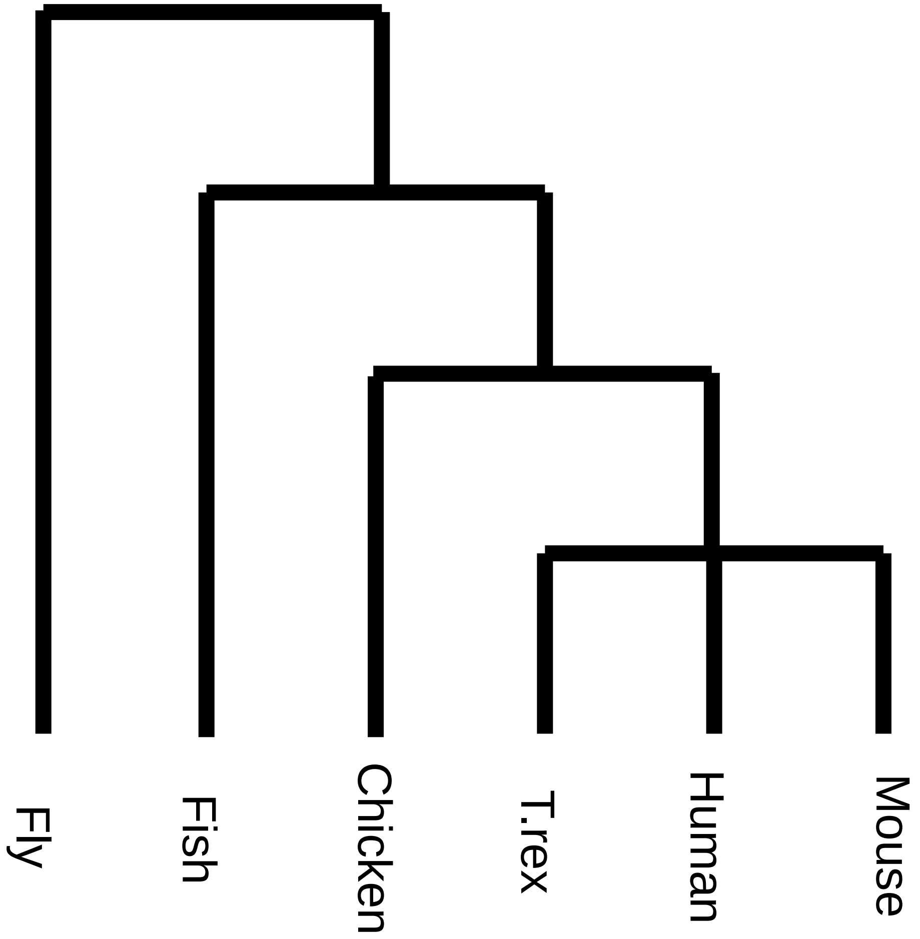
Let’s play 5-questions…

	Fly	Fish	Chicken	Human	T.rex	Mouse
Fly	0	1	-	-	-	-
Fish	1	0	-	-	-	-
Chicken	3	4	0	-	-	-
Human	5	4	2	0	-	-
T.rex	4	3	2	1	0	-
Mouse	4	3	3	1	2	0
	16	14				



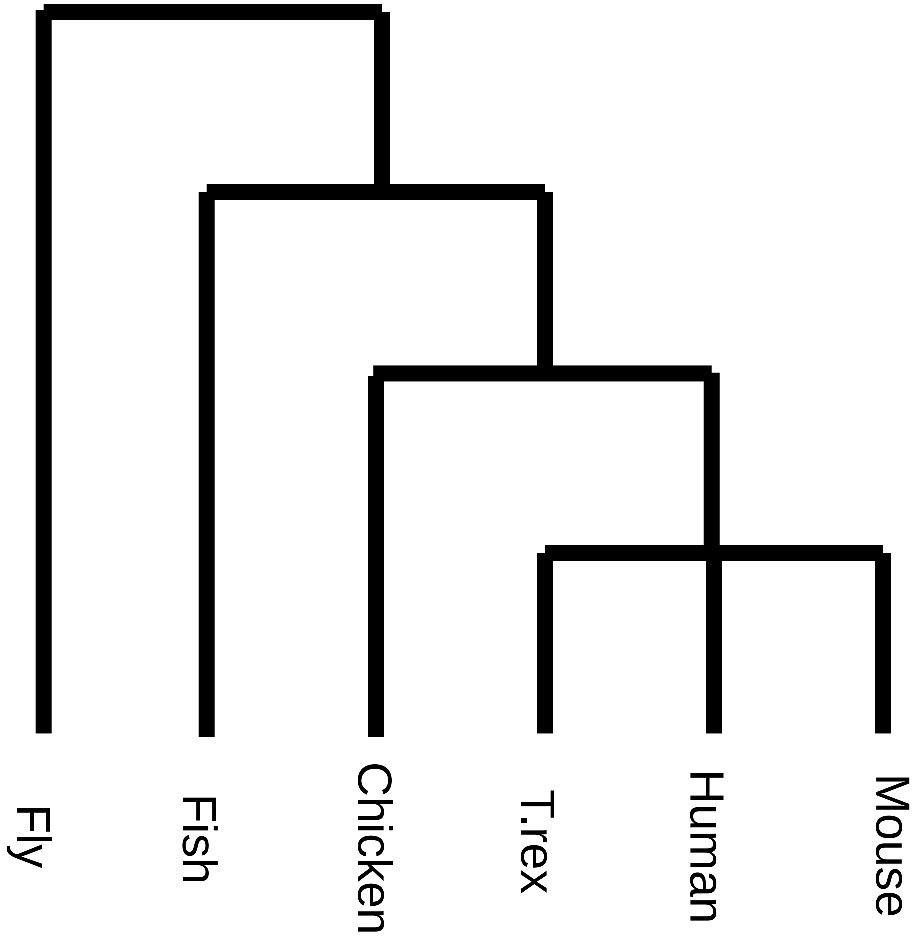
Let’s play 5-questions…

	Fly	Fish	Chicken	Human	T.rex	Mouse
Fly	0	1	-	-	-	-
Fish	1	0	-	-	-	-
Chicken	3	4	0	-	-	-
Human	5	4	2	0	-	-
T.rex	4	3	2	1	0	-
Mouse	4	3	3	1	2	0
	16	14				

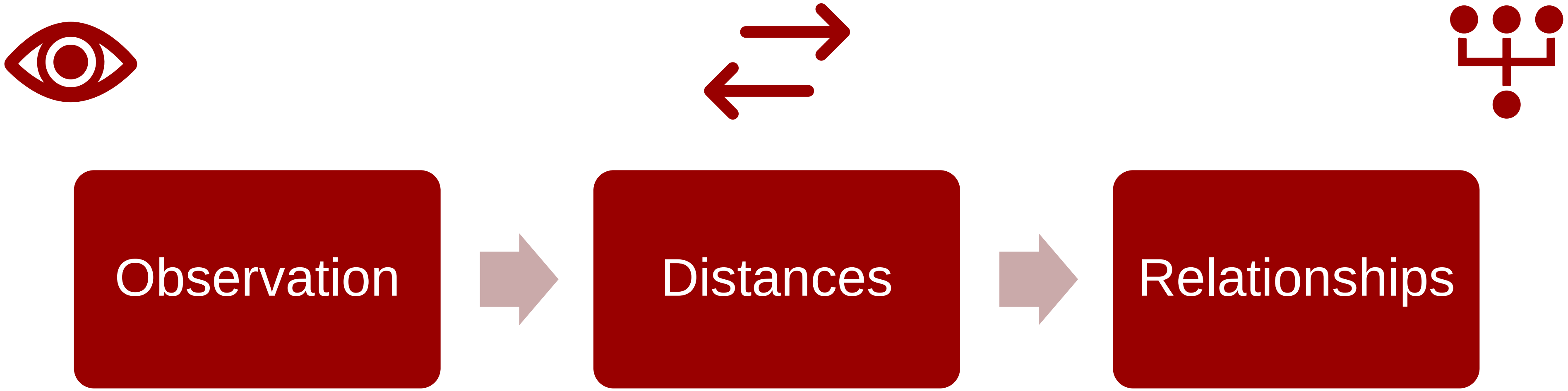


Let's play 5-questions…

- Do our trees look the same?
- Does this tree reflect nature?
- Did I ask good questions?
- Was there any difficulties with the process?
- Is there a better way?



Morphological Phylogenetics

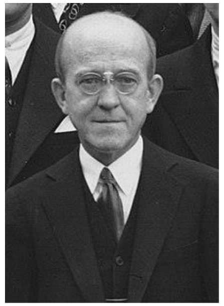


Observation

Distances

Relationships

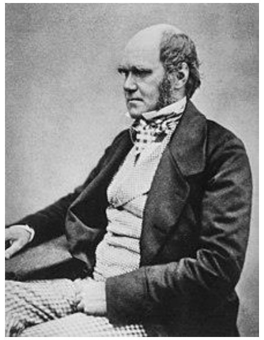
DNA is the molecule of heredity



1859
On the Origin of Species

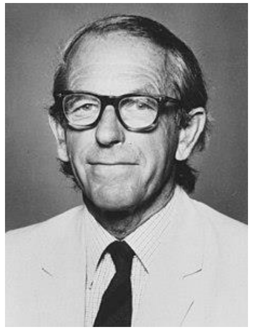
1944
Avery Experiment

2024
Practical Phylogentics

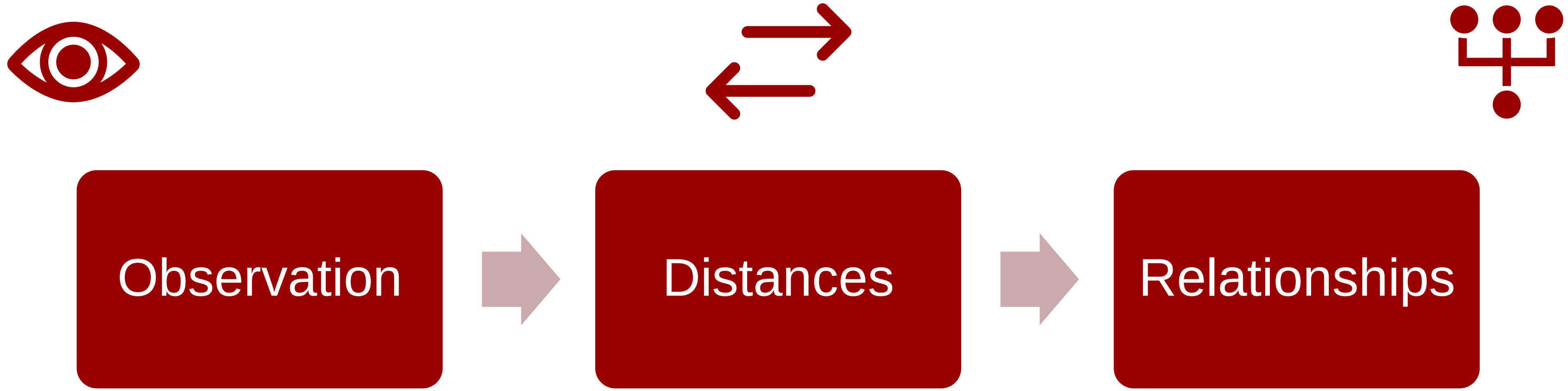


1977
Sanger Sequencing

~2005
NGS



Morphological Phylogenetics

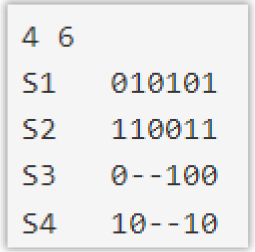


Observation

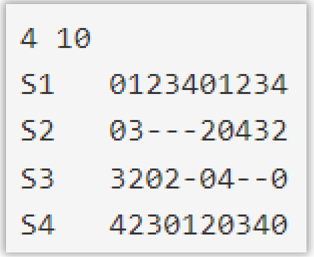
Distances

Relationships

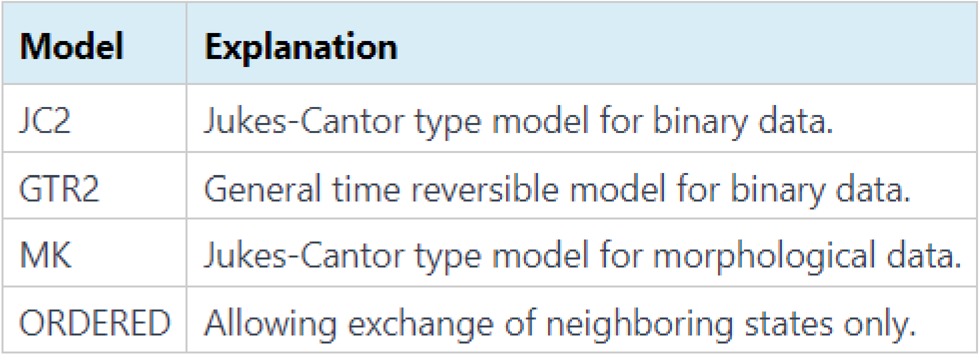
Binary, categorical or continous morphological characteristics



```
4 6
S1   010101
S2   110011
S3   0--100
S4   10--10
```

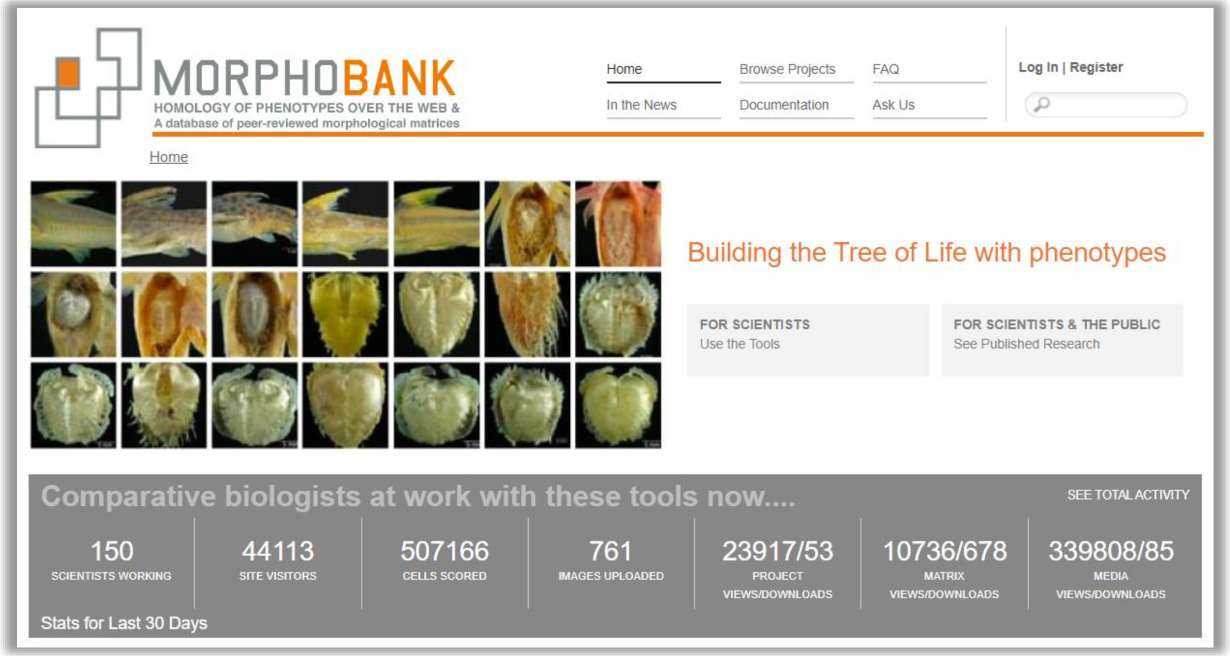


```
4 10
S1   0123401234
S2   03---20432
S3   3202-04--0
S4   4230120340
```



Model	Explanation
JC2	Jukes-Cantor type model for binary data.
GTR2	General time reversible model for binary data.
MK	Jukes-Cantor type model for morphological data.
ORDERED	Allowing exchange of neighboring states only.

Morphological Phylogenetics

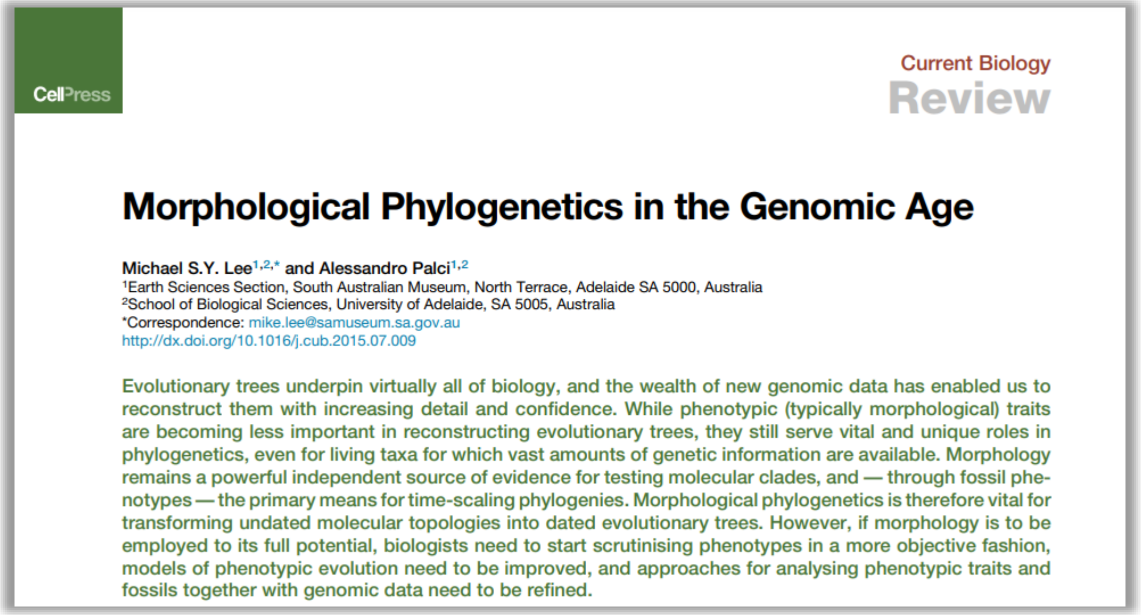


Building the Tree of Life with phenotypes

Comparative biologists at work with these tools now....

150	44113	507166	761	23917/53	10736/678	339808/85
SCIENTISTS WORKING	SITE VISITORS	CELLS SCORED	IMAGES UPLOADED	PROJECT VIEWS/DOWNLOADS	MATRIX VIEWS/DOWNLOADS	MEDIA VIEWS/DOWNLOADS

Stats for Last 30 Days



Current Biology
Review

Morphological Phylogenetics in the Genomic Age

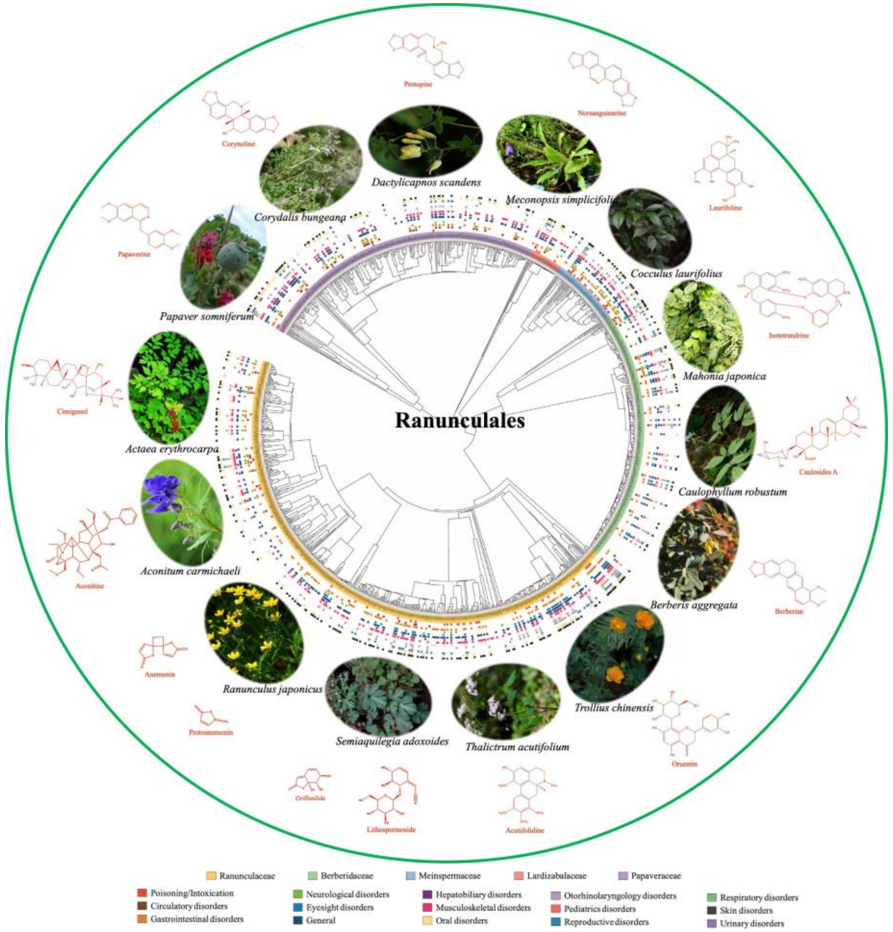
Michael S.Y. Lee[1,2,*] and Alessandro Palci[1,2]
[1]Earth Sciences Section, South Australian Museum, North Terrace, Adelaide SA 5000, Australia
[2]School of Biological Sciences, University of Adelaide, Adelaide SA 5005, Australia
*Correspondence: mike.lee@samuseum.sa.gov.au

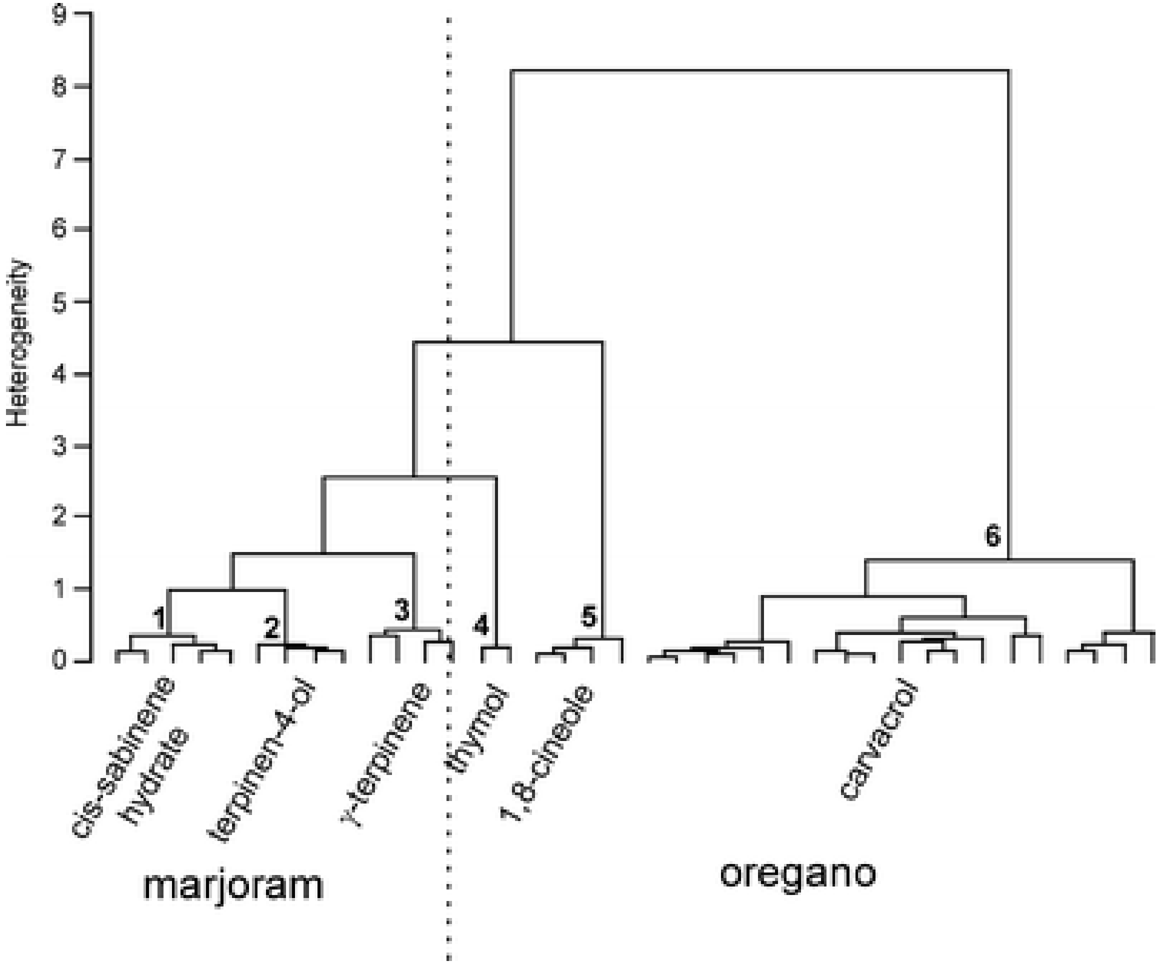
http://dx.doi.org/10.1016/j.cub.2015.07.009

Evolutionary trees underpin virtually all of biology, and the wealth of new genomic data has enabled us to reconstruct them with increasing detail and confidence. While phenotypic (typically morphological) traits are becoming less important in reconstructing evolutionary trees, they still serve vital and unique roles in phylogenetics, even for living taxa for which vast amounts of genetic information are available. Morphology remains a powerful independent source of evidence for testing molecular clades, and — through fossil phenotypes — the primary means for time-scaling phylogenies. Morphological phylogenetics is therefore vital for transforming undated molecular topologies into dated evolutionary trees. However, if morphology is to be employed to its full potential, biologists need to start scrutinising phenotypes in a more objective fashion, models of phenotypic evolution need to be improved, and approaches for analysing phenotypic traits and fossils together with genomic data need to be refined.

Lee et al., Curr Biol, 2015

Chemotaxonomy (Pharmacotaxonomy)

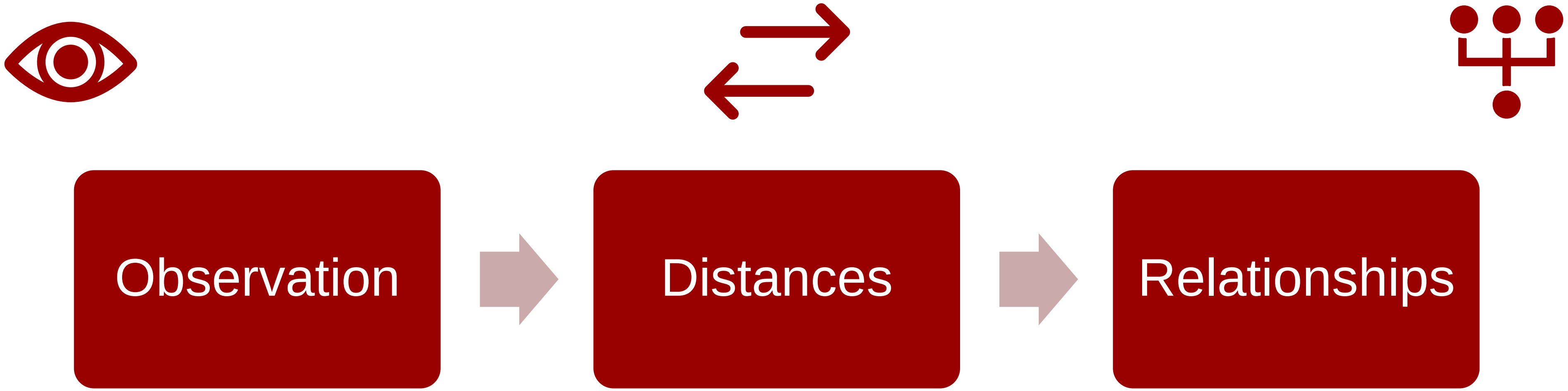




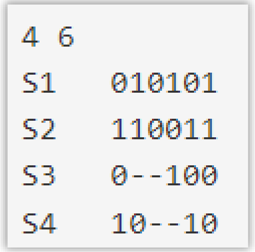
Hao et al., Front Plant Sci, *2022*

Baranska et al., *Anal & Bioanal Chem*, 2005

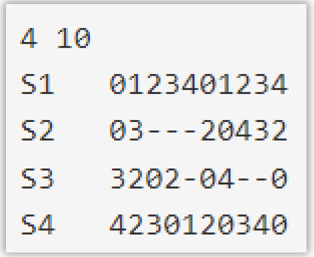
Chemotaxonomy (Pharmacotaxonomy)



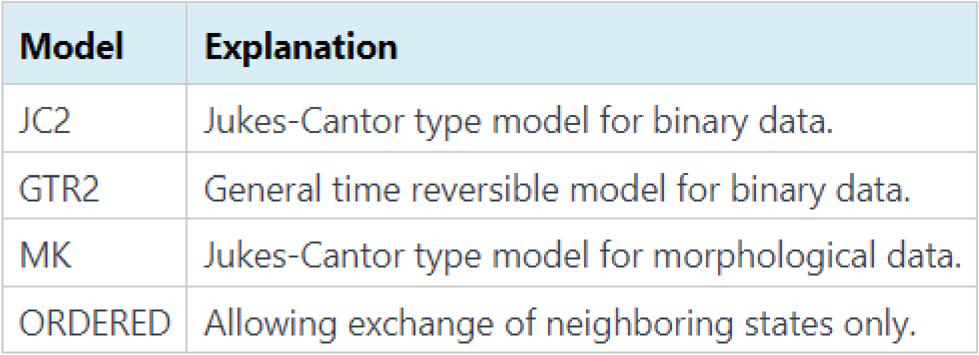
Binary, categorical or continous metabolomic characteristics



```
4 6
S1   010101
S2   110011
S3   0--100
S4   10--10
```



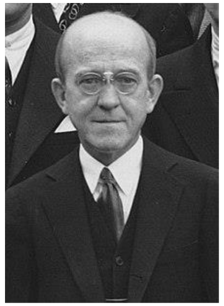
```
4 10
S1   0123401234
S2   03---20432
S3   3202-04--0
S4   4230120340
```



Model	Explanation
JC2	Jukes-Cantor type model for binary data.
GTR2	General time reversible model for binary data.
MK	Jukes-Cantor type model for morphological data.
ORDERED	Allowing exchange of neighboring states only.

Mihn et al. *Mol Biol and Evol*, 2020

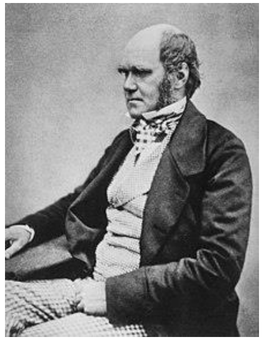
DNA is the molecule of heredity



1859
On the Origin of Species

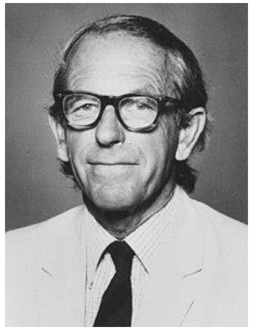
1944
Avery Experiment

2024
Practical Phylogentics

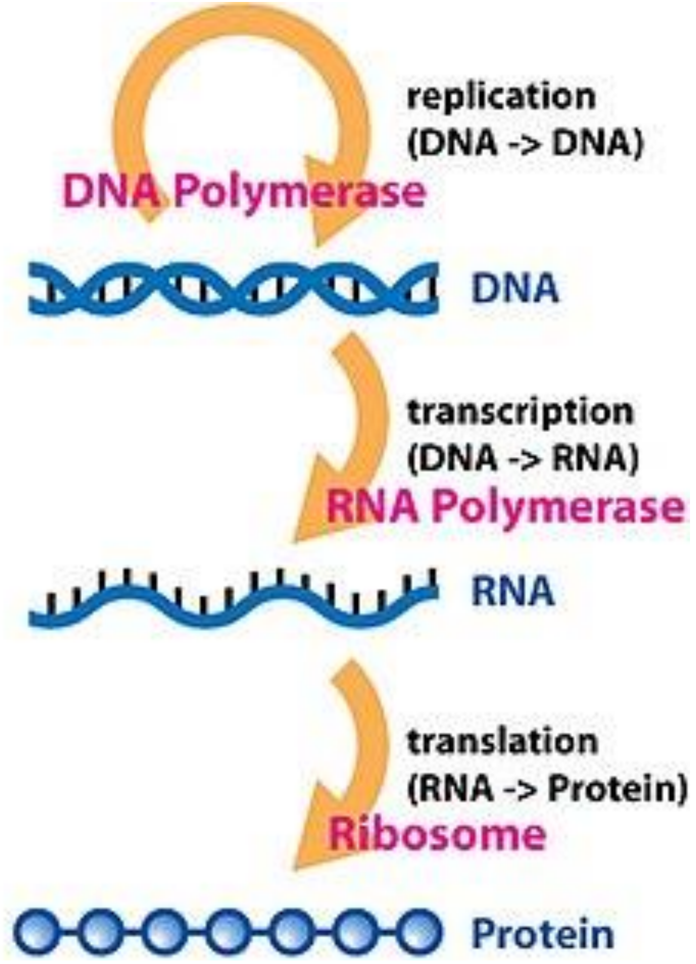


1977
Sanger Sequencing

~2005
NGS

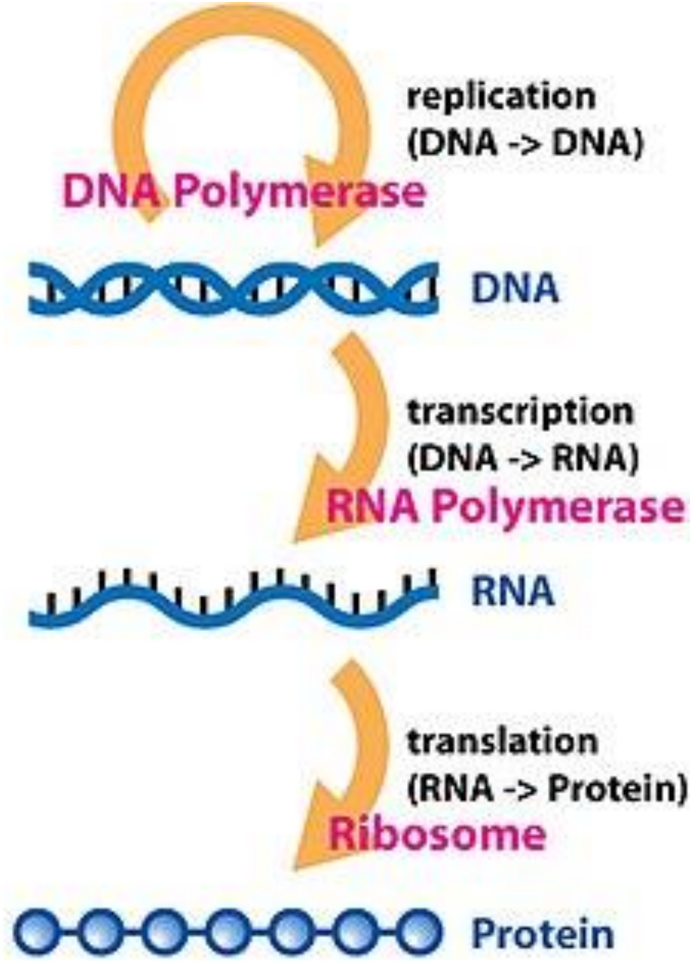


DNA is the molecule of heredity



Central Dogma!

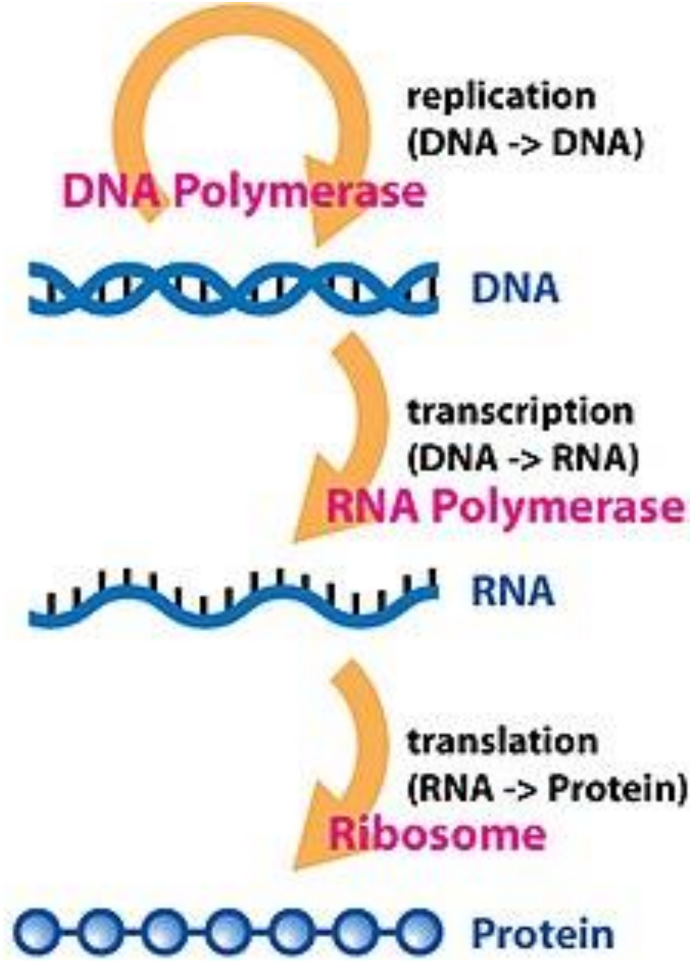
DNA is the molecule of heredity



DNA

- Direct observation of the molecule of heredity (i.e. mutations).
- Not confounded by convergence.

DNA is the molecule of heredity

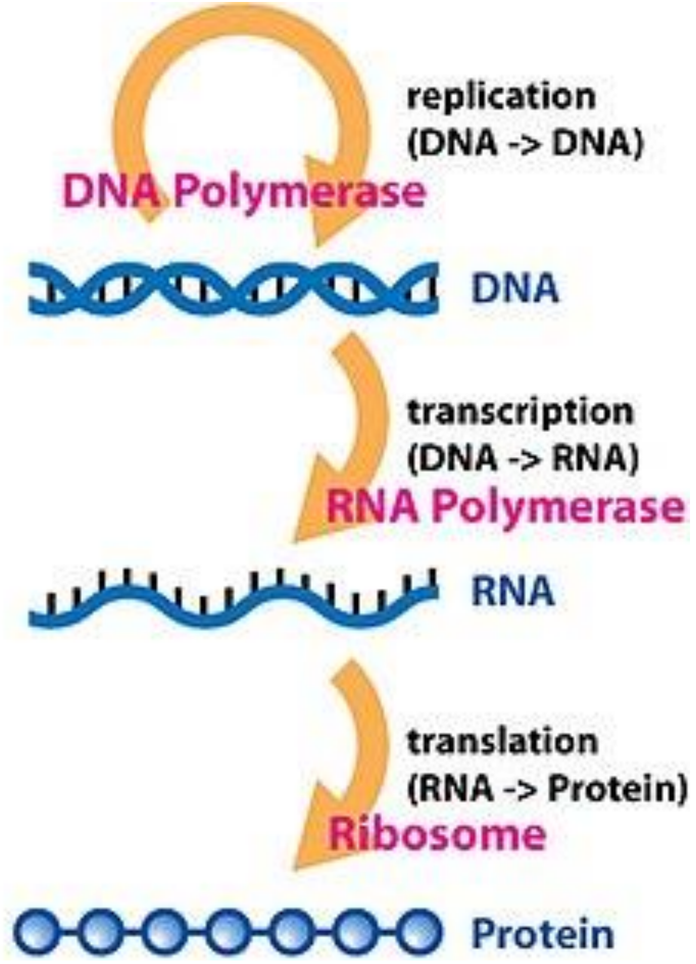


DNA

- Direct observation of the molecule of heredity (i.e. mutations).
- Not confounded by convergence.

- Compositionally biased
- Difficult to align
- Difficult to infer ‘true homology’

DNA is the molecule of heredity



DNA

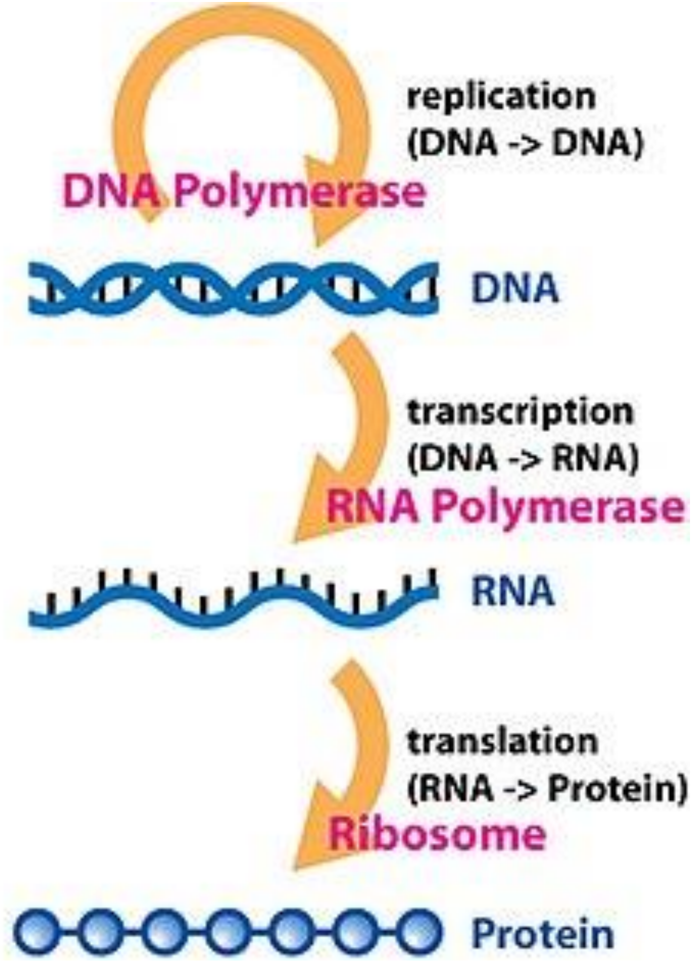
- Direct observation of the molecule of heredity (i.e. mutations).
- Not confounded by convergence.

- Compositionally biased
- Difficult to align
- Difficult to infer ‘true homology’

PROTEIN

- Directly related to the DNA sequence.
- Easier to align and infer ‘true homology’

DNA is the molecule of heredity



DNA

- Direct observation of the molecule of heredity (i.e. mutations).
- Not confounded by convergence.

- Compositionally biased
- Difficult to align
- Difficult to infer 'true homology'

PROTEIN

- Directly related to the DNA sequence.
- Easier to align and infer 'true homology'

- Hides synonymous mutations
- Can only reprisent coding sequences (duh!)

Session 1: Learning Objectives

- Explain the different types of data that can be used to infer a tree. Understand the advantages of using DNA or protein sequences.

- Create a simple distance matrix and draw simple trees by hand.

If you haven’t already…

Download data from:

https://github.com/drboothtj/practical_phylogenetics

If you don’t have an alignment editor download BioEdit:

https://thalljiscience.github.io/

Practical Phylogenetics

Session 2: From Sequences to Trees

Learning Objectives

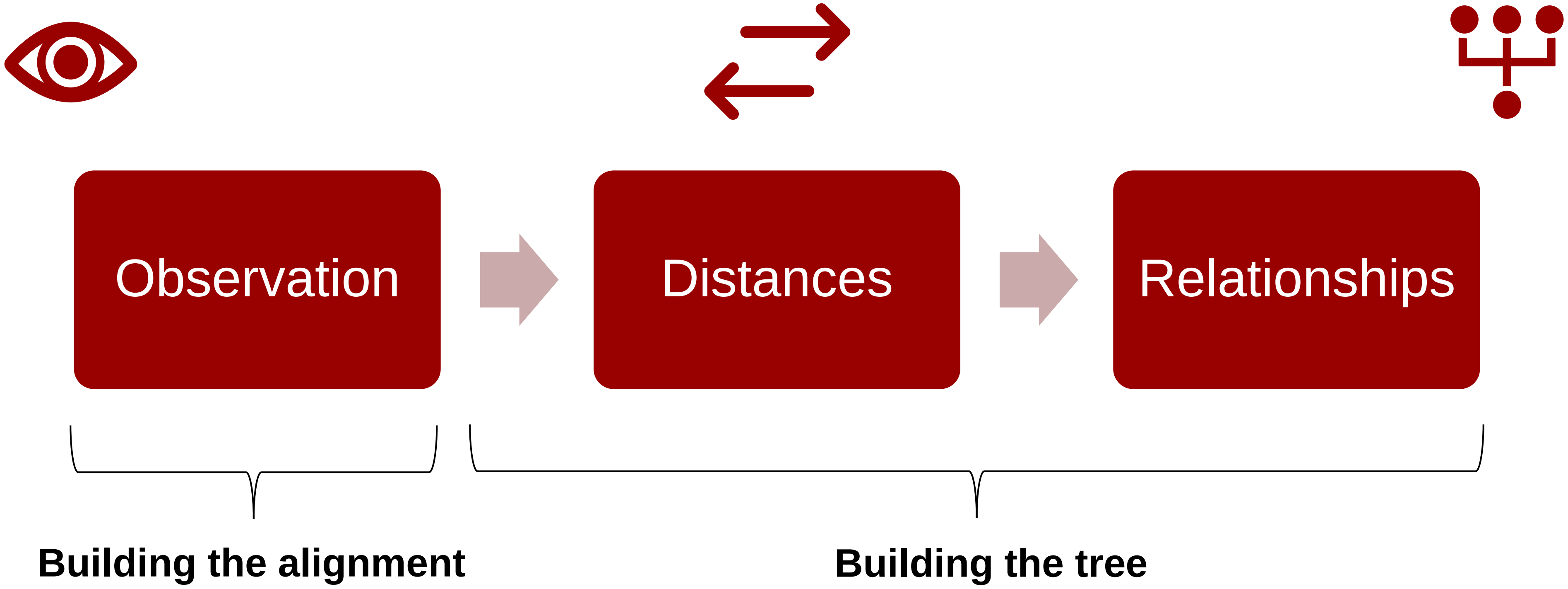
- Explain the different types of data that can be used to infer a tree. Understand the advantages of using DNA or protein sequences.

- Use web-based software to identify homologues, align sequences and infer phylogenetic trees and have a general knowledge of the command line tools available and use these techniques to infer trees for genes of interest.

- Avoid common errors that can occur during alignment and inference.

Practical Phylogenetics

Session 2a: Gathering homologues



Observation

Distances

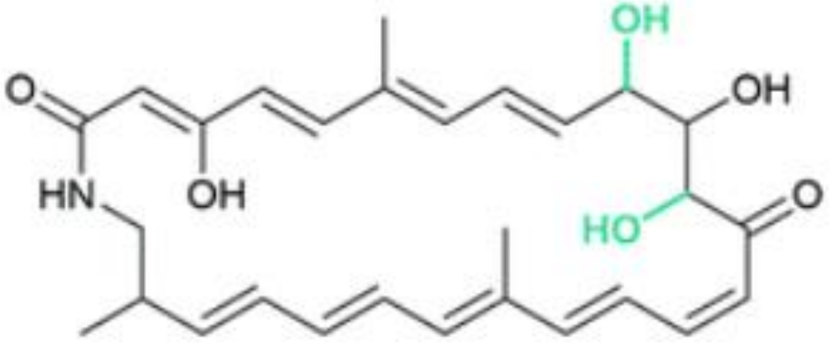
Relationships

Building the alignment

Building the tree

Today’s example

- We are interested in the biosynthesis of sceliphrolactam.
- There are two copies of a P450 hydroxylase, SceE and SceD.
- Can we identify whether or not these copies are a result of a recent gene duplication?
- https://dev.mibig.secondarymetabolites.org/repository/BGC0001770.4/index.html#r1c1



Sceliphrolactam
Streptomyces sp. SD85

Step One: Gather Homologues

A homologue is:

“Any gene that encodes a structurally similar protein with a shared evolutionary history.”

Step One: Gather Homologues

A paralogue is:

> **“Any homologue that resides within the same genome.”**

Note: the relationship between the homologues and paralogues will help us determine the history of our P450!

Thom’s Golden Rule #1

NEVER BUILD A TREE WITHOUT A HYPOTHESIS!

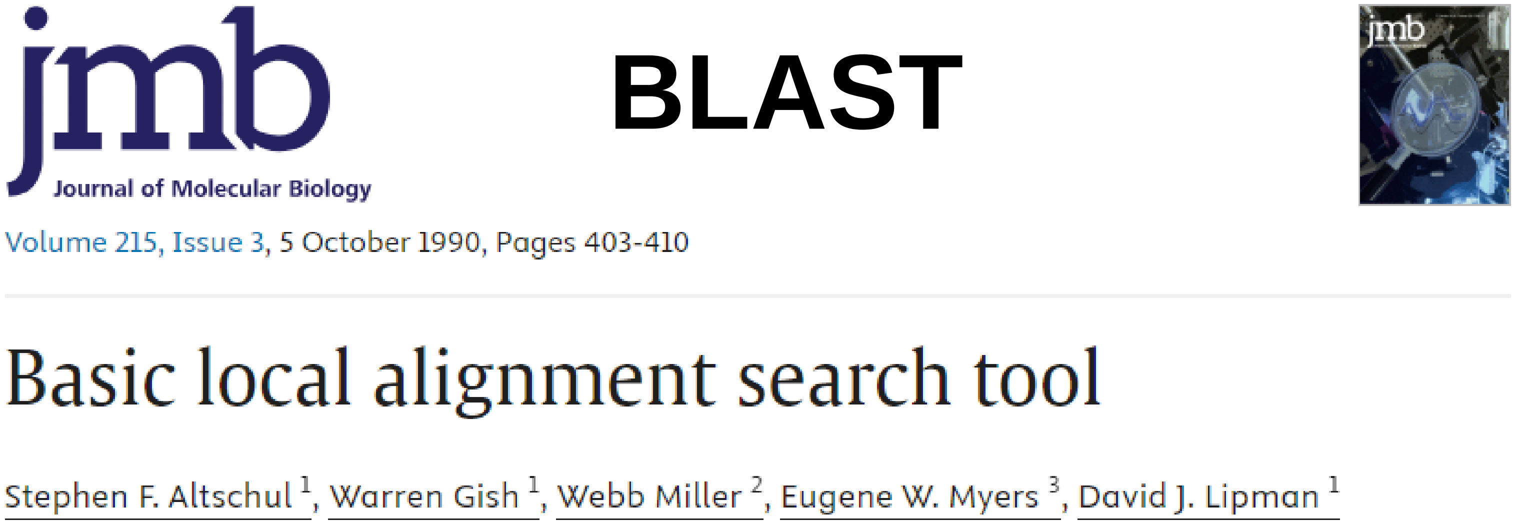
Tools and databases for finding homologues

- Sequence-based vs. Profile-based

Tools and databases for finding homologues

- **Sequence-based** vs. Profile-based

BLAST

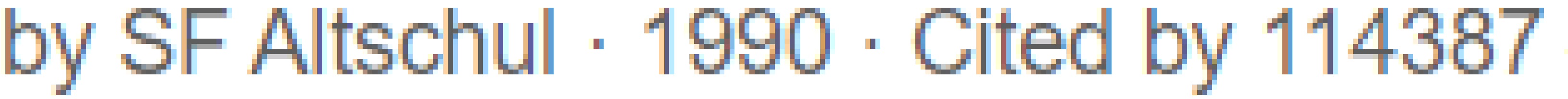


jmb Journal of Molecular Biology

Volume 215, Issue 3, 5 October 1990, Pages 403-410

Basic local alignment search tool

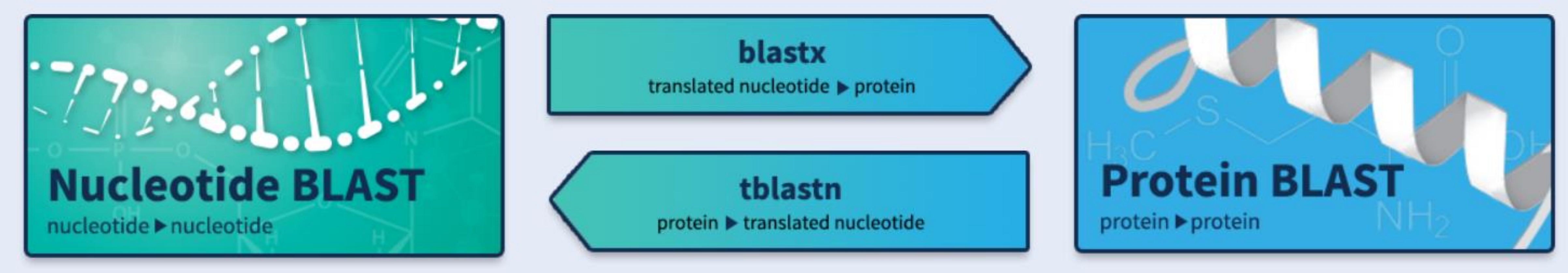
Stephen F. Altschul [1], Warren Gish [1], Webb Miller [2], Eugene W. Myers [3], David J. Lipman [1]



by SF Altschul · 1990 · Cited by 114387

Tools and databases for finding homologues

- **Sequence-based** vs. Profile-based



Nucleotide BLAST

nucleotide ▶ nucleotide

blastx

translated nucleotide ▶ protein

tblastn

protein ▶ translated nucleotide

Protein BLAST

protein ▶ protein

Tools and databases for finding homologues

- Sequence-based vs. Profile-based

Sequence-based

Pros:

- Needs only a single sequence
- Sensitive for closely related sequences

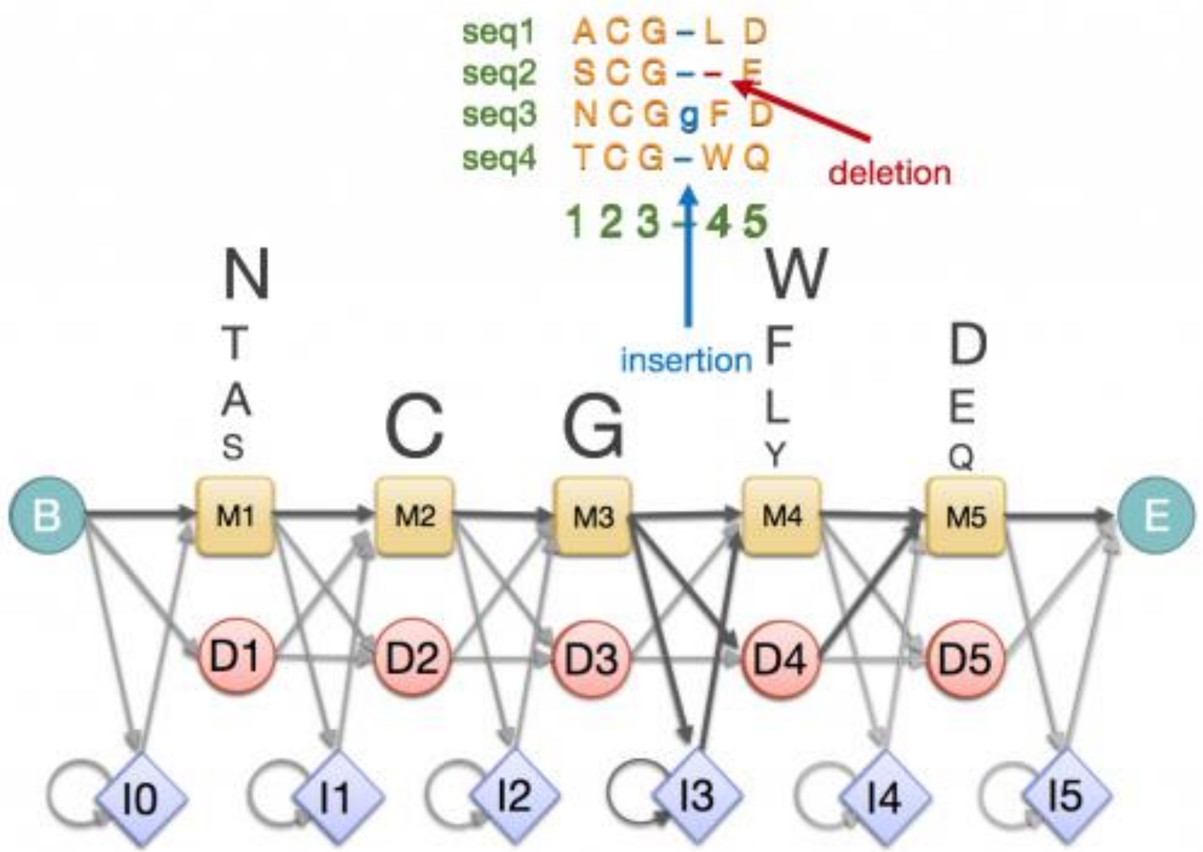
Cons:

- Only reprisents a single sequence
- Poor at identifying distant homologues
- Confoudned by low-complexity

Tools and databases for finding homologues

- Sequence-based vs. **Profile-based**

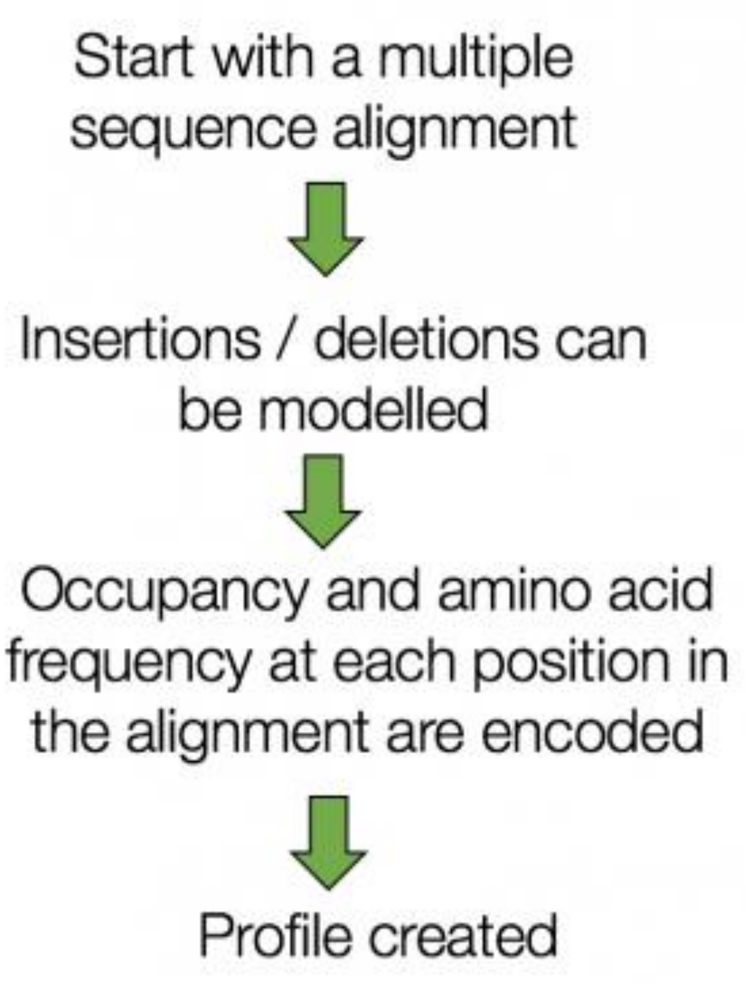
HMMER



Tools and databases for finding homologues

- Sequence-based vs. **Profile-based**

HMMER



Start with a multiple sequence alignment

Insertions / deletions can be modelled

Occupancy and amino acid frequency at each position in the alignment are encoded

Profile created

Tools and databases for finding homologues

- Sequence-based vs. Profile-based

Sequence-based

Pros:

- Needs only a single sequence
- Sensitive for closely related sequences

Cons:

- Only reprisents a single sequence
- Poor at identifying distant homologues
- Confounded by low-complexity

Profile-based

Pros:

- Can reprisent multiple sequences
- Information rich
- Can identify distant homologues

Cons:

- Dependent on the quality of the model
- Less sensitive

Exercise 1: Gather Homologues

1. Download the P450 sequences from sce.faa **1_homologues/exercise_1**

2. Familiarise yourself with the fasta format (.faa)

3. Go to NCBI and run a blastP search on these two proteins using default parameters

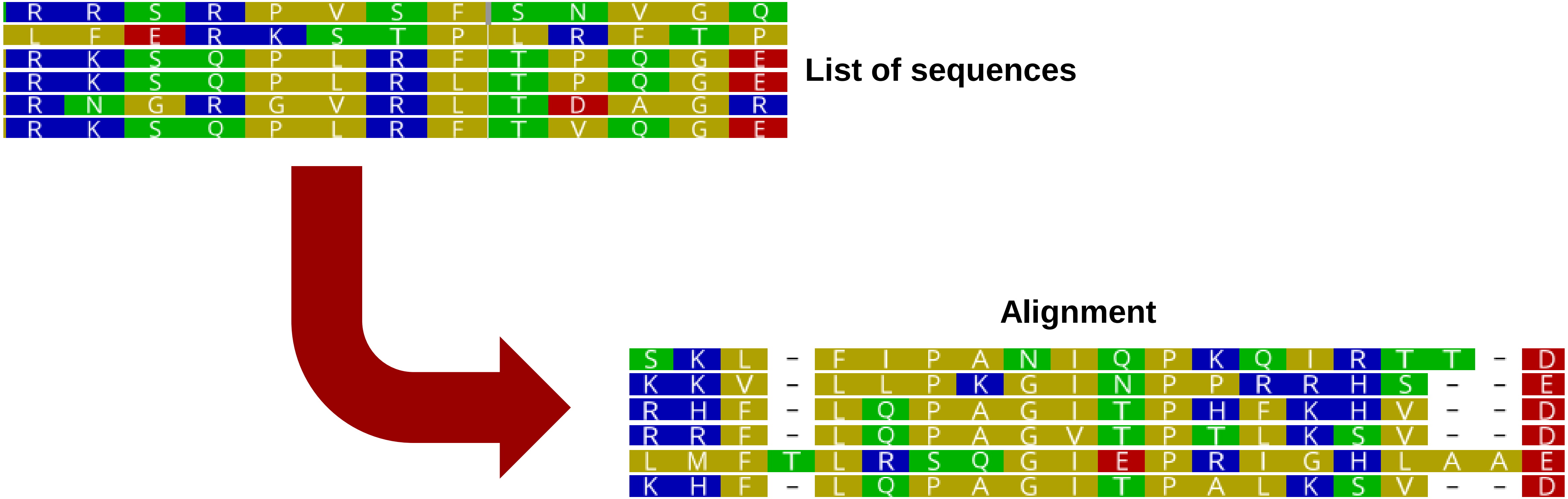
4. Download the results as an unaligned fasta file. **Can you notice any patterns?**

5. **If you have time,** try running against different databases. **What are the differences between these databases? Why might you use them?**

Practical Phylogenetics

Session 2b: Building an alignment

Step Two: Building an alignment



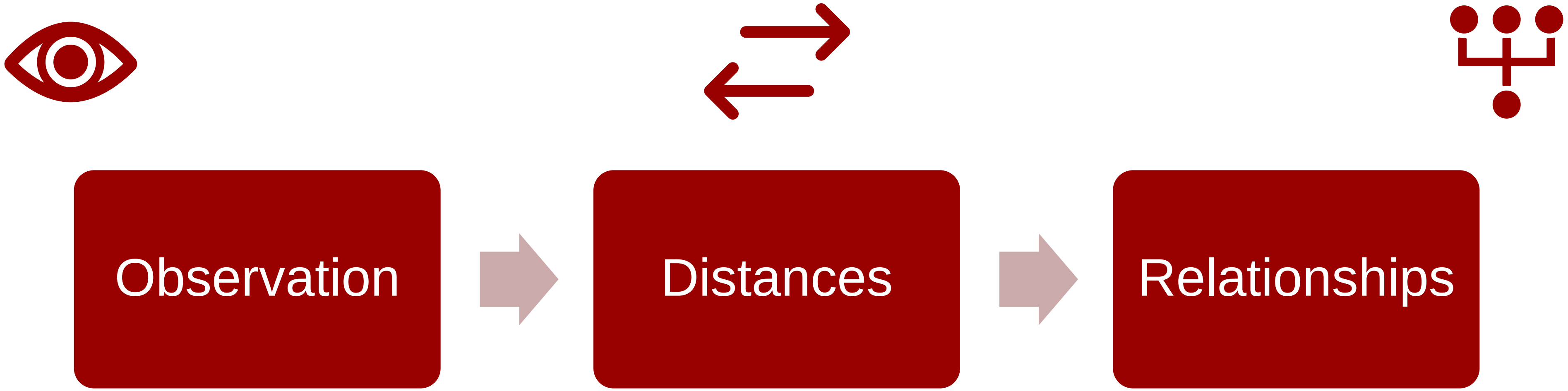
List of sequences

```
R R S R P V S F S N V G Q
L F E R K S T P L R F T P
R K S Q P L R F T P Q G E
R K S Q P L R L T P Q G E
R N G R G V R L T D A G R
R K S Q P L R F T V Q G E
```

Alignment

```
S K L - F I P A N I Q P K Q I R T T - D
K K V - L L P K G I N P P R R H S - - E
R H F - L Q P A G I T P H F K H V - - D
R R F - L Q P A G V T P T L K S V - - D
L M F T L R S Q G I E P R I G H L A A E
K H F - L Q P A G I T P A L K S V - - D
```

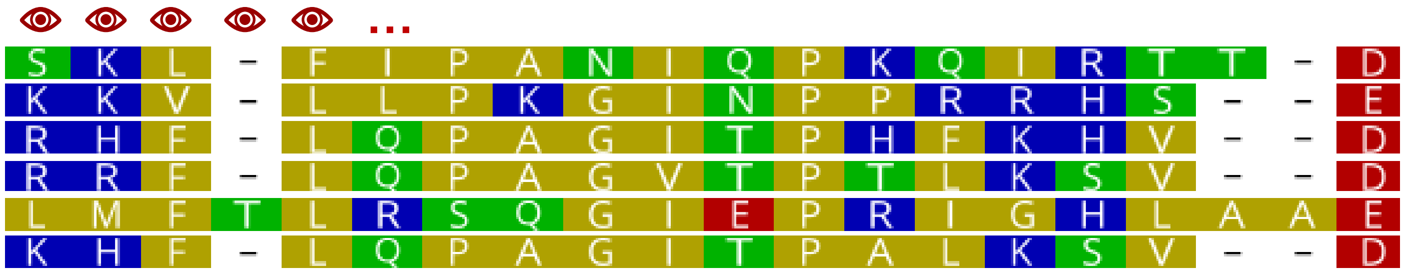
Step Two: Building an alignment



Observation

Distances

Relationships



```
...
S K L - F I P A N I Q P K Q I R T T - D
K K V - L L P K G I N P P R R H S - - E
R H F - L Q P A G I T P H F K H V - - D
R R F - L Q P A G V T P T L K S V - - D
L M F T L R S Q G I E P R I G H L A A E
K H F - L Q P A G I T P A L K S V - - D
```

Step Two: Building an alignment

There are many tools avaliable for alignment.

Some popular tools:

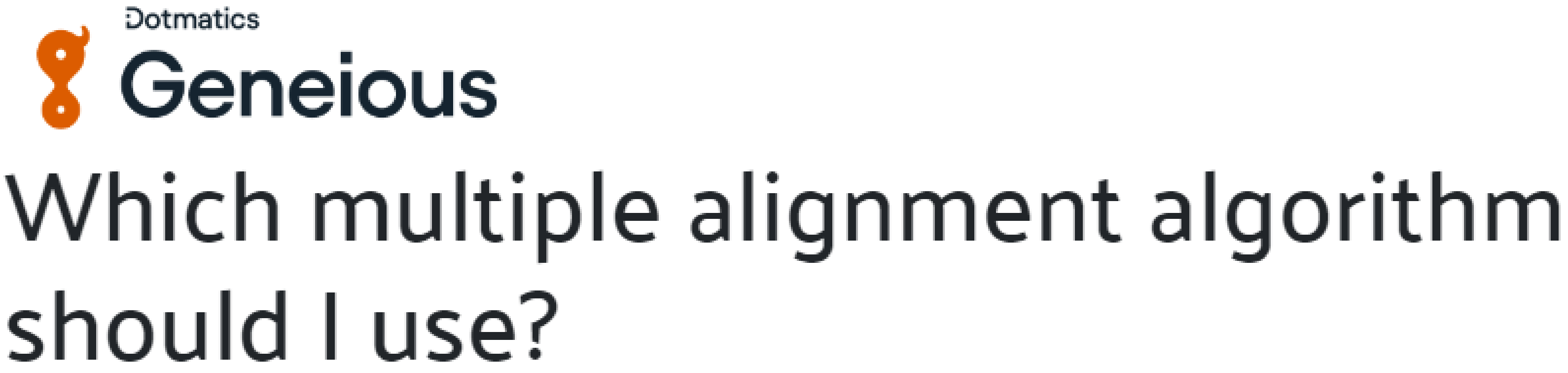
- Clustal
- MUSCLE
- Maft
- …

Step Two: Building an alignment

There are many tools avaliable for alignment.

Some popular tools:

- Clustal
- MUSCLE
- Maft
- …



Dotmatics Geneious

Which multiple alignment algorithm should I use?

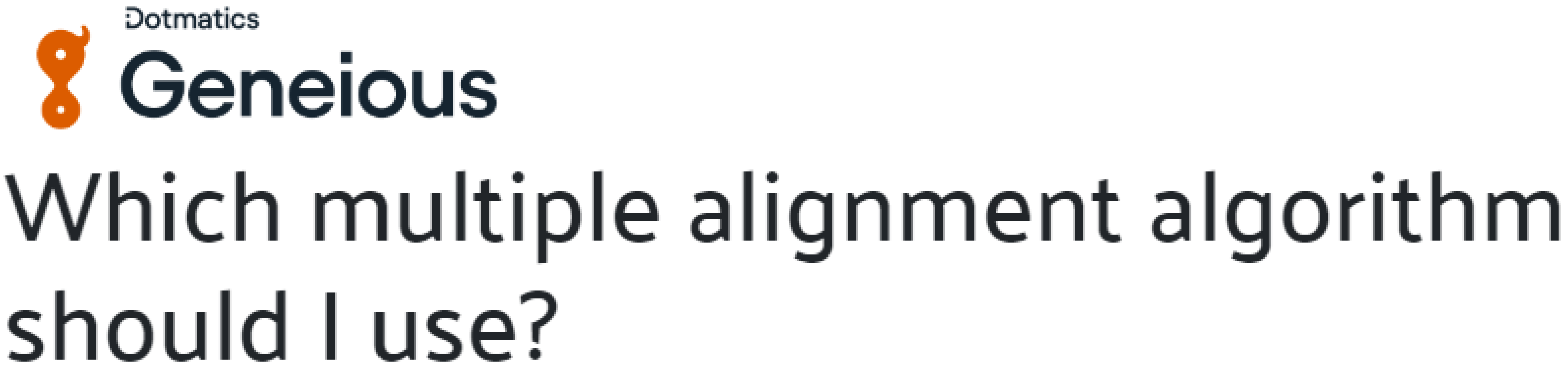
https://help.geneious.com/hc/en-us/articles/360044627712-Which-multiple-alignment-algorithm-should-I-use

Step Two: Building an alignment

There are many tools avaliable for alignment.

Some popular tools:

- Clustal
- MUSCLE
- Maft
- …



Dotmatics

Geneious

Which multiple alignment algorithm should I use?

https://help.geneious.com/hc/en-us/articles/360044627712-Which-multiple-alignment-algorithm-should-I-use

Who knows…

Exercise Two: Building an alignment

1. Combined the data from both files and the outgroup_p450s.faa in **1_homologues/exercise_2**

2. Go to: https://www.genome.jp/tools-bin/clustalw and set the input to ‘protein’ and the output to ‘fasta’ and run. **Do you get an error? What does this error mean for our study?**

3. Solve all the errors to finally get the alignment – or cheat when you get bored and download combined_clean.faa from **1_homologues/combined** and submit that to the server.

Thom’s Golden Rule #2

ALWAYS CHECK YOUR ALIGNMENTS!

Step Two: Building an alignment

The common problems:

1. **Incorrect input**

   E.g. wrong sequences, missing/incomplete sequences, sequences in the wrong complement etc. You are falible!

2. **Poor alignment**

   Especially in low complexity regions – particularly a problem with DNA!

3. **Bad trimming**

   The beginning and ends of sequences are often problematic!

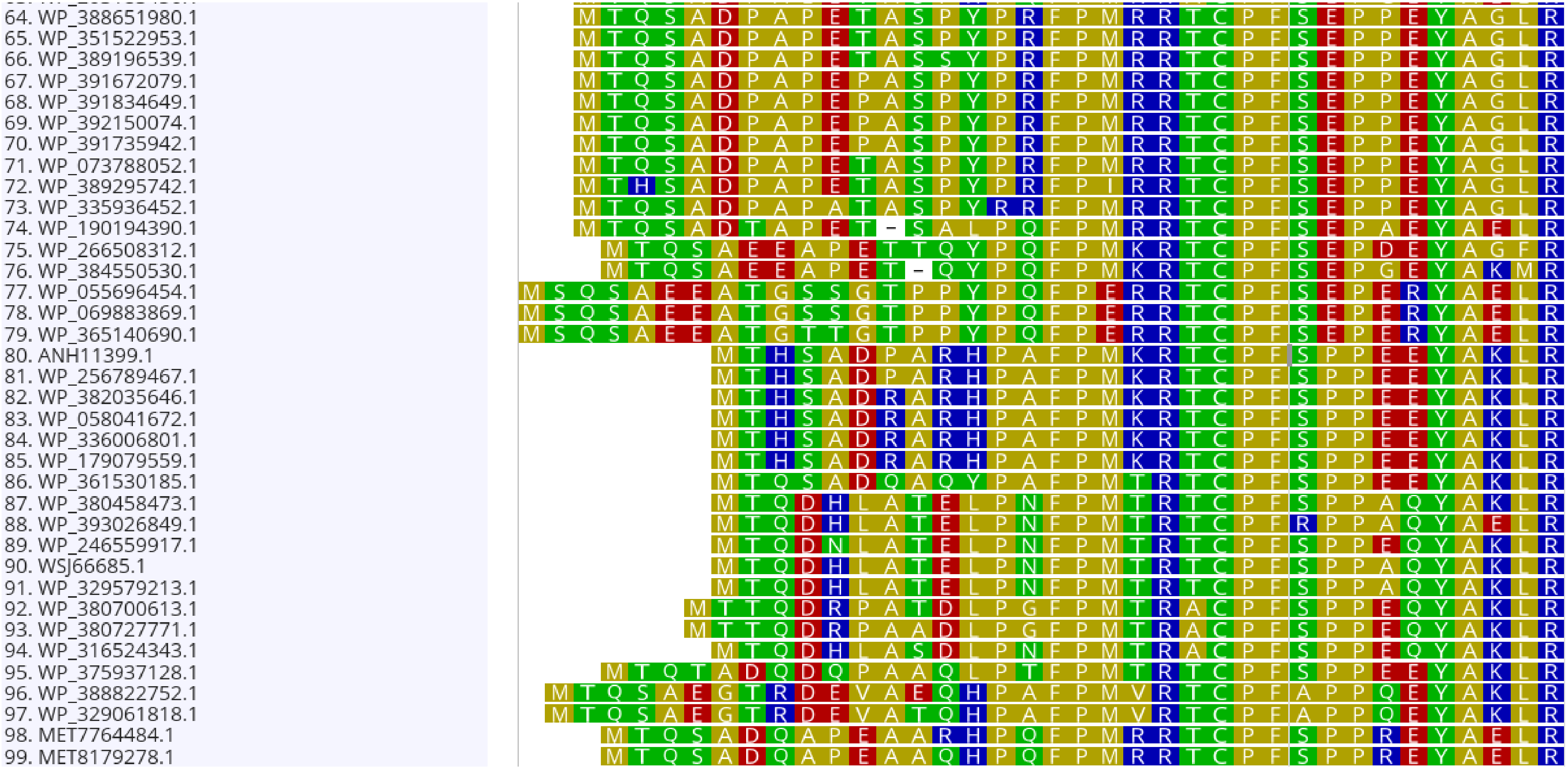
4. **Non-homologous regions**

   Be careful when studying multi-domain proteins!

Exercise Three: Building an alignment

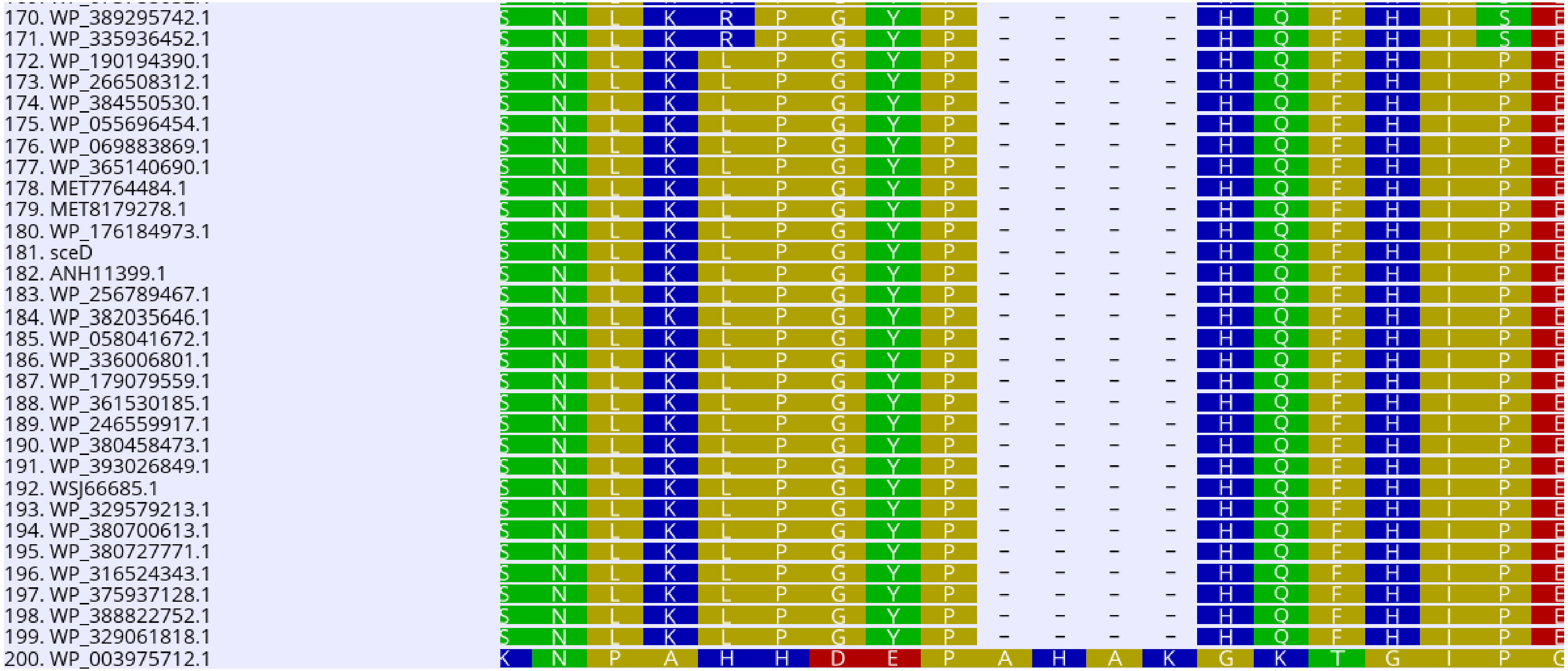
1. Go to: https://alignmentviewer.org/ and import your alignment
2. Pay particular attention to the start and the end of the sequences. **Can you spot any instances where the alignment does not reflect homology?**

Step Two: Building an alignment



```
64. WP_388651980.1    MTQSADPAPETASPYPRFPMRRTCPFSEPPEYAGLR
65. WP_351522953.1    MTQSADPAPETASPYPRFPMRRTCPFSEPPEYAGLR
66. WP_389196539.1    MTQSADPAPETASSYPRFPMRRTCPFSEPPEYAGLR
67. WP_391672079.1    MTQSADPAPEPASPYPRFPMRRTCPFSEPPEYAGLR
68. WP_391834649.1    MTQSADPAPEPASPYPRFPMRRTCPFSEPPEYAGLR
69. WP_392150074.1    MTQSADPAPEPASPYPRFPMRRTCPFSEPPEYAGLR
70. WP_391735942.1    MTQSADPAPEPASPYPRFPMRRTCPFSEPPEYAGLR
71. WP_073788052.1    MTQSADPAPETASPYPRFPMRRTCPFSEPPEYAGLR
72. WP_389295742.1    MTHSADPAPETASPYPRFPIRRTCPFSEPPEYAGLR
73. WP_335936452.1    MTQSADPAPATASPYRRFPMRRTCPFSEPPEYAGLR
74. WP_190194390.1    MTQSADTAPET-SALPQFPMRRTCPFSEPAEYAELR
75. WP_266508312.1     MTQSAEEAPETTQYPQFPMKRTCPFSEPDEYAGFR
76. WP_384550530.1     MTQSAEEAPET-QYPQFPMKRTCPFSEPGEYAKMR
77. WP_055696454.1  MSQSAEEATGSSGTPPYPQFPERRTCPFSEPERYAELR
78. WP_069883869.1  MSQSAEEATGSSGTPPYPQFPERRTCPFSEPERYAELR
79. WP_365140690.1  MSQSAEEATGTTGTPPYPQFPERRTCPFSEPERYAELR
80. ANH11399.1             MTHSADPARHPAFPMKRTCPFSPPEEYAKLR
81. WP_256789467.1         MTHSADPARHPAFPMKRTCPFSPPEEYAKLR
82. WP_382035646.1         MTHSADRARHPAFPMKRTCPFSPPEEYAKLR
83. WP_058041672.1         MTHSADRARHPAFPMKRTCPFSPPEEYAKLR
84. WP_336006801.1         MTHSADRARHPAFPMKRTCPFSPPEEYAKLR
85. WP_179079559.1         MTHSADRARHPAFPMKRTCPFSPPEEYAKLR
86. WP_361530185.1         MTQSADQAQYPAFPMTRTCPFSPPEEYAKLR
87. WP_380458473.1         MTQDHLATELPNFPMTRTCPFSPPAQYAKLR
88. WP_393026849.1         MTQDHLATELPNFPMTRTCPFRPPAQYAELR
89. WP_246559917.1         MTQDNLATELPNFPMTRTCPFSPPEQYAKLR
90. WSJ66685.1             MTQDHLATELPNFPMTRTCPFSPPAQYAKLR
91. WP_329579213.1         MTQDHLATELPNFPMTRTCPFSPPAQYAKLR
92. WP_380700613.1        MTTQDRPATDLPGFPMTRACPFSPPEQYAKLR
93. WP_380727771.1        MTTQDRPAADLPGFPMTRACPFSPPEQYAKLR
94. WP_316524343.1         MTQDHLASDLPNFPMTRACPFSPPEQYAKLR
95. WP_375937128.1     MTQTADQDQPAAQLPTFPMTRTCPFSPPEEYAKLR
96. WP_388822752.1   MTQSAEGTRDEVAEQHPAFPMVRTCPFAPPQEYAKLR
97. WP_329061818.1   MTQSAEGTRDEVATQHPAFPMVRTCPFAPPQEYAKLR
98. MET7764484.1       MTQSADQAPEAARHPQFPMRRTCPFSPPREYAELR
99. MET8179278.1       MTQSADQAPEAAQHPQFPMRRTCPFSPPREYAELR
```

Step Two: Building an alignment



```
170. WP_389295742.1   S N L K R P G Y P - - - - H Q F H I S E
171. WP_335936452.1   S N L K R P G Y P - - - - H Q F H I S E
172. WP_190194390.1   S N L K L P G Y P - - - - H Q F H I P E
173. WP_266508312.1   S N L K L P G Y P - - - - H Q F H I P E
174. WP_384550530.1   S N L K L P G Y P - - - - H Q F H I P E
175. WP_055696454.1   S N L K L P G Y P - - - - H Q F H I P E
176. WP_069883869.1   S N L K L P G Y P - - - - H Q F H I P E
177. WP_365140690.1   S N L K L P G Y P - - - - H Q F H I P E
178. MET7764484.1     S N L K L P G Y P - - - - H Q F H I P E
179. MET8179278.1     S N L K L P G Y P - - - - H Q F H I P E
180. WP_176184973.1   S N L K L P G Y P - - - - H Q F H I P E
181. sceD             S N L K L P G Y P - - - - H Q F H I P E
182. ANH11399.1       S N L K L P G Y P - - - - H Q F H I P E
183. WP_256789467.1   S N L K L P G Y P - - - - H Q F H I P E
184. WP_382035646.1   S N L K L P G Y P - - - - H Q F H I P E
185. WP_058041672.1   S N L K L P G Y P - - - - H Q F H I P E
186. WP_336006801.1   S N L K L P G Y P - - - - H Q F H I P E
187. WP_179079559.1   S N L K L P G Y P - - - - H Q F H I P E
188. WP_361530185.1   S N L K L P G Y P - - - - H Q F H I P E
189. WP_246559917.1   S N L K L P G Y P - - - - H Q F H I P E
190. WP_380458473.1   S N L K L P G Y P - - - - H Q F H I P E
191. WP_393026849.1   S N L K L P G Y P - - - - H Q F H I P E
192. WSJ66685.1       S N L K L P G Y P - - - - H Q F H I P E
193. WP_329579213.1   S N L K L P G Y P - - - - H Q F H I P E
194. WP_380700613.1   S N L K L P G Y P - - - - H Q F H I P E
195. WP_380727771.1   S N L K L P G Y P - - - - H Q F H I P E
196. WP_316524343.1   S N L K L P G Y P - - - - H Q F H I P E
197. WP_375937128.1   S N L K L P G Y P - - - - H Q F H I P E
198. WP_388822752.1   S N L K L P G Y P - - - - H Q F H I P E
199. WP_329061818.1   S N L K L P G Y P - - - - H Q F H I P E
200. WP_003975712.1   K N P A H H D E P A H A K G K T G I P G
```

Exercise 4: Building an alignment

1. Before you forget… trim your alignment!

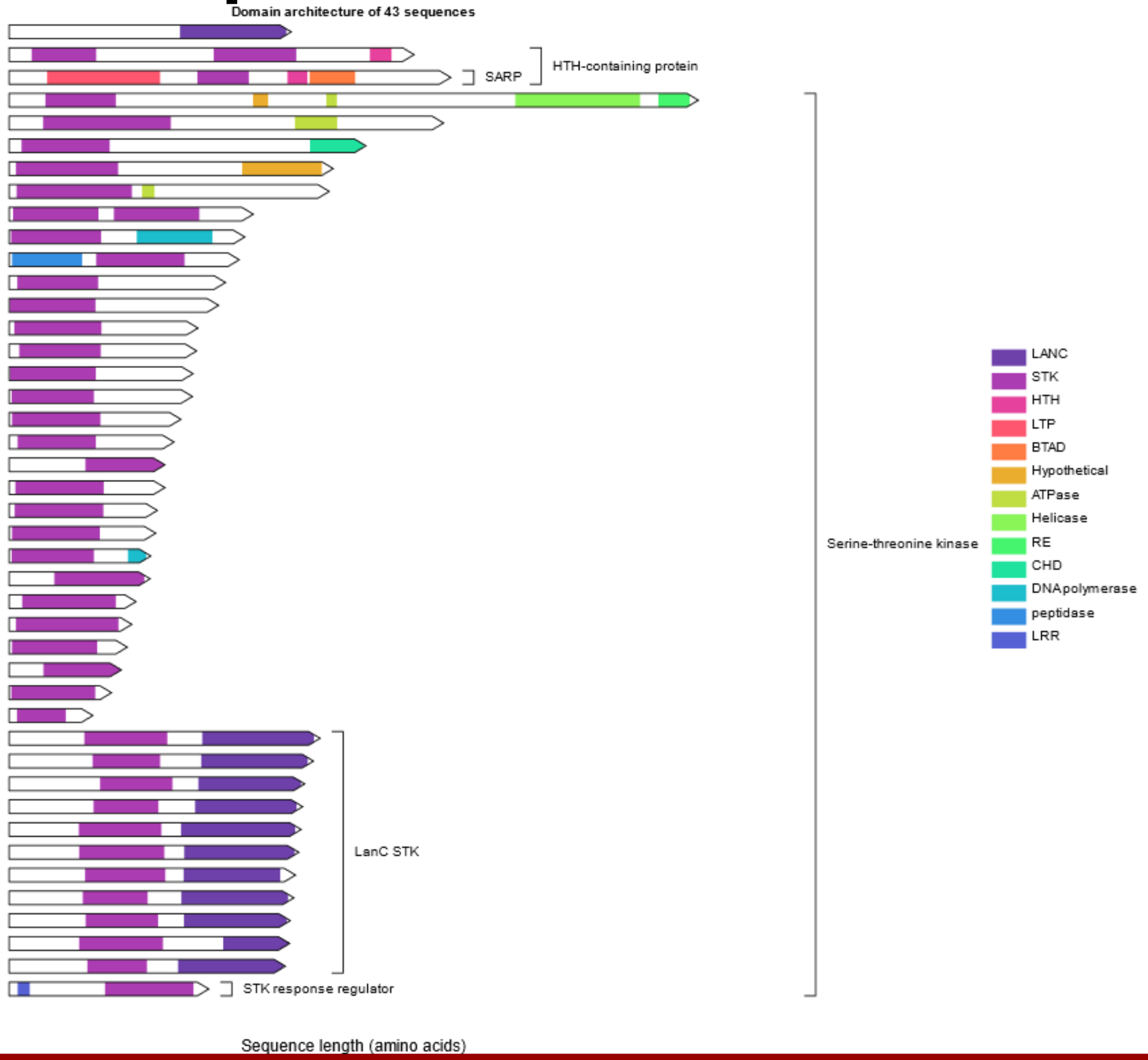
2. Now for a different example, download **lanc.faa** from **2_aligments/exercise_4**.

3. Check the alignment? **Does it look well aligned? Why/ why not?**

4. Go to NCBI's CDD search: https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi. Pick a few sequences and submit them. **What domains are each sequence made up from? Are these proteins homologous (be careful with your answer!!!)?**

5. **Is an alignment of these proteins a suitable way to address their evolutionary history? Why/ why not? How could we proceed with these proteins?**

A note on protein domains



Thom’s Golden Rule #3

ONLY ALIGN PROTEINS IF ALL DOMAINS ARE HOMOLOGOUS!

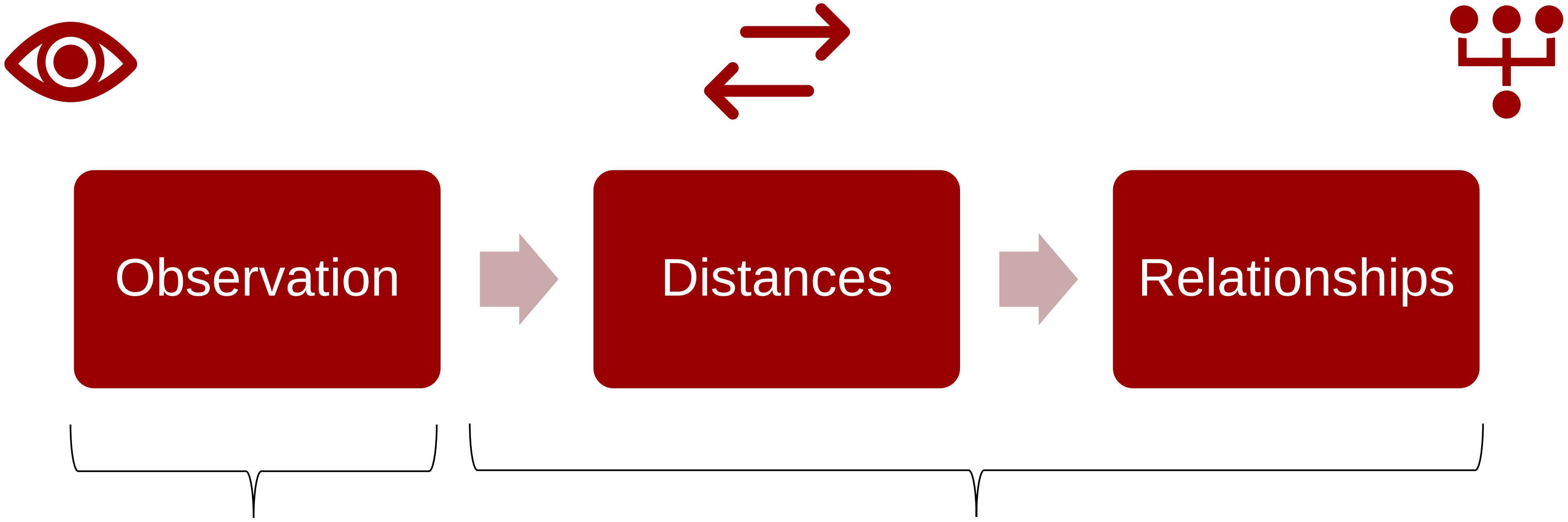
Step Two: Build an alignment

1. Check your alignment.
2. Trim if necissary.
3. When you are happy with you alignment save it for the next step.

Practical Phylogenetics

Session 2c: Building a Tree

Step 3: Building Trees



Observation

Distances

Relationships

Building the alignment

Building the tree

P	K	Q	I	R	T	T	–	D
P	P	R	R	H	S	–	–	E
P	H	F	K	H	V	–	–	D
P	T	L	K	S	V	–	–	D
P	R	I	G	H	L	A	A	E
P	A	L	K	S	V	–	–	D

Step 3: Building Trees

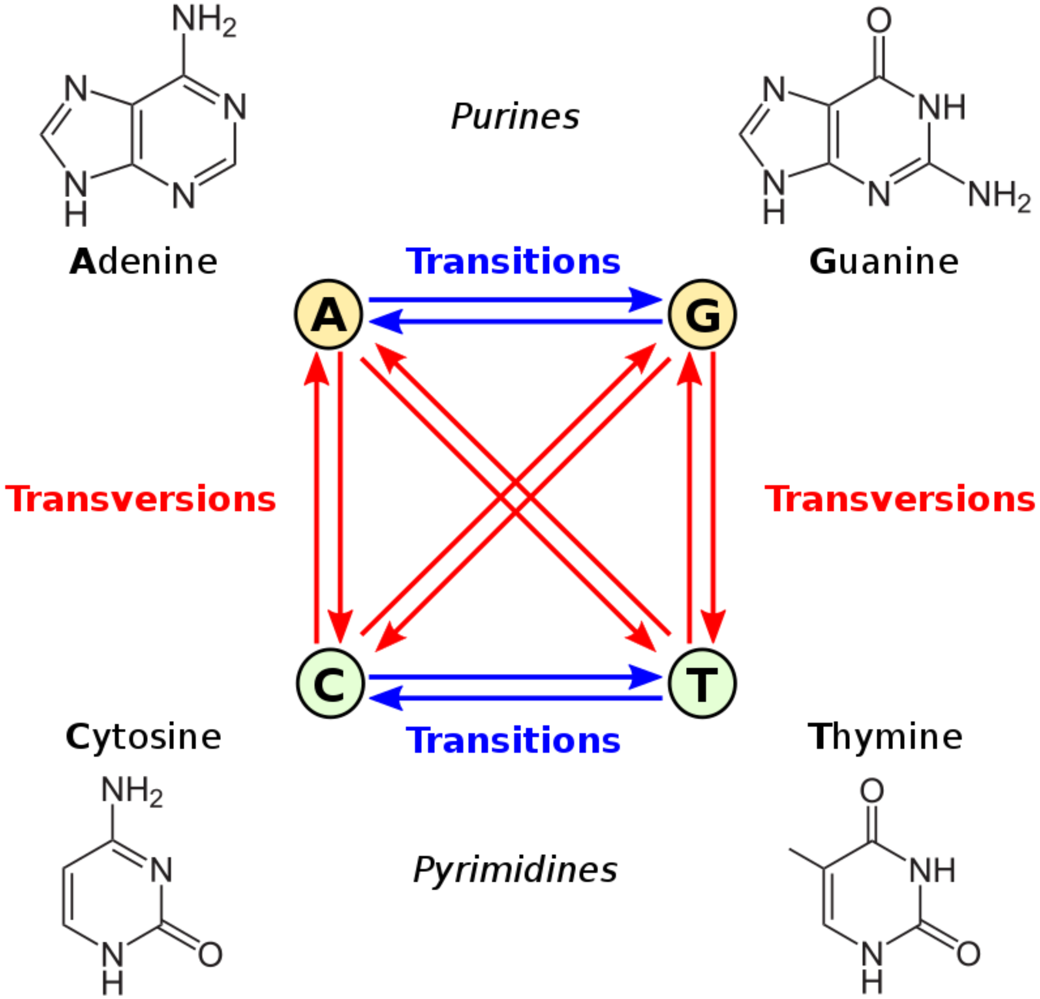
Distance – measure the genetic distance (under an evolutionary model)

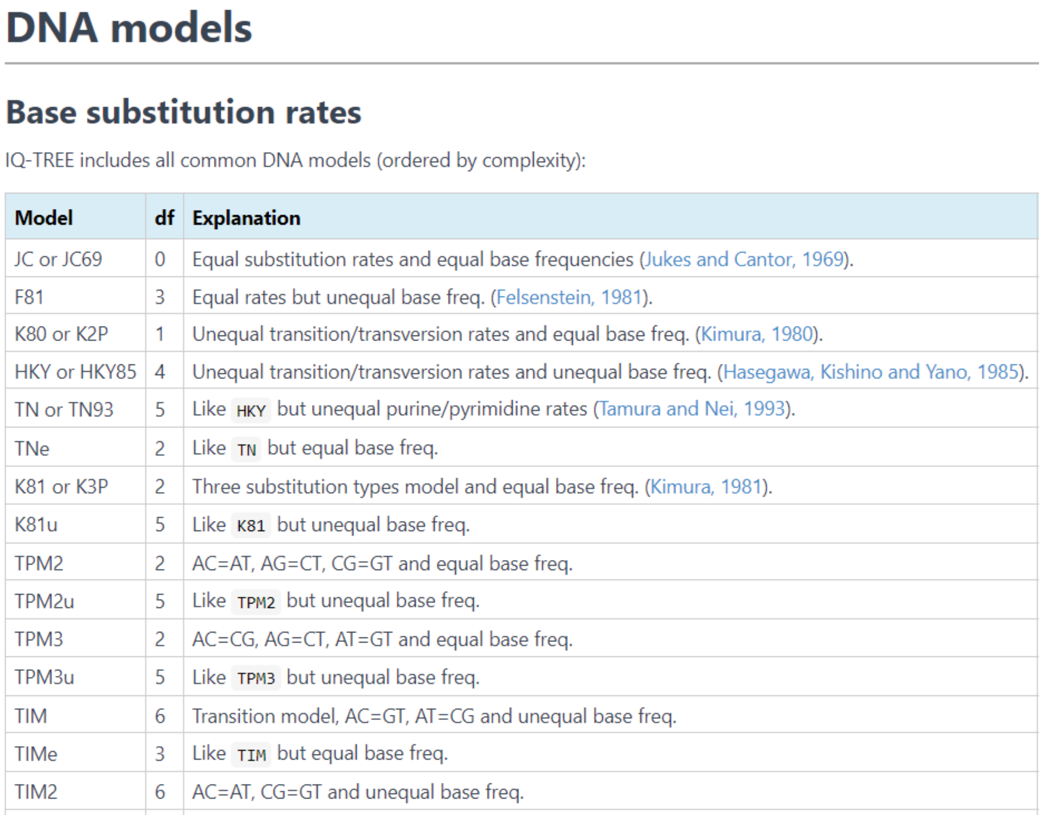
Maximum parsimony – build the shortest tree (the smallest required number of changes)

Maximum Likelihood – find the tree that has the highest liklihood to reprisent the alignment

Bayesian – find the tree that has the best ‘posterior probability’

Step 3: Building Trees





DNA models

Base substitution rates

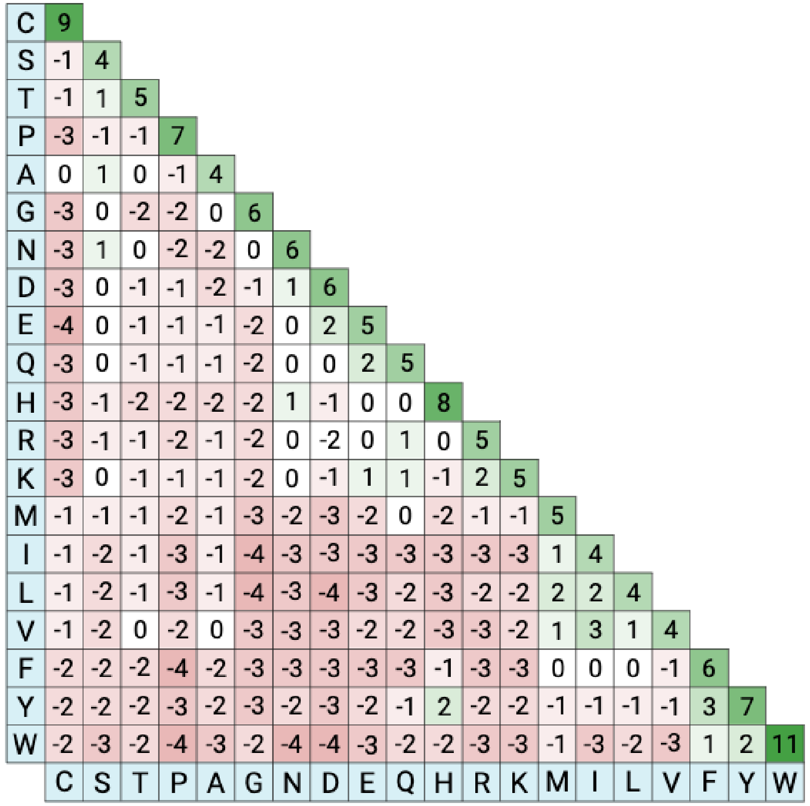
IQ-TREE includes all common DNA models (ordered by complexity):

Model	df	Explanation
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).
TN or TN93	5	Like `HKY` but unequal purine/pyrimidine rates (Tamura and Nei, 1993).
TNe	2	Like `TN` but equal base freq.
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).
K81u	5	Like `K81` but unequal base freq.
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.
TPM2u	5	Like `TPM2` but unequal base freq.
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.
TPM3u	5	Like `TPM3` but unequal base freq.
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.
TIMe	3	Like `TIM` but equal base freq.
TIM2	6	AC=AT, CG=GT and unequal base freq.

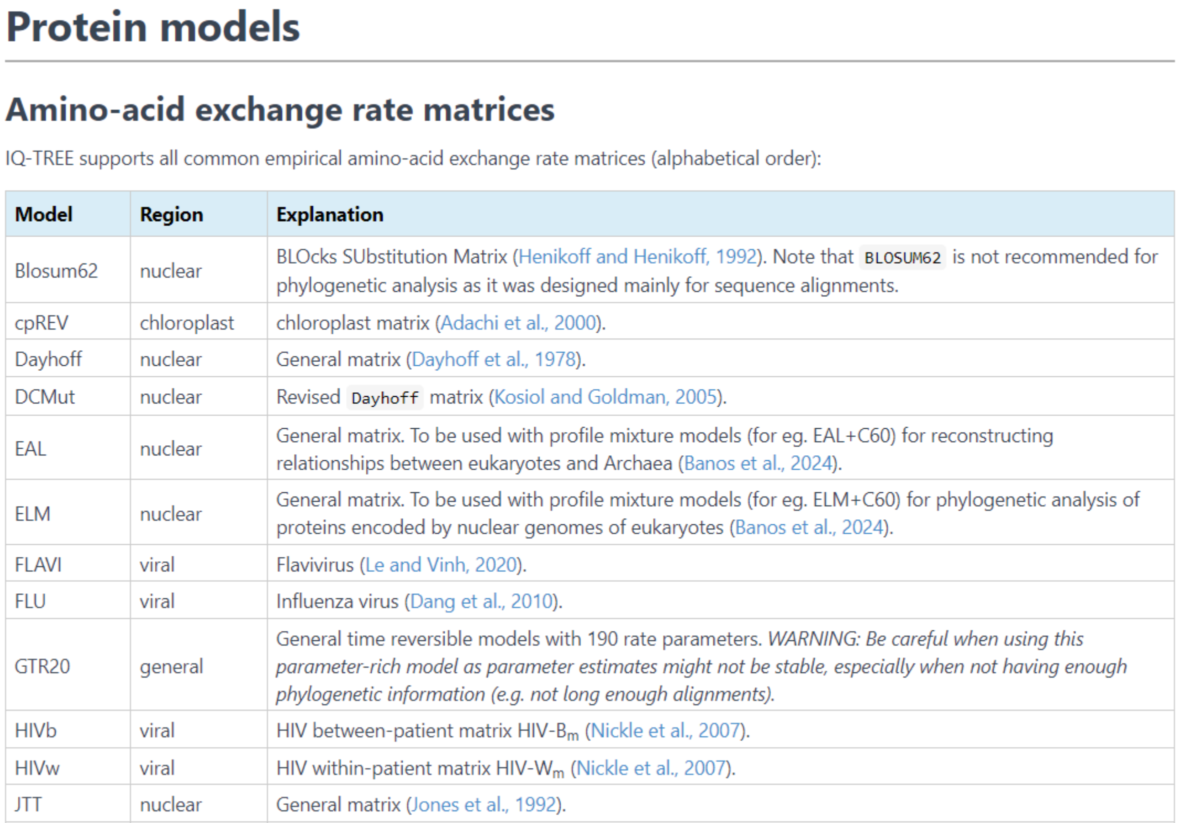
Mihn et al. *Mol Biol and Evol*, 2020

Step 3: Building Trees

BLOSUM 62 Amino Acid Substitution Matrix



C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W



Protein models

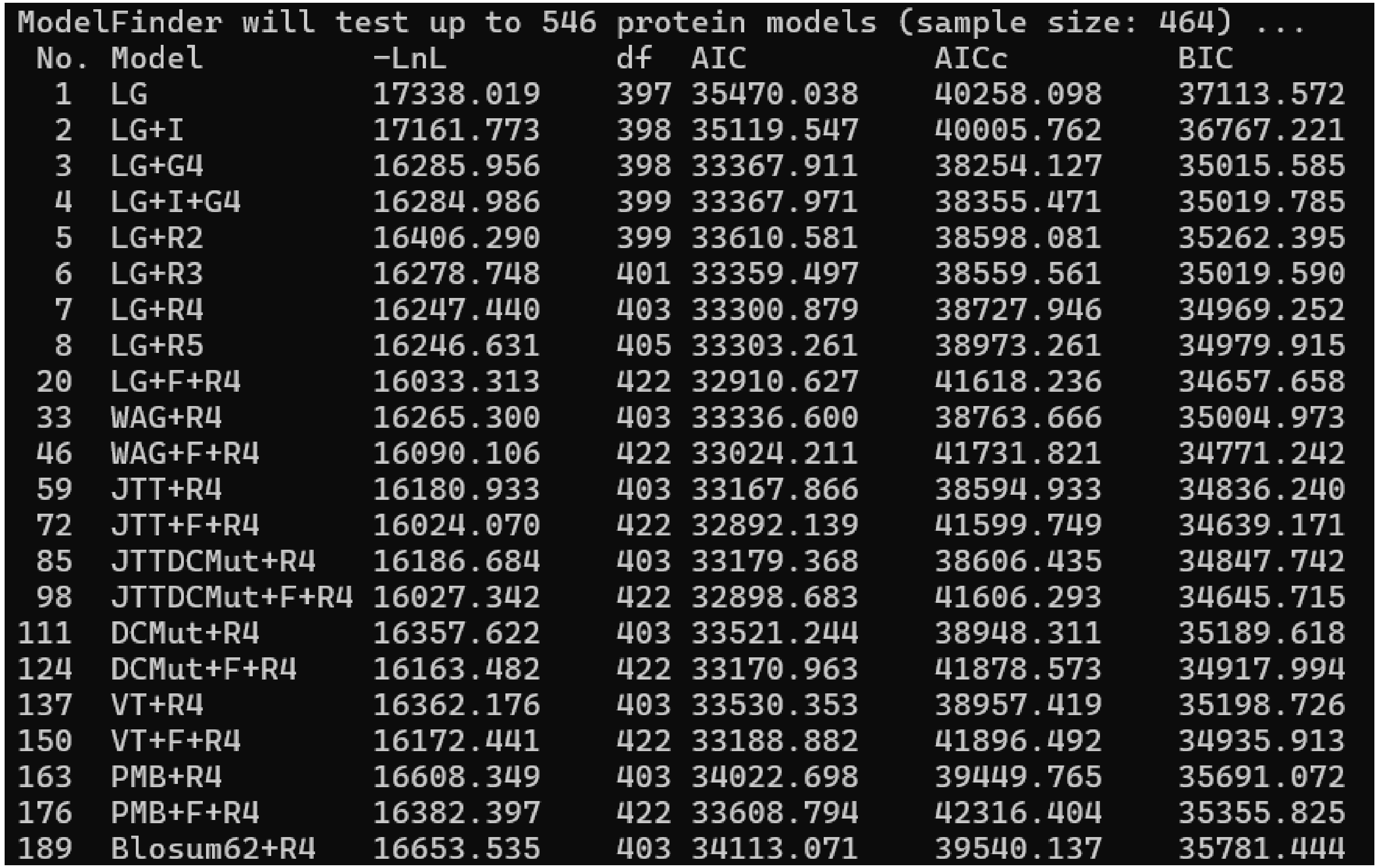
Amino-acid exchange rate matrices

IQ-TREE supports all common empirical amino-acid exchange rate matrices (alphabetical order):

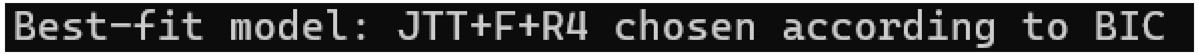
Model	Region	Explanation
Blosum62	nuclear	BLOcks SUbstitution Matrix (Henikoff and Henikoff, 1992). Note that `BLOSUM62` is not recommended for phylogenetic analysis as it was designed mainly for sequence alignments.
cpREV	chloroplast	chloroplast matrix (Adachi et al., 2000).
Dayhoff	nuclear	General matrix (Dayhoff et al., 1978).
DCMut	nuclear	Revised `Dayhoff` matrix (Kosiol and Goldman, 2005).
EAL	nuclear	General matrix. To be used with profile mixture models (for eg. EAL+C60) for reconstructing relationships between eukaryotes and Archaea (Banos et al., 2024).
ELM	nuclear	General matrix. To be used with profile mixture models (for eg. ELM+C60) for phylogenetic analysis of proteins encoded by nuclear genomes of eukaryotes (Banos et al., 2024).
FLAVI	viral	Flavivirus (Le and Vinh, 2020).
FLU	viral	Influenza virus (Dang et al., 2010).
GTR20	general	General time reversible models with 190 rate parameters. *WARNING: Be careful when using this parameter-rich model as parameter estimates might not be stable, especially when not having enough phylogenetic information (e.g. not long enough alignments).*
HIVb	viral	HIV between-patient matrix HIV-B$_m$ (Nickle et al., 2007).
HIVw	viral	HIV within-patient matrix HIV-W$_m$ (Nickle et al., 2007).
JTT	nuclear	General matrix (Jones et al., 1992).

Mihn et al. *Mol Biol and Evol*, 2020

Step 3: Building Trees



```
ModelFinder will test up to 546 protein models (sample size: 464) ...
 No. Model         -LnL         df  AIC          AICc         BIC
  1  LG            17338.019    397 35470.038    40258.098    37113.572
  2  LG+I          17161.773    398 35119.547    40005.762    36767.221
  3  LG+G4         16285.956    398 33367.911    38254.127    35015.585
  4  LG+I+G4       16284.986    399 33367.971    38355.471    35019.785
  5  LG+R2         16406.290    399 33610.581    38598.081    35262.395
  6  LG+R3         16278.748    401 33359.497    38559.561    35019.590
  7  LG+R4         16247.440    403 33300.879    38727.946    34969.252
  8  LG+R5         16246.631    405 33303.261    38973.261    34979.915
 20  LG+F+R4       16033.313    422 32910.627    41618.236    34657.658
 33  WAG+R4        16265.300    403 33336.600    38763.666    35004.973
 46  WAG+F+R4      16090.106    422 33024.211    41731.821    34771.242
 59  JTT+R4        16180.933    403 33167.866    38594.933    34836.240
 72  JTT+F+R4      16024.070    422 32892.139    41599.749    34639.171
 85  JTTDCMut+R4   16186.684    403 33179.368    38606.435    34847.742
 98  JTTDCMut+F+R4 16027.342    422 32898.683    41606.293    34645.715
111  DCMut+R4      16357.622    403 33521.244    38948.311    35189.618
124  DCMut+F+R4    16163.482    422 33170.963    41878.573    34917.994
137  VT+R4         16362.176    403 33530.353    38957.419    35198.726
150  VT+F+R4       16172.441    422 33188.882    41896.492    34935.913
163  PMB+R4        16608.349    403 34022.698    39449.765    35691.072
176  PMB+F+R4      16382.397    422 33608.794    42316.404    35355.825
189  Blosum62+R4   16653.535    403 34113.071    39540.137    35781.444
```



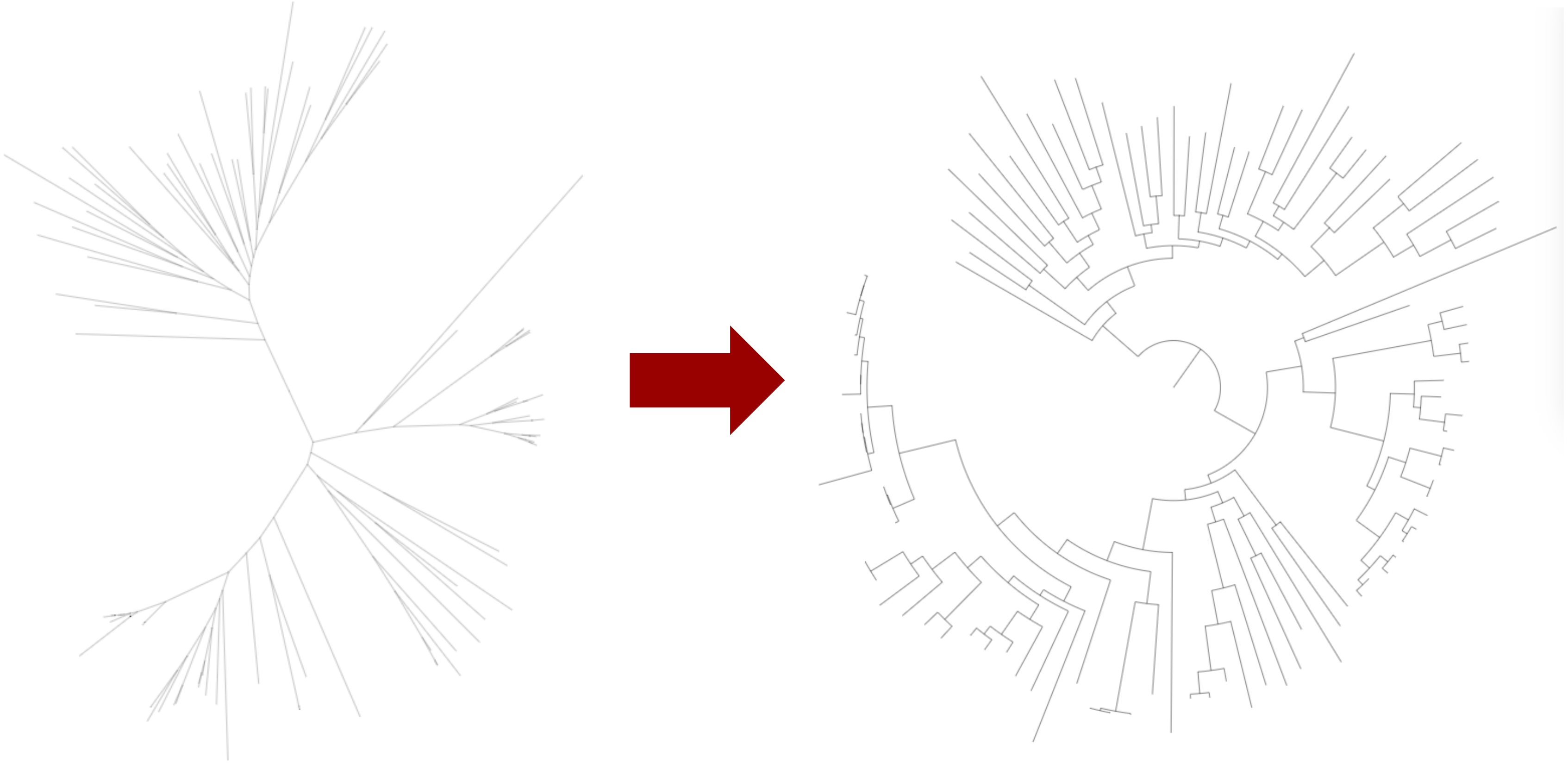
```
Best-fit model: JTT+F+R4 chosen according to BIC
```

Thom’s Golden Rule #4

MORE MATHS DOES NOT MEAN A BETTER TREE!

Step 3: Building Trees

An **outgroup** is a selection taxa that you know are the earliest diverging. This is important if you want to root your tree. Most software do not require an outgroup but correct rooting is necissary if you want to infer the order of evolutionary events.



Step Three: Inferring our tree

1. Go to http://iqtree.cibiv.univie.ac.at/
2. Input your data, select the data type and hit run.

Learning Objectives

- Explain the different types of data that can be used to infer a tree. Understand the advantages of using DNA or protein sequences.

- Use web-based software to identify homologues, align sequences and infer phylogenetic trees and have a general knowledge of the command line tools available and use these techniques to infer trees for genes of interest.
- Avoid common errors that can occur during alignment and inference.

Practical Phylogenetics

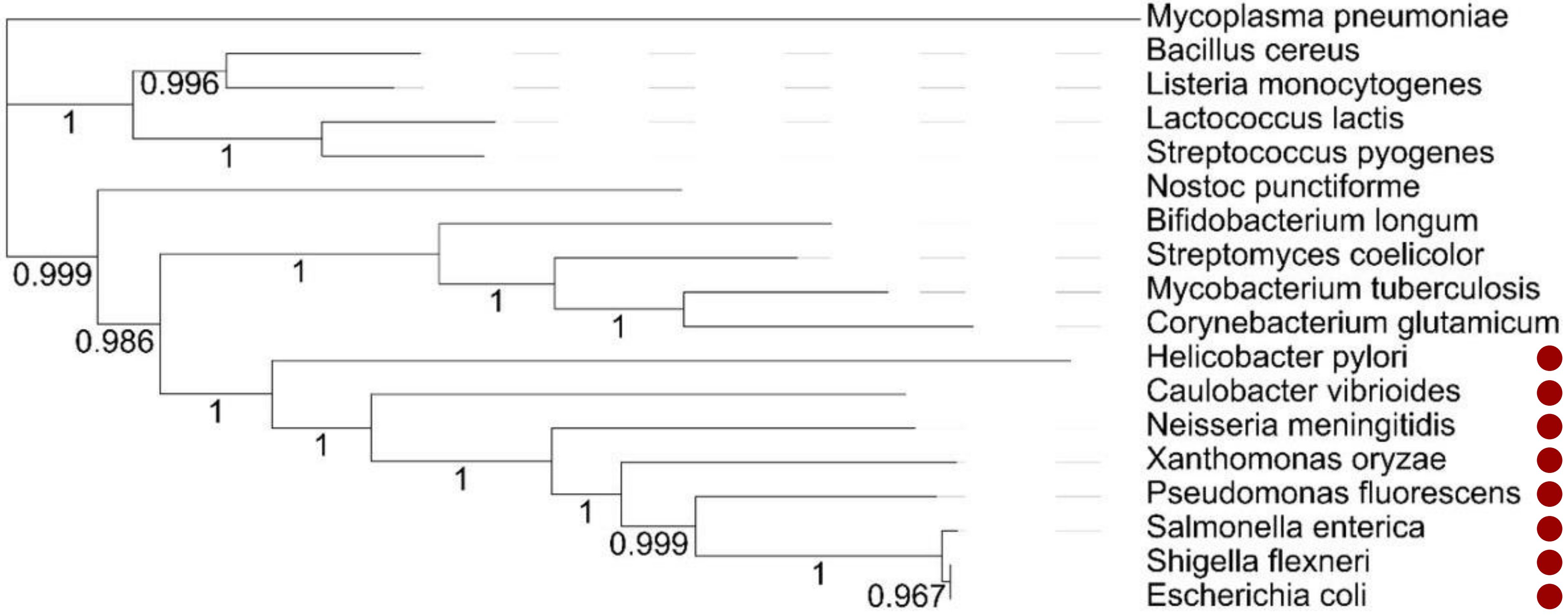
Session 3: Interpreting Trees

Session 3: Learning Objectives

- Confidently identify common phylogenetic patterns such as: polyphyly, paraphyly, monophyly, polytomy and understand the differences between orthologues, paralogues and homologues.
- Critically examine the quality of phylogenetic trees, including an understanding of the importance of outgroups and the difference between a rooted and unrooted tree understand the differences between likelihood, bootstrap, and consensus scores.

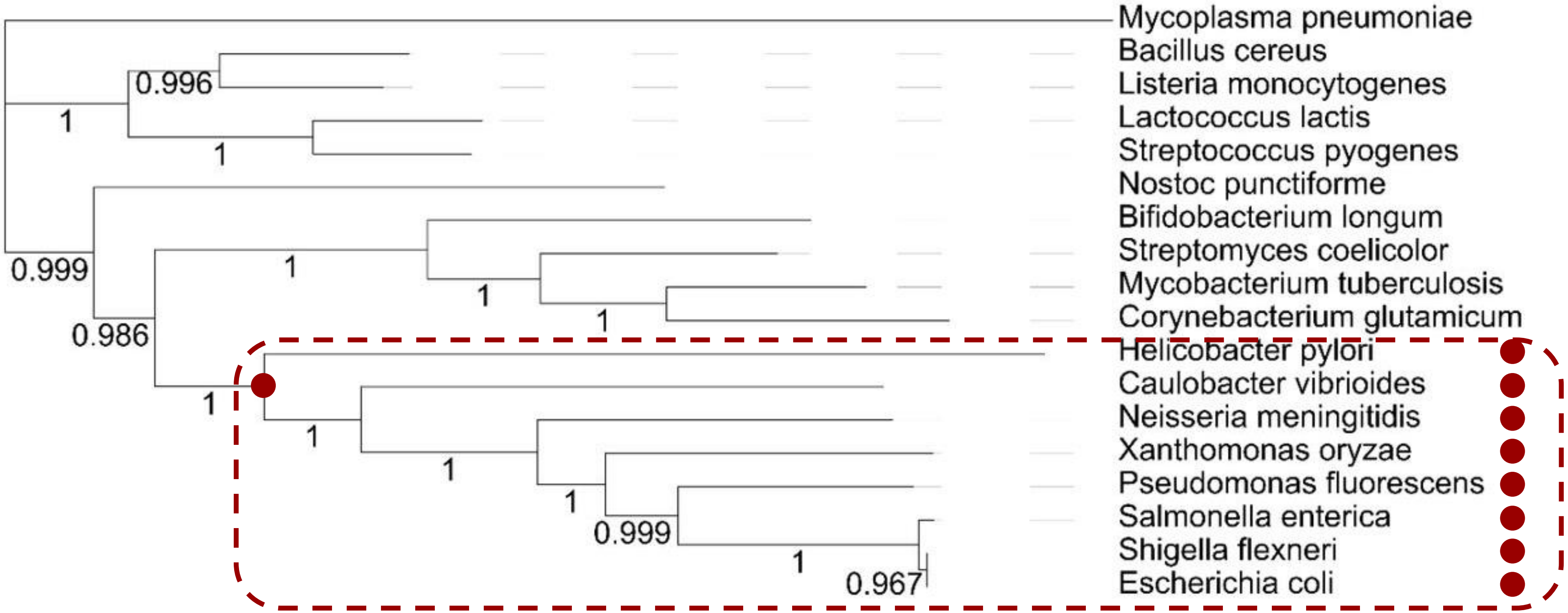
Monophyly

- A group is considered monophyletic if the descendants of the most recent common ancestor (MRCA) are reprisented by all members of the group.



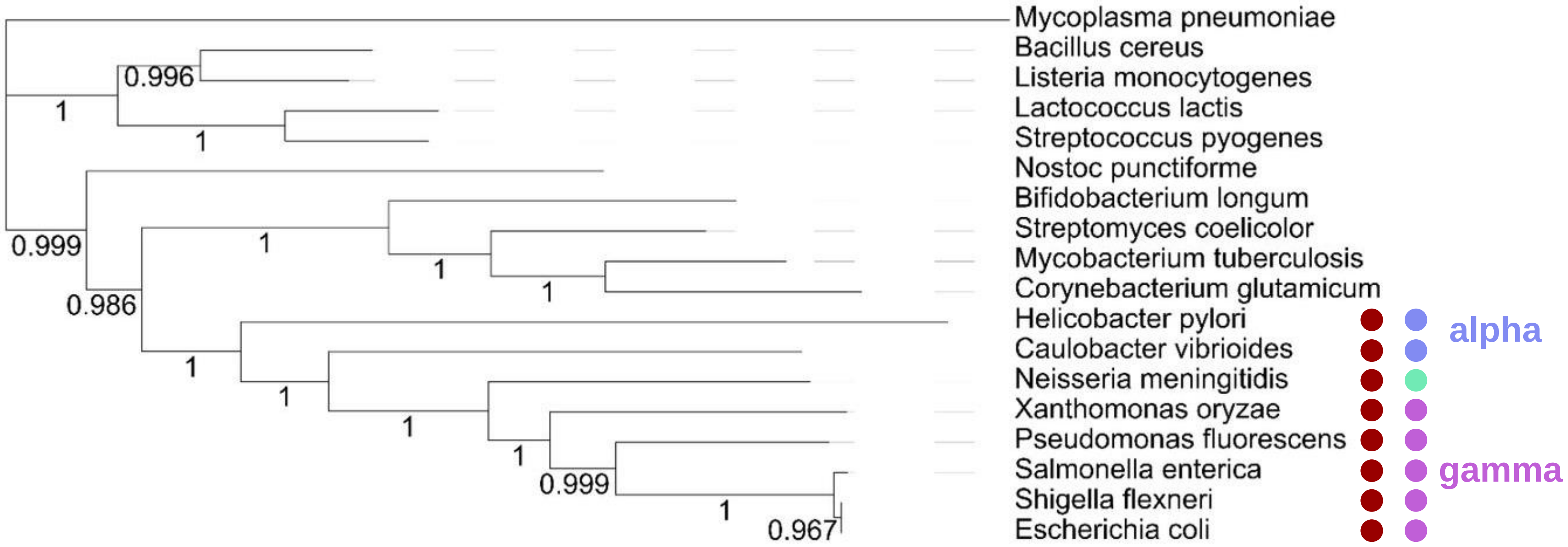
Monophyly

- A group is considered monophyletic if the descendants of the most recent common ancestor (MRCA) are reprisented by all members of the group.



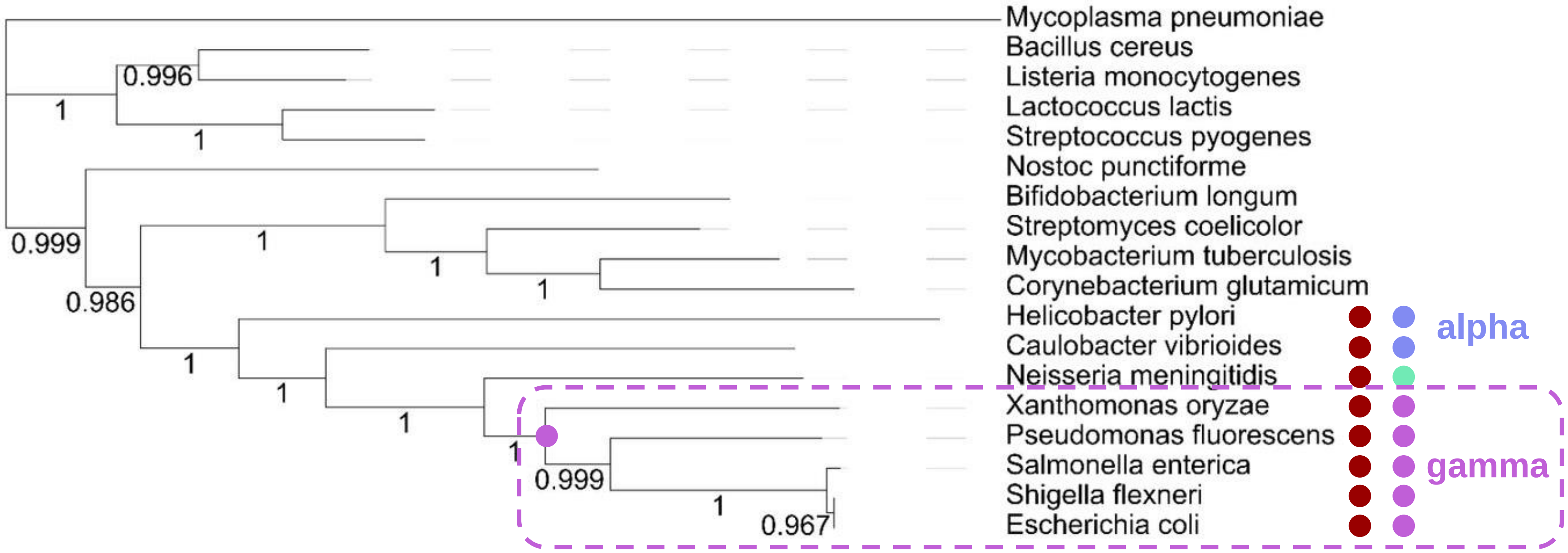
Monophyly

- A group is considered monophyletic if the descendants of the most recent common ancestor (MRCA) are reprisented by all members of the group.



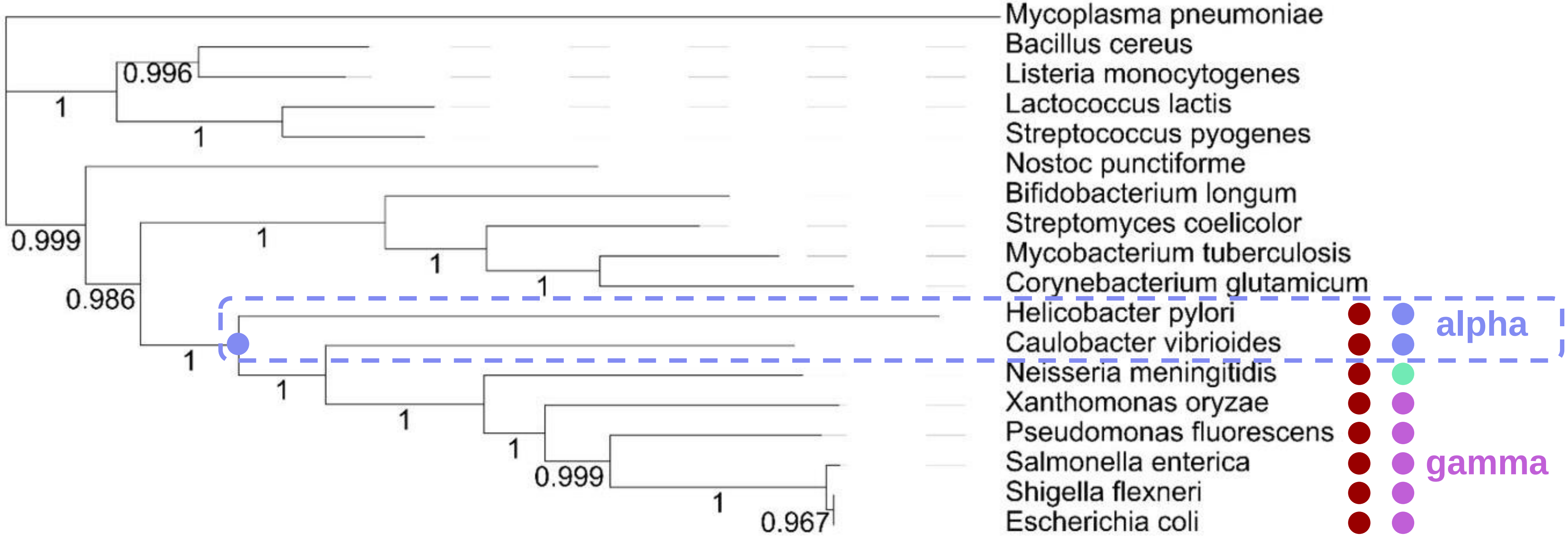
Monophyly

- A group is considered monophyletic if the descendants of the most recent common ancestor (MRCA) are reprisented by all members of the group.



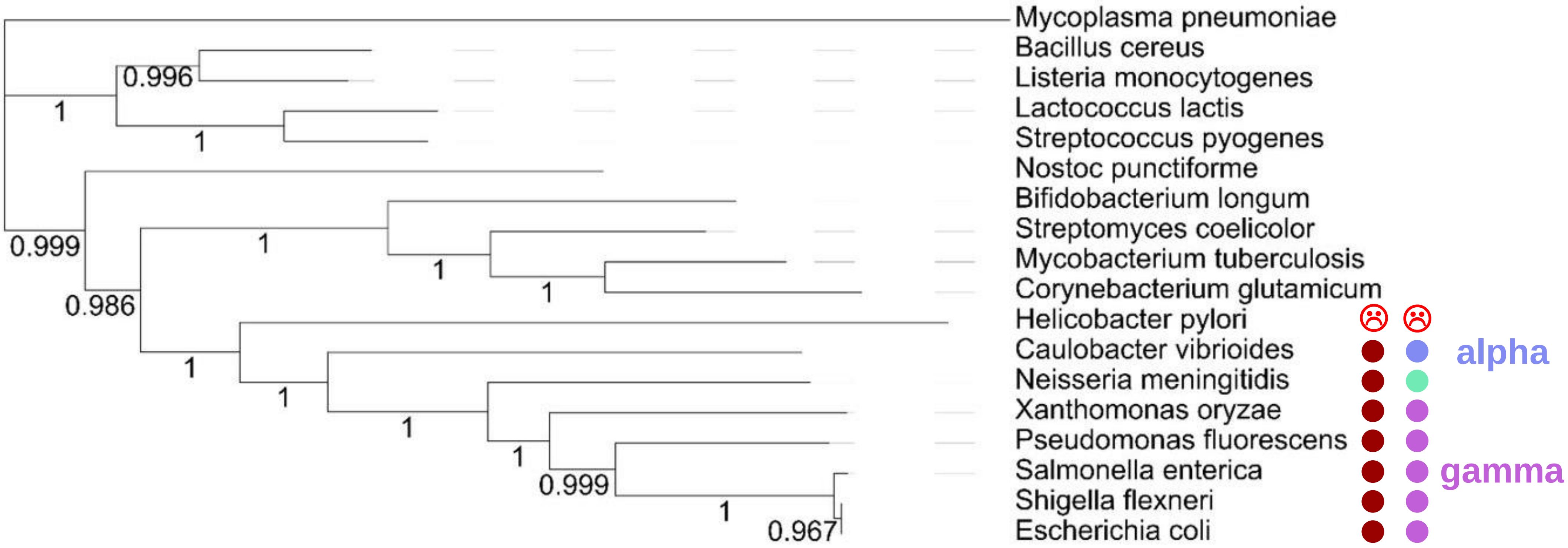
Monophyly

- A group is considered monophyletic if the descendants of the most recent common ancestor (MRCA) are reprisented by all members of the group.

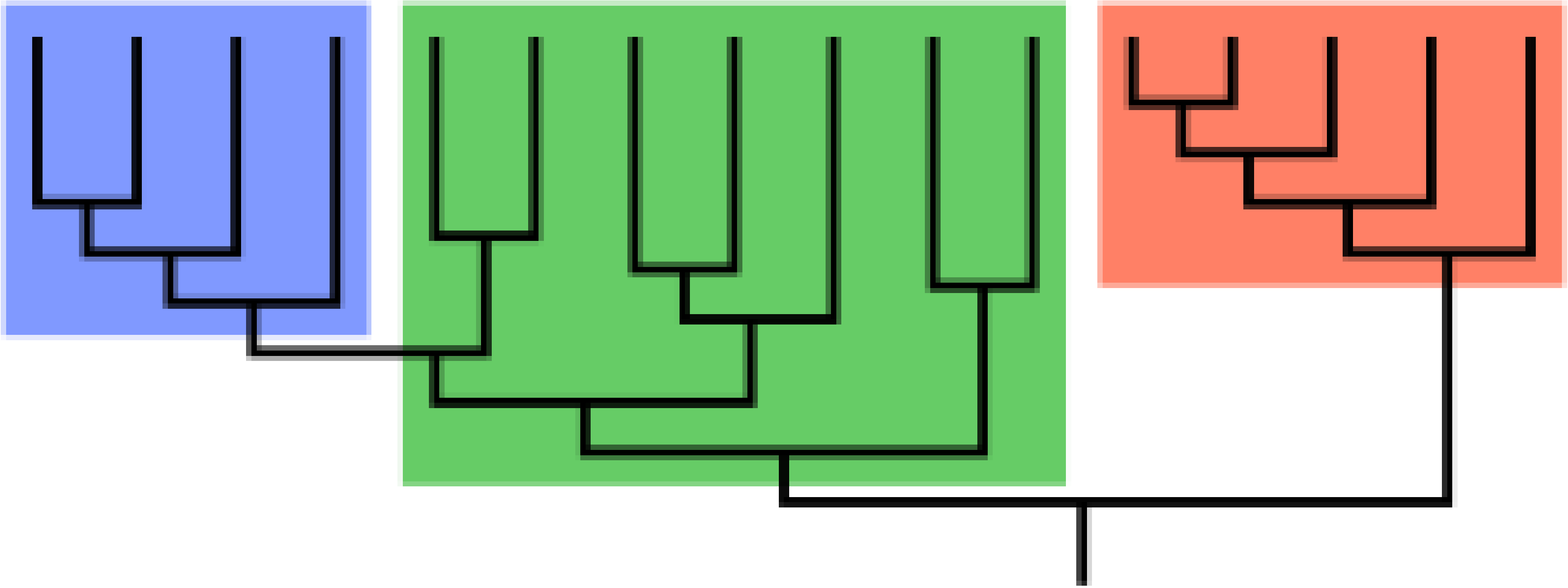


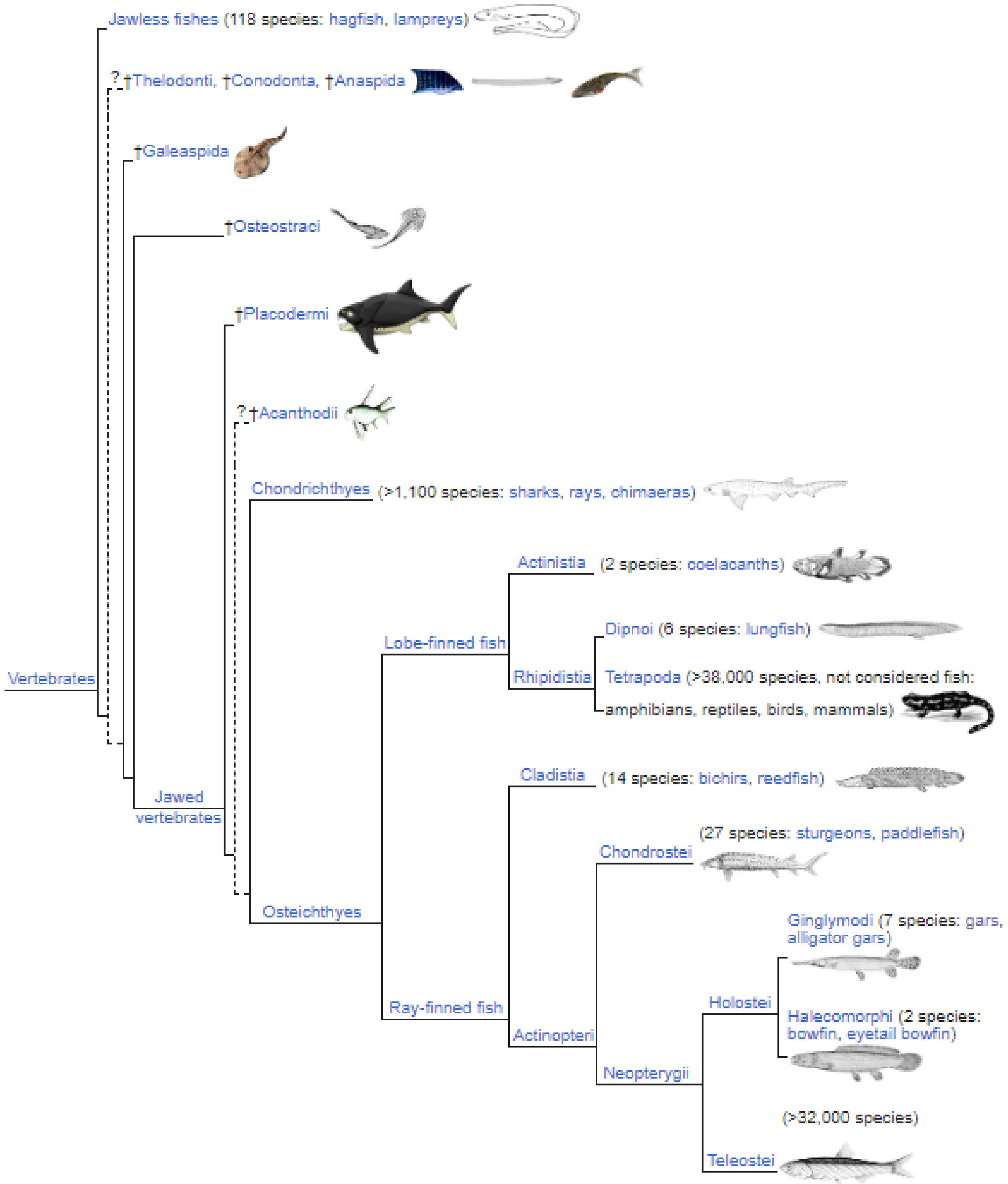
Paraphyly

- A group is considered paraphyletic if the descendants of the most recent common ancestor (MRCA) are not reprisented by all members of the group.



There is no such thing as a fish…



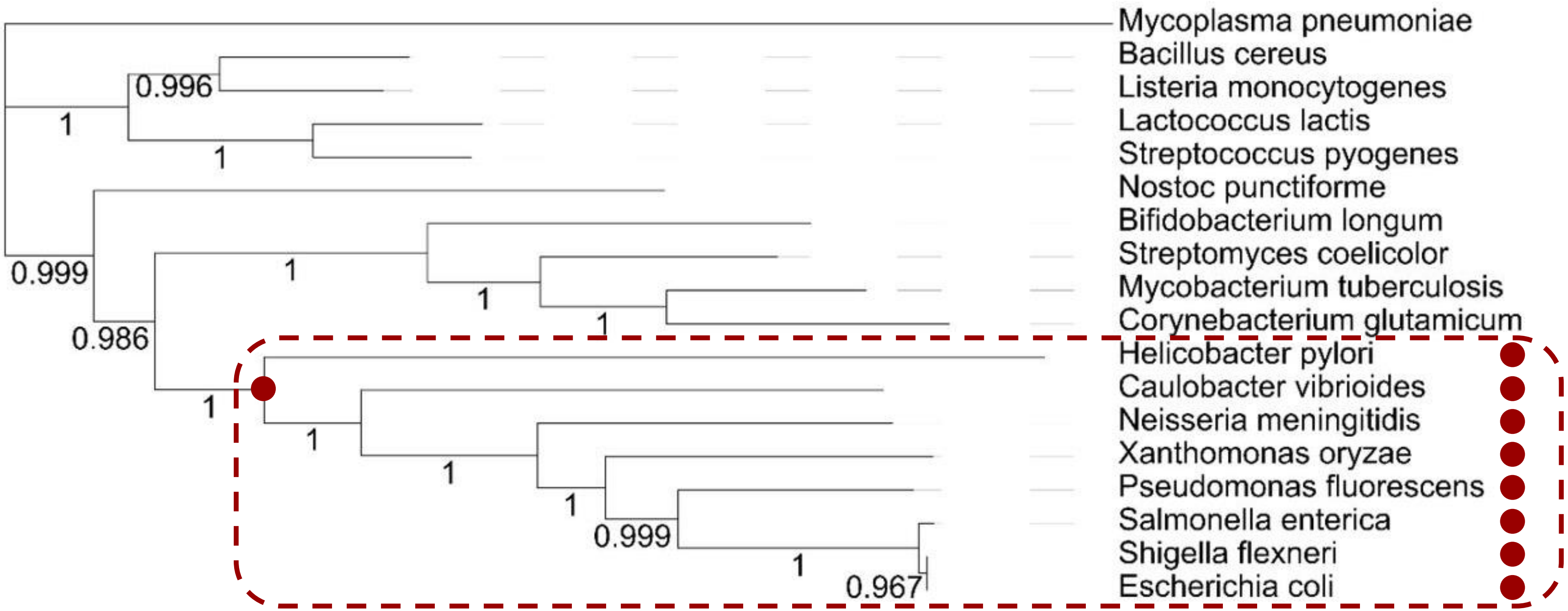

Jawless fishes (118 species: hagfish, lampreys)
?†Thelodonti, †Conodonta, †Anaspida
†Galeaspida
†Osteostraci
†Placodermi
?†Acanthodii
Chondrichthyes (>1,100 species: sharks, rays, chimaeras)
Actinistia (2 species: coelacanths)
Dipnoi (6 species: lungfish)
Lobe-finned fish
Rhipidistia
Tetrapoda (>38,000 species, not considered fish: amphibians, reptiles, birds, mammals)
Vertebrates
Cladistia (14 species: bichirs, reedfish)
Jawed vertebrates
(27 species: sturgeons, paddlefish)
Chondrostei
Osteichthyes
Ginglymodi (7 species: gars, alligator gars)
Holostei
Ray-finned fish
Halecomorphi (2 species: bowfin, eyetail bowfin)
Actinopteri
Neopterygii
(>32,000 species)
Teleostei

sh…

- Why do people say there is no such thing as a fish?
- Sharks belong to Chondrichthyes. Salmon belong to Actinopteri. Are you more related to a shark or a salmon?
- **Are humans apes? Are humans archea!?**

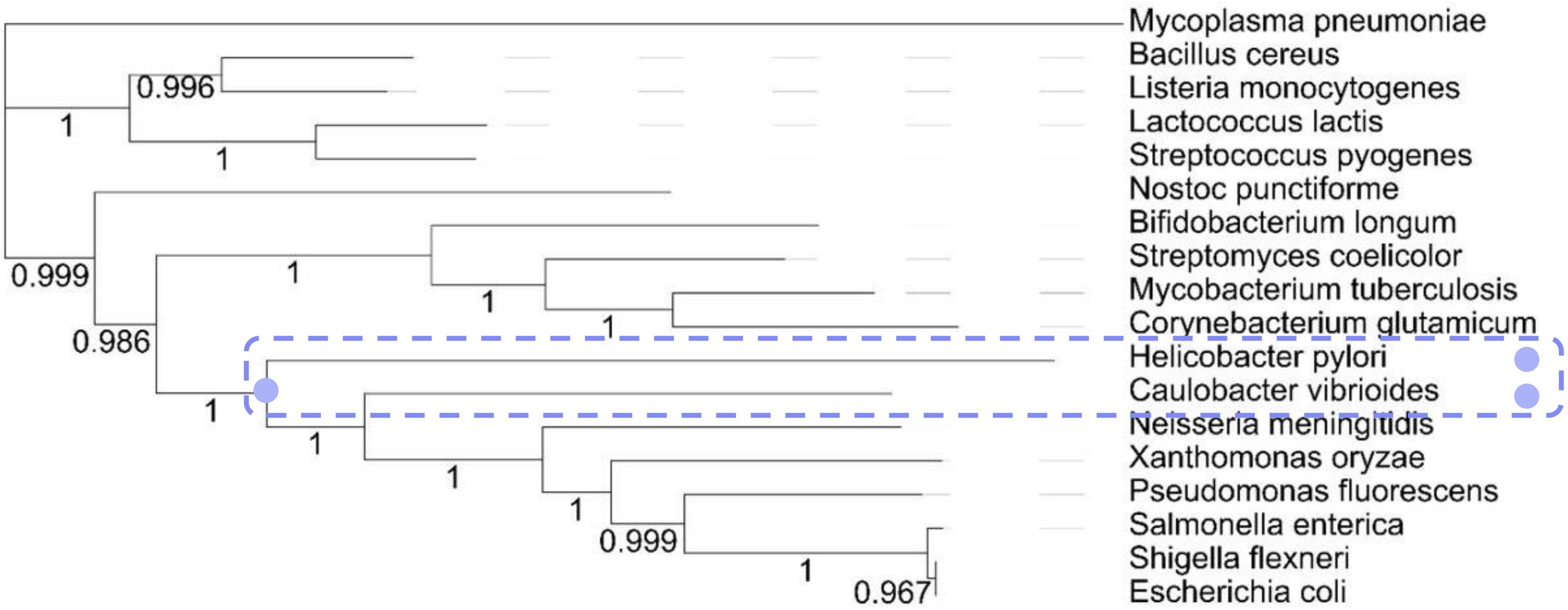
Monophyly

- A group is considered monophyletic if the descendants of the most recent common ancestor (MRCA) are reprisented by all members of the group. **And it contains the MRCA!**



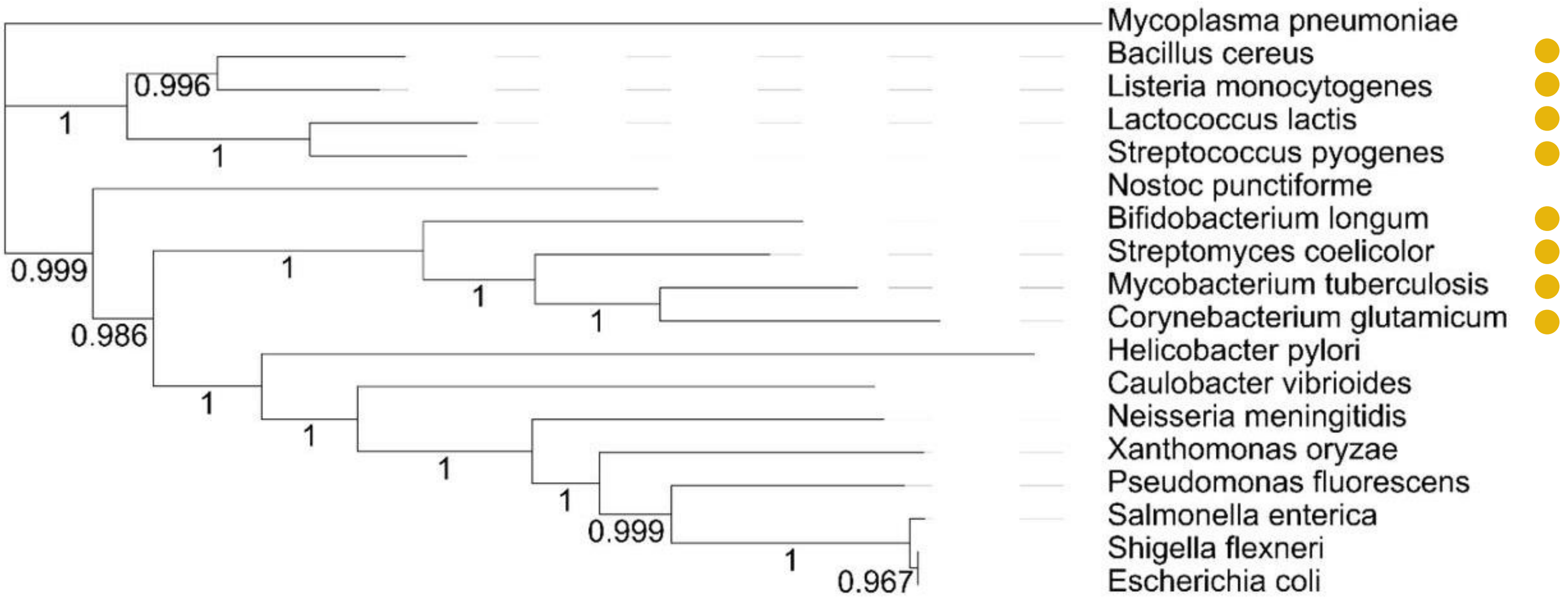
Paraphyly

- A group is considered paraphyletic if the descendants of the most recent common ancestor (MRCA) are not reprisented by all members of the group. **And it contains the MRCA!**



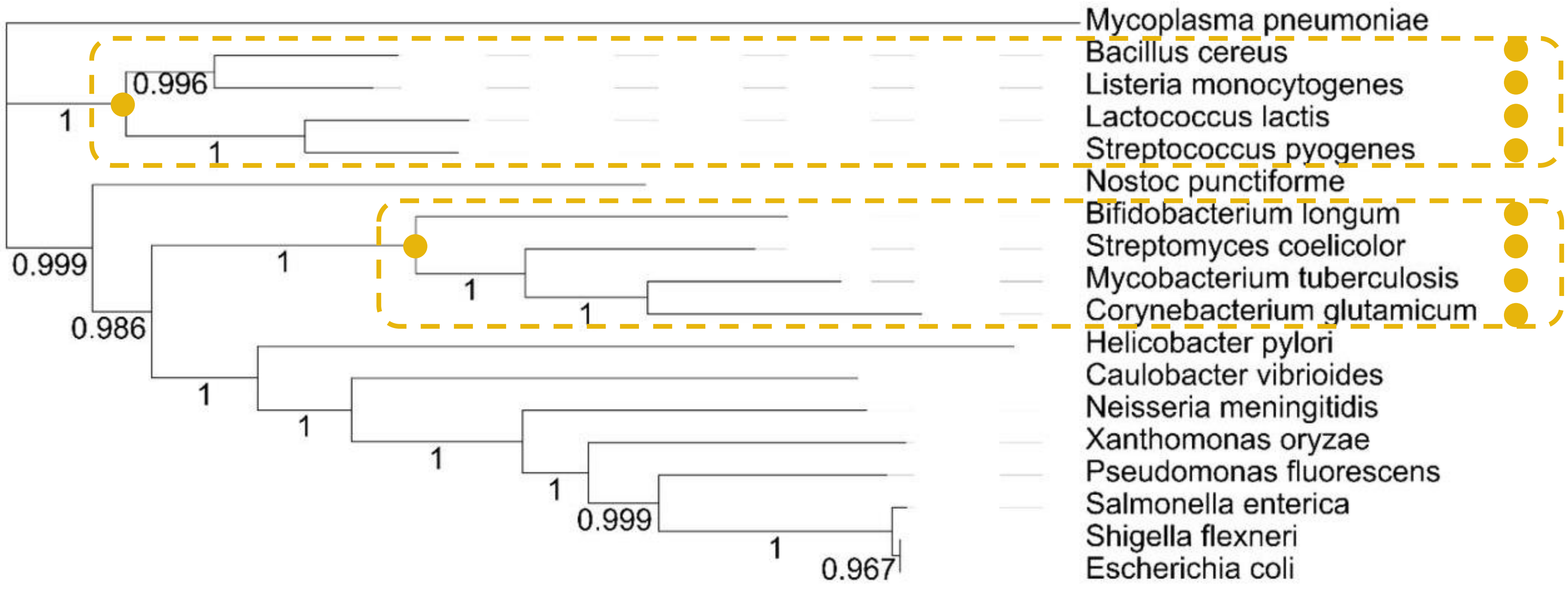
Polyphyly

- A group is considered polyphyletic it does not contain the most recent common ancestor (MRCA) of all members.



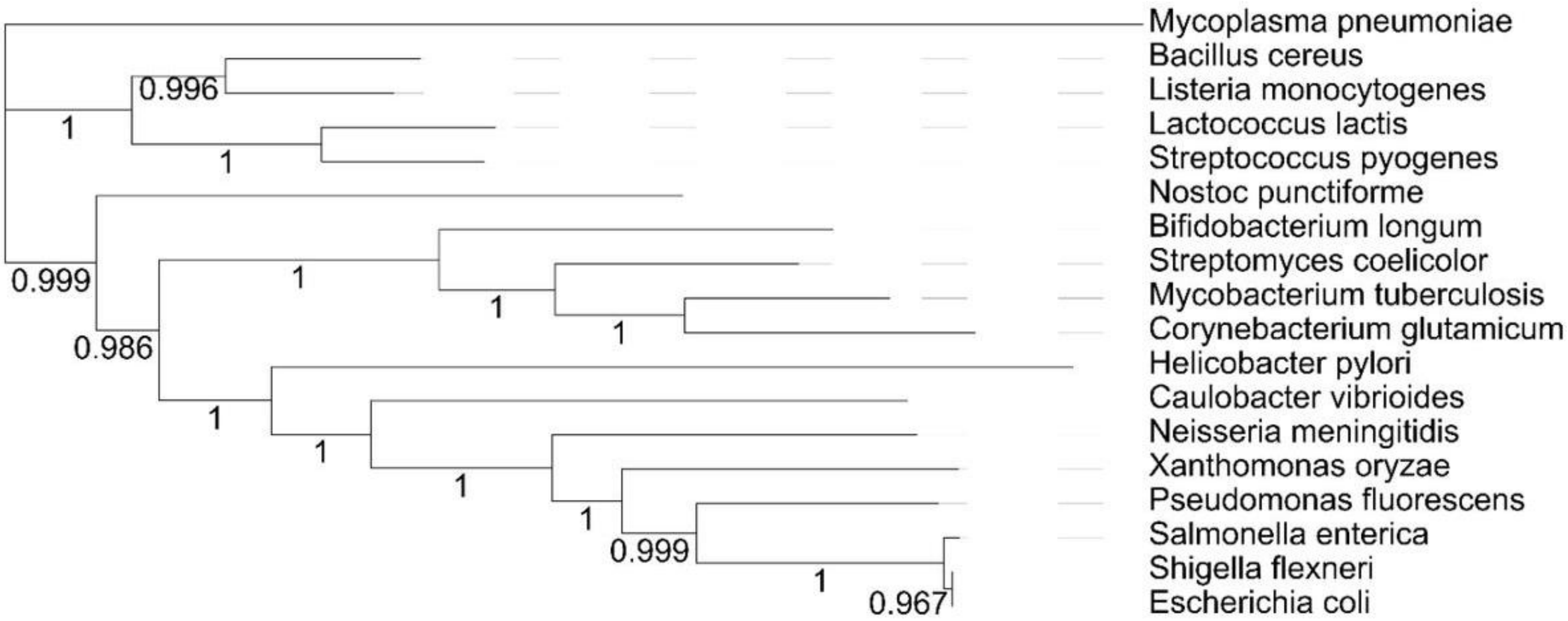
Polyphyly

- A group is considered polyphyletic it does not contain the most recent common ancestor (MRCA) of all members.



Interpreting Tree Quality

- **Is this a well supported tree? Is this a good tree?**



Scores

- Branch lengths
- Likelihoods
- Bootstraps
- Consensus

Scores

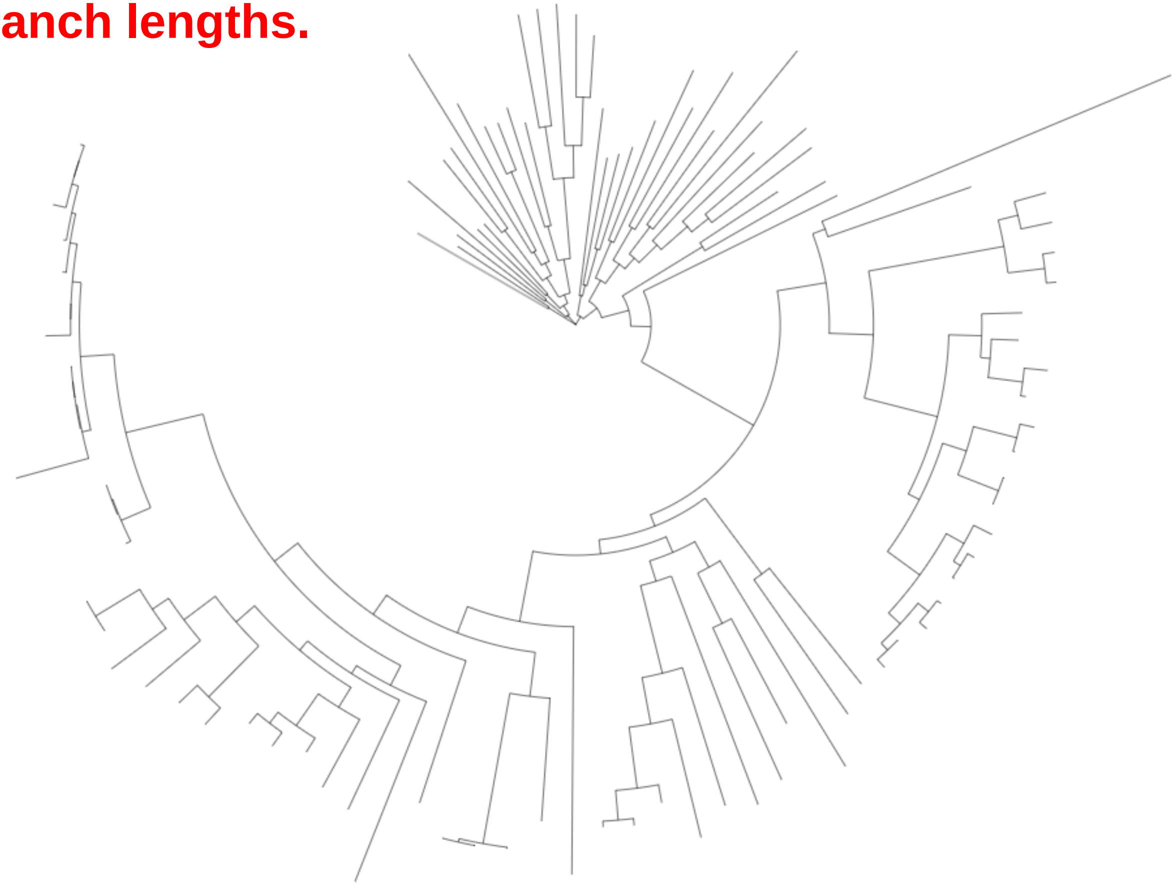
- Branch lengths
- Likelihoods
- Bootstraps
- Consensus

Branch lengths explain the relationships between your taxa. Typically they are measured in substituions per site. Although they are not a direct indication of quality you need to beware of long branch lengths (long branch attraction) or unusually distributed branch lengths.

Scores

- **Branch lengths**
- **Likelihoods**
- **Bootstraps**
- **Consensus**

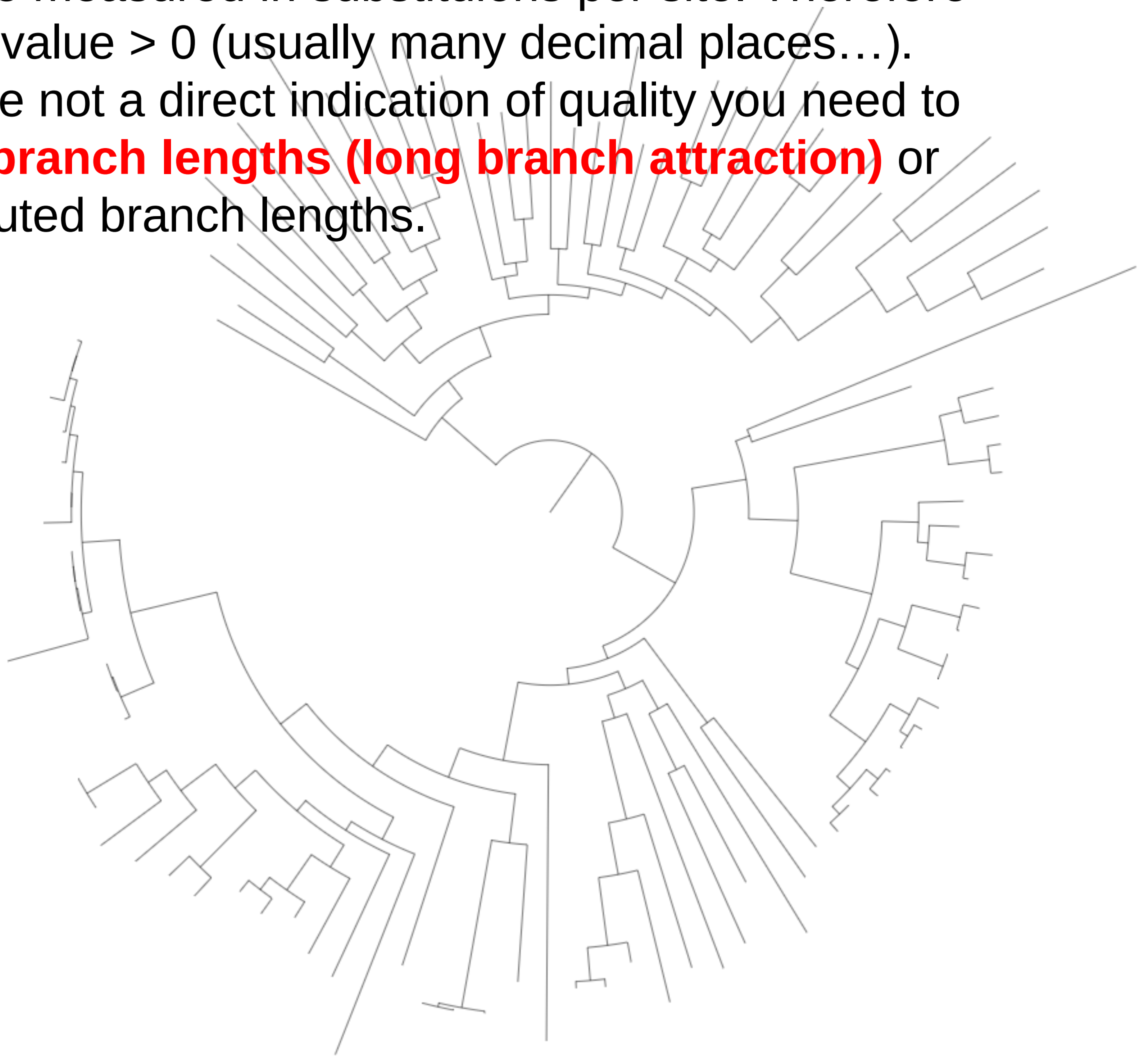
Branch lengths explain the relationships between your taxa. Typically they are measured in substituions per site. Although they are not a direct indication of quality you need to beware of long branch lengths (long branch attraction) or **unusually distributed branch lengths.**



Scores

- Branch lengths
- Likelihoods
- Bootstraps
- Consensus

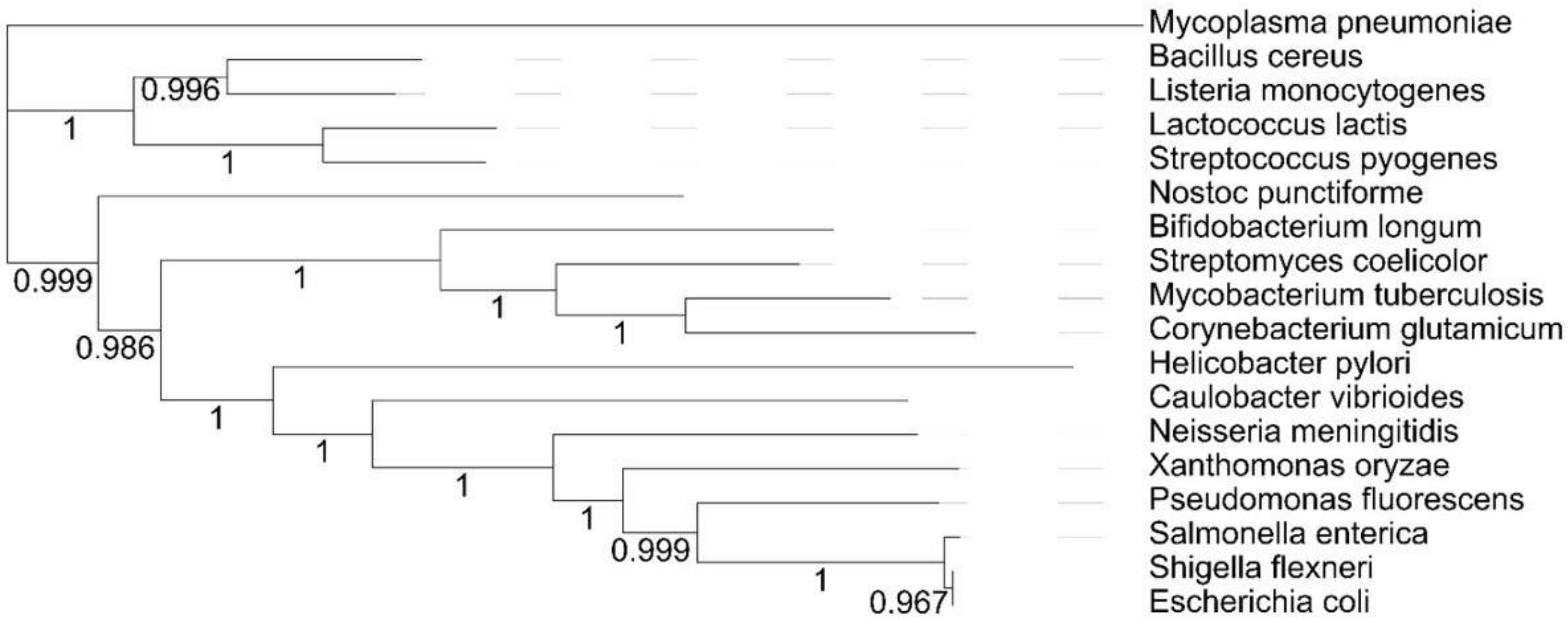
Branch lengths explain the relationships between your taxa. Typically they are measured in substituions per site. Therefore they can be any value > 0 (usually many decimal places…). Although they are not a direct indication of quality you need to beware of **long branch lengths (long branch attraction)** or unusually distributed branch lengths.



Scores

- **Branch lengths**
- **Likelihoods**
- **Bootstraps**
- **Consensus**

Likelihoods are a measure of support. They explain how well the tree fits the alignment. In other words the **likelihood** a given branch reprisents the input data. They are measured between 0 (unsopported) and 1 (maximum support). As with other statistical measures what is considered 'good' is arbitary but in general > 0.9 is considered very good support.



Scores

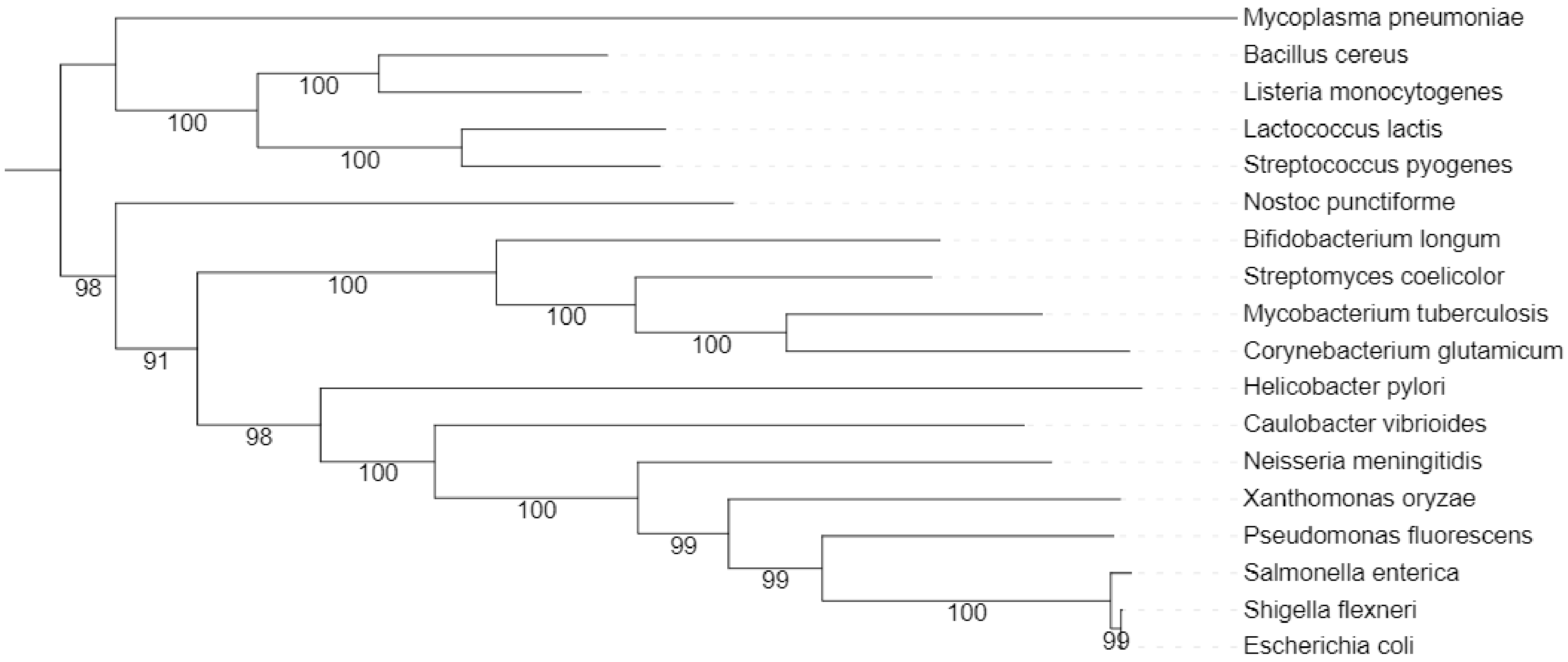
- **Branch lengths**
- **Likelihoods**
- **Bootstraps**
- **Consensus**

Bootstrapping is a measure of resampling. In a case with 1000 bootstrap replicates, the tree is built 1000 times from different subsets of the alignment. We then plot how many times that branch appeared across all 1000 trees. Sometimes plotted as a raw figure but often expressed as a percentage. Be careful using/interpreting bootstraps they can tell us more information about our data but can also obscure important information.

Scores

- Branch lengths
- Likelihoods
- Bootstraps
- Consensus

Bootstrapping is a measure of resampling. In a case with 1000 bootstrap replicates, the tree is built 1000 times from different subsets of the alignment. We then plot how many times that branch appeared across all 1000 trees. Sometimes plotted as a raw figure but often expressed as a percentage. Be careful using/interpreting bootstraps they can tell us more information about our data but can also obscure important information.



Thom’s Golden Rule #5

NEVER SHOW BOOTSTRAP VALUES ALONE!

Scores

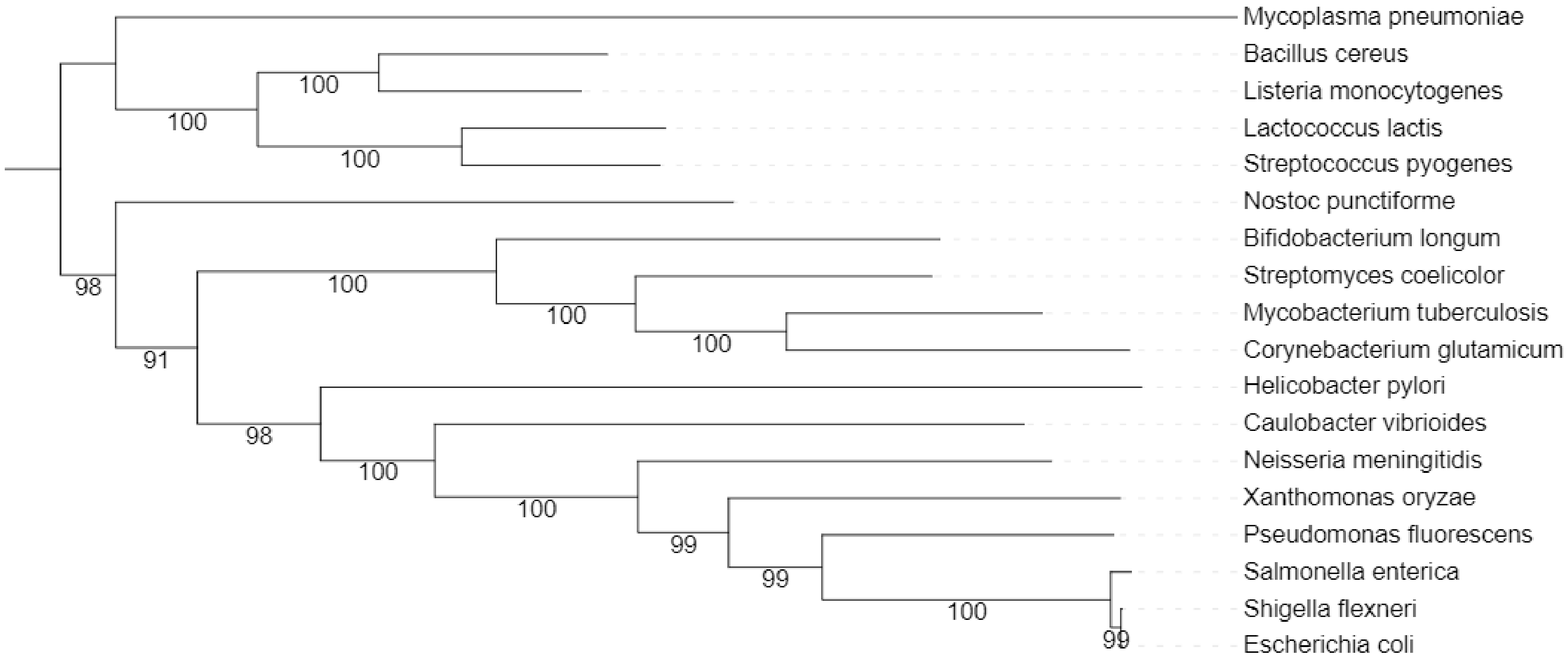
- **Branch lengths**
- **Likelihoods**
- **Bootstraps**
- **Consensus**

Similar to bootstrapping, **consensus** is a reprisentation of the agreement between multiple trees. It is typically reprisented as a percentage. There are many, many reasons to build a consensus tree and they can be a very powerful, if often undertilised, tool.

Scores

- **Branch lengths**
- **Likelihoods**
- **Bootstraps**
- **Consensus**

Similar to bootstrapping, **consensus** is a reprisentation of the agreement between multiple trees. It is typically reprisented as a percentage. There are many, many reasons to build a consensus tree and they can be a very powerful, if often undertilised, tool.



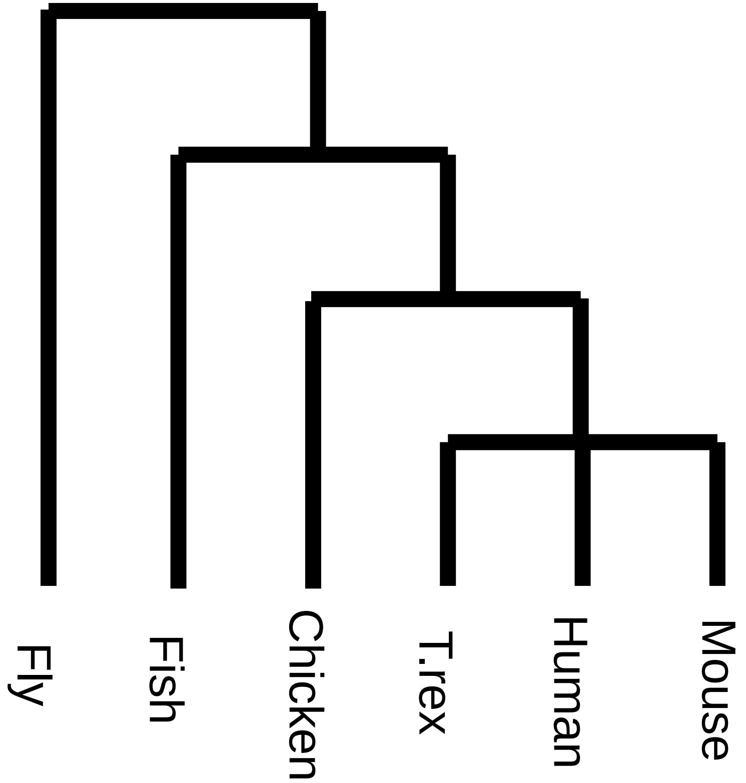
Bootstraps are expressed as a consensus!

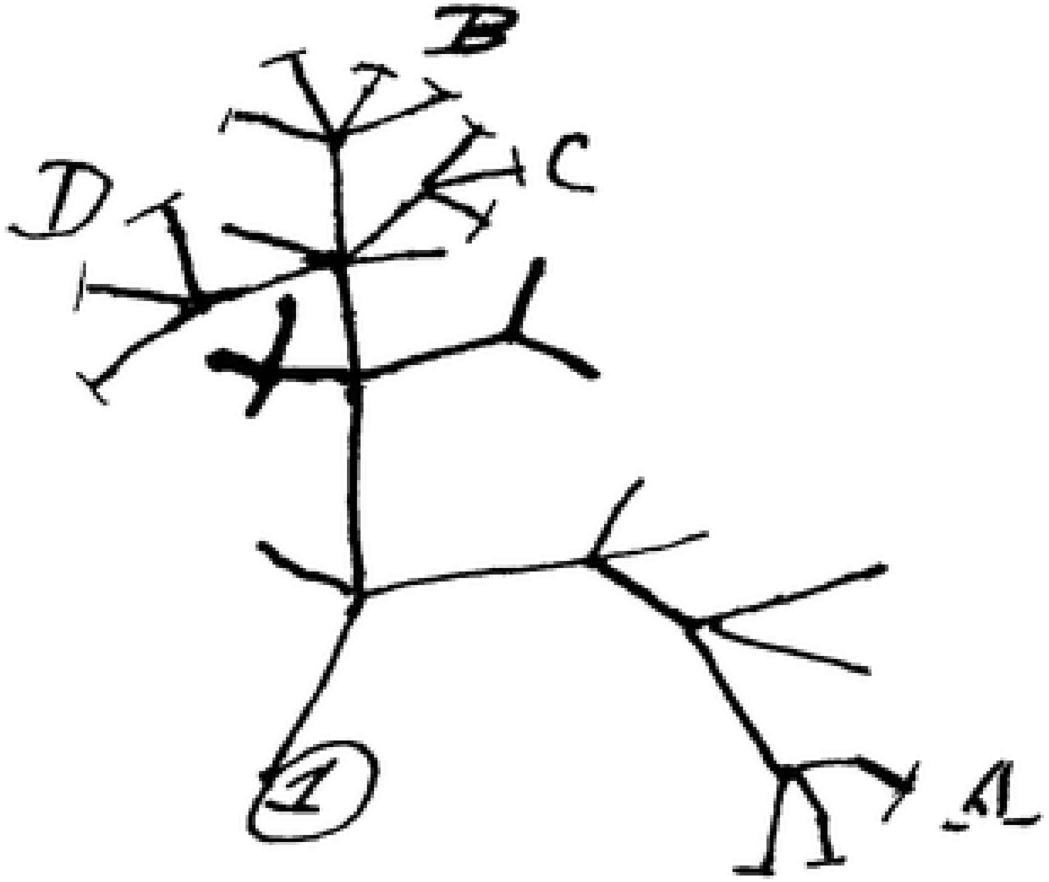
Thom’s Golden Rule #6

NEVER BE ASHAMED OF POOR BRANCH SUPPORT!

Polytomy

A polytomy is when more than two lineages descend from a single node i.e. multifurcation not bifurcation.





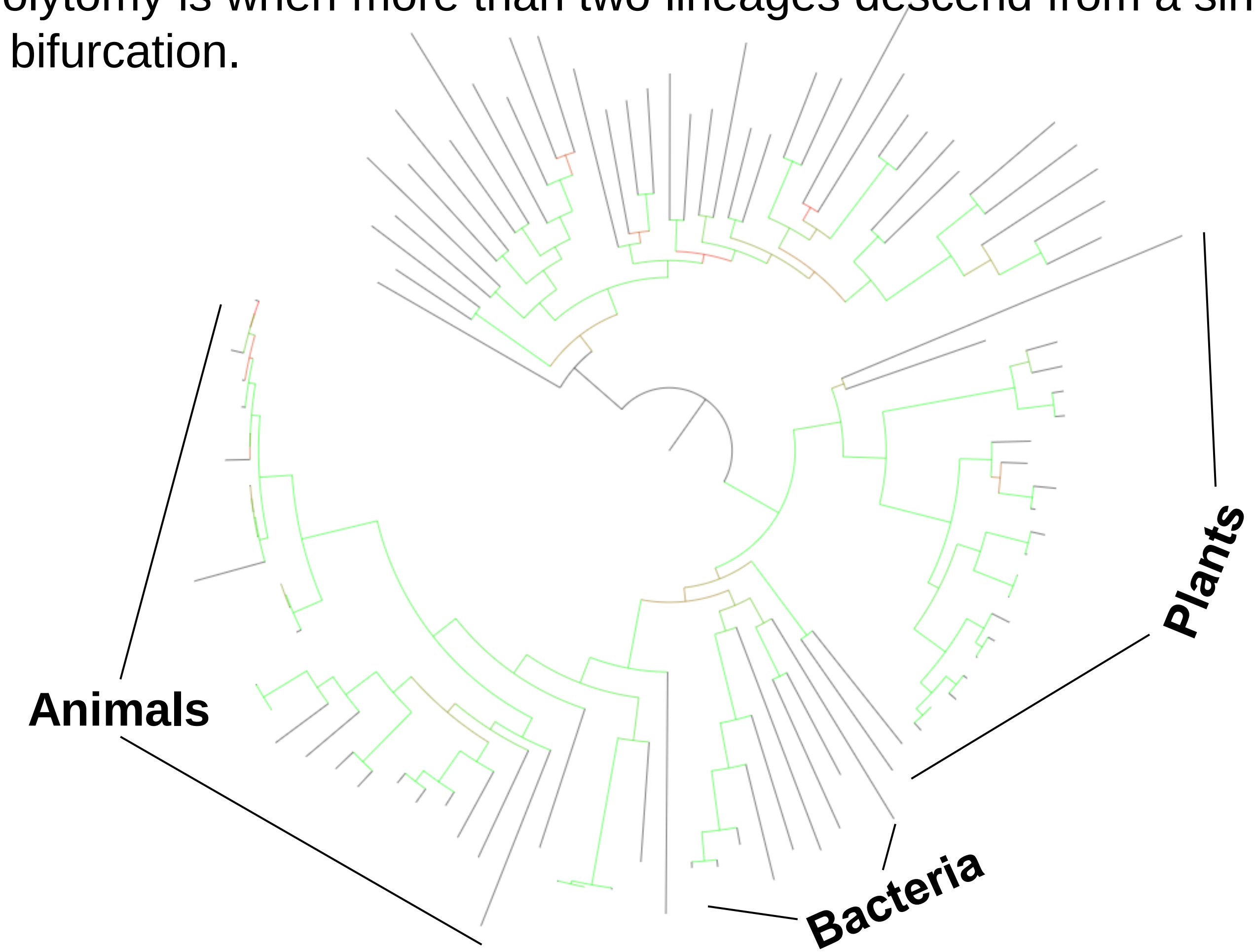
‘Hard polytomies’

A **hard polytomy** exists reprisents a ‘true’ multifurcation i.e. a lineage splits into three in a single generation. These are **very rare** in biological examples.

A **soft polytomy** results where data is inssufficent to resolve the relationship. These are **common** in biological examples.

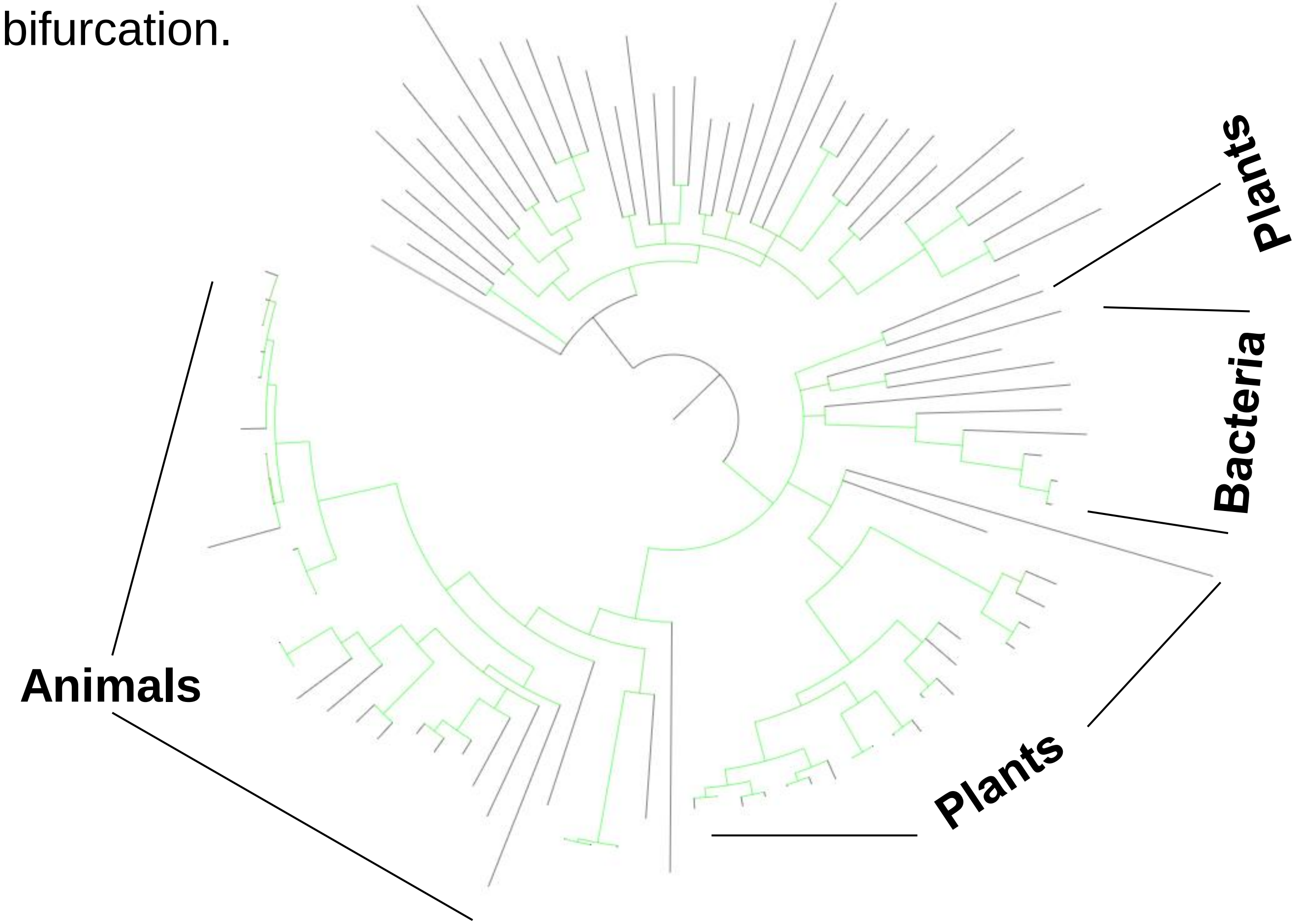
Polytomy

A polytomy is when more than two lineages descend from a single node i.e. multifurcation not bifurcation.



Polytomy

A polytomy is when more than two lineages descend from a single node i.e. multifurcation not bifurcation.



Step Three: Infer a tree

1. Go to the interactive tree of life (IToL) https://itol.embl.de/ and upload your tree. If the analysis hasn’t finsihed use mine (**3_trees/ p450s.tree**)

2. Right click and find ‘root tree by midpoint’. **Do your outgroups form a monophyletic clade? Are your outgroups the earliest diverging sequences?**

3. Do: ”Advanced> Branch metadata display > bootstraps/metadata > text”. **Is the tree well suported overall? Are your non-outgroup taxa monophyletic and strongly supported?**

4. Find sceE and sceD in the tree. Find the MRCA of the two proteins. **Are they closely related? Do you think these homologues are the result of a recent gene duplication?**

5. **How could we improve this tree? What other information can we use to interrogate this relationship? What further analysis could we do to elucidate this relationship further?**

COURSEWORK

To pass this course you must complete a 3 – 5 minute presentation on the phylogenetics of a gene/protein/organism of interest.

The presentation must cover:

1. Breif background of your sequence of interest
2. The evolutionary hypothesis you wanted to test
3. The methods you used to find the homologues, make the alignment and build the tree
4. A description of the quality of the resulting tree
5. Whether the tree supported or confirmed your hypothesis