Likelihood in Phylogenetics

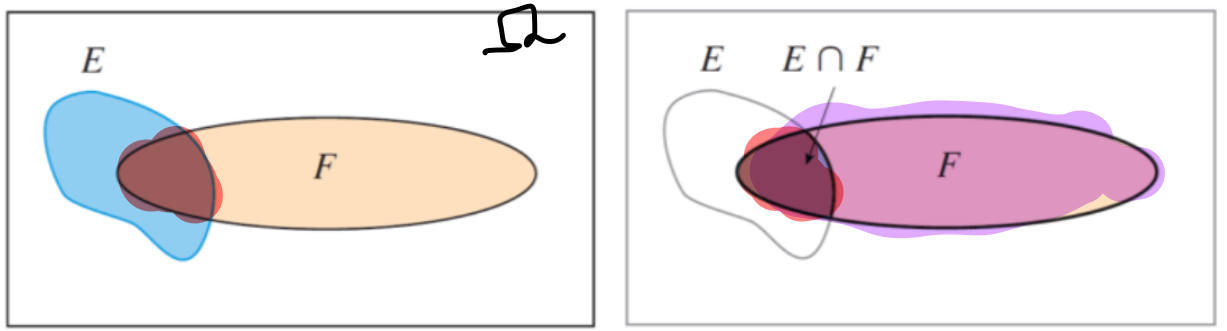
What matters in probability is almost never one single probability field, but rather interrelatedness of many probability fields (Freudenthal 1973, pp. 613).

Rosana Zenil-Ferguson
Assistant Professor. University of Kentucky
MOLE at Woods Hole workshop 2026

Background Knowledge

- Conditional probability
- Total probability
- Phylogenetic tree notation (Newick/Computer)

Conditional probability



Definition of **conditional probability** or Bayes' Theorem

If $P(F) > 0$ then

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

given that

$E - \Omega$

$F - \Omega$

$(E \cap F) \rightarrow \sigma\text{-algebra}$

and

Example: Color Blindness

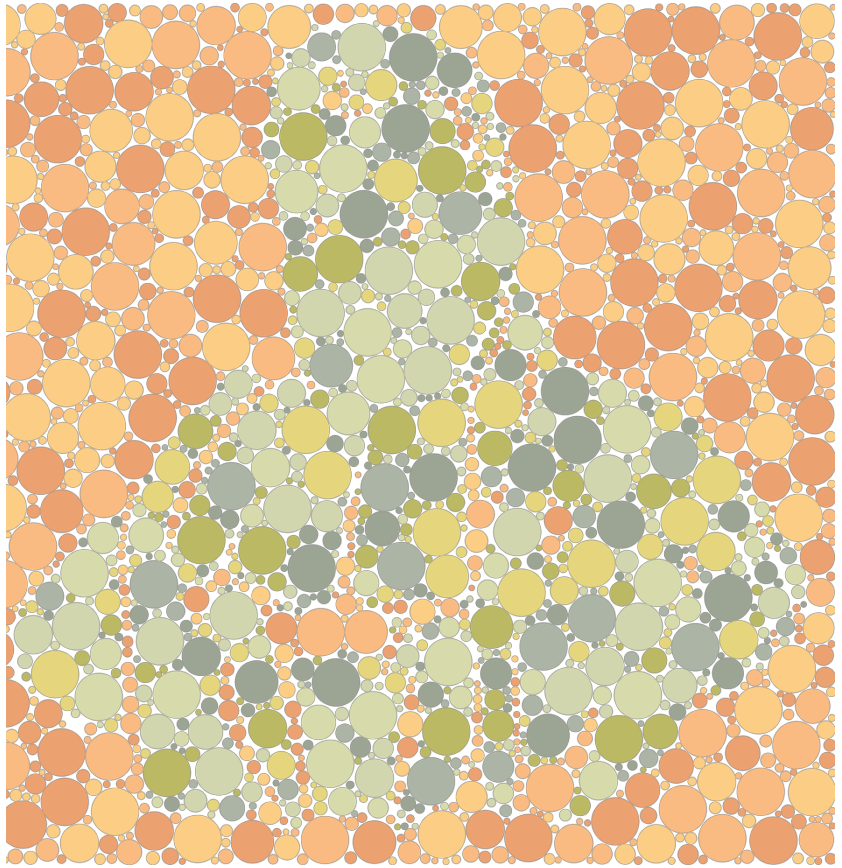
Color blindness is caused by a genetic mutation and results in people not being able to distinguish certain colors.

$$P(S=1, C=1) = 0.05$$

- The probability of being a male and having color blindness is 0.05
- The probability of being a female and having color blindness is 0.005

$$P(S=0, C=1) = 0.005$$

$$S = \begin{cases} 0 & \text{female} \\ 1 & \text{male} \end{cases} \qquad C = \begin{cases} 0 & \text{color friendly} \\ 1 & \text{color blind} \end{cases}$$



Using conditional probability to reflect knowledge

$$P(S=0) = P(S=1) = 0.5$$

- Assuming that there is an equal number of males and females in the population, what is the probability of an individual being color blind, given that the individual is a male?

$$P(C=1 \mid S=1) = \frac{P(C=1, S=1)}{P(S=1)} = \frac{0.05}{0.5} = 0.1$$

- Assuming that there is an equal number of males and females in the population, what is the probability of an individual being color blind, given that the individual is a female?

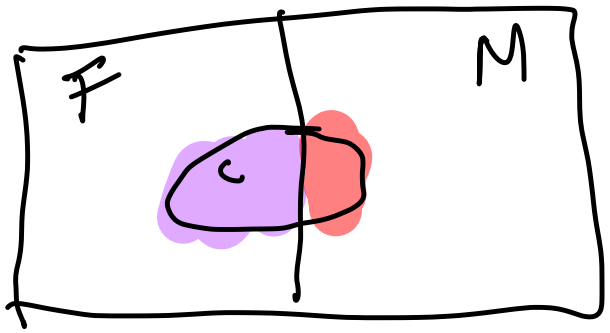
$$0.01$$

Total Probability

- Assuming that there is an equal number of males and females in the population, what is the probability that a randomly chosen individual will be color blind?

$$P(C=1) = P(C=1 \mid S=0)P(S=0) + P(C=1 \mid S=1)P(S=1)$$

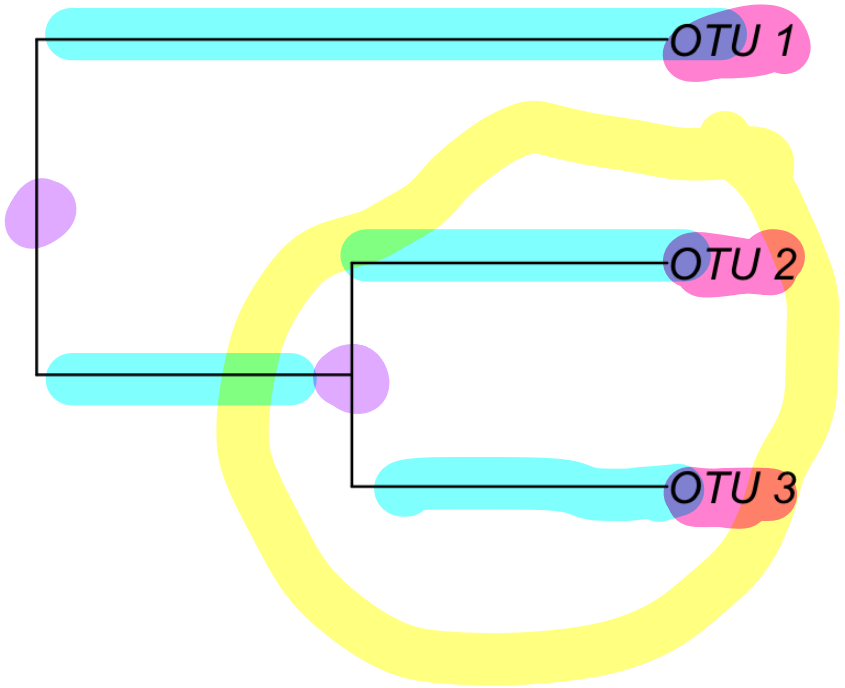
or



$$= (0.01) \times 0.5 + 0.1 \times 0.5$$

$$= 0.055$$

Phylogenetic tree notation



Node: a point in a phylogeny where a lineage splits
- one speciation event
- common ancestor of branches that extend from it

Branches: lineages evolving through time
- the stuff between speciation events

Tips: terminal ends representing species, molecules, or populations being compared

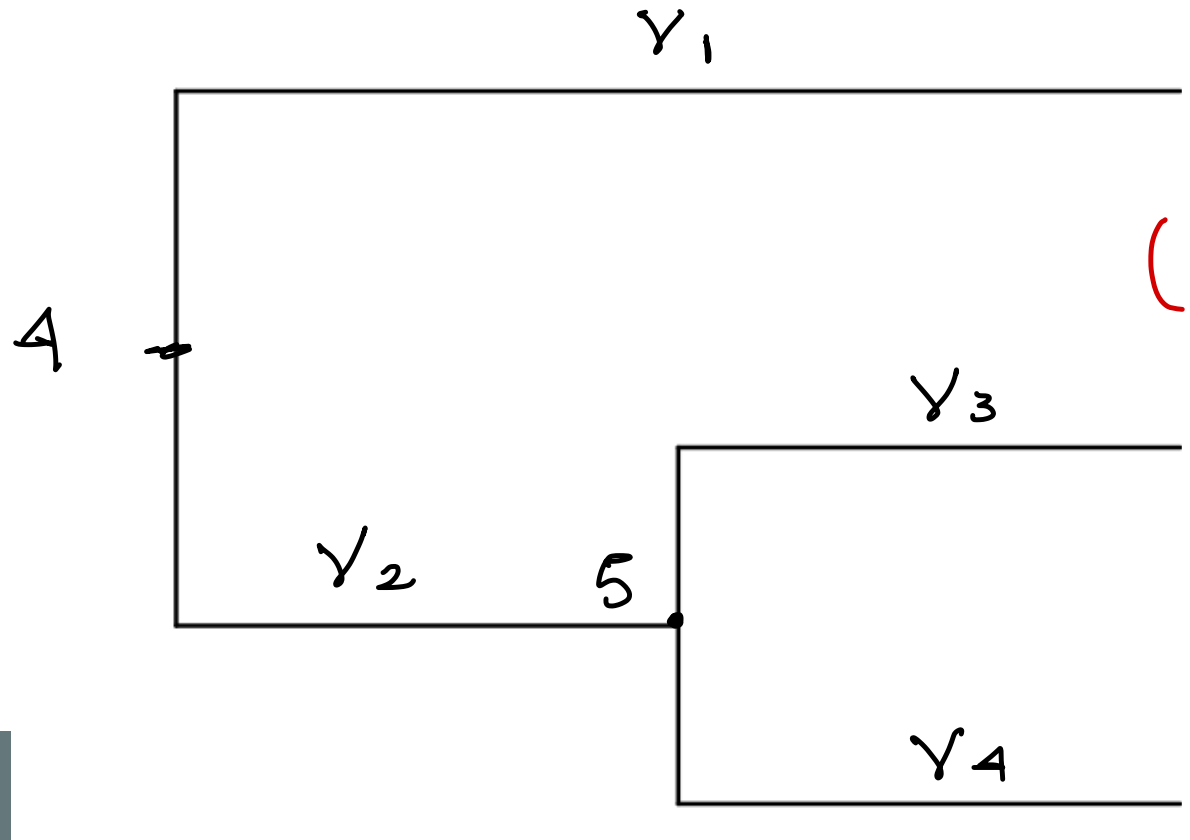
Clade: An ancestor and all its descendants

Rooted tree: includes a distantly related "outgroup" species to polarize changes and show which node is common ancestor to all lineages in ingroup

When interpreting trees, pay most attention to nodes (= common ancestors)!!

Newick format

ν → Notation for branch lengths "nu"



Newick Notation

((2, 3), 1);

Likelihood for a phylogenetic tree

Three essential ingredients

1. Molecular Data

2. Model for nucleotide evolution

3. A phylogenetic tree: (a) the relationships between taxa, (b) branch lengths.

Historical data: A morphological character matrix

Morphological data matrix for Carnivora

Taxa	Character-state scoring											
	# 1	2	3	4	5	6	7	8	9	10	11	12
Lemur (outgroup)	0	0	0	0	0	0	0	0	0	0	0	0
Cat	0	1	0	1	0	0	1	1	1	0	0	0
Hyena	0	1	0	1	0	0	1	0	1	0	0	0
Civet	0	1	0	0	0	0	0	0	1	0	0	0
Dog	1	0	0	0	1	0	0	0	0	0	0	0
Raccoon	1	0	0	0	1	0	0	0	0	0	0	0
Bear	1	0	0	0	1	1	0	0	0	1	0	0
Otter	1	0	0	0	1	0	0	0	0	1	0	0
Seal	1	0	1	0	1	1	0	0	0	1	1	1
Walrus	1	0	1	0	1	1	0	0	0	1	1	1
Sea lion	1	0	1	0	1	1	0	0	0	1	0	0

All character states in outgroup are ancestral

This does NOT mean that the outgroup lacks derived traits!

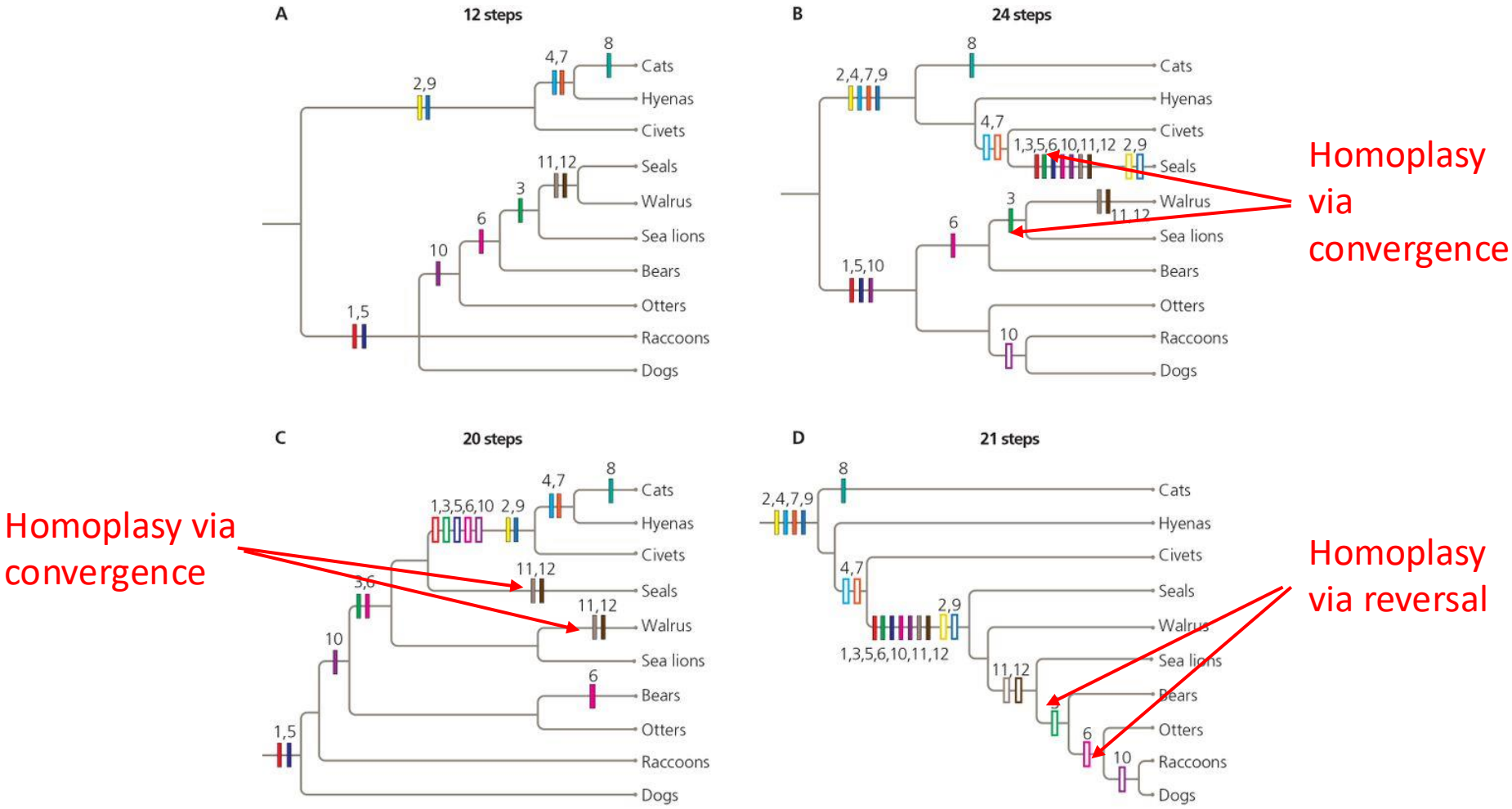
Lots of these…

1 1 1 1 0 0 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 1

Not informative because they are not synapomorphies in ingroup.

Outgroups help us identify **shared derived** states (**synapomorphies)**

How to choose a tree that best explains the data



Bars = synapomorphies (shared, derived traits)

Open bars = reversion to ancestral-like state

Parsimony analysis in practice…

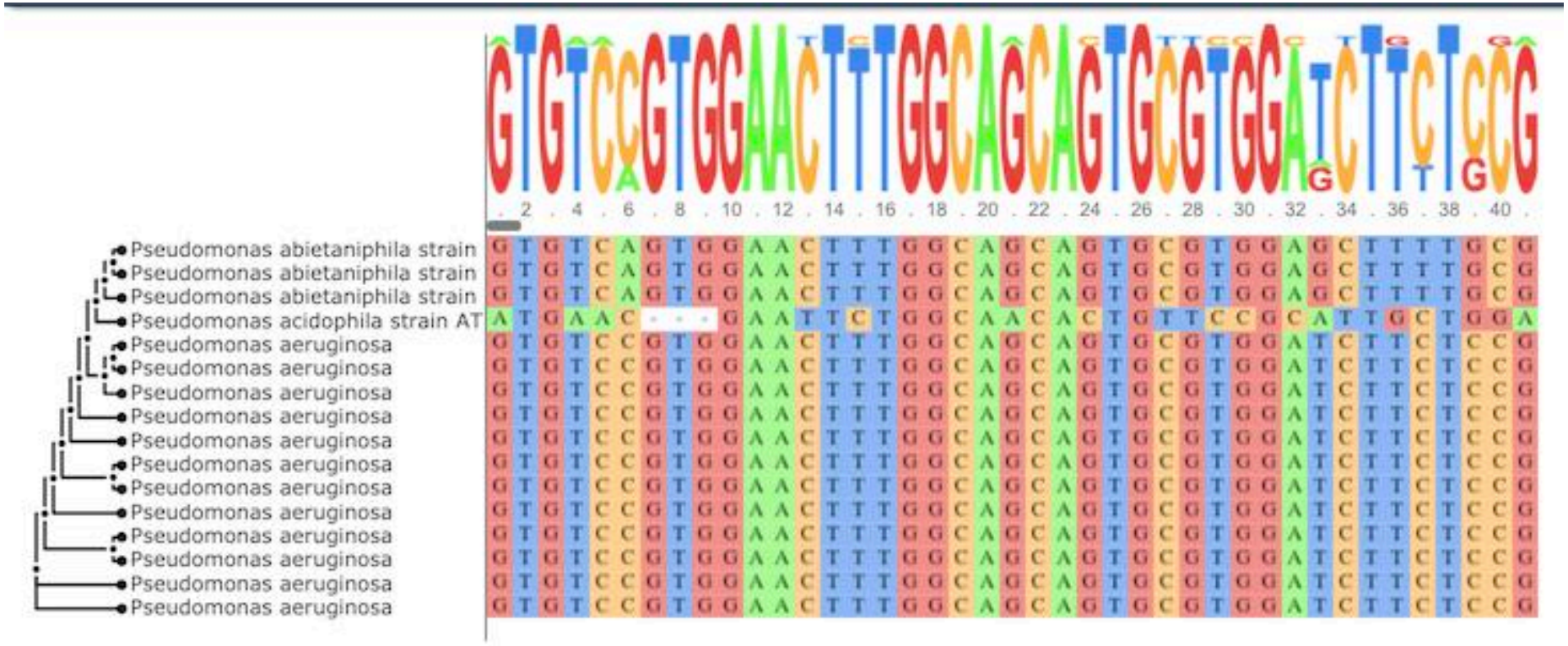
TABLE 9.1 The Huge Number of Possible Tree Topologies

# of Taxa	# Unrooted trees	# Rooted trees
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
11	34,459,425	654,729,075
12	654,729,075	13,749,310,575
13	13,749,310,575	316,234,143,225

Even with computers, trees cannot be exhaustively searched for most analyses so these programs employ algorithms to efficiently search “tree space”

The modern era molecular phylogenetics

1. Molecular Data

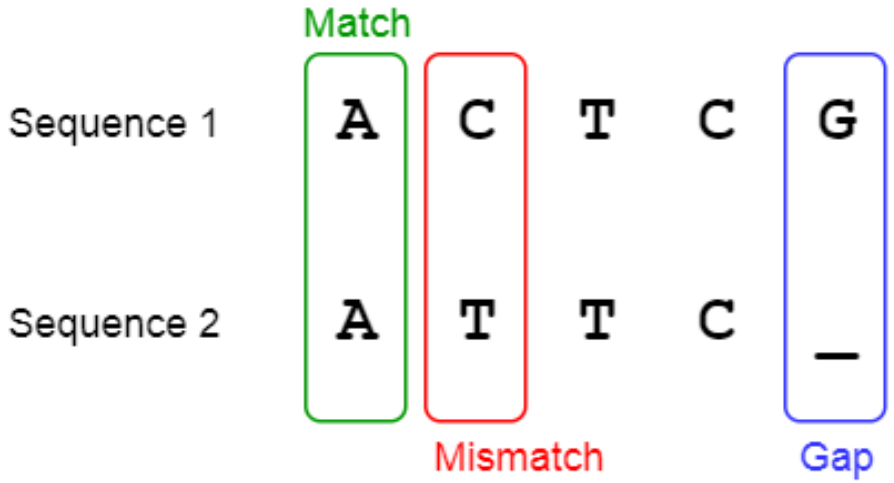


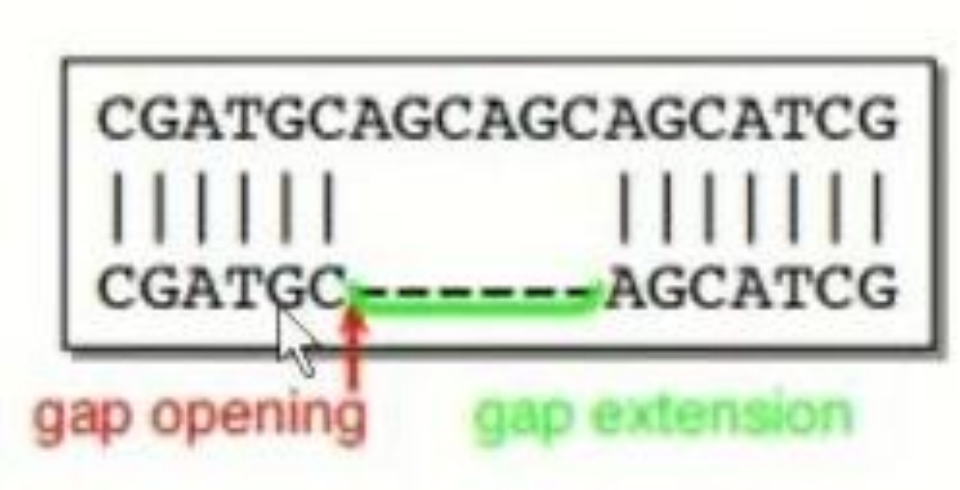
Multiple sequence alignment

- Multiple sequence alignment (MSA) is important for phylogenetic estimation or model-based inference of evolutionary processes
- The goal of MSA is to introduce gaps into sequences so that columns of an aligned matrix contain character states that are homologous
- Homology cannot be directly observed but can be inferred

Inferring Homology

- Placing gaps in a sequence is penalized too
- Introducing a new gap usually has a higher cost than extending an existing gap





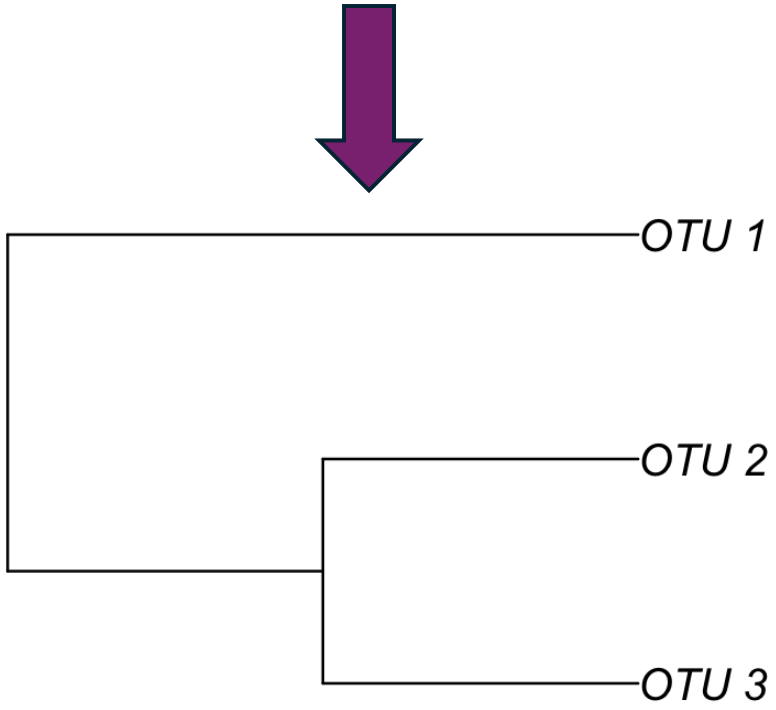
Generic Alignment Scoring Parameters

match = +5
mismatch = -3
gap open = -5
gap extension = -2

	Score	
Example 1		
Alignment 1 AGTTCCCTG AGTTA--TG	20	In example 1, the first alignment has a higher score for minimizing gap openings
Alignment 2 AGTTCCCTG AGTT-A-TG	17	
Example 2		
Alignment 1 AGTTCCACTG AGTTA---TG	18	In example 2, the score imporovement from an extra match outwieghs the cost of an extra gap opening
Alignment 2 AGTTCCACTG AGTT--A-TG	23	

Likelihood in a phylogenetic tree

Taxon	Site 1	Site 2	Site 3	Site 4
OTU 1	G	G	G	G
OTU 2	G	A	A	G
OTU 3	A	A	G	A



How do we go from sequences to trees?

2. Model for nucleotide substitutions
Probability Matrix (Conditional Probability)

Conditional Probability Distribution

X = nucleotide value over the branches of the phyla

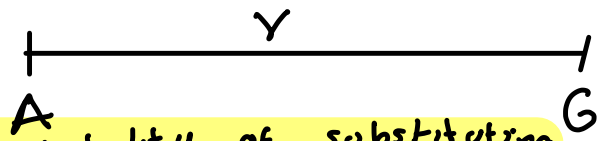
$X = \{A, G\}$

$$P = \begin{matrix} & A & G \\ A & P_{AA} & P_{AG} \\ G & P_{GA} & P_{GG} \end{matrix} \begin{matrix} \leftarrow \text{finish} \\ = 1 \\ = 1 \end{matrix}$$

start (rows)

$$P_{AG} = P(X(v) = G \mid X(0) = A)$$

given that



the probability of substituting a G for an A in a branch length v

Assumptions

(1)

Taxon	Site 1
OTU 1	G
OTU 2	G
OTU 3	A

Site 1 ⊥ Site 2

independent

Taxon	Site 2
OTU 1	G
OTU 2	A
OTU 3	A

(2)

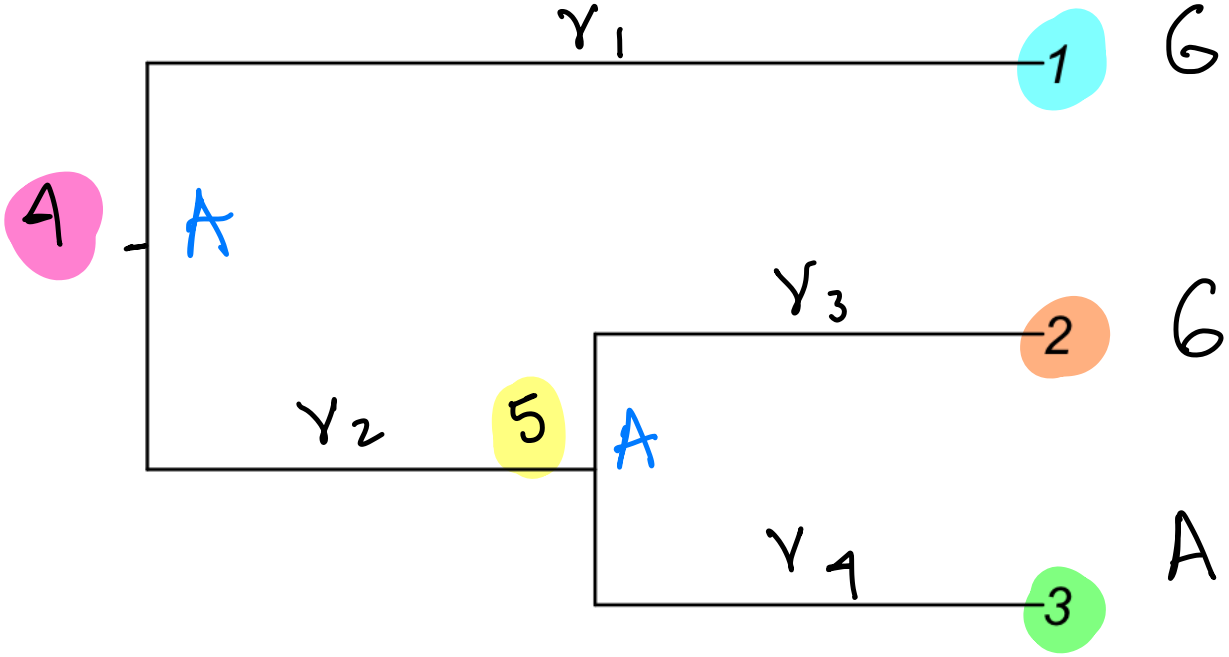
Conditional independence

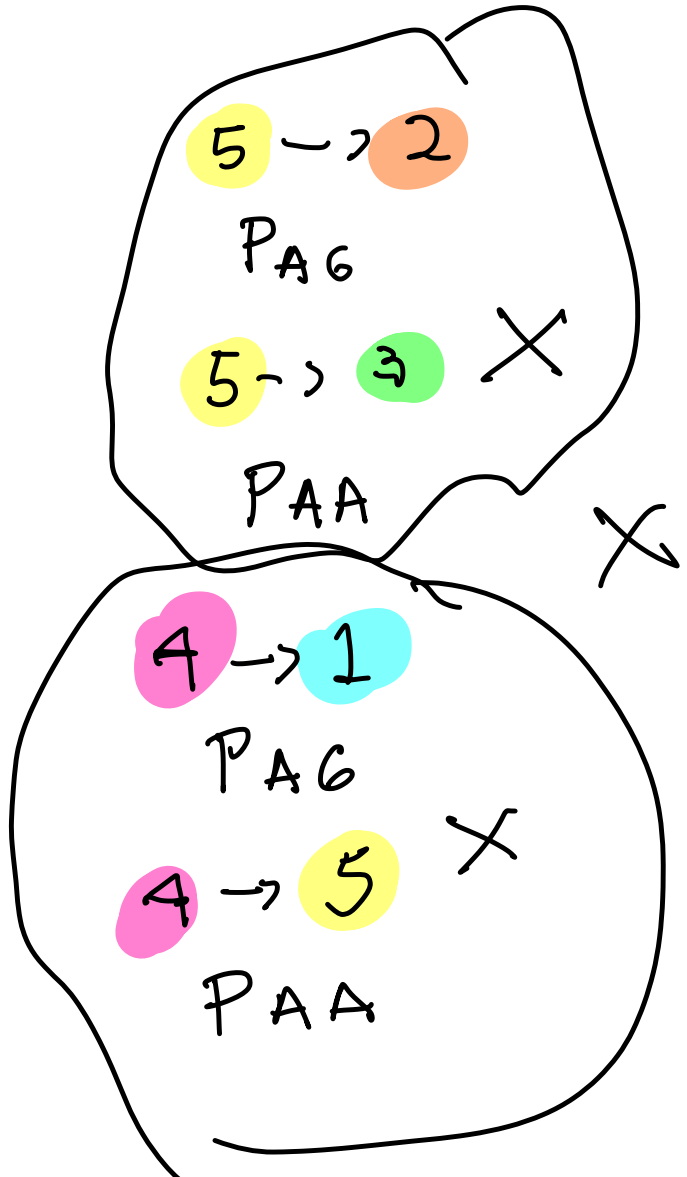
When conditioning over the most recent common ancestor two lineages are independent

One probable story for Site 1

Taxon	Site 1
OTU 1	G
OTU 2	G
OTU 3	A

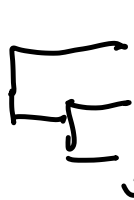
Assumption
4 has A
and 5 has A





All the probable stories for Site 1

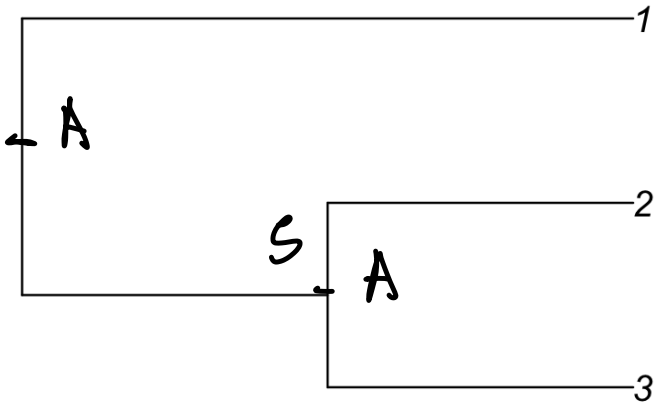
$p_0(A)$, $p_0(G)$ } probability of starting with a nucleotyde

$L($  ; Site 1, P $)$

Ψ

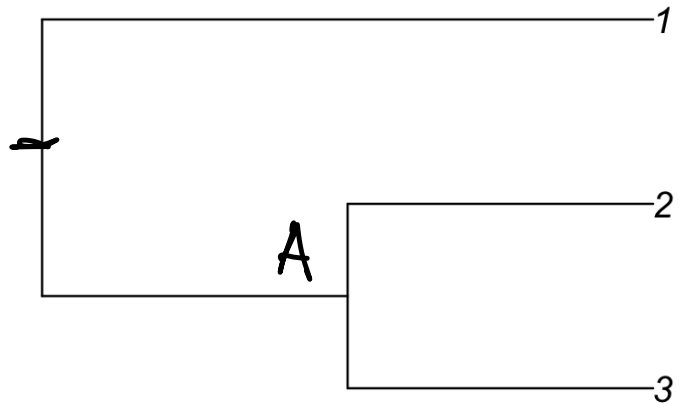
psi

$p_{AA}^{v4} \times p_{AG}^{v3} \times p_{AA}^{v2} \times p_{AG}^{v1} \times p_0(A)$



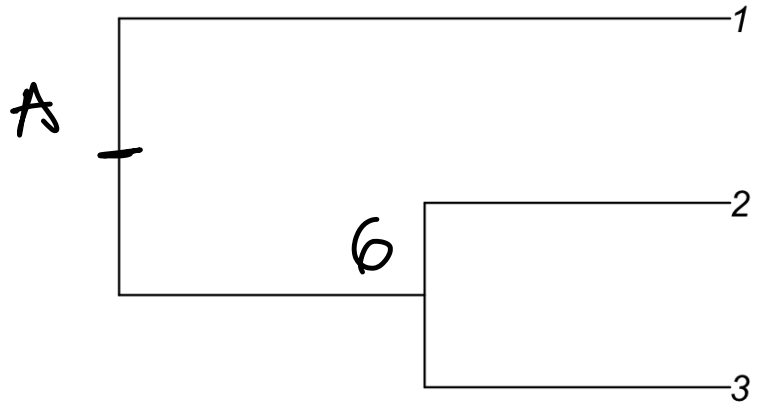
+

$p_{GA}^{v4} \times p_{GG}^{v3} \times p_{AG}^{v2} \times p_{AG}^{v1} \times p_0(G)$



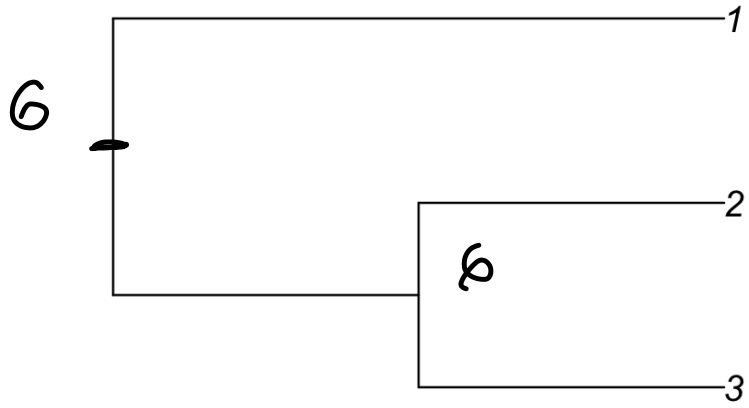
+

$p_{AA}^{v4} \times p_{AG}^{v3} \times p_{GA}^{v2} \times p_{GG}^{v1} \times p_0(A)$



+

$p_{GA}^{v4} \times p_{GG}^{v3} \times p_{GG}^{v2} \times p_{GG}^{v1} \times p_0(G)$



Probability for a single site givem one tree and model

one tree but all the stories

$$P(\text{Site } 1 | \psi_1, \text{Model}) = \sum_{j=A}^{G} \left(\sum_{i=A}^{G} (p_{iA}^{v4} \times p_{iG}^{v3}) \times p_{ij}^{v2} \right) \times p_{jG}^{v1} \times p_0(j)$$

only 1 site

$P = \begin{pmatrix} P_{AA} & P_{AG} \\ P_{GA} & P_{GG} \end{pmatrix}$

putting all together

Likelihood for all sites and one tree under one model

$$\mathcal{L}(\psi_1; \text{Data}, \text{Model (P)}) = \prod_{k=1}^{4} \text{P}(\text{Site k} | \psi_1, \text{Model (P)})$$

a single tree

multiply all the sites because we assume they are independent

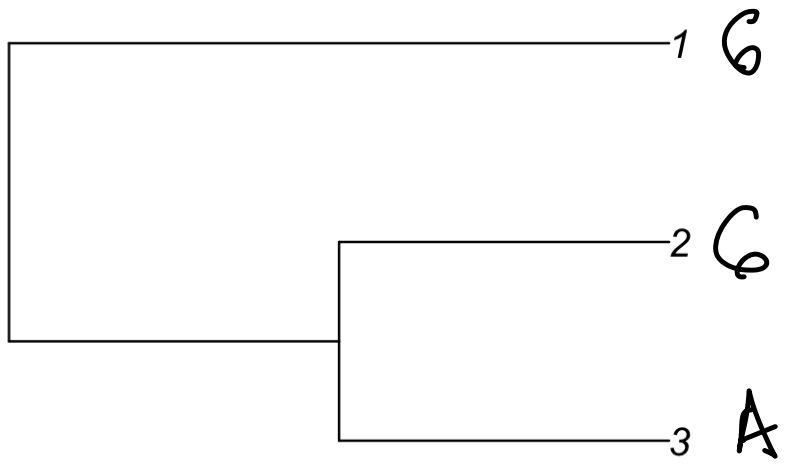
Maximum likelihood phylogenetic tree (MLE)

Which tree maximizes the likelihood function?

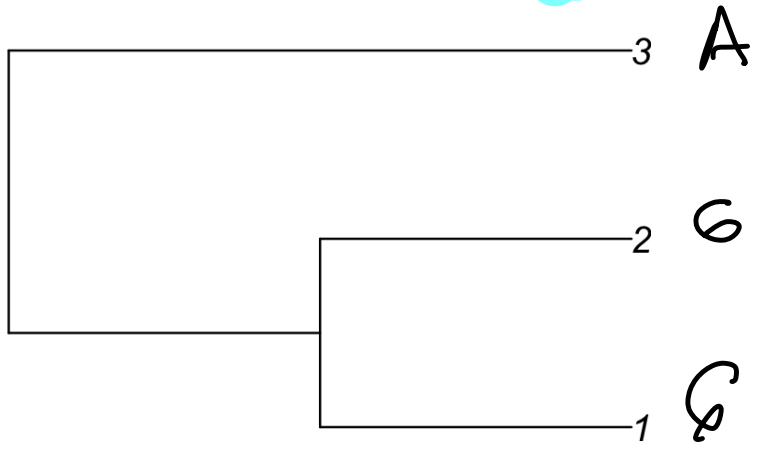
$$\max_{k=1,2,3} (\mathcal{L}(\text{Data; Model (P)}, \psi_k))$$

→ check now on multiple trees

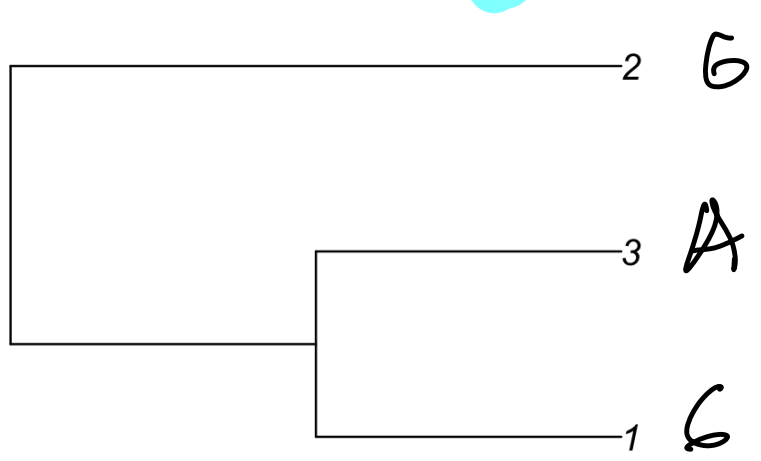
$\mathcal{L}(\text{Data; Model (P)}, \psi_1)$



$\mathcal{L}(\text{Data; Model (P)}, \psi_2)$



$\mathcal{L}(\text{Data; Model (P)}, \psi_3)$



What is changing in this likelihood?

$$P = \begin{bmatrix} 5/6 & 1/6 \\ 1/6 & 5/6 \end{bmatrix}$$

$$P_0(A) = p_0(G) = \frac{1}{2}$$

Inferring a phylogeny using the likelihood function

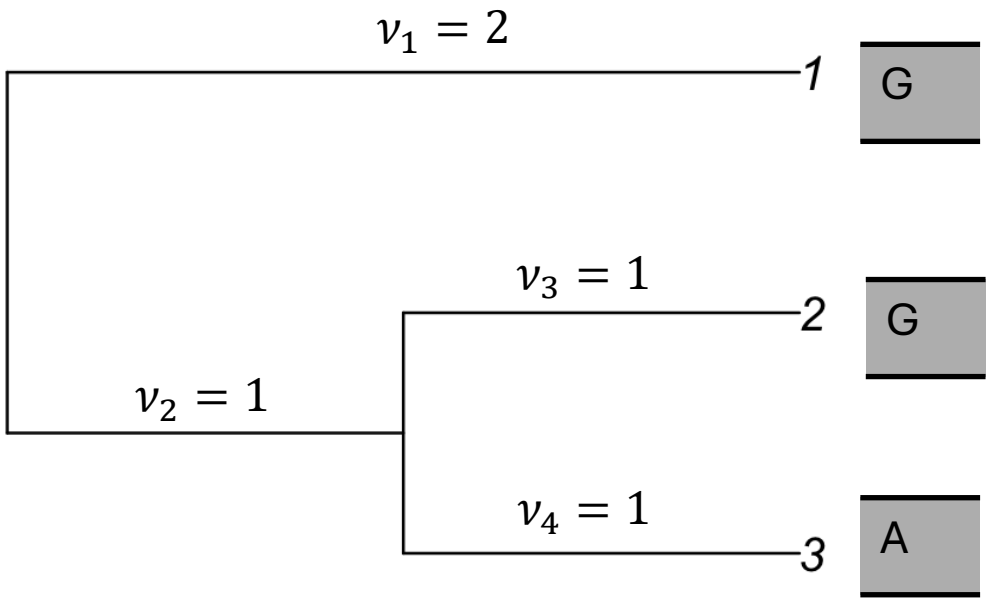
Model for molecular evolution

A mechanism to propose trees

Assumption about branch lengths

Exercise:

Calculate the probability of this site given this phylogenetic tree when the model changes by branch length.



$$P(1) = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix}$$

$$P(2) = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$$