

# Introduction to phylogenetic analysis

- how to prepare, build, interpret...
- and love a phylogenetic tree



CCB skills seminar, 11/18/2025

Ryuichi Ono

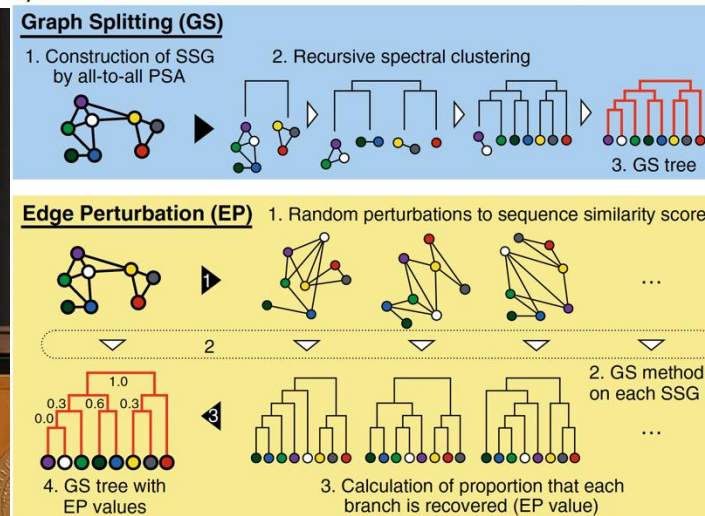
Cress Lab



# Acknowledgements



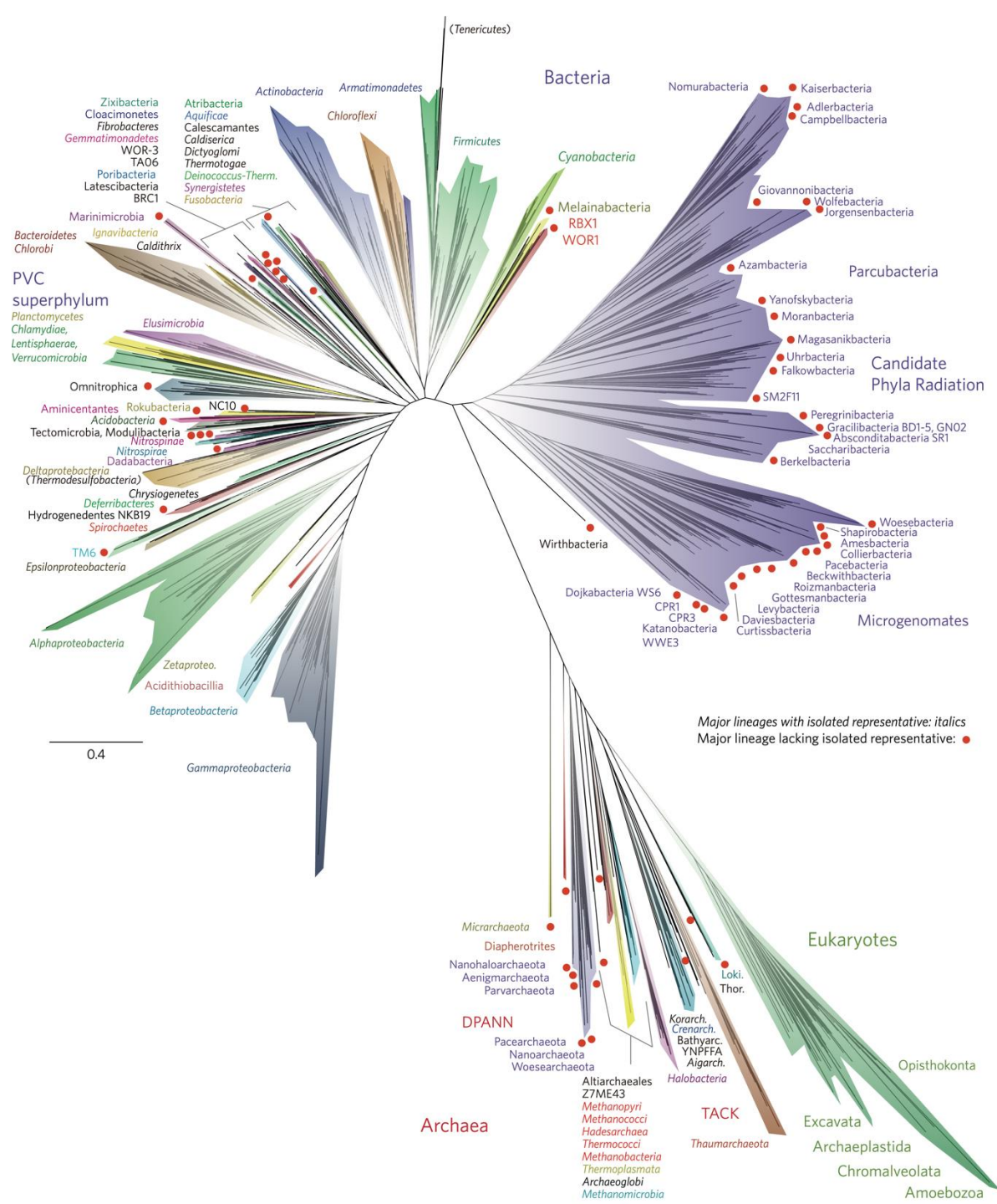
**Shout out to Dr. Brady Cress, Jack Demaray, CCB team, and friends in Cress/Rubin lab**



**Huge thanks to Dr. Motomu Matsui for the feedbacks on this presentation**

Matsui and Iwasaki. Systematic Biology (2020)

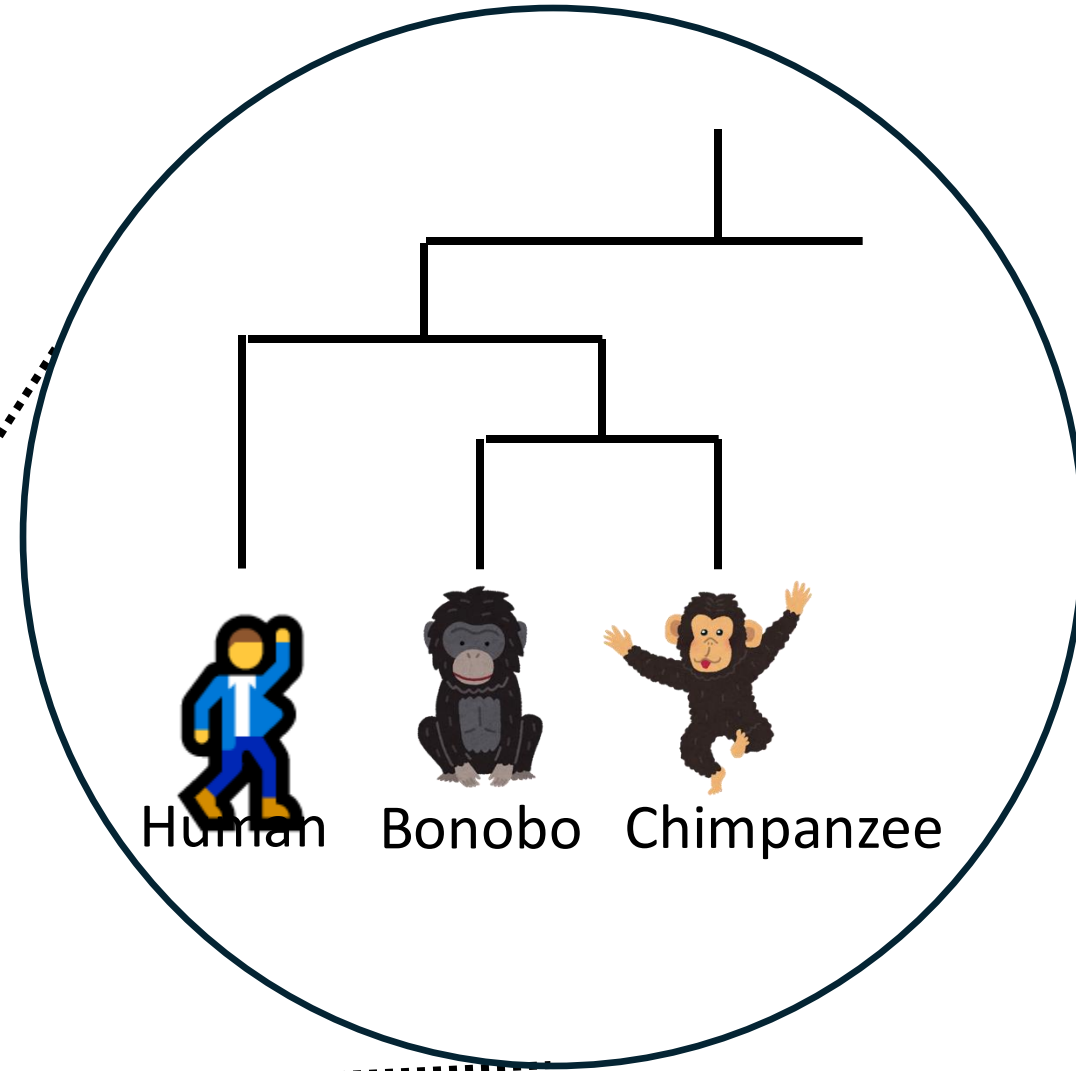
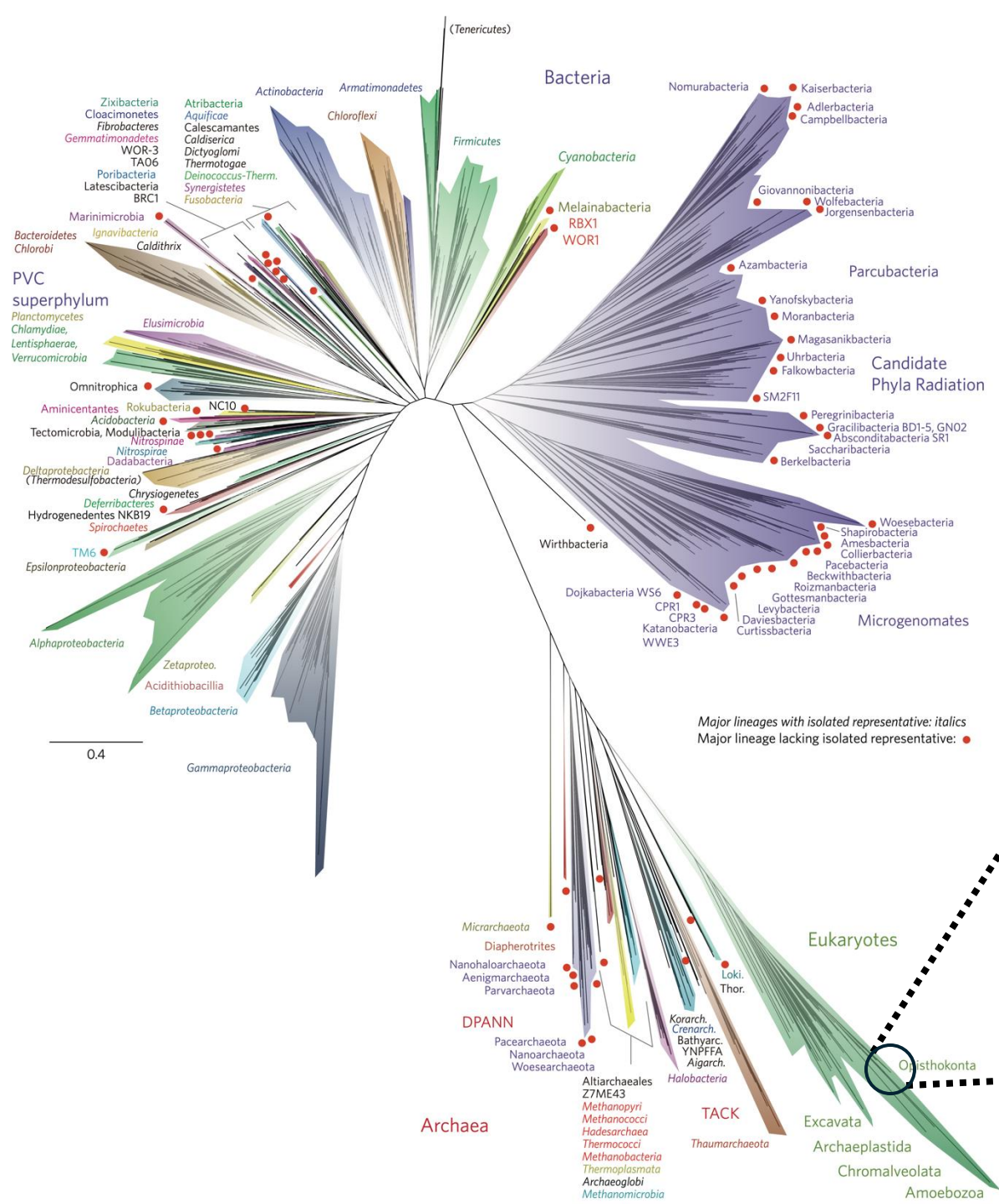
# What can we learn from tree?





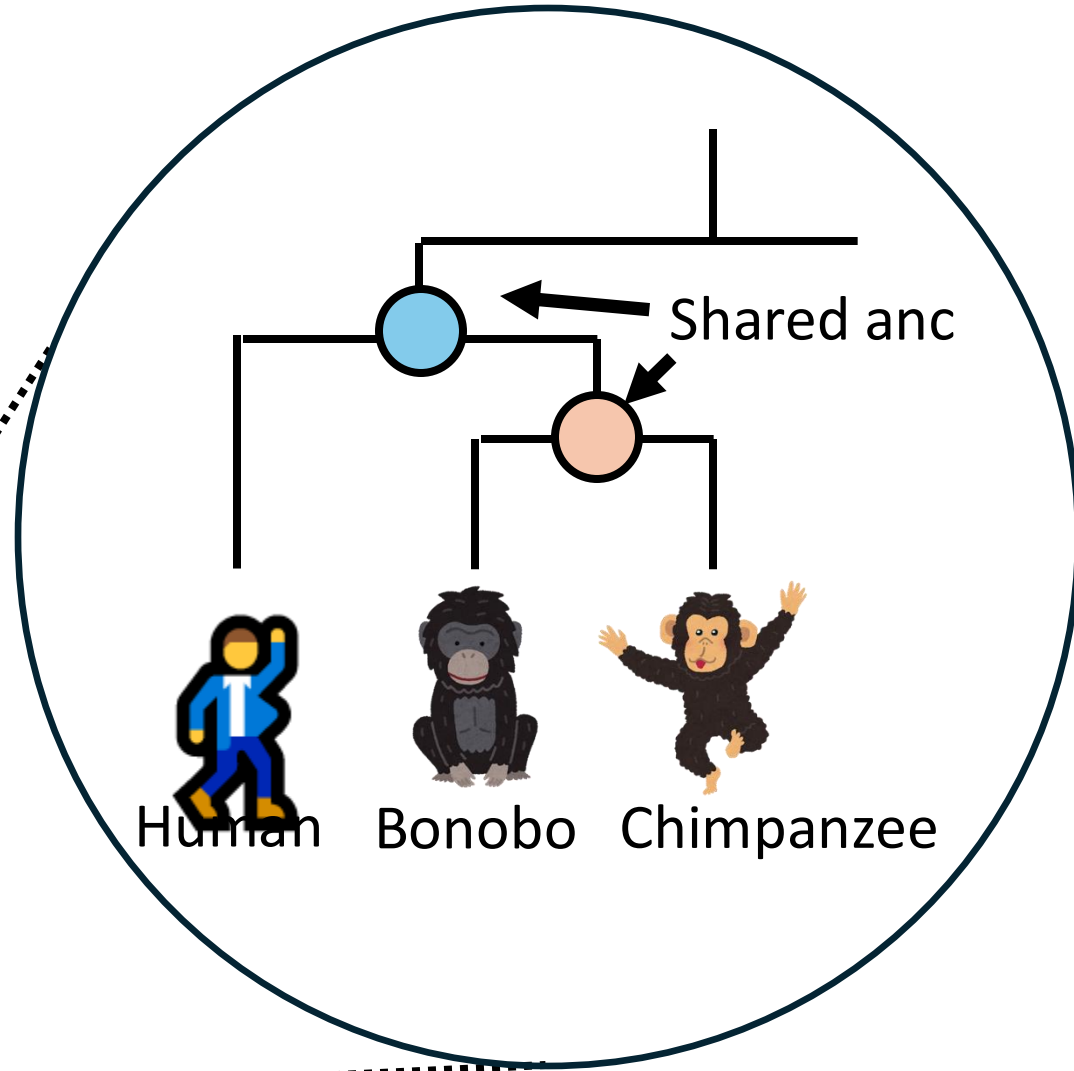
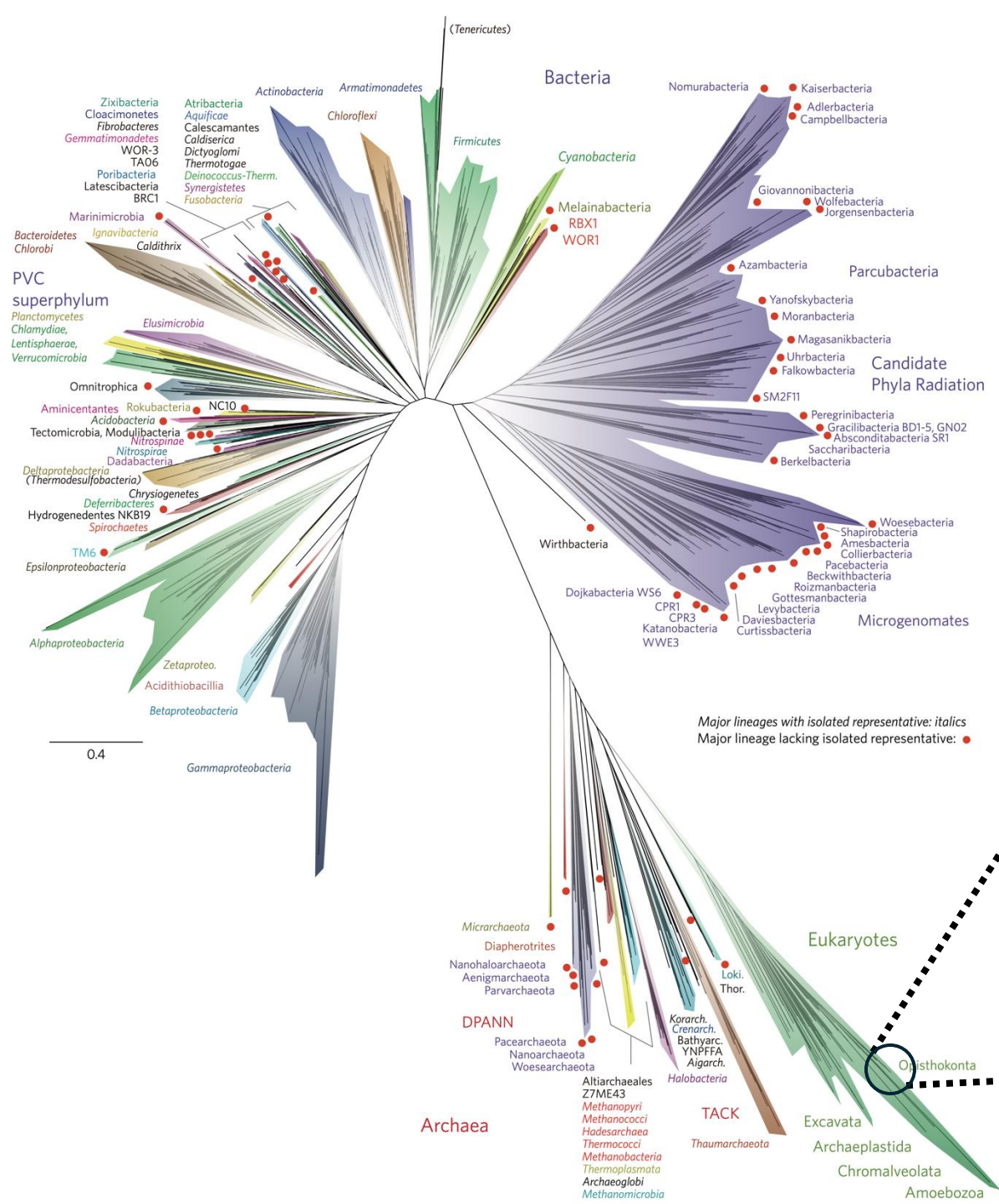


# What can we learn from tree?

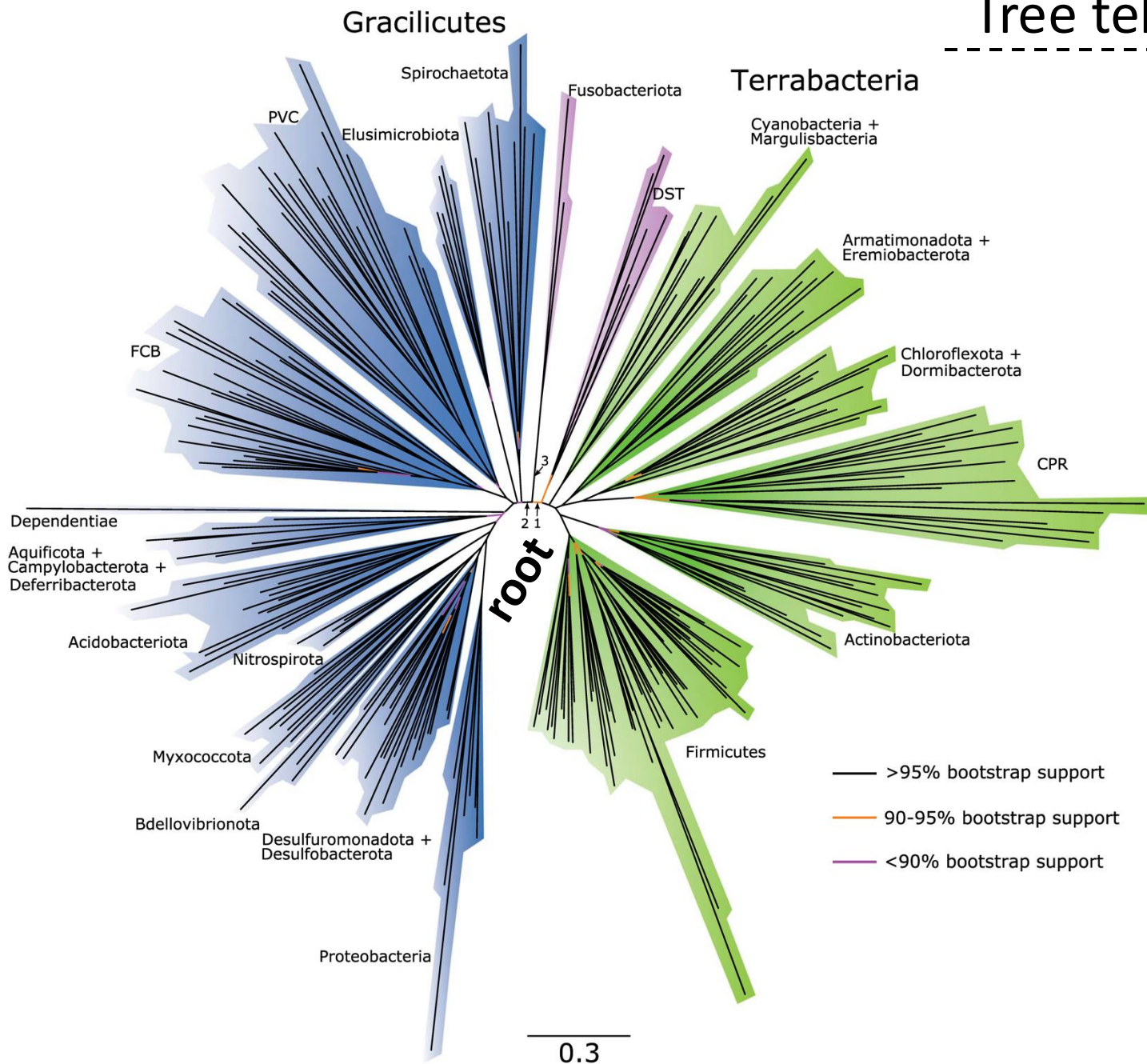




# What can we learn from tree?

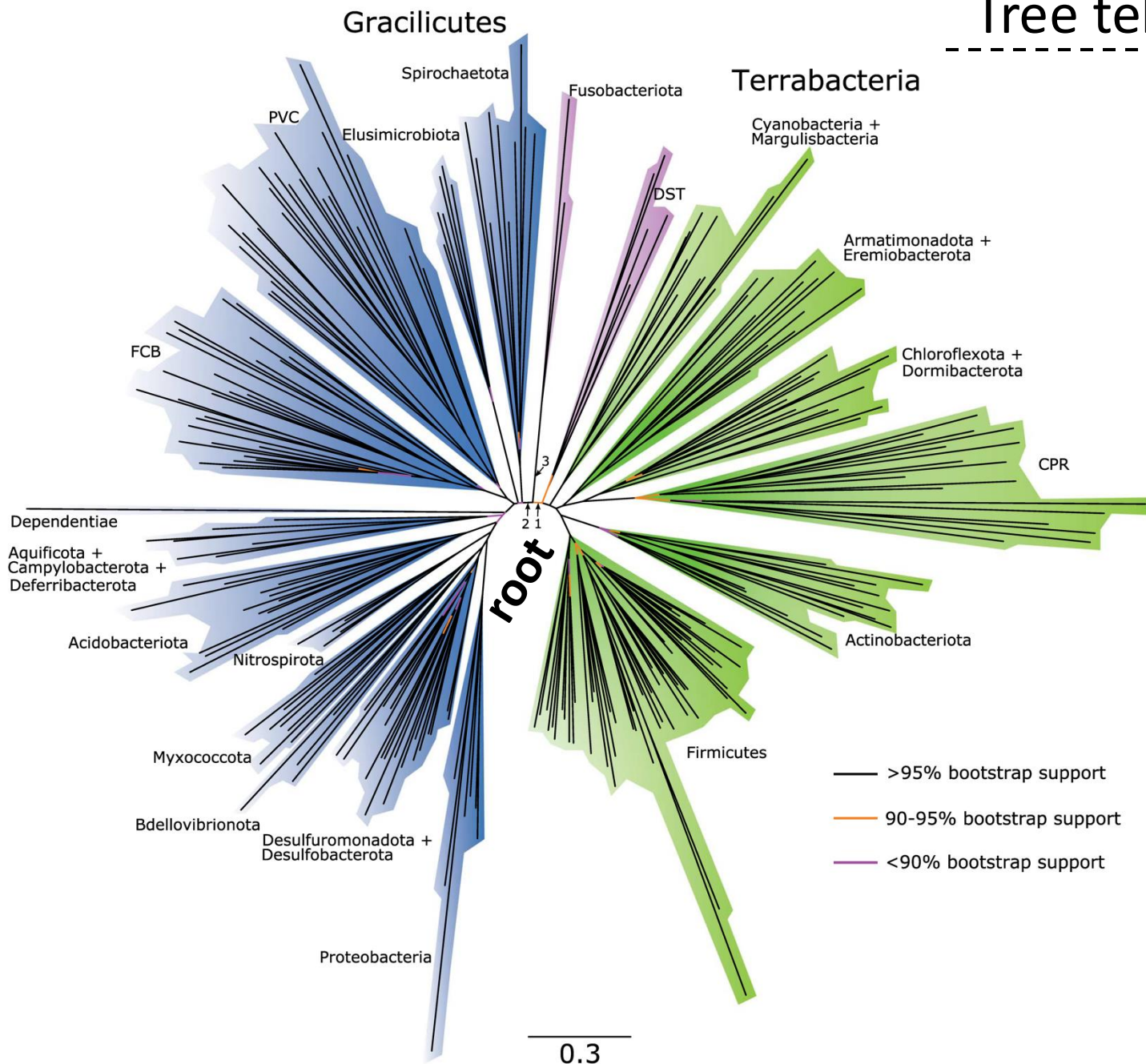


# Tree tells us the ultimate ancestor of life

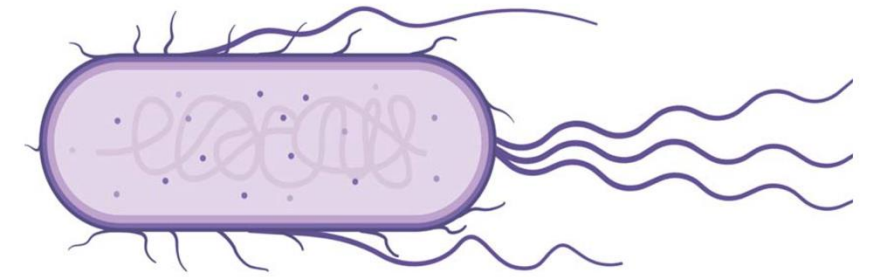




# Tree tells us the ultimate ancestor of life



## Last Bacterial Common Ancestor

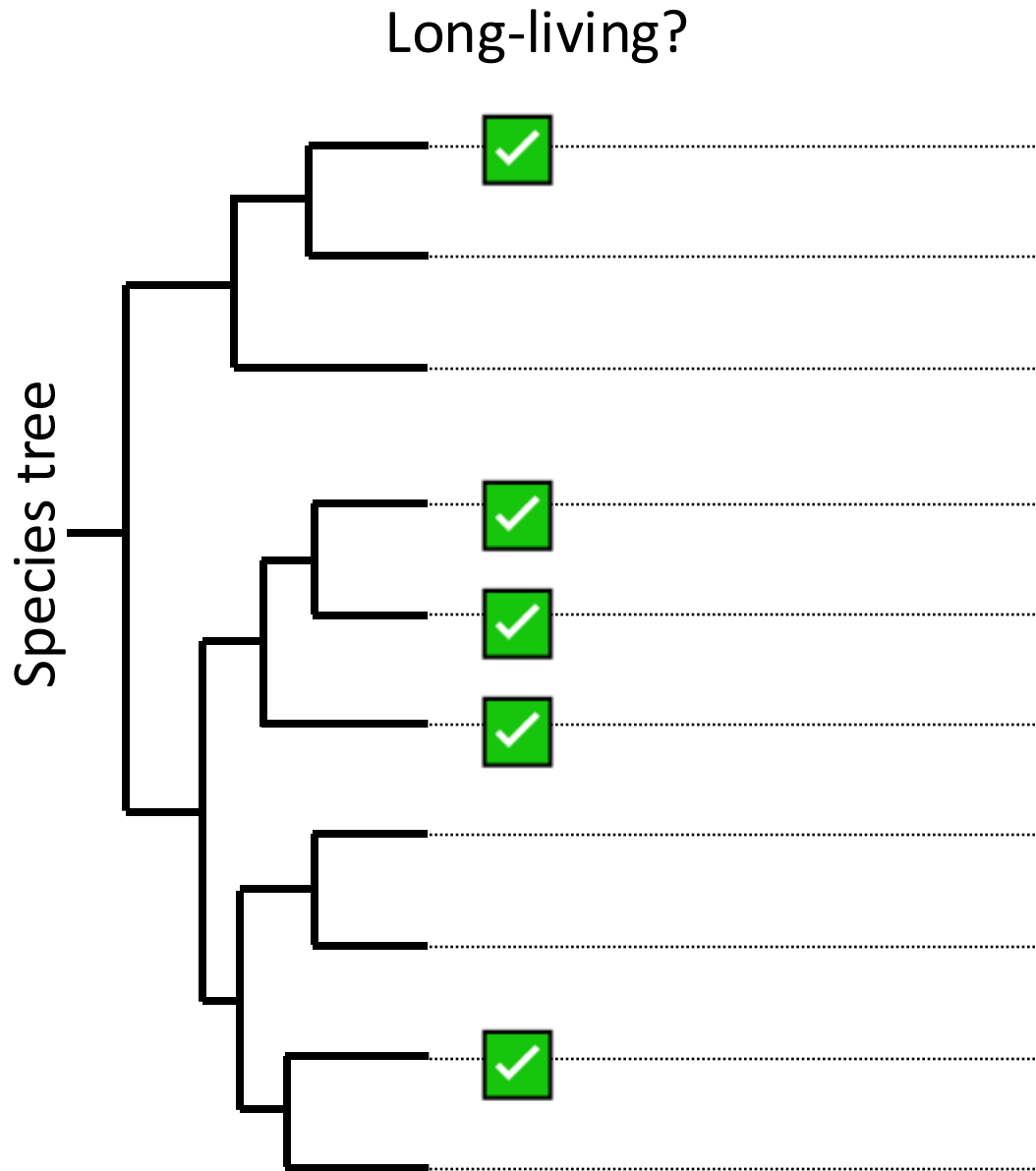


● Ribosomes ● DNA/nucleoid

- Rod-shaped cell
- Double membrane
- Can move
- Can sense chemicals
- CRISPR-Cas



## Tree can also be a starting point of analysis

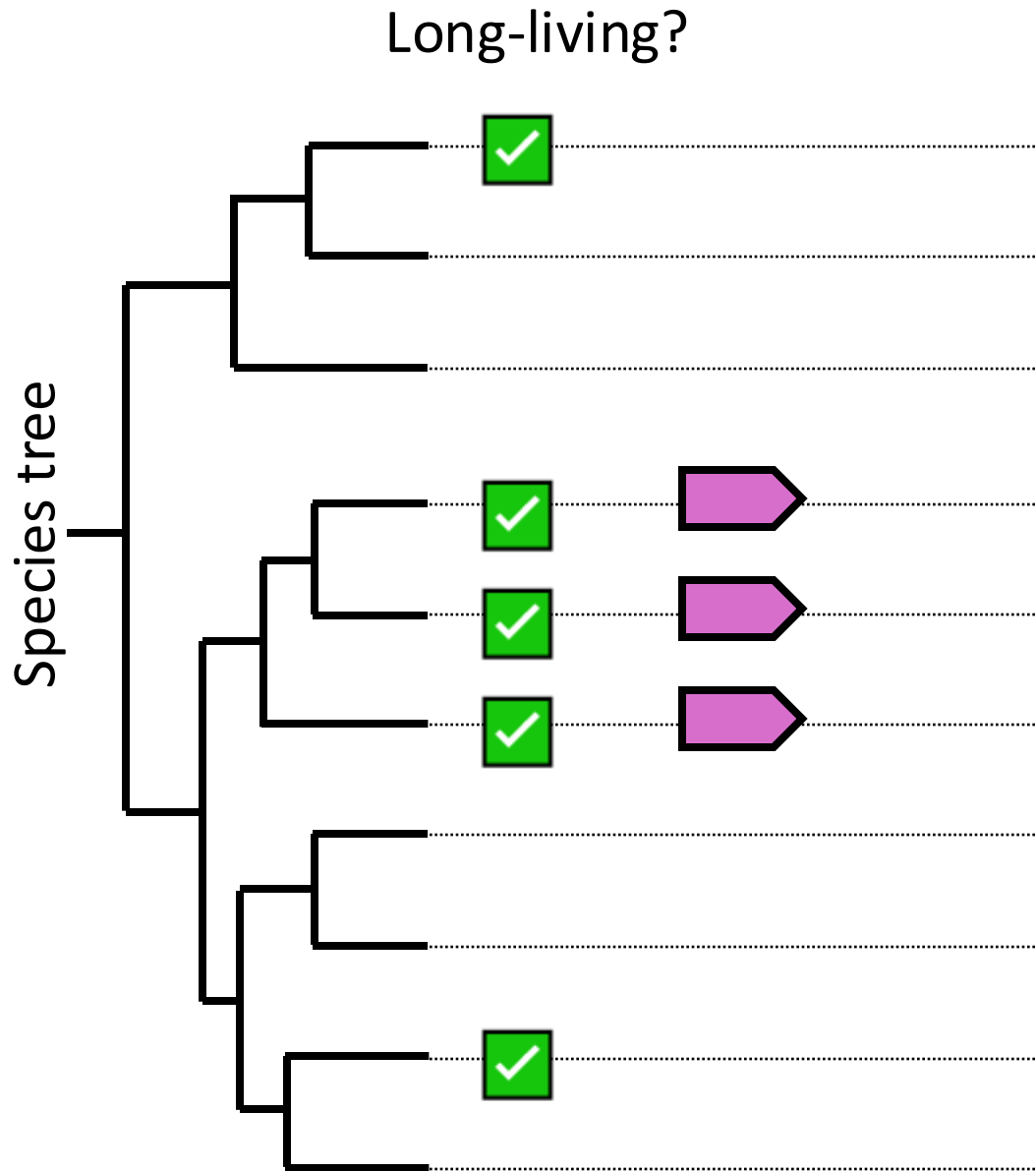


Let's say you found that species which possess **gene A** tend to be long-living...



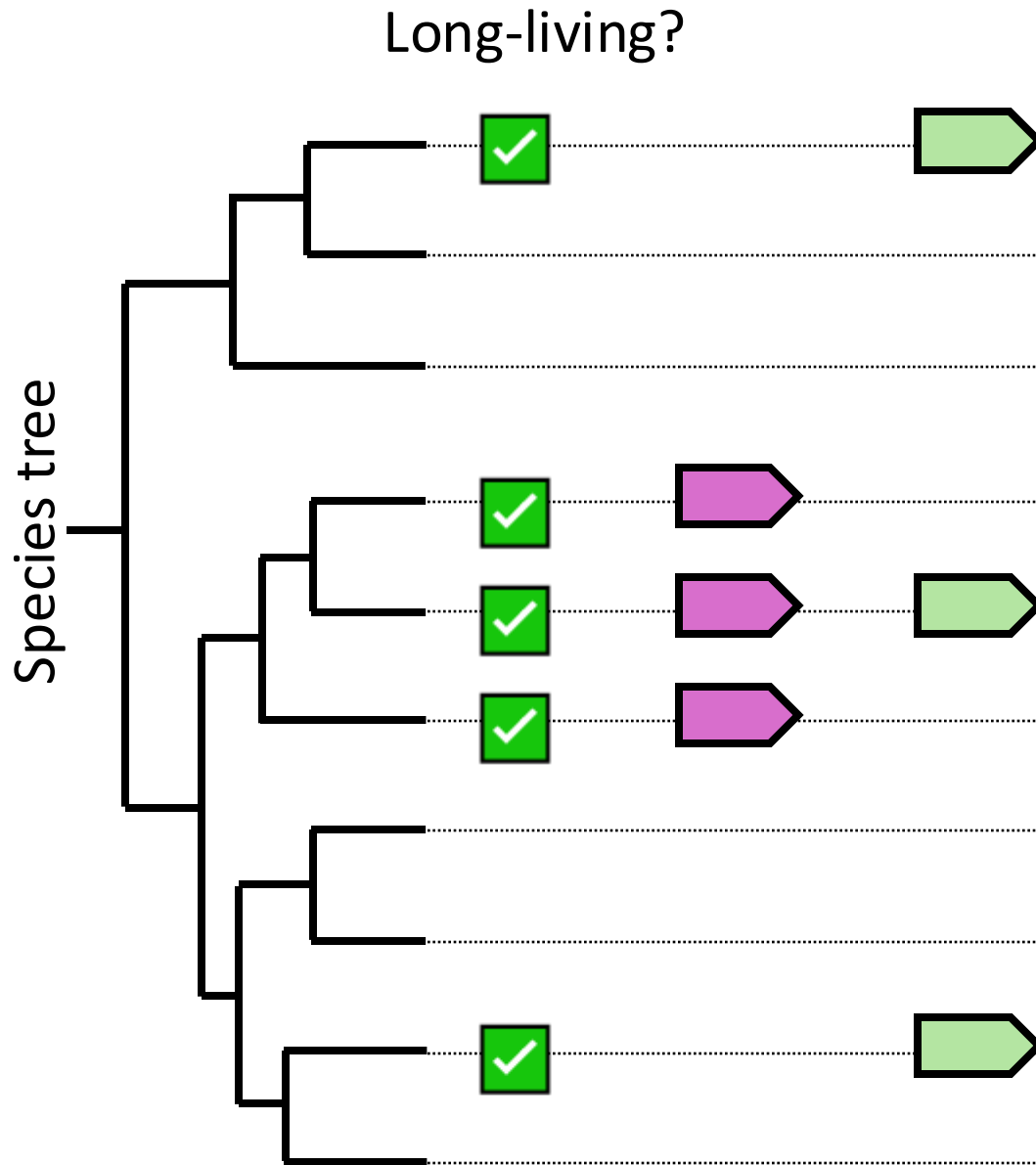


## Tree can also be a starting point of analysis





## Tree can also be a starting point of analysis



Let's say you found that species which possess **gene A** tend to be long-living...



Now, will you believe that gene A is involved in the extended lifespan?

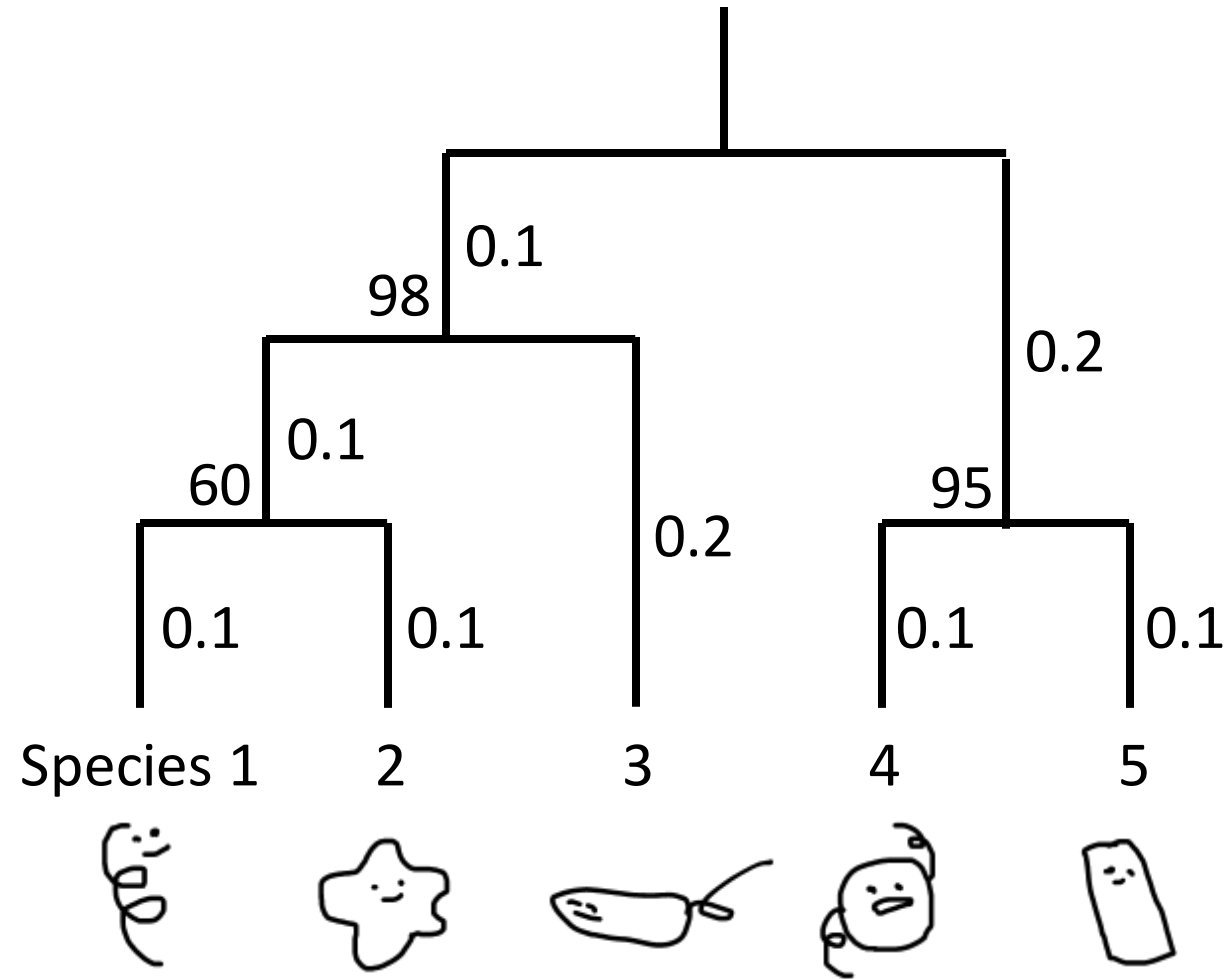
Considering the **phylogenetic signal**, **gene B** is more likely to be involved





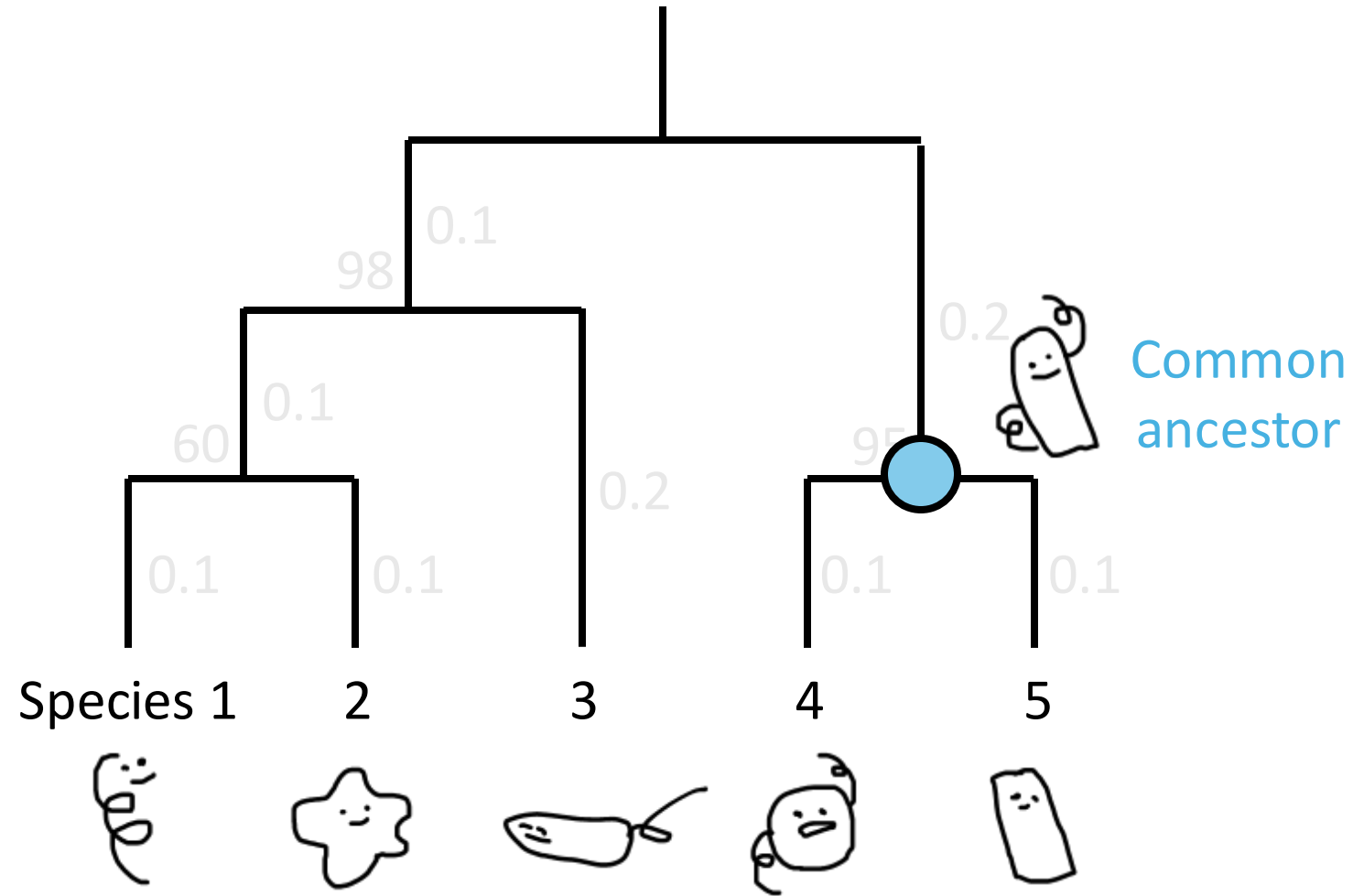
# Tree 101 – how to interpret labels on tree

---



# Tree 101 – how to interpret labels on tree

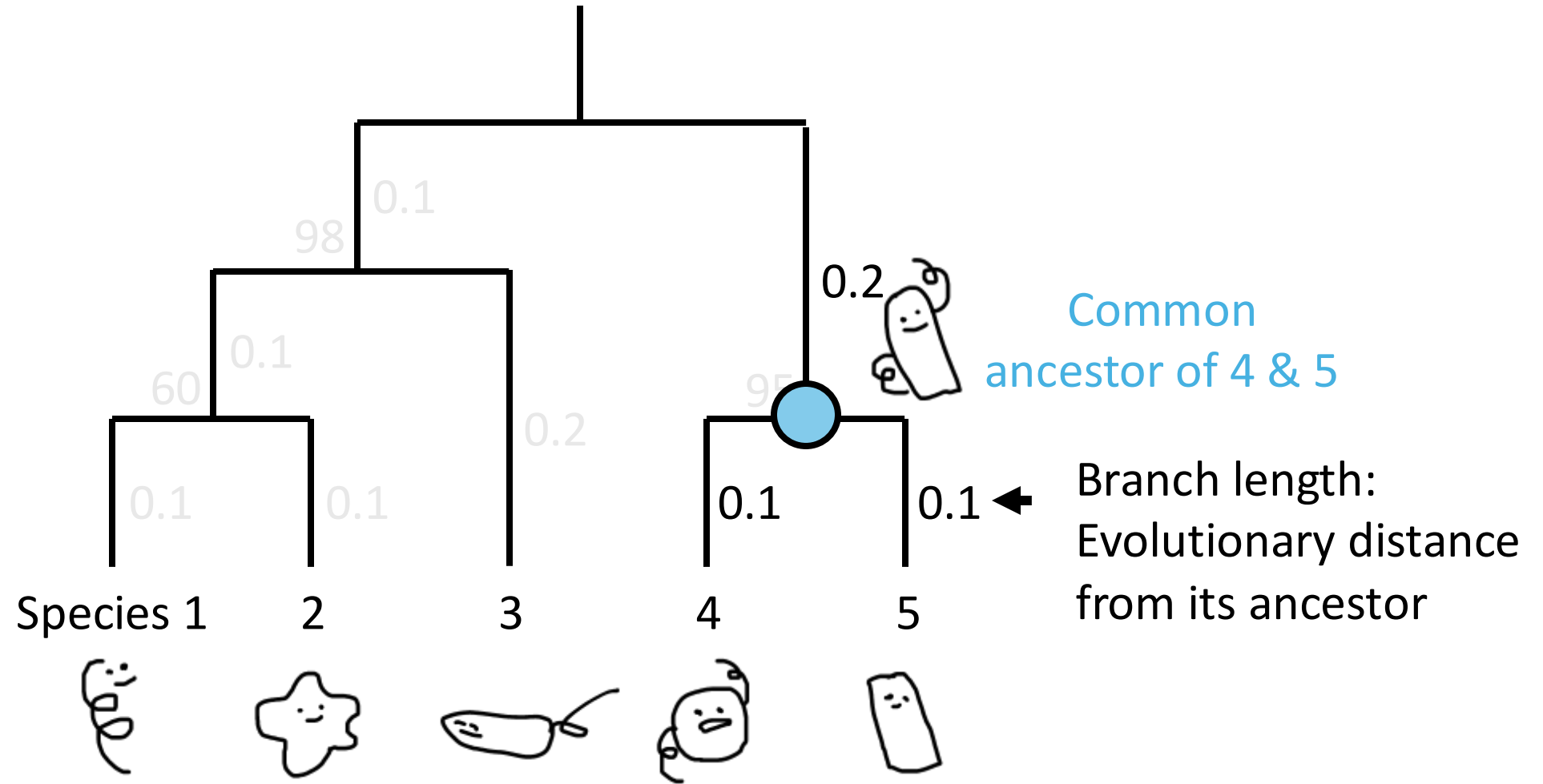
---





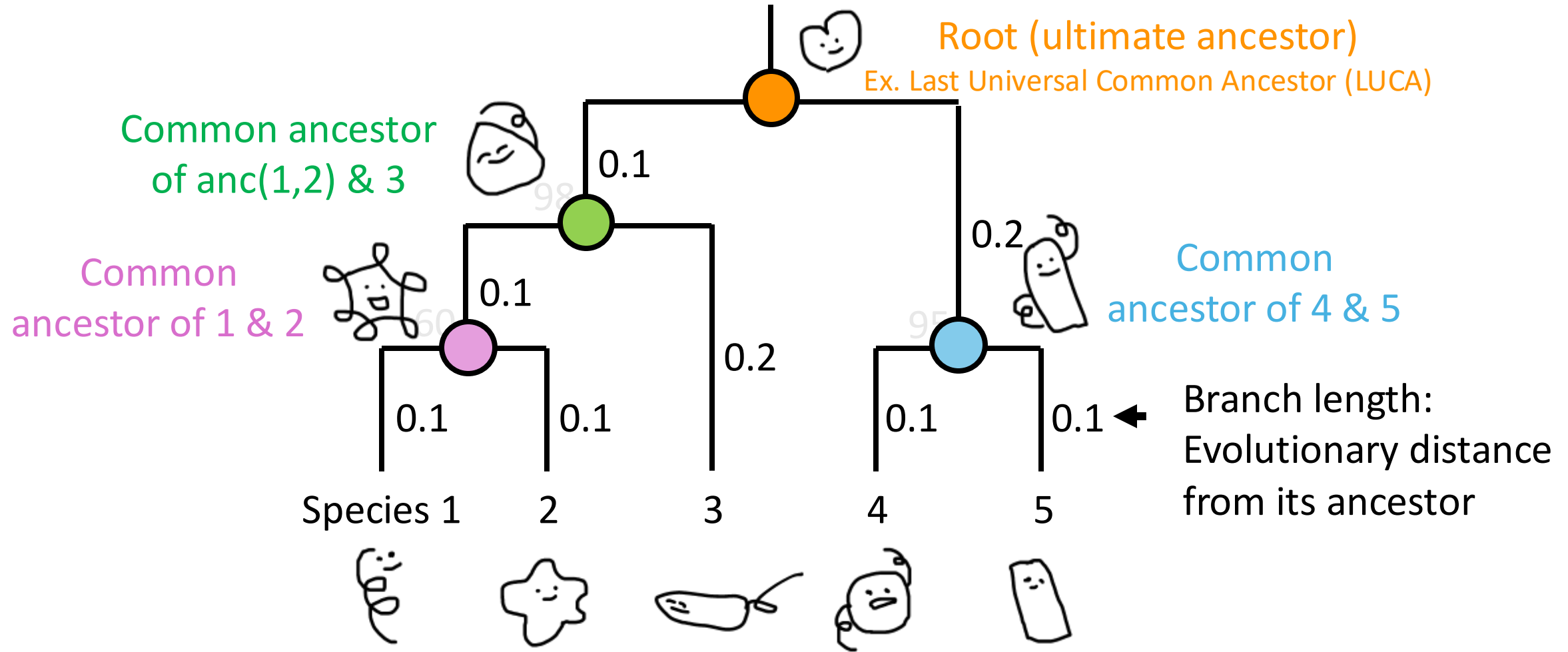
# Tree 101 – how to interpret labels on tree

---



# Tree 101 – how to interpret labels on tree

---

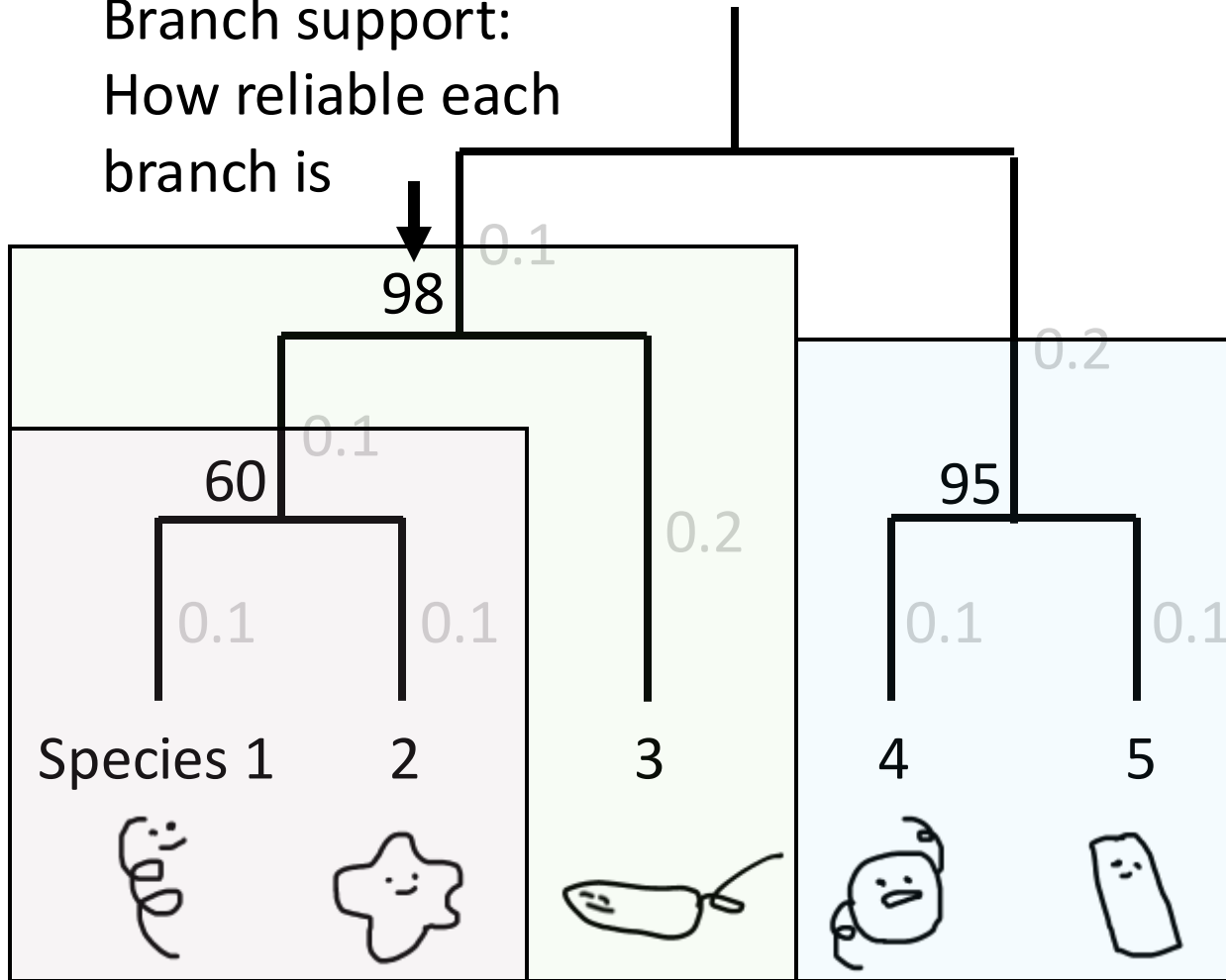




# Tree 101 – how to interpret labels on tree

---

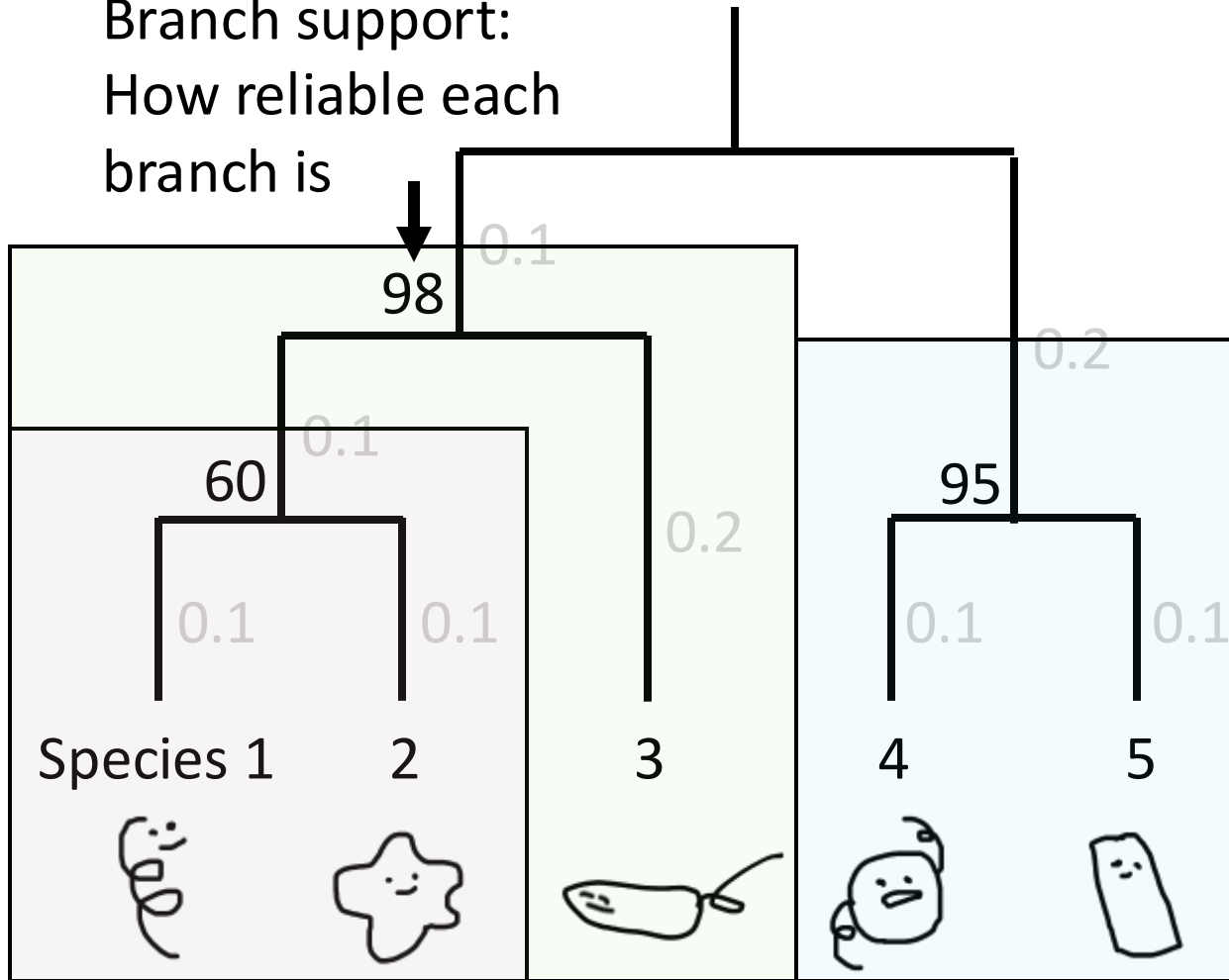
Branch support:  
How reliable each  
branch is



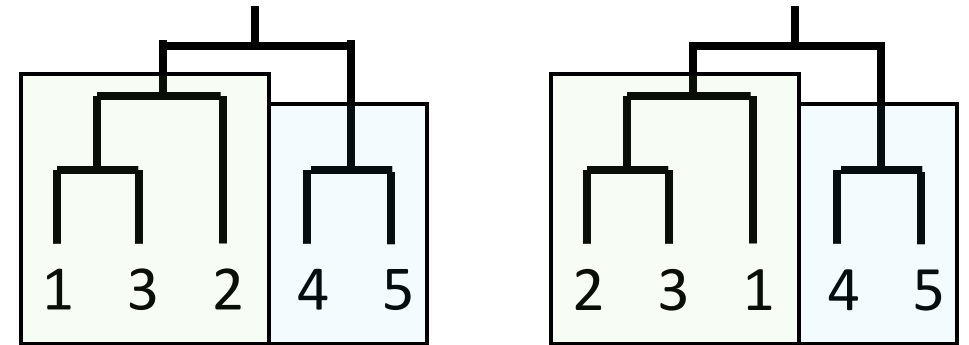
# Tree 101 – how to interpret labels on tree

---

Branch support:  
How reliable each  
branch is



Other possibilities





# Workflow to build a tree – alignment

---

①

Sequence alignment

②

Model selection

③

Tree inference

④

Assessment

# Workflow to build a tree – alignment

---

①

Sequence alignment

②

Model selection

③

Tree inference

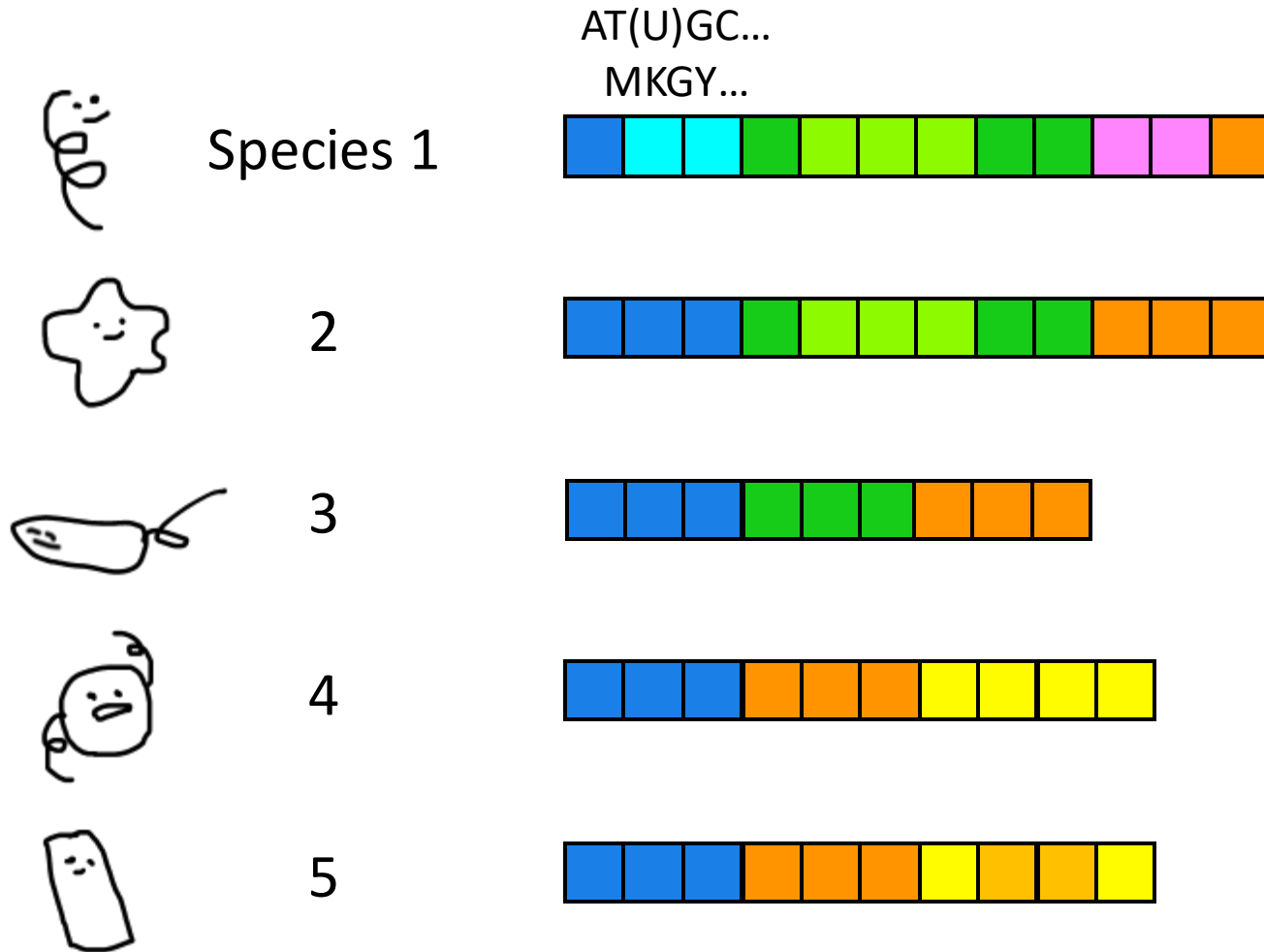
④

Assessment



# Workflow to build a tree – alignment

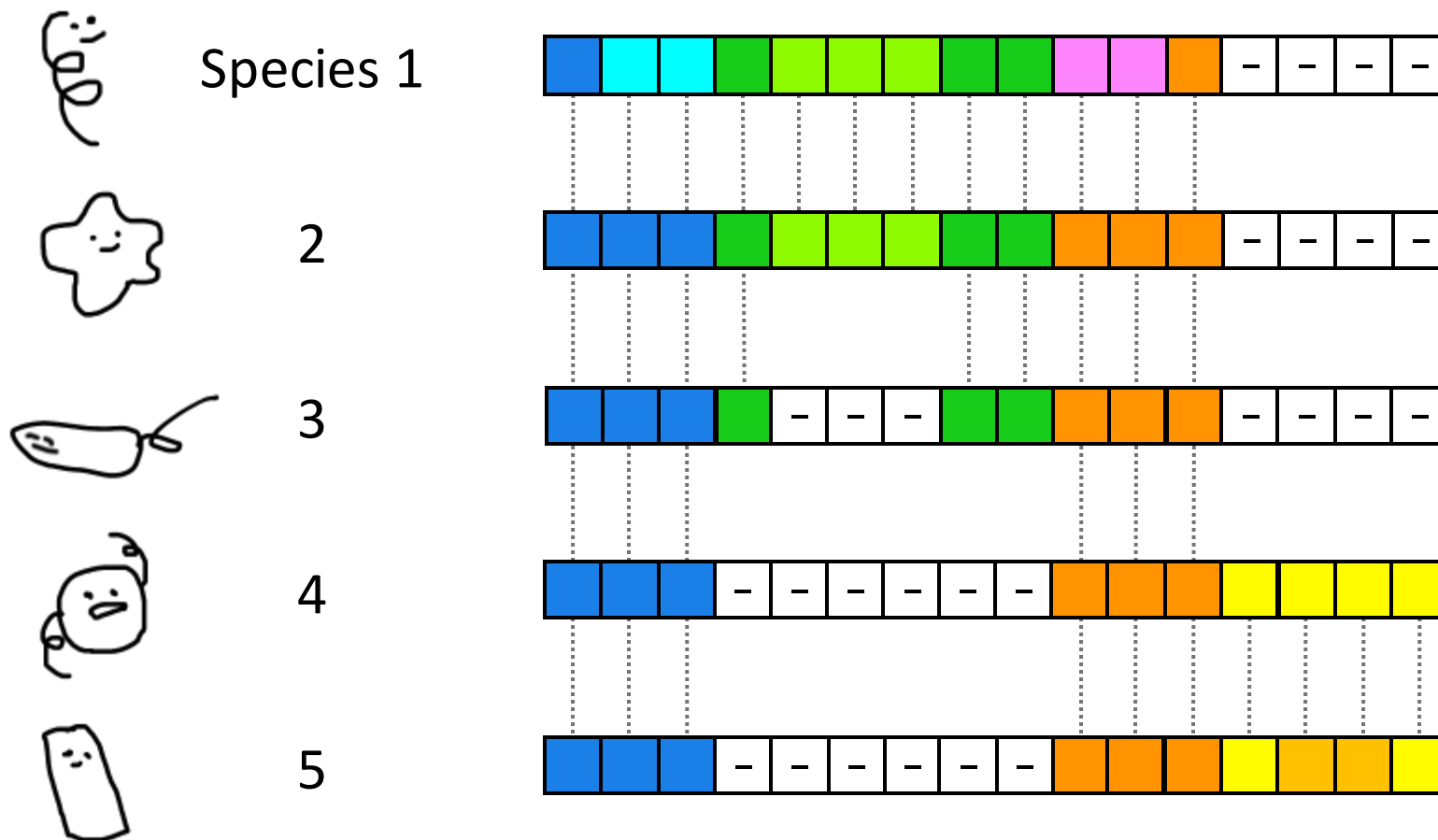
---



- 16S rRNA
- ribosomal protein(s)
- Other marker gene(s)

# Workflow to build a tree – alignment

---

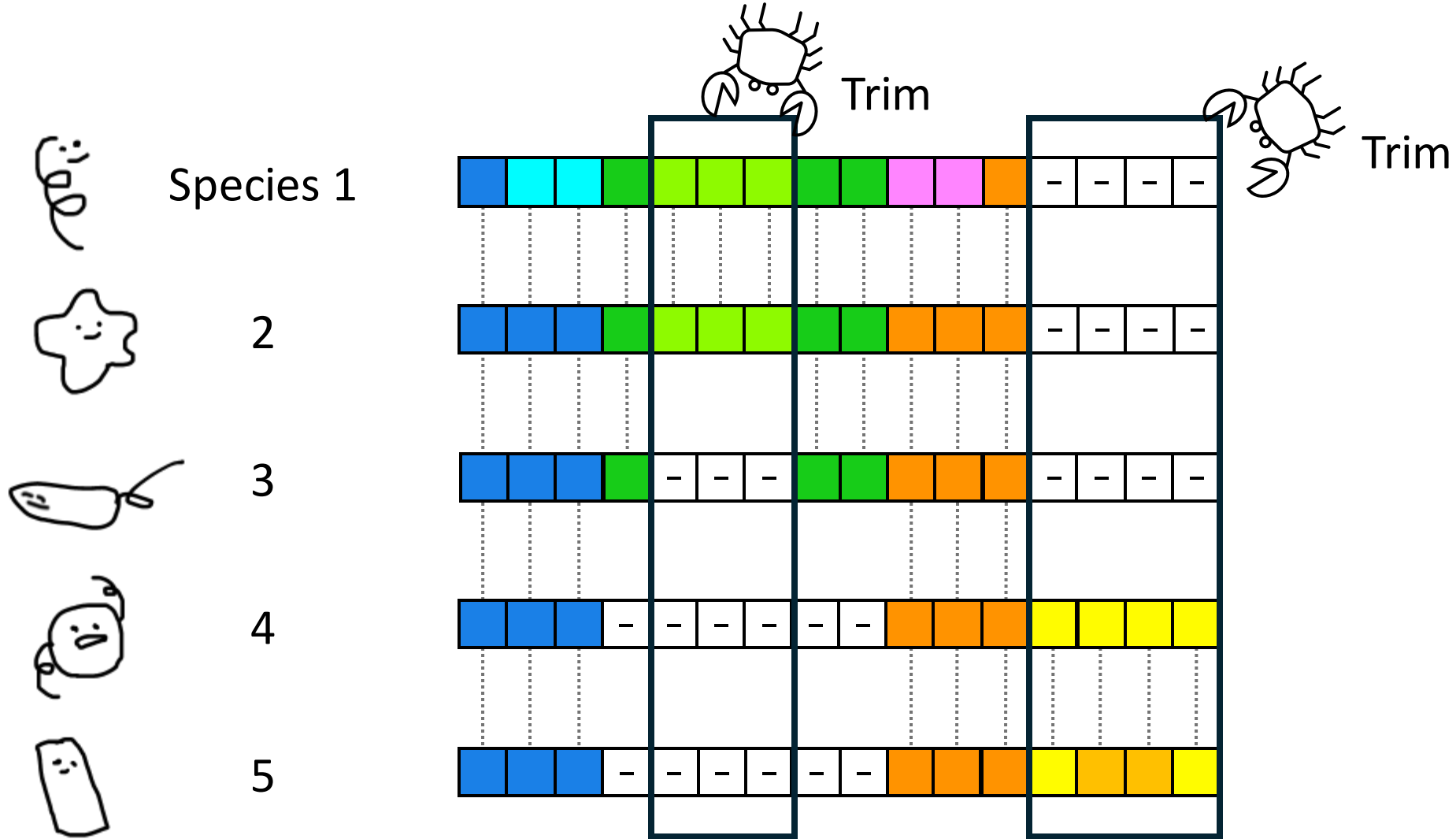


Example: MAFFT,  
Muscle, T-coffee, etc.



# Please be careful when you trim alignments

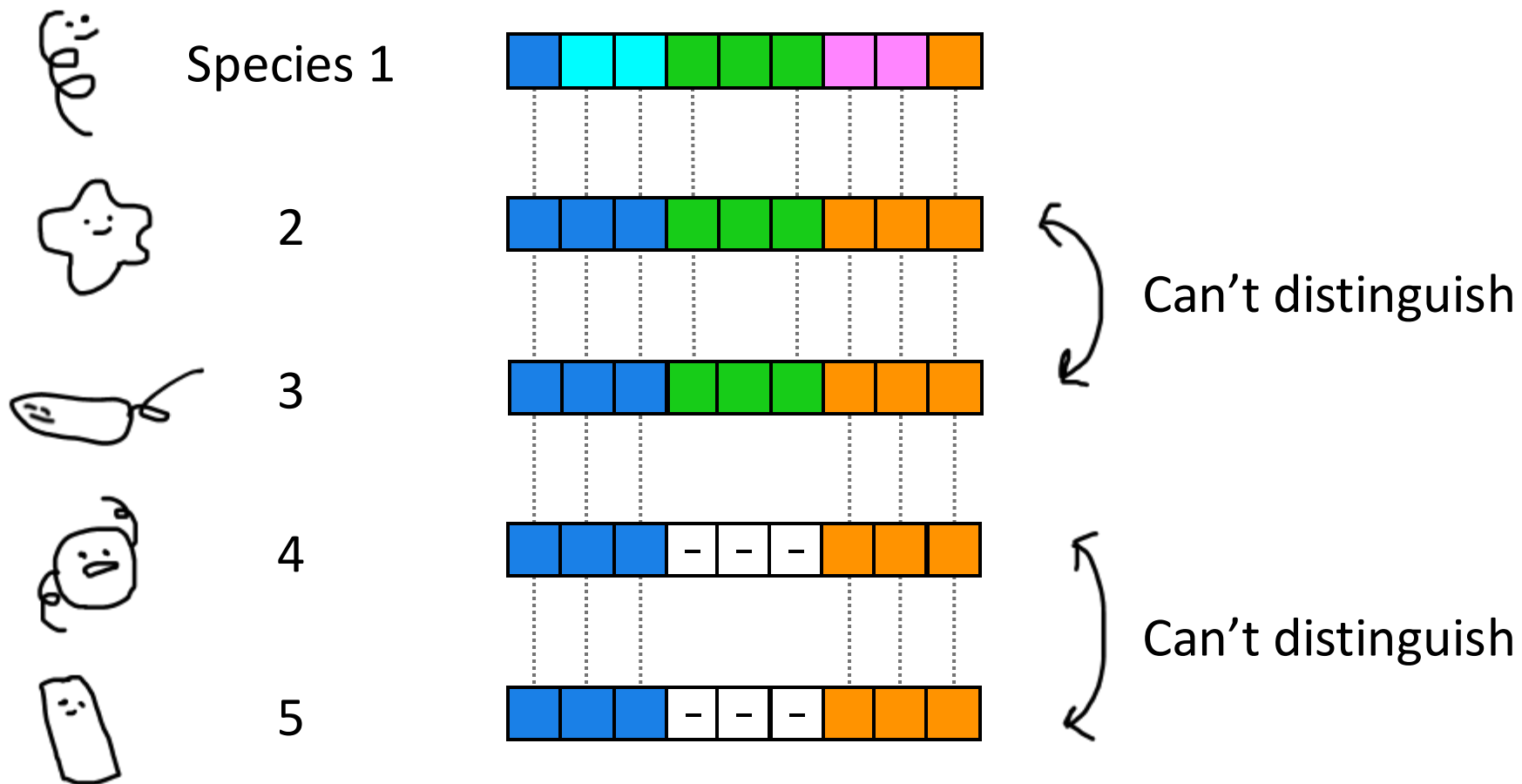
---



Example: TrimAl,  
HmmCleaner, ClipKIT,  
etc.

# Please be careful when you trim alignments

---



Example: TrimAl,  
HmmCleaner, ClipKIT,  
etc.

# Workflow to build a tree – model selection

---

①

Sequence alignment

②

Model selection

③

Tree inference

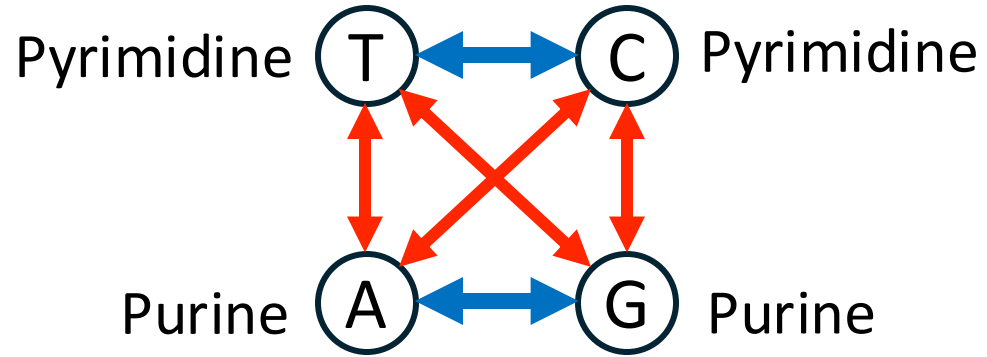
④

Assessment



## Overview of “evolutionary model”

---

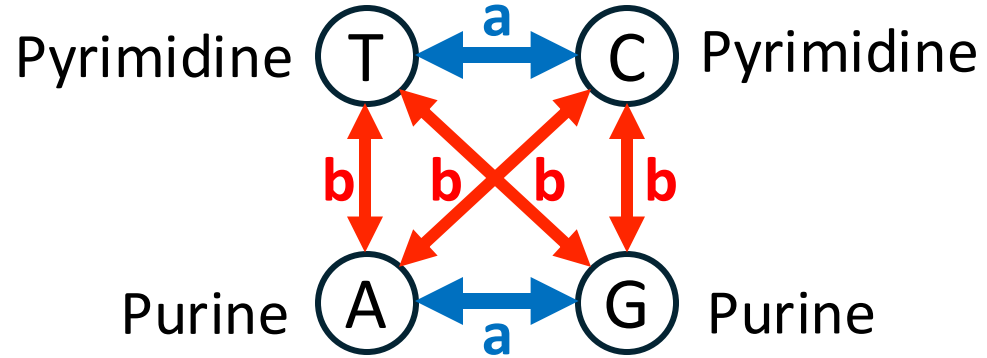


**Transition** is much more likely  
to occur than **transversion**

# Overview of “evolutionary model”

---

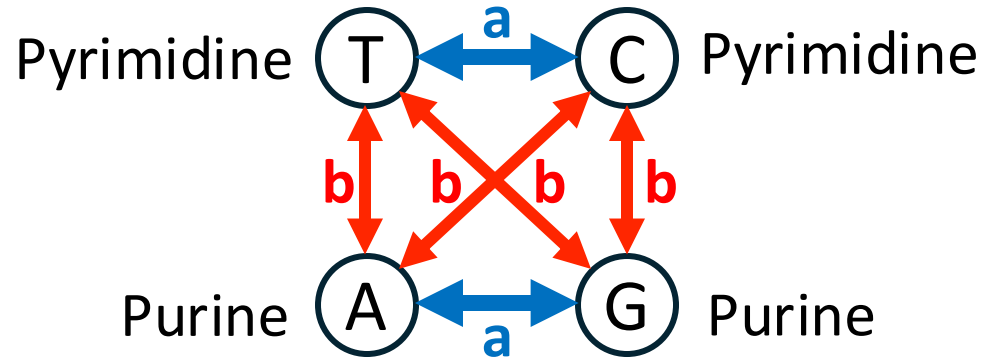
$a, b$  = Mutation rate



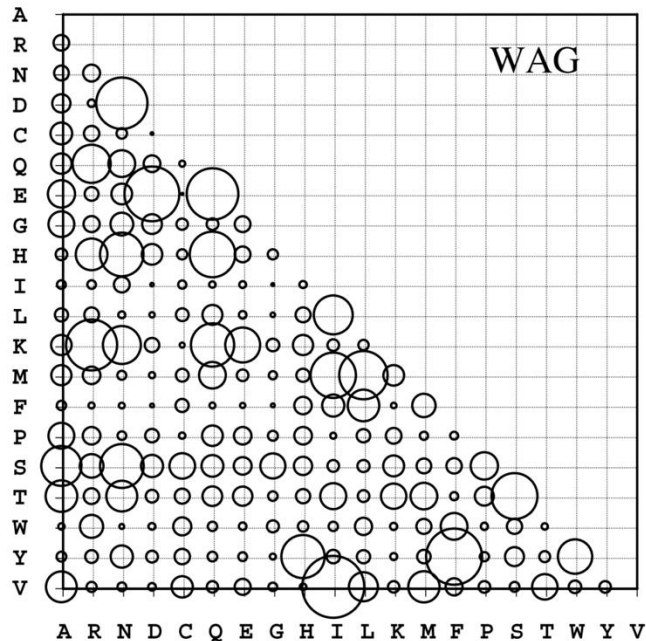
**Transition** is much more likely to occur than **transversion**

# Overview of “evolutionary model”

$a, b$  = Mutation rate



**Transition** is much more likely to occur than **transversion**



For amino acid substitutions,  
Empirical models are widely used



# Workflow to build a tree – tree inference

---

①

Sequence alignment

②

Model selection

③

Tree inference

④

Assessment

# Workflow to build a tree – tree inference

---

When you have N leaves, possible  
tree topologies =  $(2N-5)!!$

**Five species =  $5 * 3 * 1 = 15$**



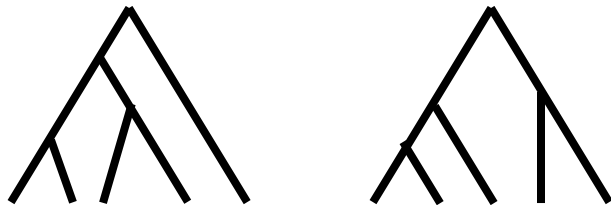
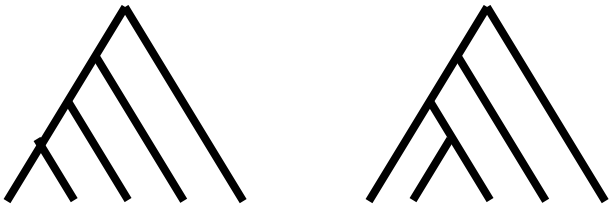
Etc... ldk

# Workflow to build a tree – tree inference

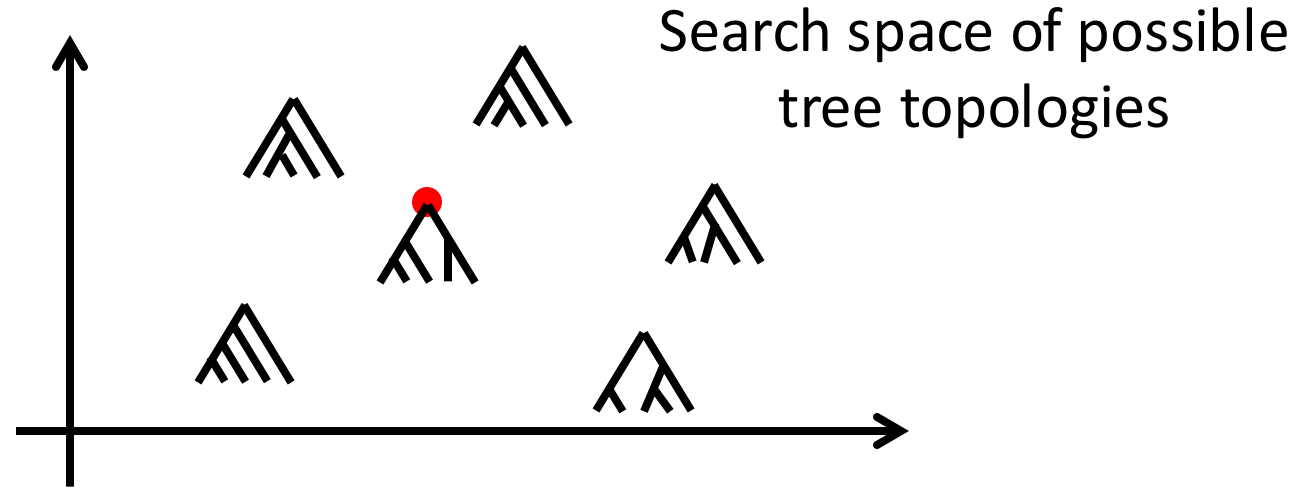
---

When you have N leaves, possible tree topologies =  $(2N-5)!!$

**Five species =  $5 * 3 * 1 = 15$**



Etc... ldk



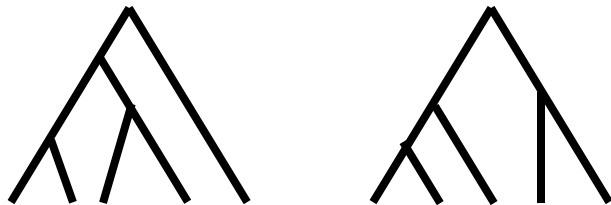
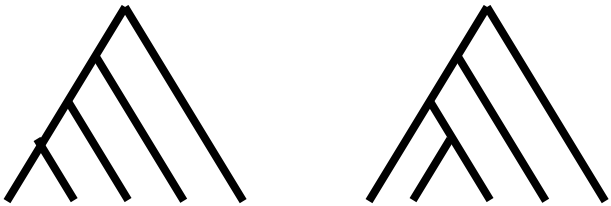


# Workflow to build a tree – tree inference

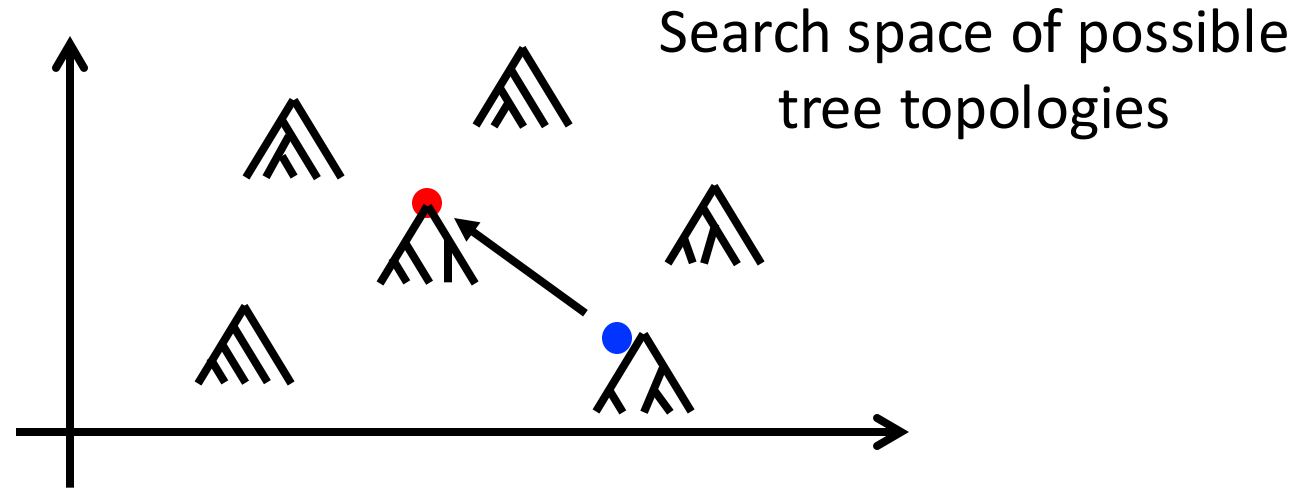
---

When you have N leaves, possible tree topologies =  $(2N-5)!!$

**Five species =  $5 * 3 * 1 = 15$**



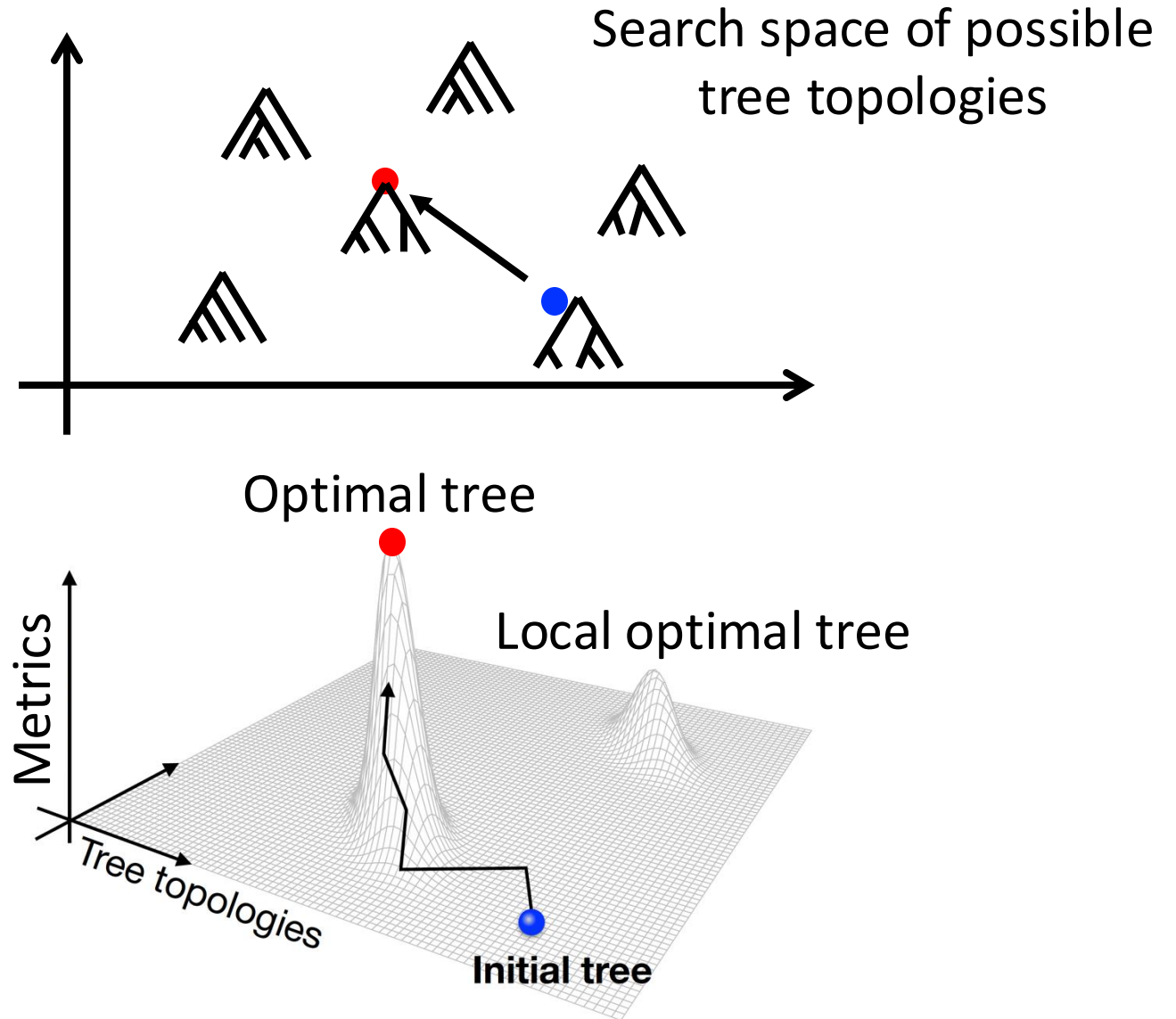
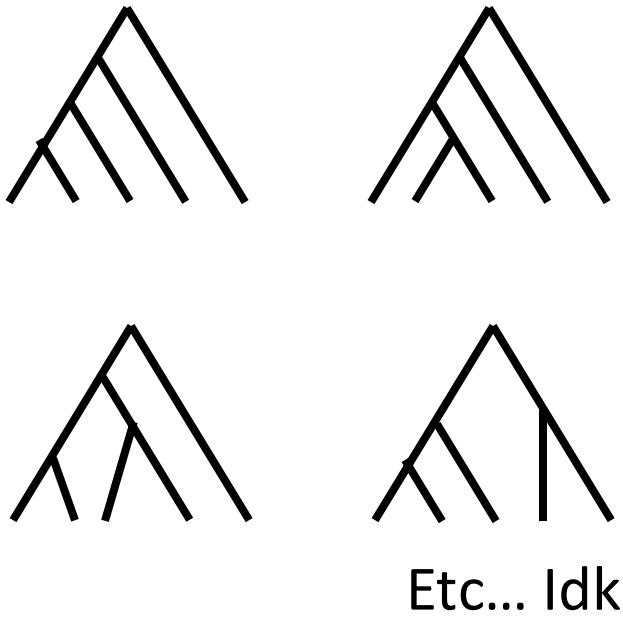
Etc... ldk



# Workflow to build a tree – tree inference

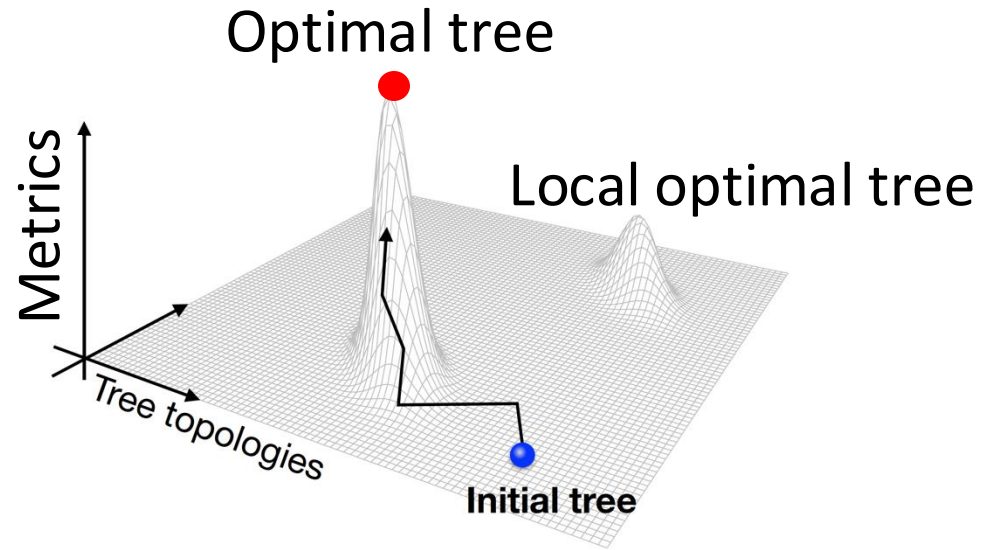
When you have N leaves, possible tree topologies =  $(2N-5)!!$

**Five species =  $5 * 3 * 1 = 15$**



# Workflow to build a tree – tree inference

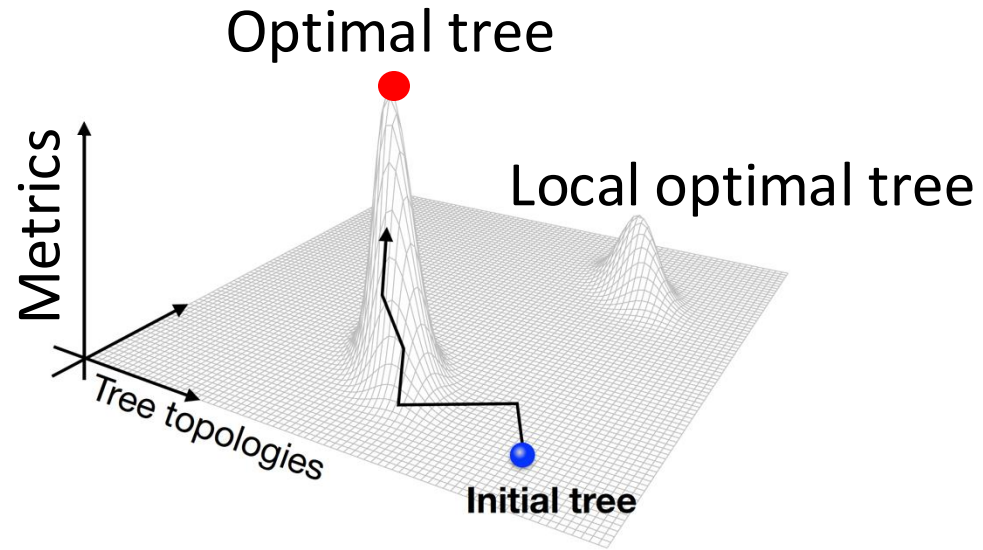
---



- Neighbor joining
- Maximum Parsimony
- Maximum likelihood
- Bayesian inference

# Workflow to build a tree – tree inference

---

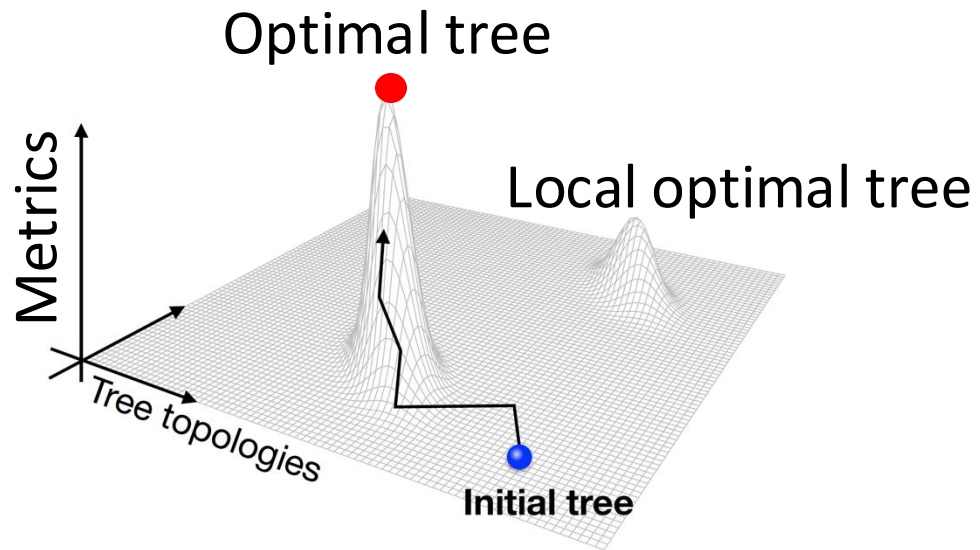


- Neighbor joining
- Maximum Parsimony
- **Maximum likelihood**
- Bayesian inference

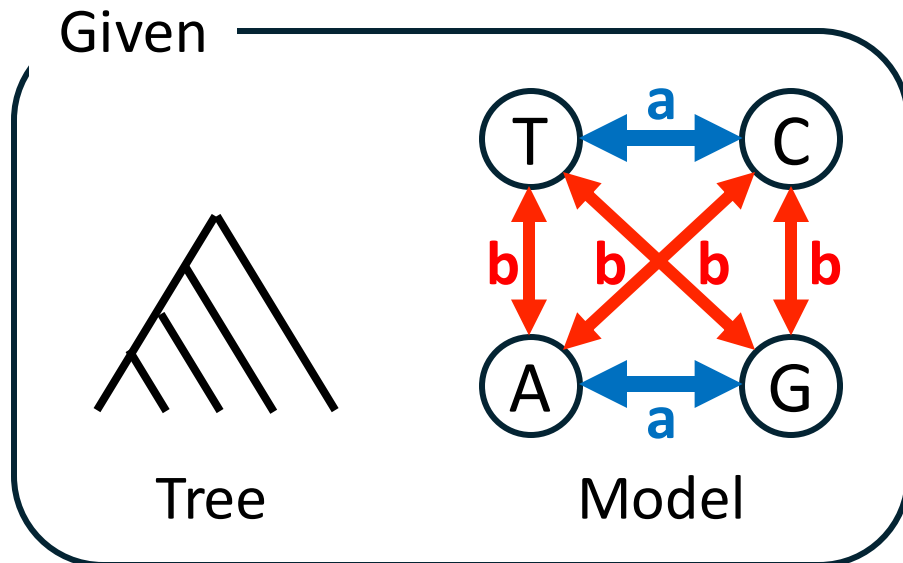
Example: IQ-Tree2



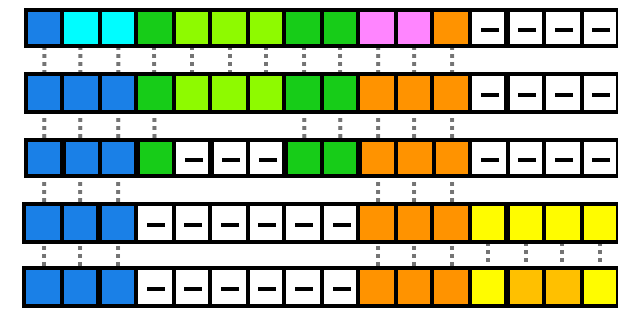
# Workflow to build a tree – tree inference



- Neighbor joining
- Maximum Parsimony
- Maximum likelihood
- Bayesian inference



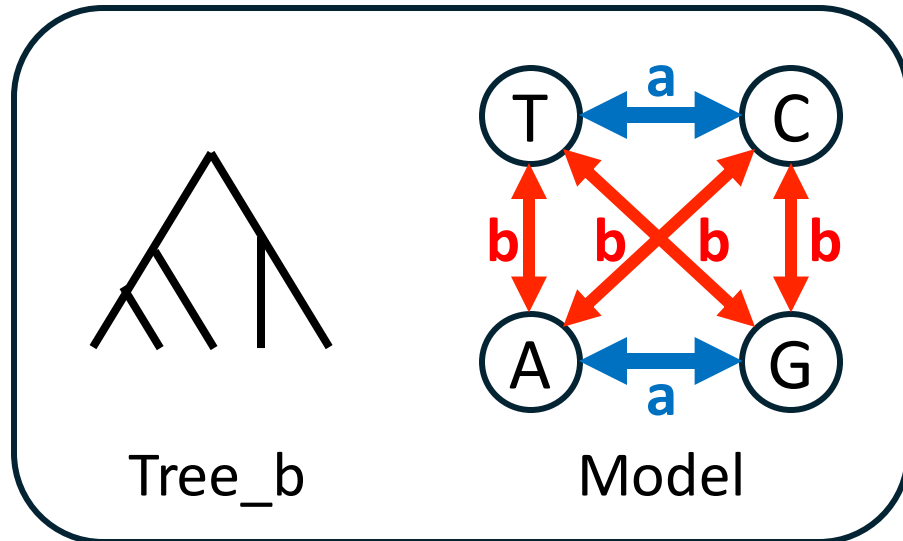
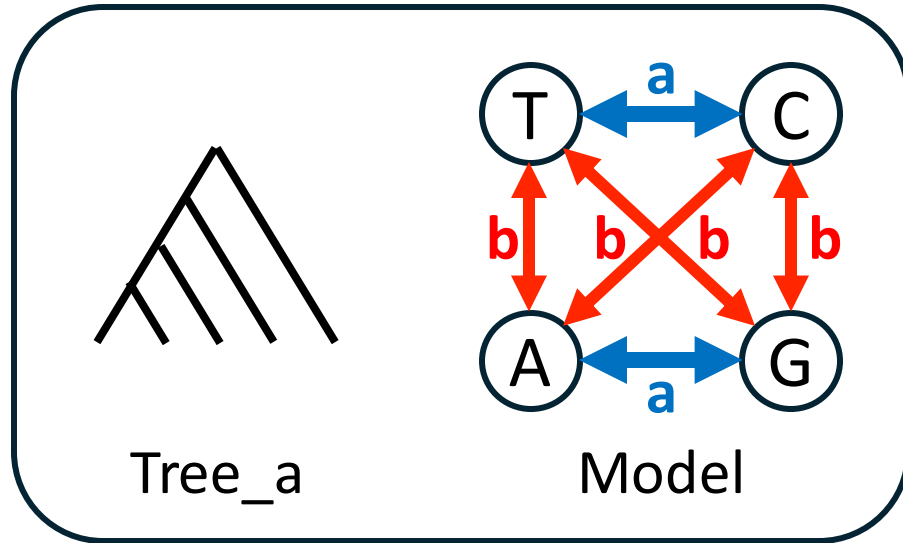
Probability = ?



Input alignment

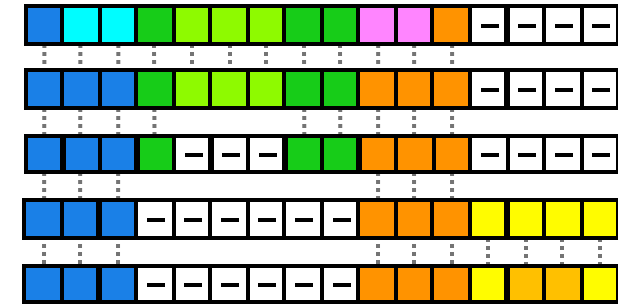
# Workflow to build a tree – tree inference

---



Prob =  $e^{-100}$

Prob =  $e^{-5}$



Input alignment

# Workflow to build a tree – assessment

---

①

Sequence alignment

②

Model selection

③

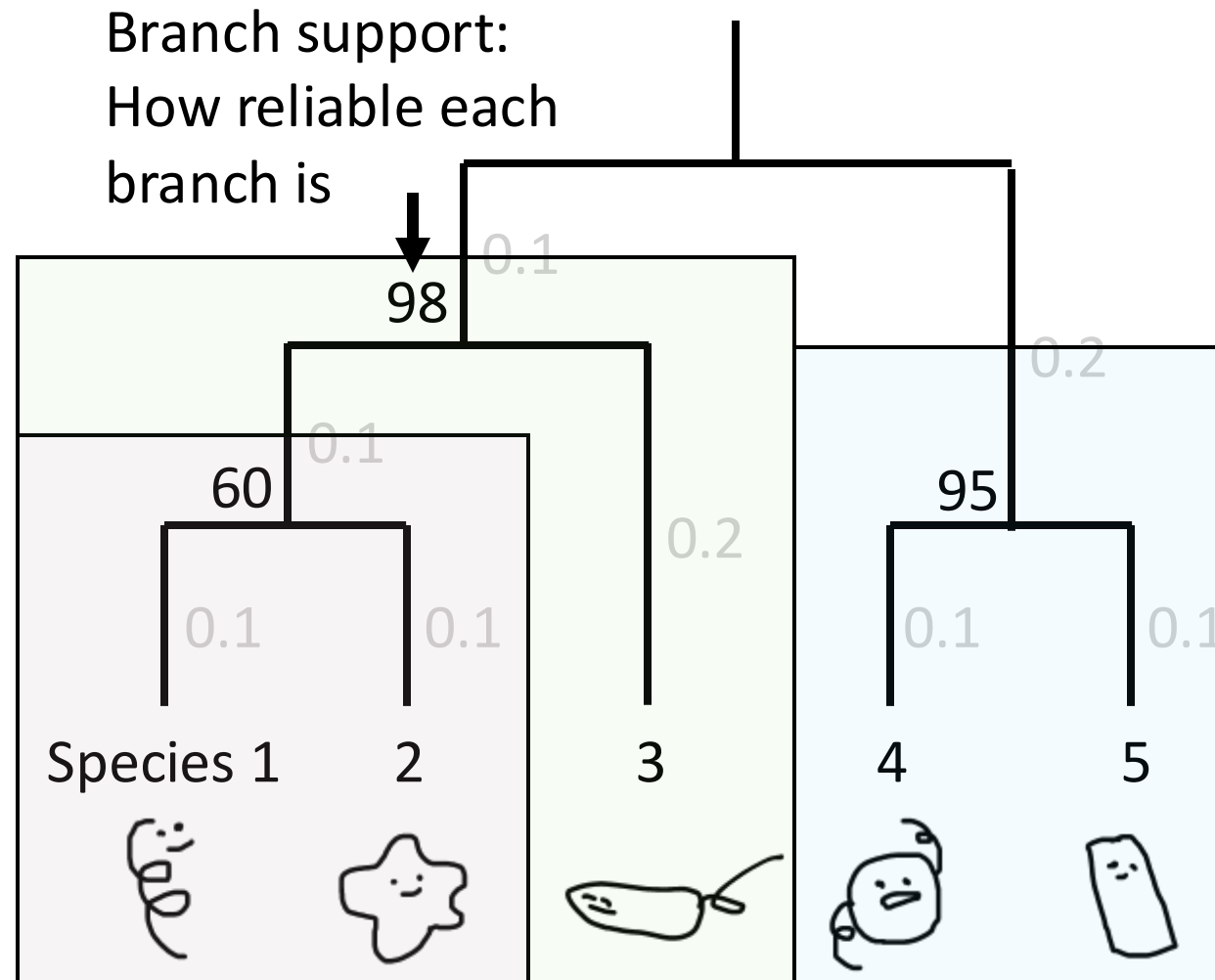
Tree inference

④

Assessment

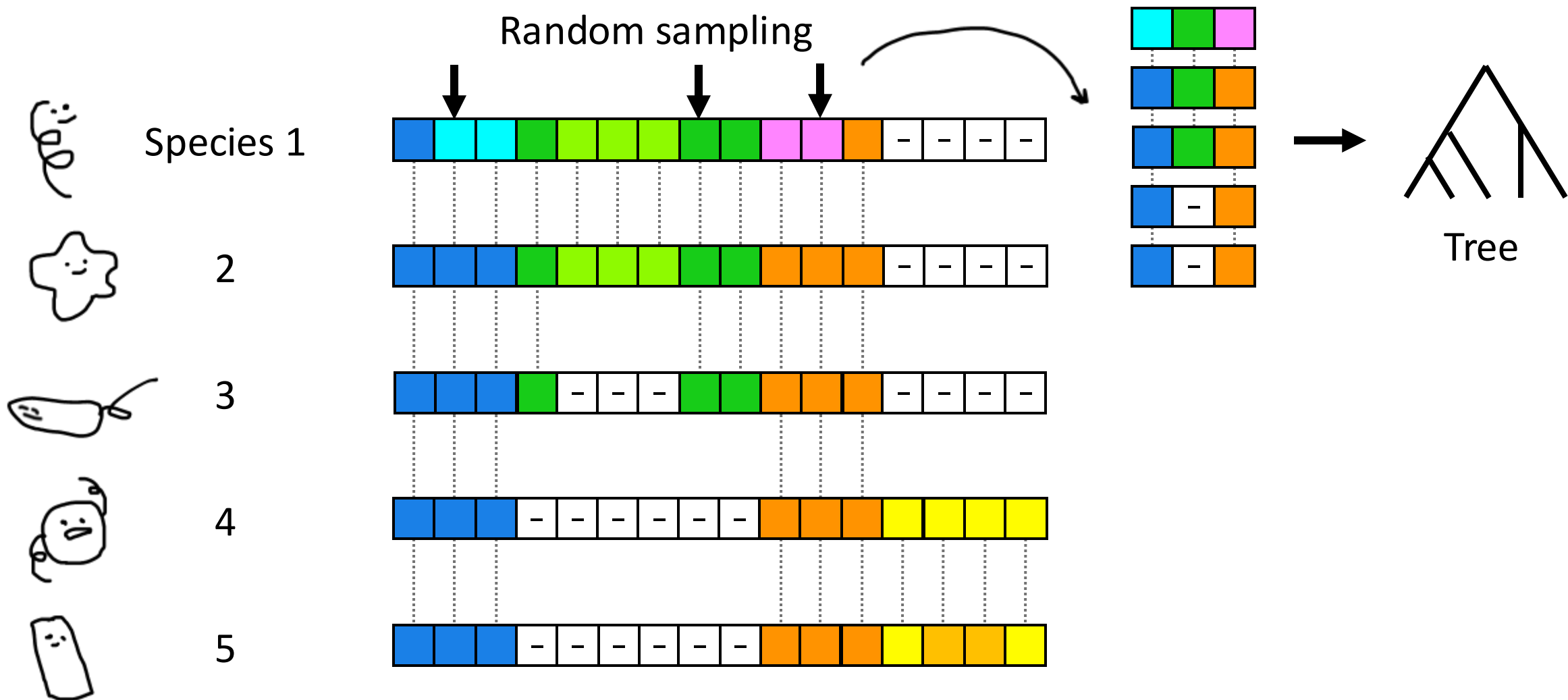
# Workflow to build a tree – assessment

---

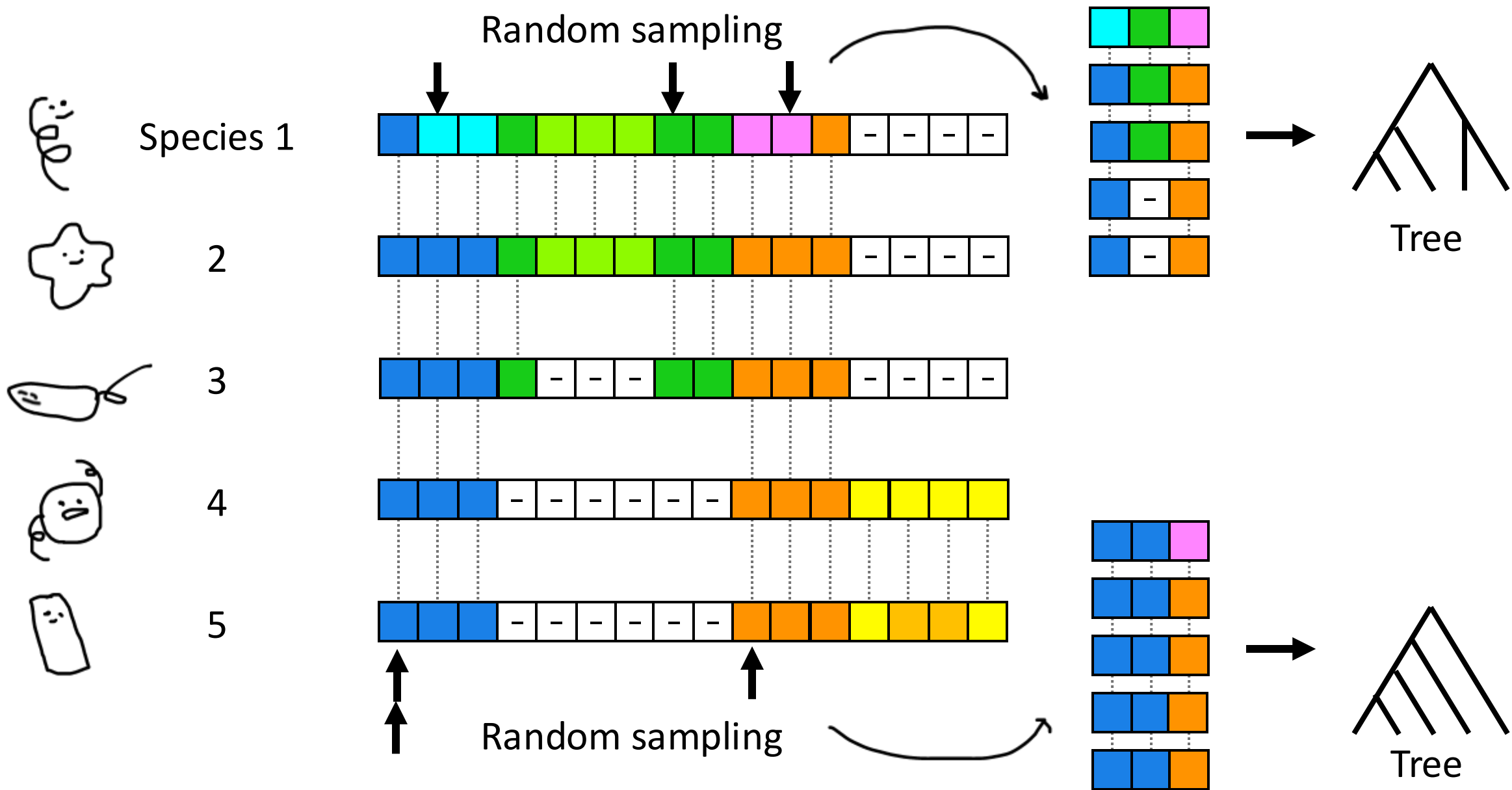




# How bootstrap works

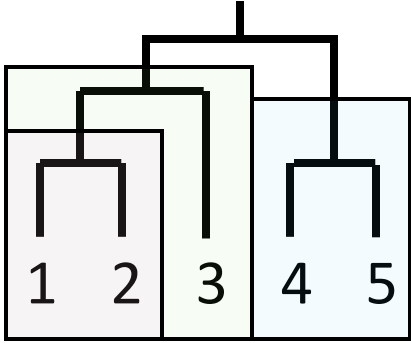
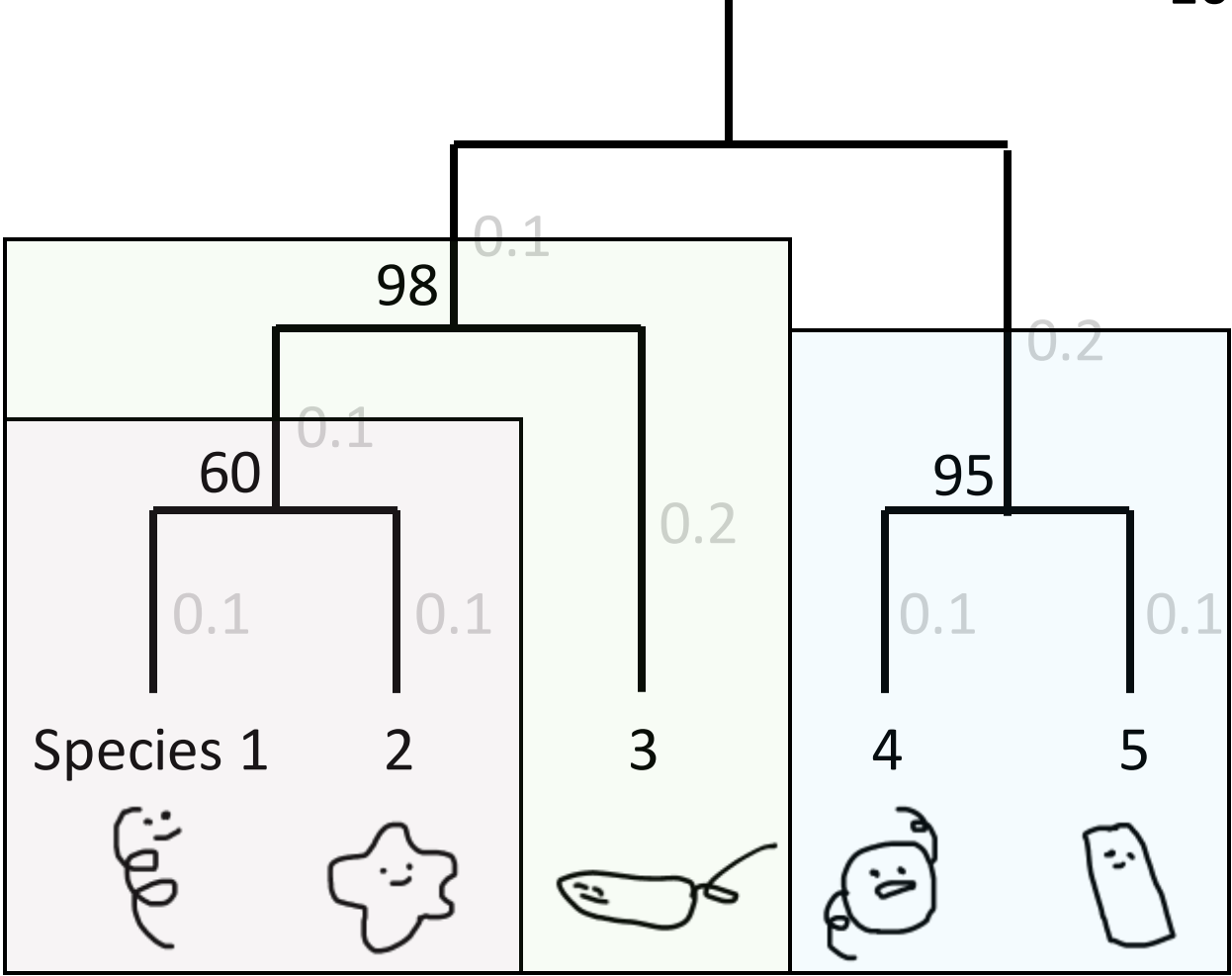


# How bootstrap works

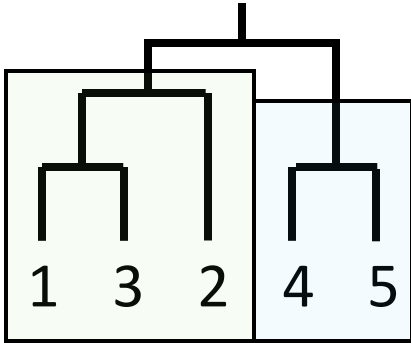


# How bootstrap works

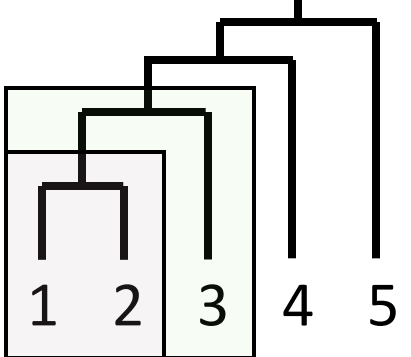
100 Iterations ->



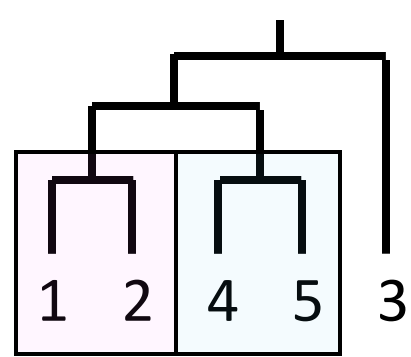
53 times



40 times



5 times



2 times

How do I look?

