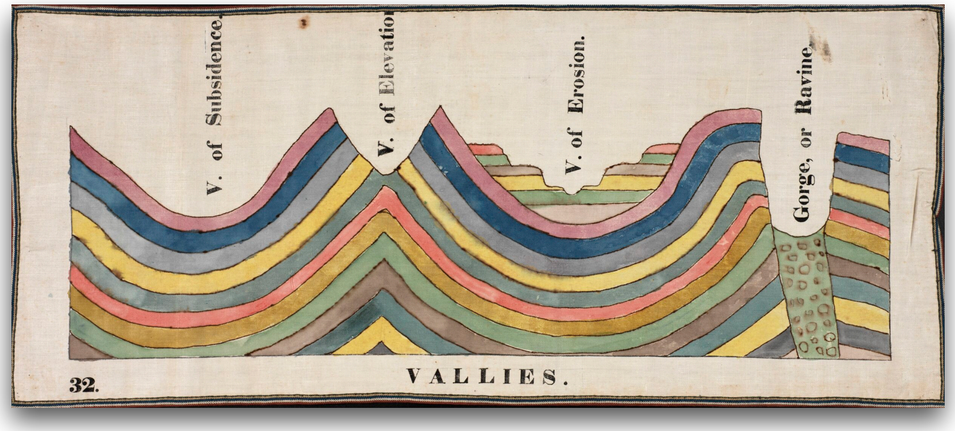
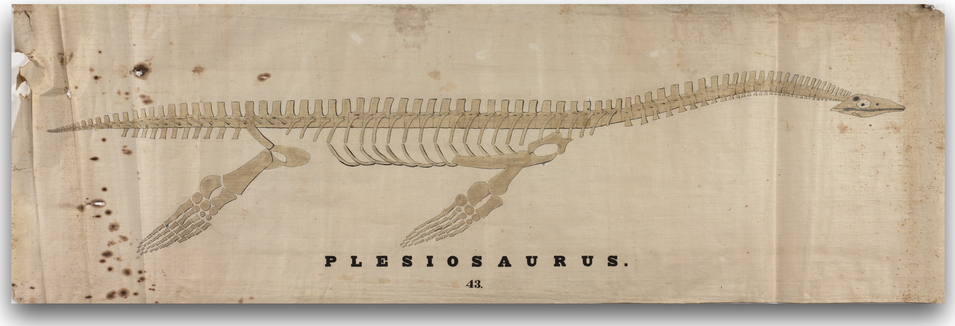
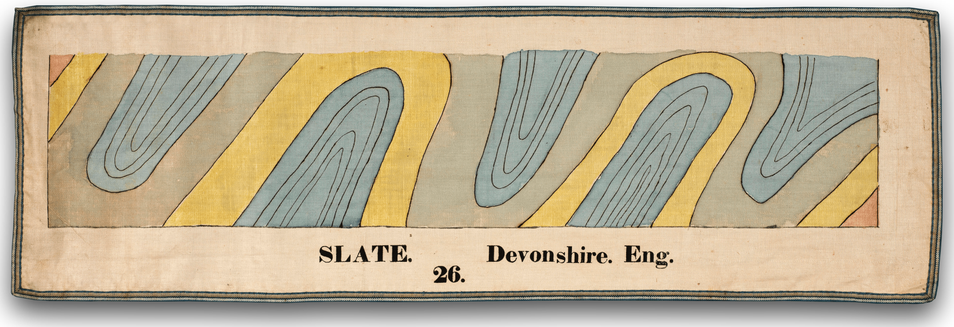






Phylogenetics 101

part A - tree thinking

3
1
2

Synapomorphies (Shared Derived Traits)

Synapomorphies are evolutionary traits that are **shared by two or more taxa** and inherited from their most recent common ancestor. These traits define **monophyletic groups (clades)** and provide strong evidence of evolutionary relationships. **Example:** The presence of feathers in birds is a synapomorphy that distinguishes them from other reptiles.

Autapomorphies (Unique Derived Traits)

Autapomorphies traits **unique to a single taxon** and not shared with any other group. While they help in distinguishing individual lineages, they do not provide information about shared ancestry among multiple taxa. **Example:** The elongated neck of giraffes (*Giraffa camelopardalis*) is an autapomorphy, as it is not found in closely related species.

Symplesiomorphies (Shared Ancestral Traits)

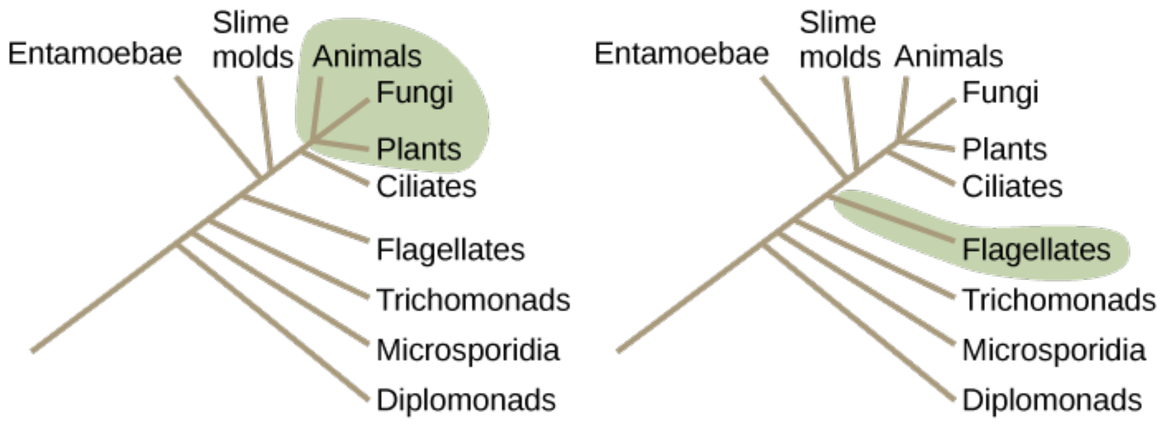
Symplesiomorphies are **ancestral traits shared by multiple taxa** that were inherited from a common ancestor predating the focal clade. These traits are misleading for phylogenetic classification, as they do not indicate close relationships among the taxa that possess them. **Example:** four limbs in mammals, reptiles, and amphibians is a symplesiomorphy inherited from the common ancestor of tetrapods.

Plesiomorphies (Ancestral Traits)

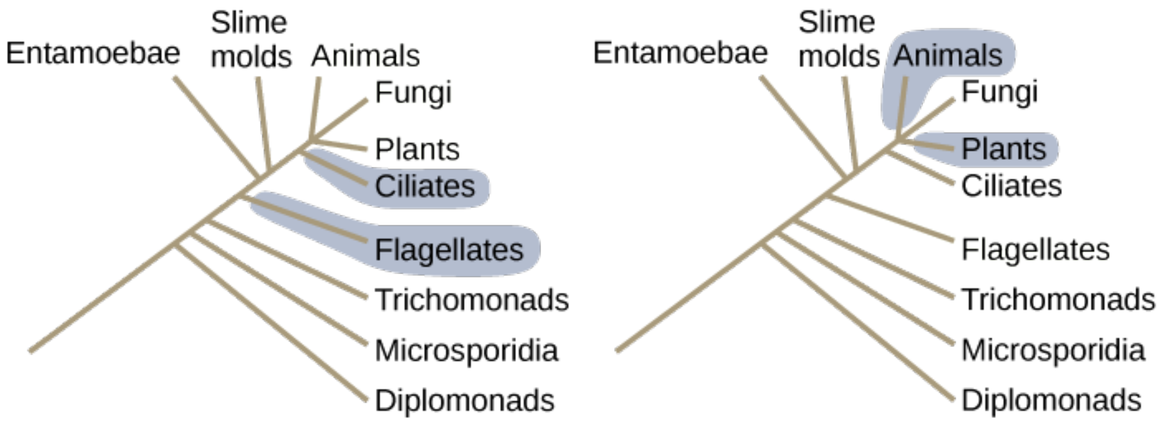
Plesiomorphic traits **originated in a distant ancestor** and are retained by multiple taxa, but they do not define a specific clade. It is a symplesiomorphy discussed in relation to a more derived state. **Example:** vertebrae in mammals is a plesiomorphy because it was inherited from early vertebrates and is also found in fish, amphibians, and reptiles.

Synapomorphies define clades

Clades



Not Clades



Cladistics? Phylogenetics? Systematics? 😱

- **Systematics** the broadest field ≈ including taxonomy (identification and classification) and phylogenetics (study of evolutionary relationships).

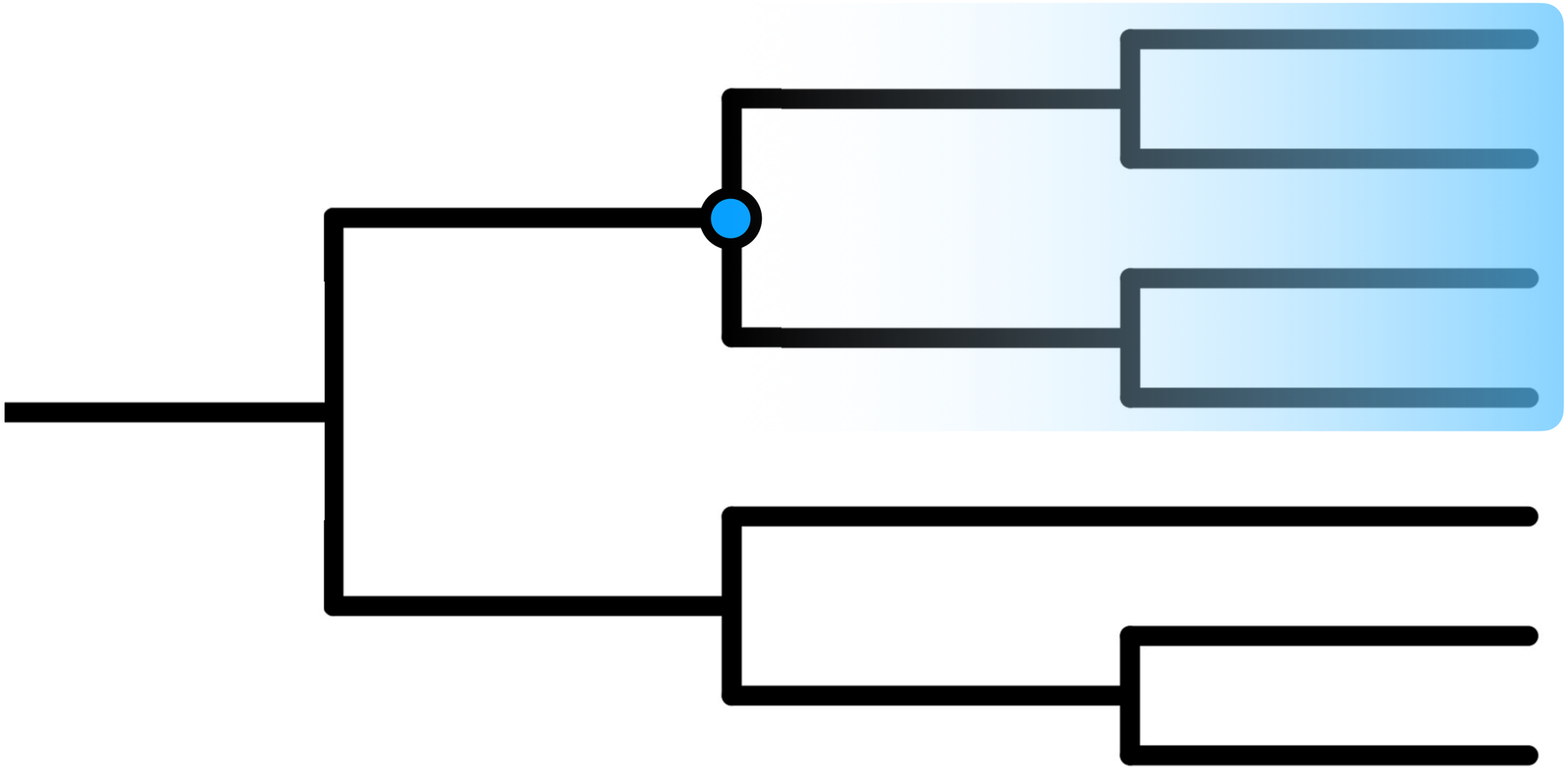
- **Taxonomy** is theory and practice of identifying, describing, naming, and classifying organisms.

- **Phylogenetics** is a subfield within systematics, specifically focused on reconstructing the evolutionary history and relationships among species.

- **Cladistics** is a specific approach within phylogenetics that focuses on common ancestry and shared derived traits.

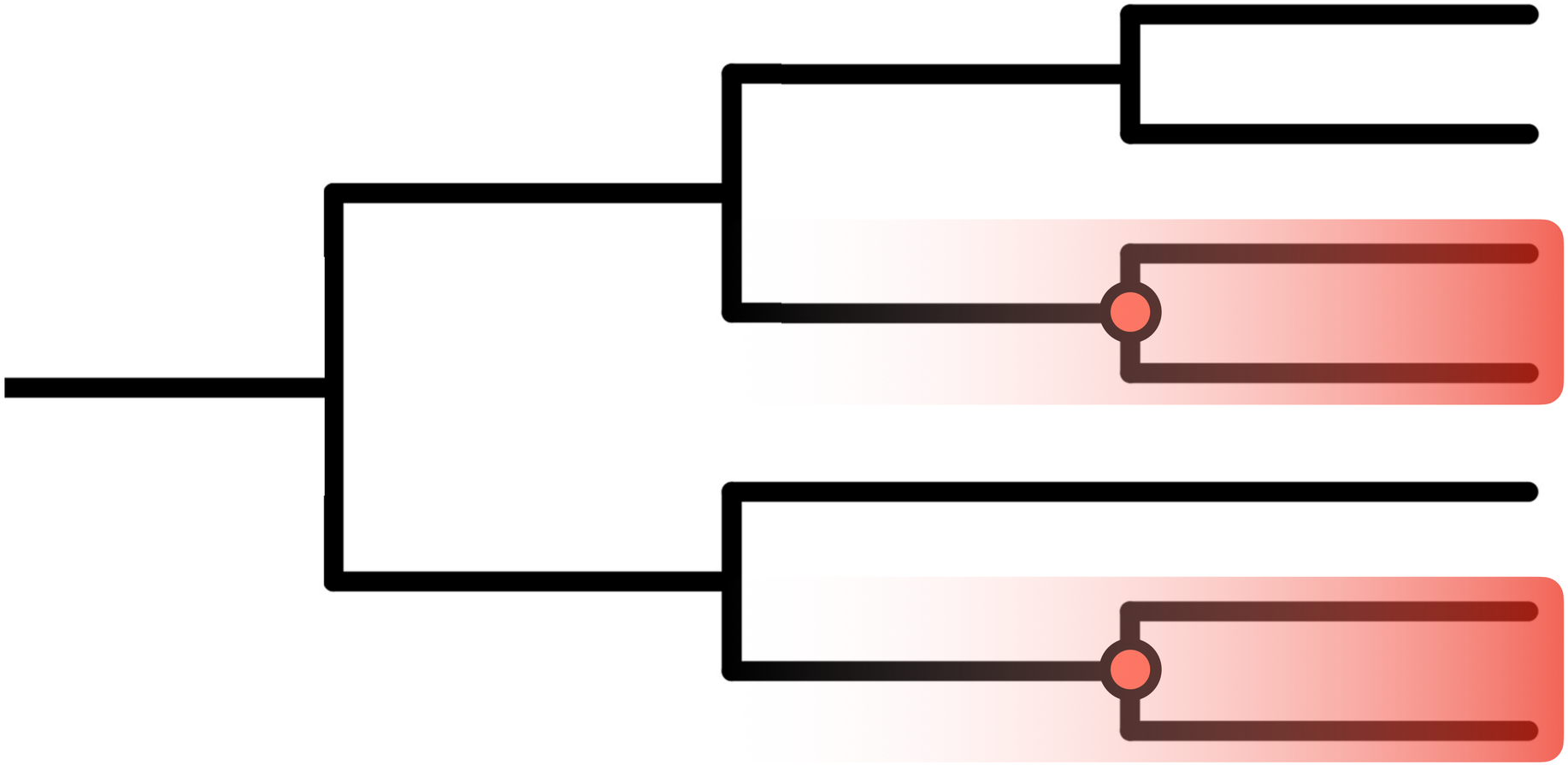
So, in a way … **Systematics > Phylogenetics > Cladistics**

Cladistics is an approch within phylogenetics, which is part of systematics. All cladistics is phylogenetics, but not all phylogenetics is cladistics.



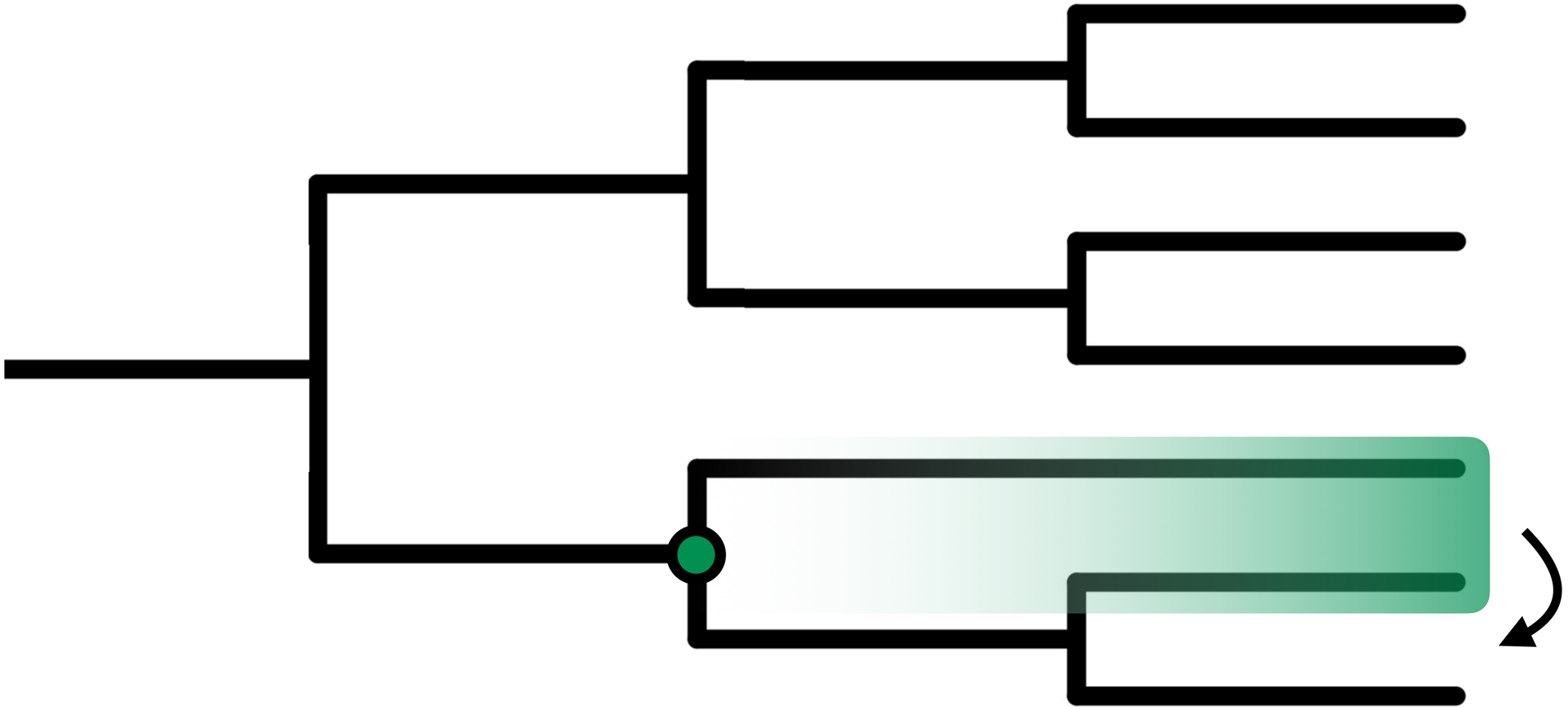
monophily

A grouping of organisms which contains: **(1)** its own most recent common ancestor, i.e. excludes non-descendants of that common ancestor; **(2)** all the descendants of that common ancestor, without exception. The condition of a taxonomic grouping being a **clade**.



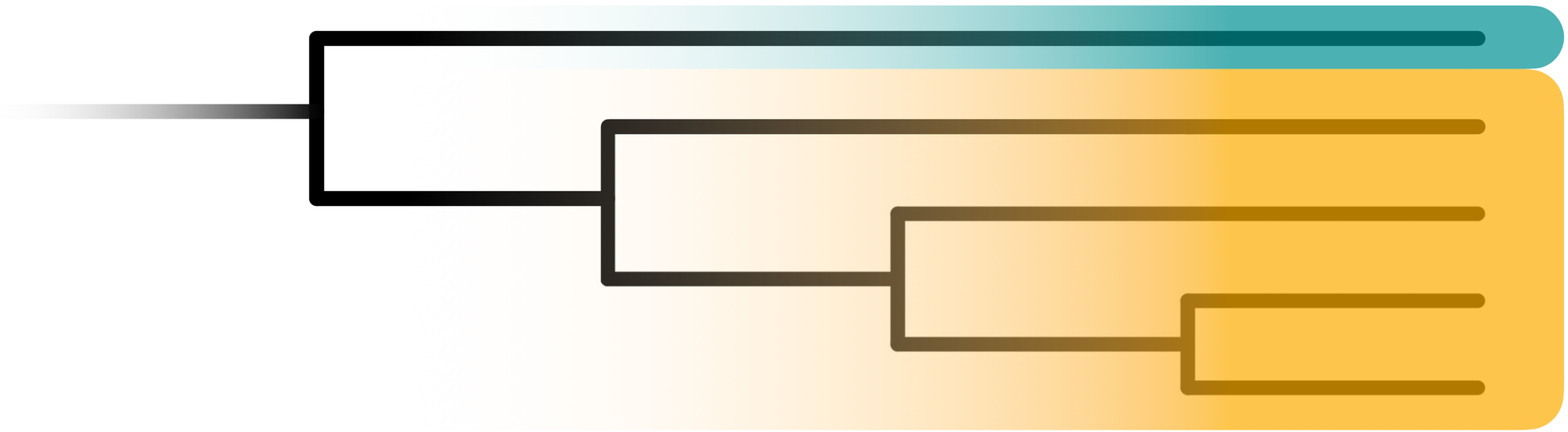
polyfily

A polyphyletic group is an assemblage that includes two or more **separate groups**, each with a **separate common ancestor**. The most recent common ancestor of the species in the polyphyletic assemblage would be the ancestor also of species not included in them.



parafyly

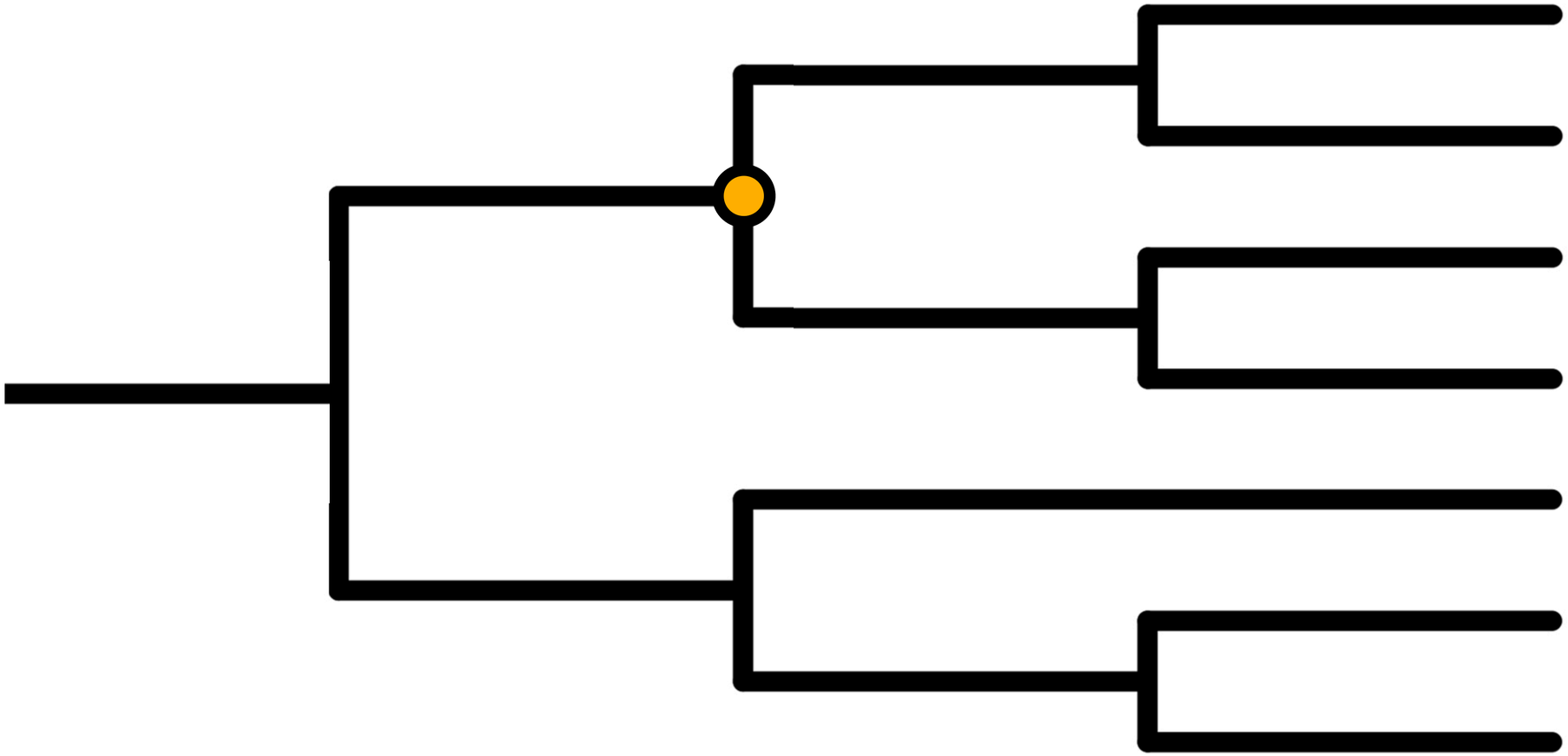
A group that consists of the grouping's **last common ancestor and some but not all of its descendant** lineages. The grouping is said to be paraphyletic with respect to the excluded subgroups.



ingroup and outgroup

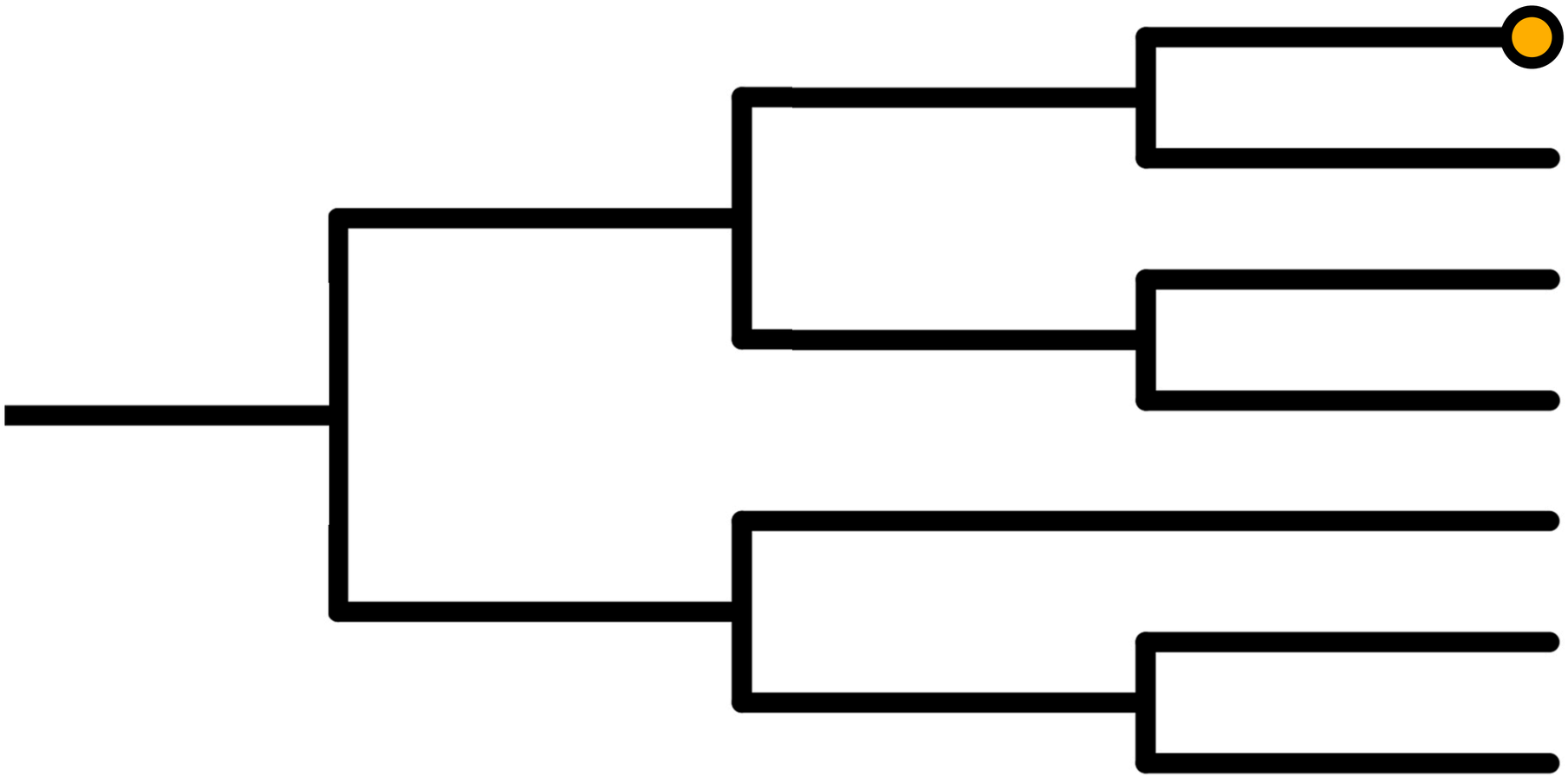
- **ingroup**: a group of tips assumed a priori to represent a clade, and the focus of our phylogenetic analyses.

- **outgroup**: a group of tips assumed a priori to lie outside the monophyly of the focal clade. It also serves to give a direction to our tree.



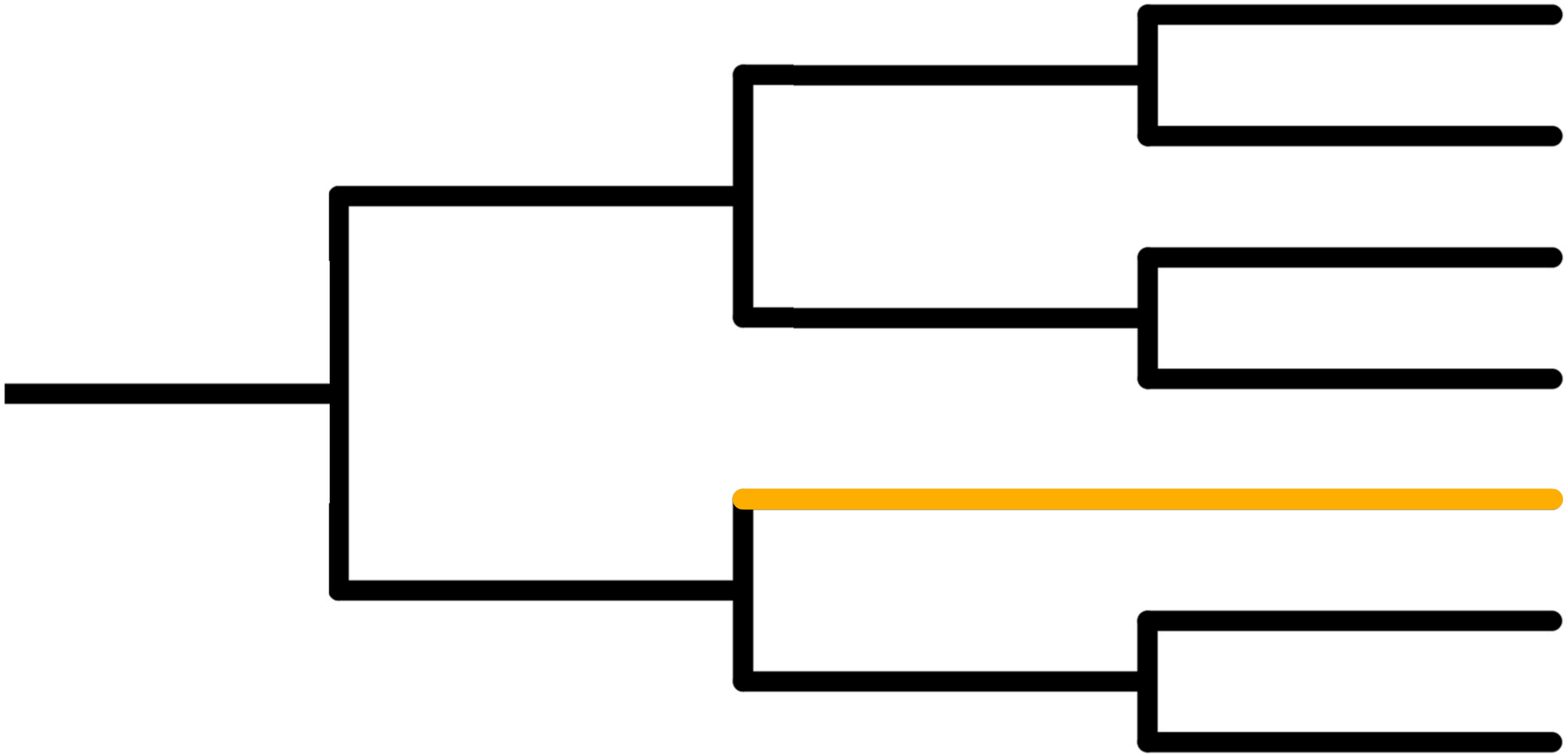
node

Nodes are the points at the ends of branches which represent **real sequences** or **hypothetical sequences** at various points in evolutionary history. In a tree of species an internal node in a tree is, biologically, an **ancestor**. You can also thin to nodes as **bipartitions or splits**.



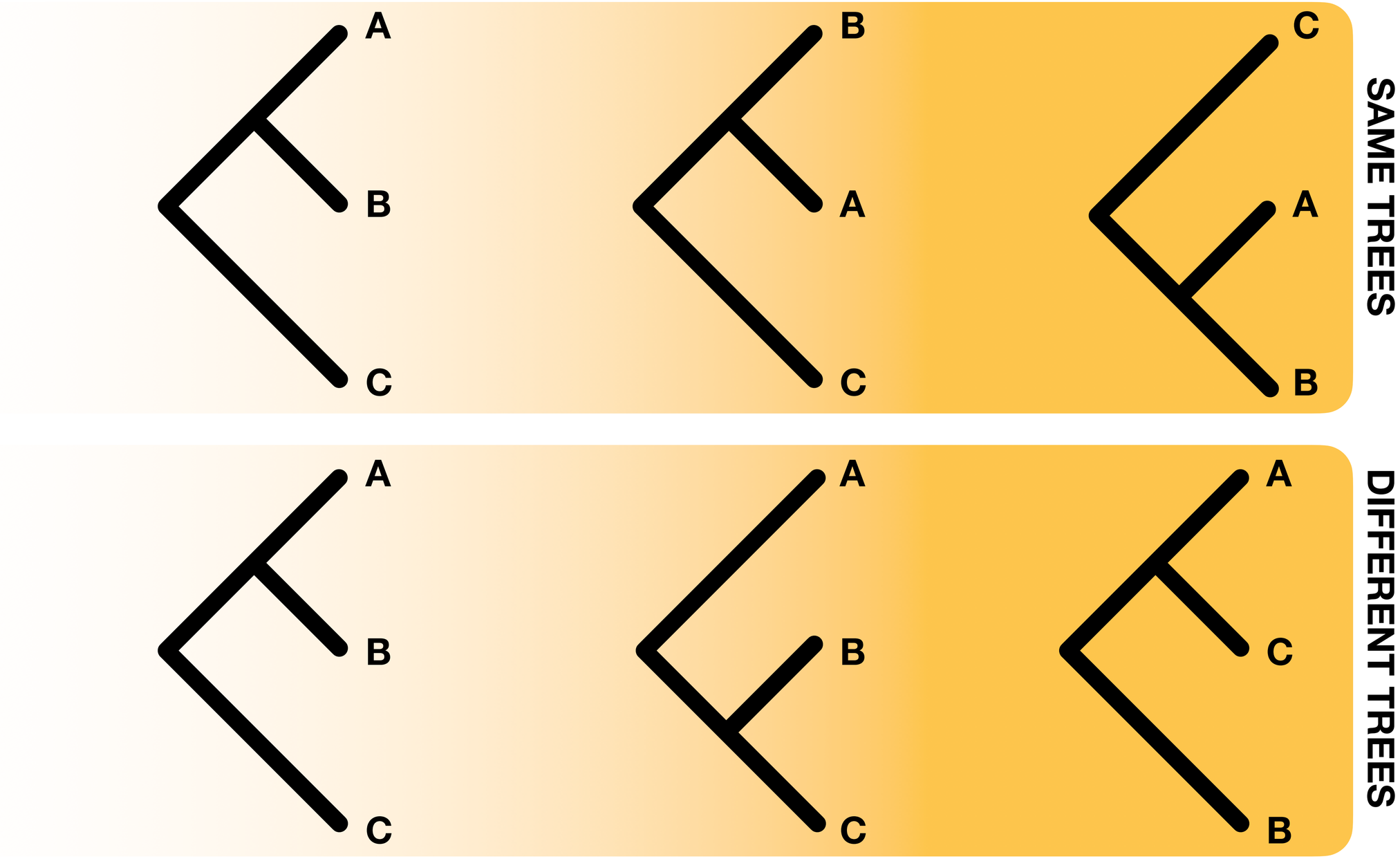
nodes (terminal or external)

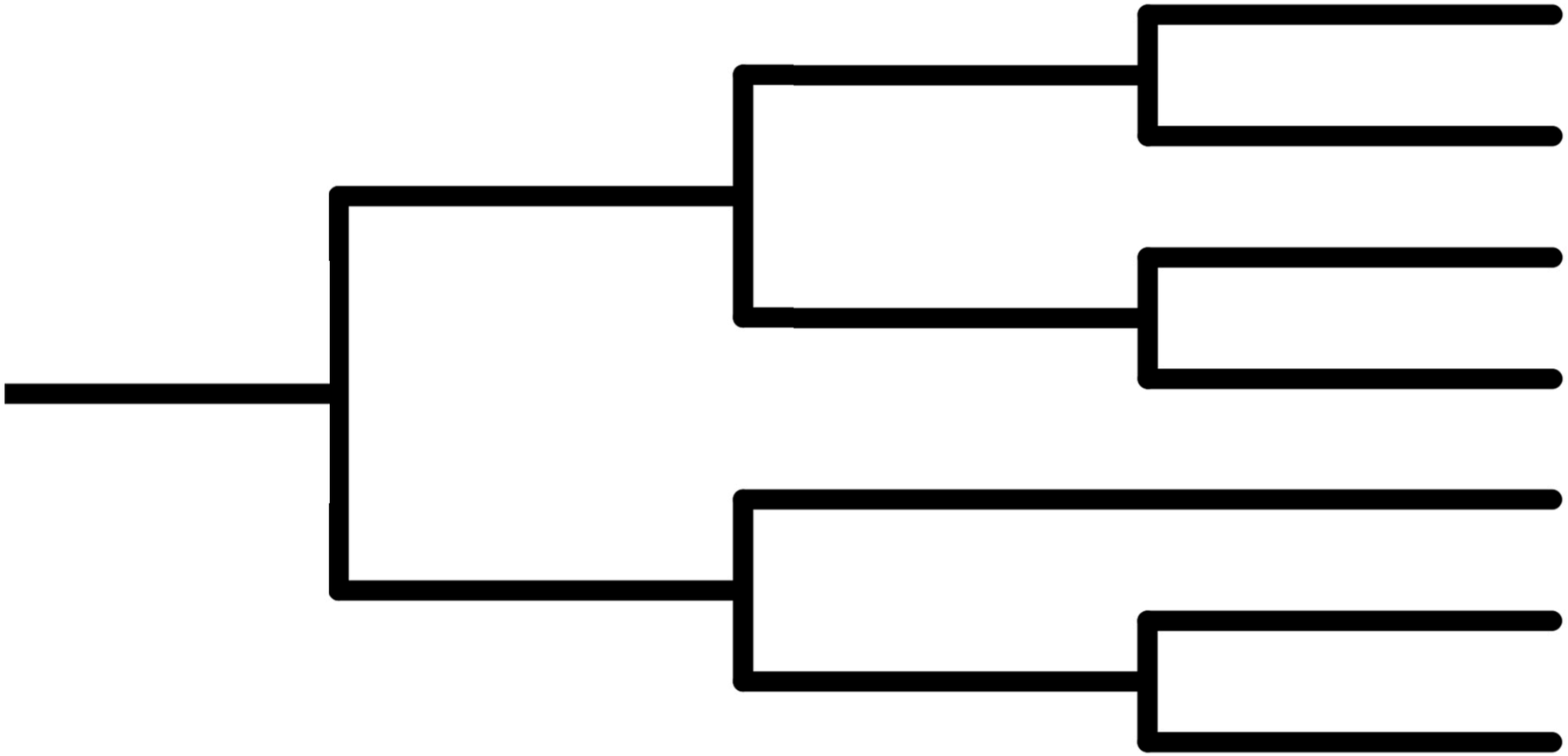
The sequence that we sampled and used to construct our phylogeny occur on single terminal branches, known as **tips or leafs**. Often referred to as **OTU**, which stands for Operational Taxonomic Unit, a noncommittal term used for the objects of study (be they species, populations or individuals).



branch

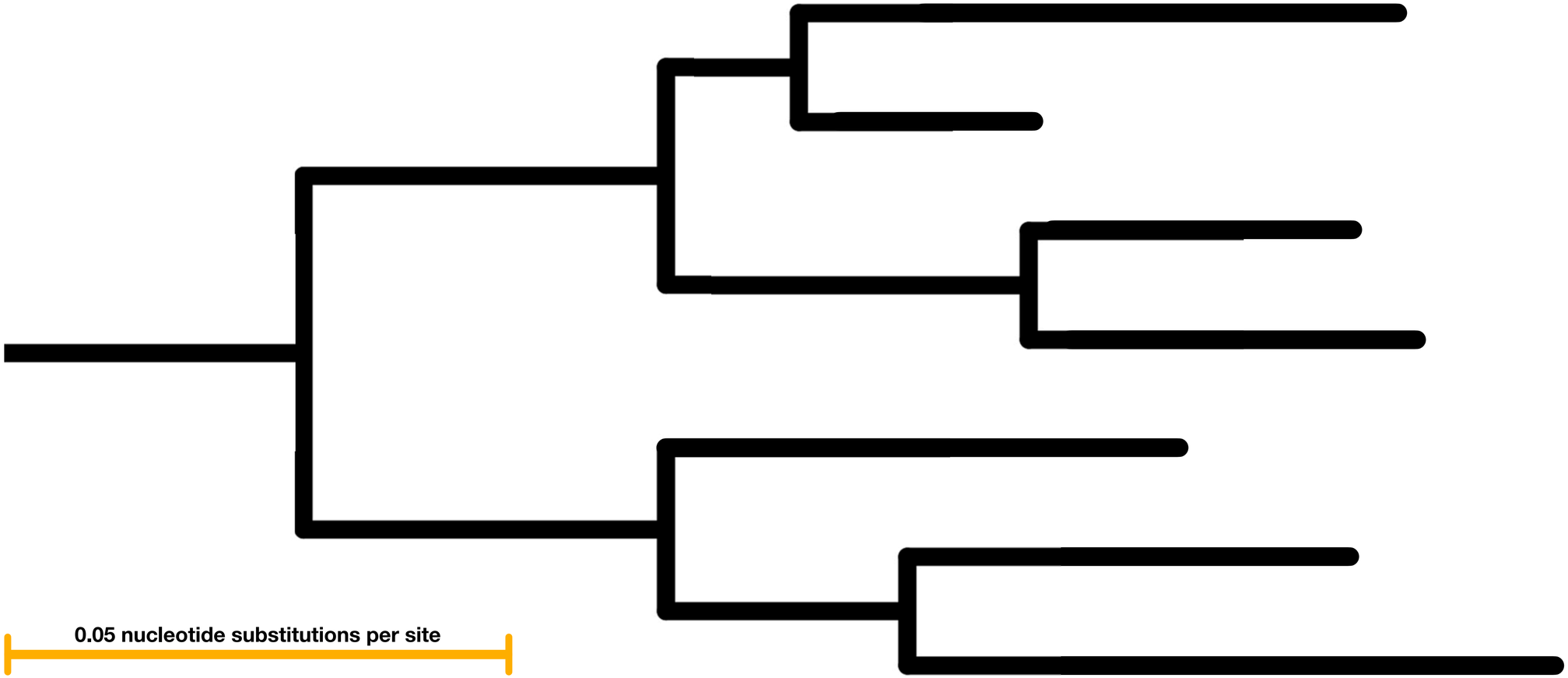
The branching structure of the tree is its **topology**. It is of particular significance because it indicates **patterns of relatedness** among taxa. Beware: any internal node can be rotated and the tree is the same.


SAME TREES
A
B
C
B
A
C
C
A
B
DIFFERENT TREES
A
B
C
A
B
C
A
C
B



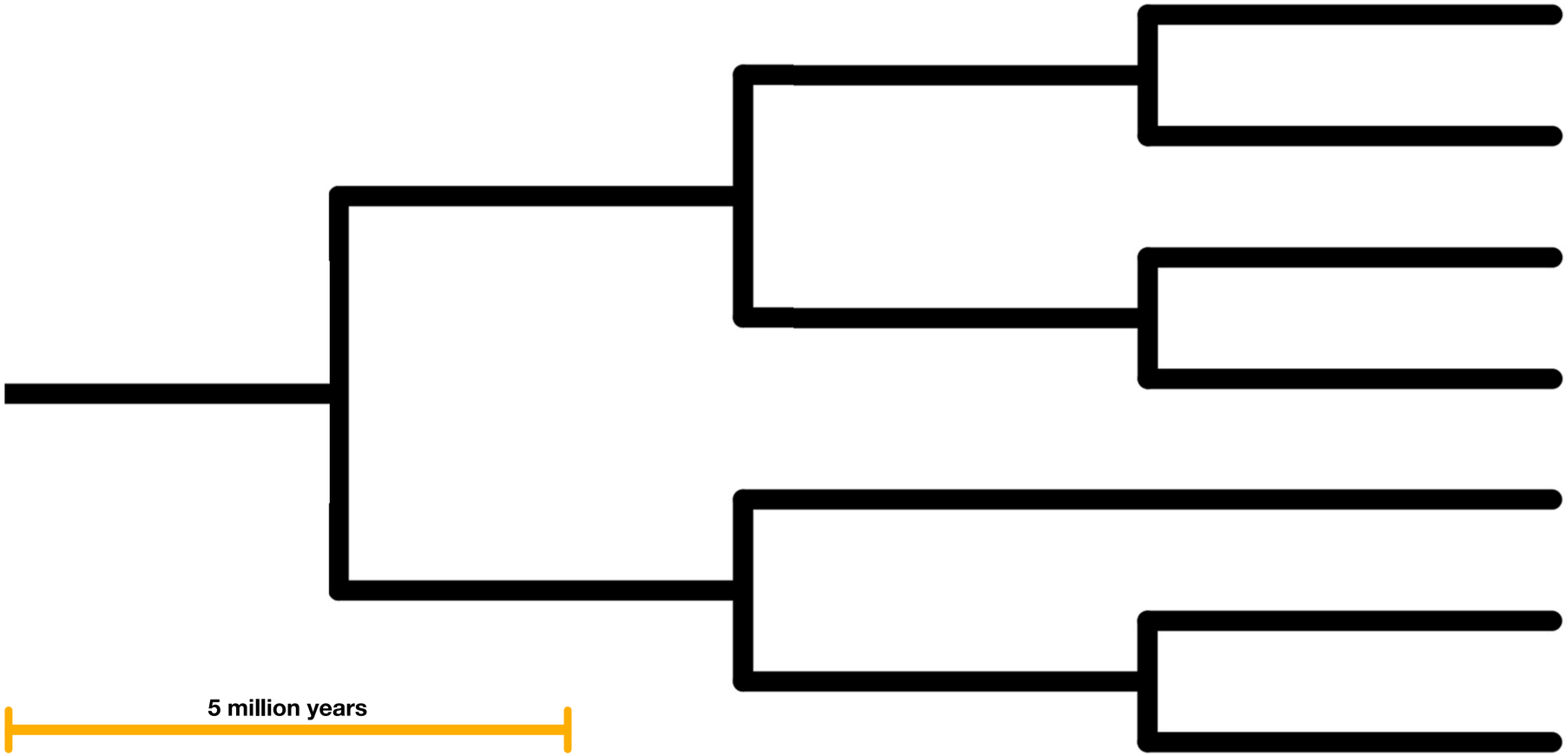
cladogram

A cladogram only represents a branching pattern; **branch lengths do not have a meaning** and do not represent either time or relative amount of character change. They are just a function of the order of branching on the tree.



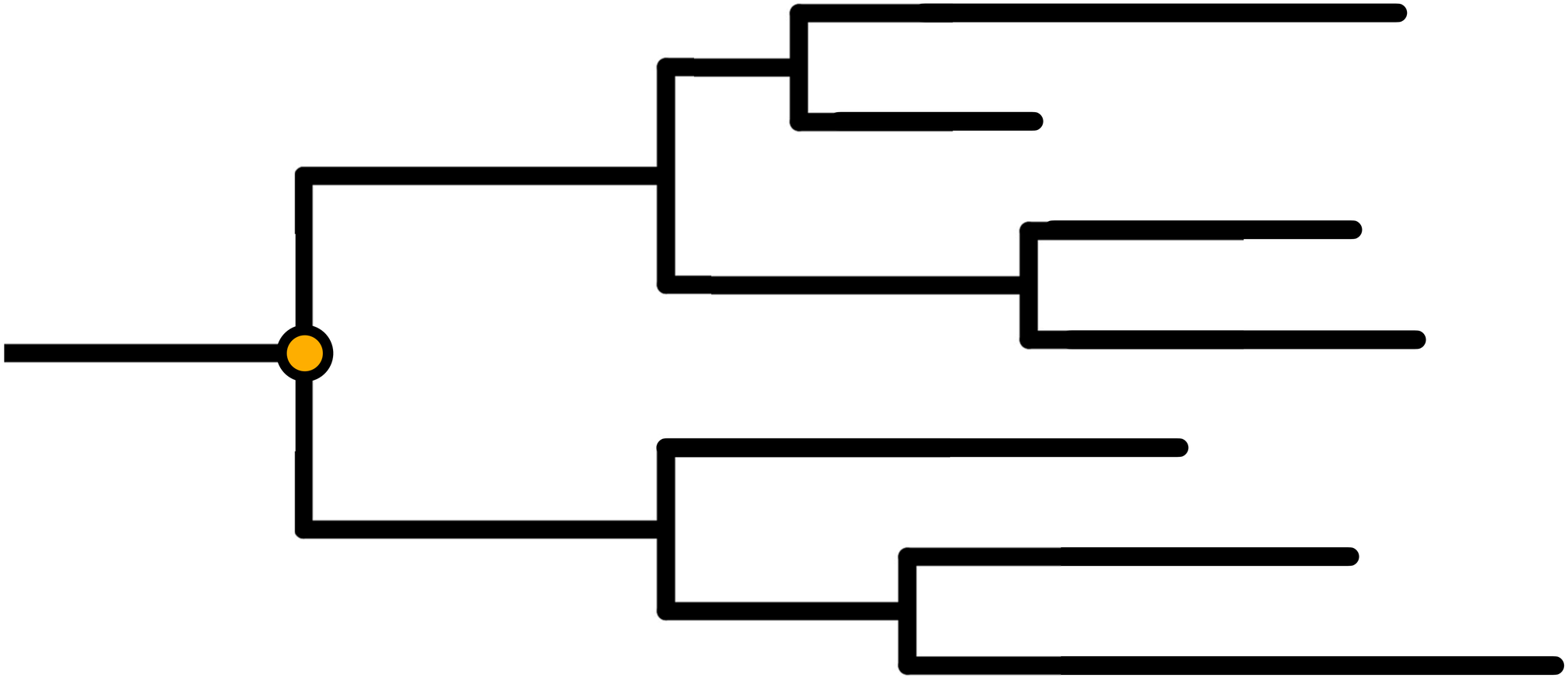
scale: evolutionary change

The line segment shows the length of **branch that represents an amount of change** of 0.05. The vertical dimension in this figure has no meaning and is used simply to lay out the tree visually with the labels evenly spaced vertically. This tree is a **phylogram**.



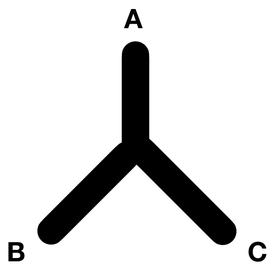
scale: time

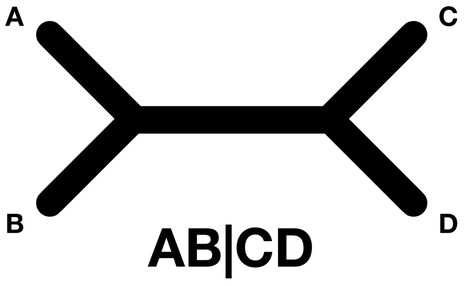
A chronogram is a phylogenetic tree that explicitly represents **time through branch lengths**. This tree is a **chronogram** or **timetree**. The process of obtaining such a phylogeny is often called a **divergence time analysis**.

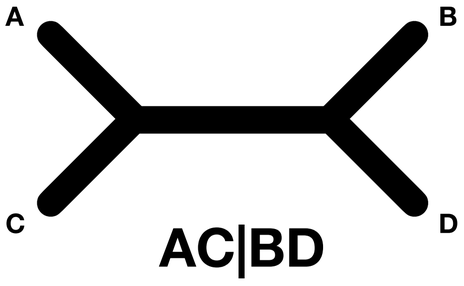


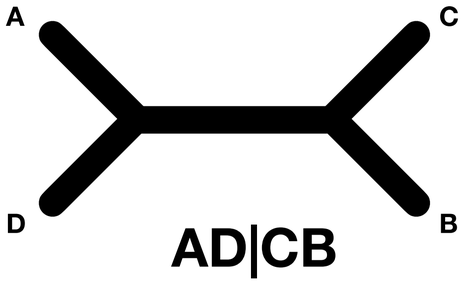
root

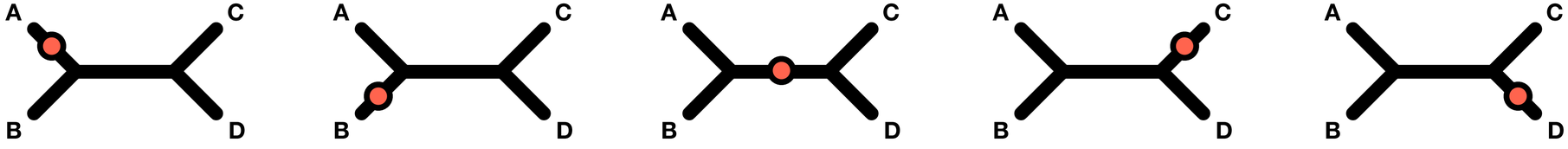
The root is a specific internal node representing **the most recent common ancestor** of all tips in the tree. It is therefore the oldest part of the tree and tells us the direction of evolution. There are two main approaches that we can use to root a tree: **outgroup rooting** and **midpoint rooting**.





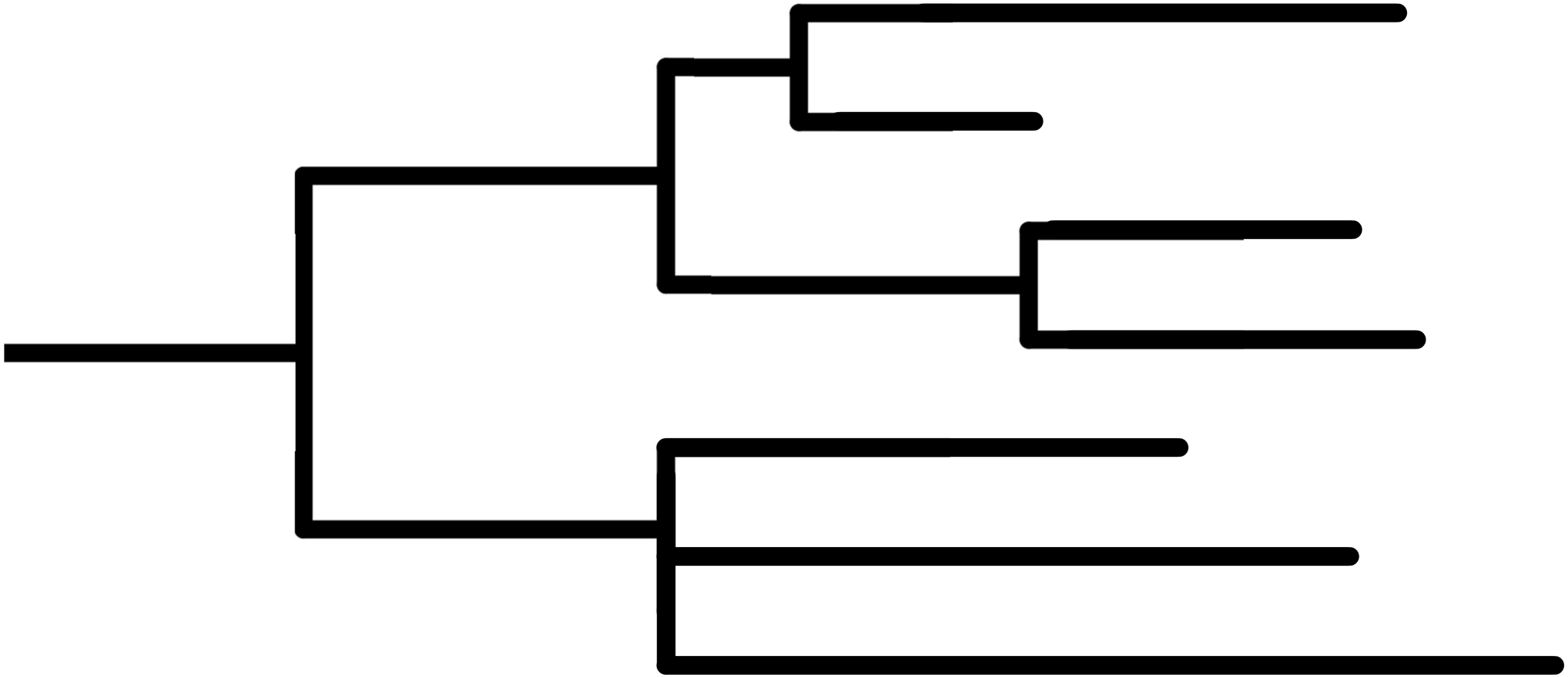






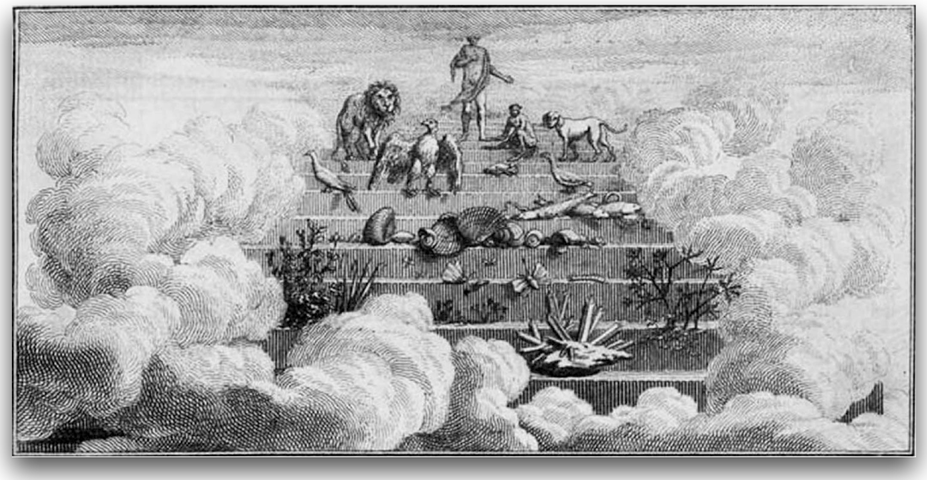
root?

The simplest ambiguous tree is a **quartet**, the unambiguous one a triplet. Trees may be rooted or unrooted. In a rooted tree, each node with some descendants represents the most recent common ancestor of those descendants. In the practical we will see that complexity grows quickly 😱.



polytomy

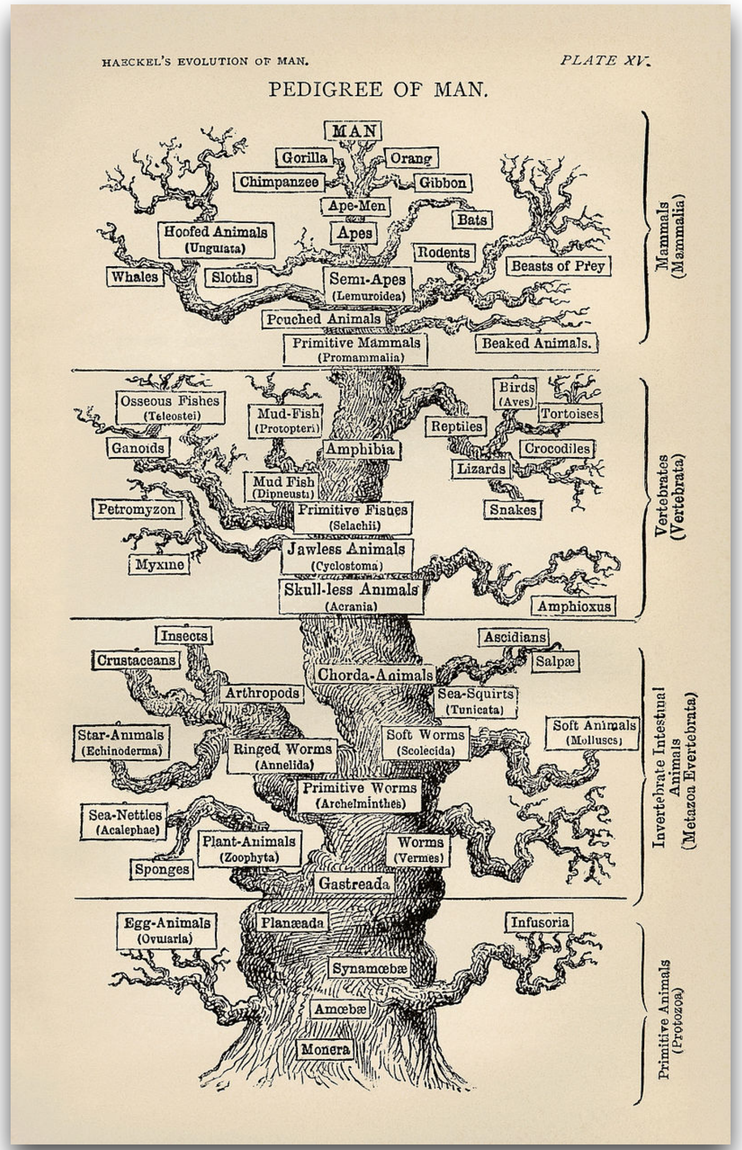
Trees can either be **bifurcating** (fully dicototmous) or **multifurcating** (containing a politomy). A polytomy is a node in a phylogenetic tree where **more than two descendant lineages** emerge, indicating an unresolved evolutionary relationship.



Kette der Wesen, Scala Naturae by Charles Bonnet (1781).

trees are not ladders

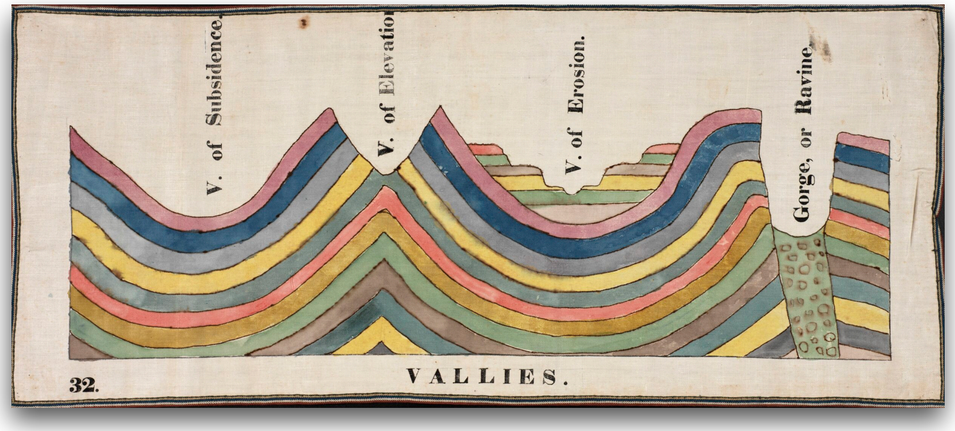
The ladder concept (Scala Naturae) is an old view that tend to organize ancestry based on complexity or some putative evolutionary superiority. Terms as “primitive”, “ancient” or “lower” referred to a lineage are misleading as they imply ladder-thinking 😑.

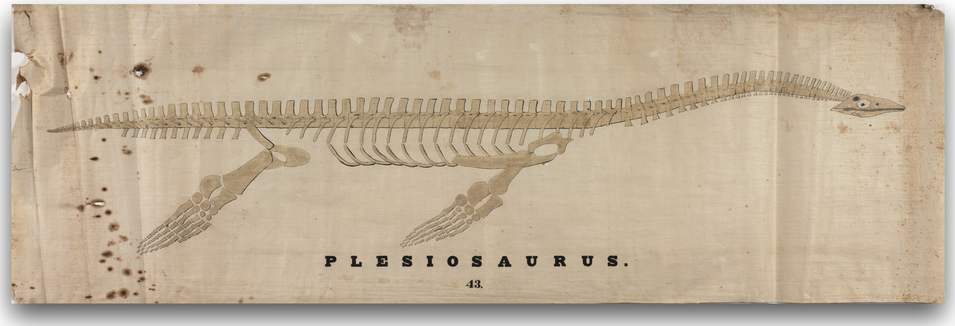


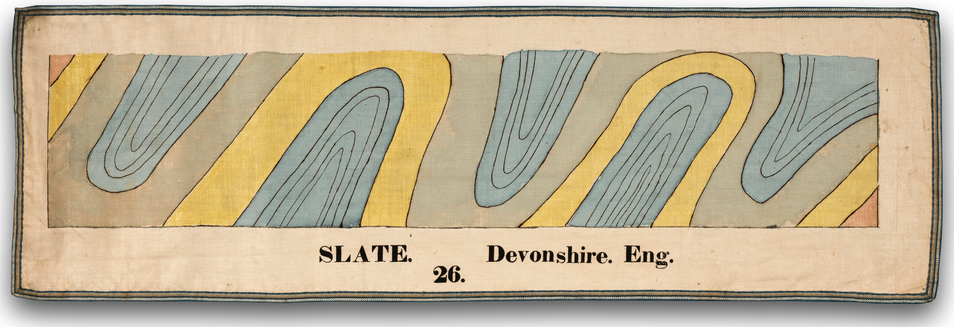
Ernst Haeckel's tree from the The Evolution of Man (1879).

IMPORTANT

- A phylogenetic tree is always an **hypothesis** about how species are related through evolution.
- Evolution describes events occurring in the **past** and that cannot be directly observed.
- Phylogenetic trees are **inferred** from data and are subject to revision as new evidence emerges.







Phylogenetics 101

part B - molecular phylogenetics

Phylogenetics using any kind of character

- **Morphological characters**: bones, teeth, floral structures.
- **Molecular characters**: DNA, RNA, and protein sequences.
- **Behavioral traits**: mating behaviors, and ecological interactions.
- **Genomic structural features**: Gene order, synteny, and presence/absence of mobile elements.
- **Cultural and linguistic data**: Used in disciplines like anthropological phylogenetics.

… just to name a few ..

Which criteria should characters posses?

Homologous: Characters among taxa must be due to shared ancestry.

Non-homoplasious: While some level of homoplasy is often unavoidable, characters that are less prone to these phenomena are preferred because they better reflect true evolutionary relationships.

Heritabile: The trait or character must be genetically or developmentally inherited. This heritability is essential so that the character's distribution reflects historical descent.

Independent: Each character should vary independently from others. If characters are correlated or redundant, they may overemphasize certain signals and bias the analysis.

Informative: A good phylogenetic character should be neither completely invariant (providing no discriminatory power) nor so variable that the signal is overwhelmed by noise.

Clear: The character should be defined in a way that allows for unambiguous coding into discrete states (e.g., "present"/"absent" or multiple clearly defined categories).

Homoplasy:

Traits that appear similar but are not derived from a common ancestor are considered to result from homoplasy. This process attains similar outcomes **independent evolution.**

- Convergent evolution: independent evolution of similar trait
- Evolutionary reversals: reversal back to an ancestral character state

Impact on Phylogenetic Inference:

Homoplasious traits and usually, are **incongruent with each another**.

in the absence of homoplasy there would be no character conflict and nferring the phylogeny would be easy. Unlukily homoplasy can be common and **mislead the reconstruction of phylogenies**.

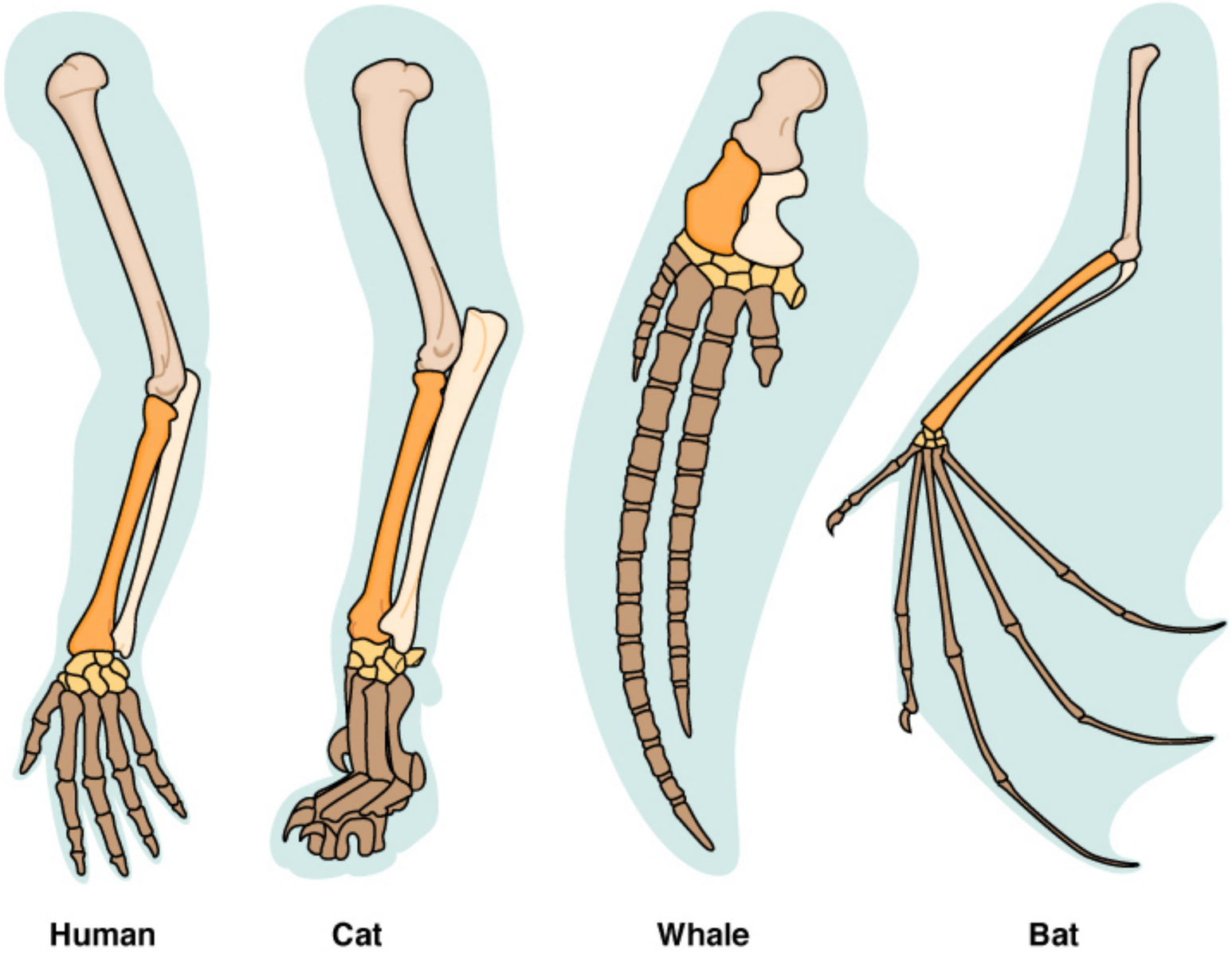
Homoplasious traits can be misleading in reconstruction of phylogenies.

- **Wings**: Birds, bats, and insects all have wings for flight, yet these structures evolved independently. While they serve the same function, the underlying anatomy and developmental origins differ.

- **Body streamlining in aquatic lifestyle:** Sharks (cartilaginous fish), dolphins (mammals) exhibit similar streamlined body shapes that help reduce drag in water. These similarities are convergent adaptations to an aquatic lifestyle.

- **Camera-Type Eyes:** Both cephalopods (like octopuses) and vertebrates have complex, image-forming eyes. Nonetheless these eyes have different anatomical structures.

- **Echolocation:** Although both bats and whales use echolocation for navigation and hunting, they evolved this ability indipendently.

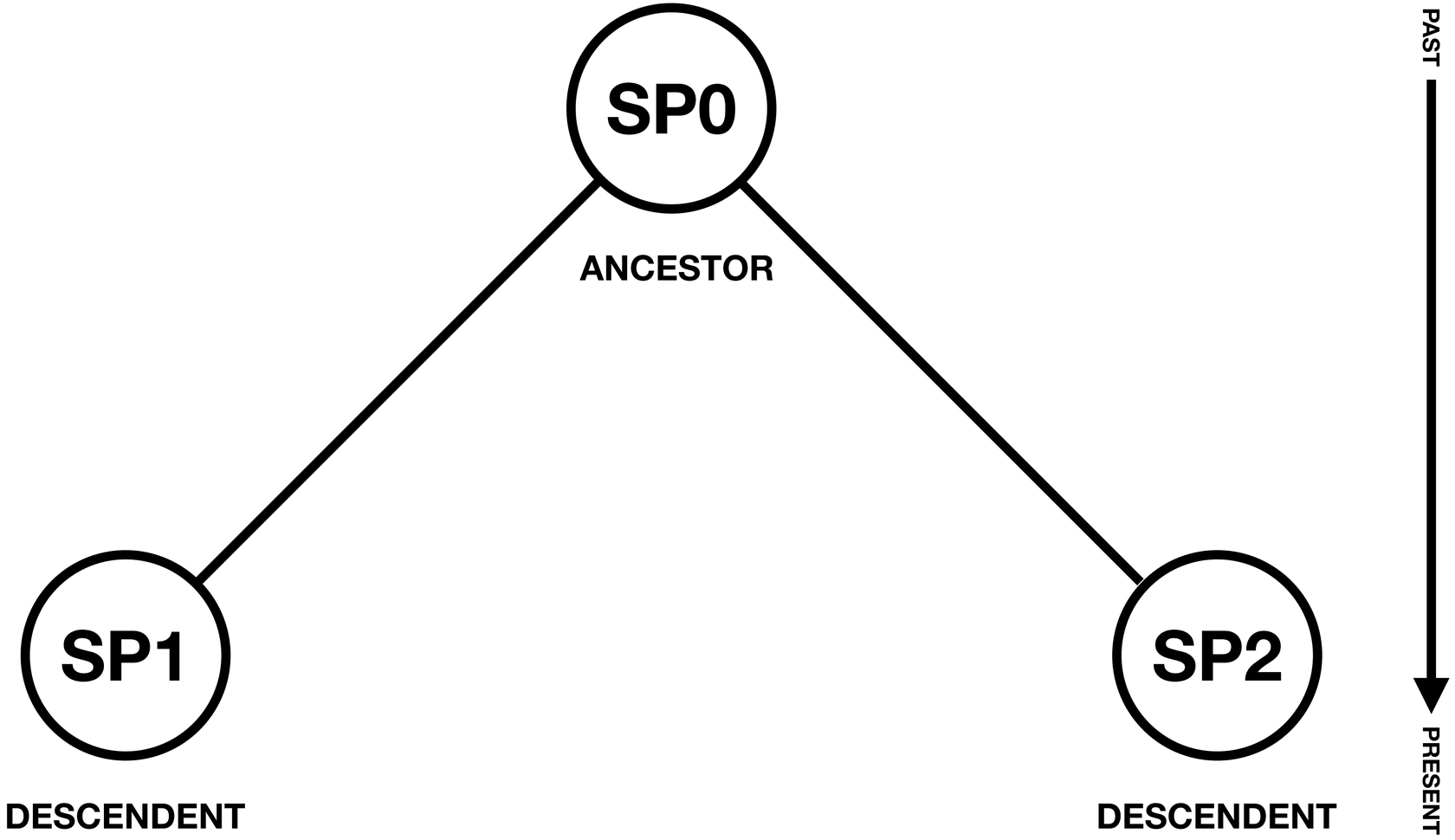

Human
Cat
Whale
Bat

Why molecular phylogenetics?

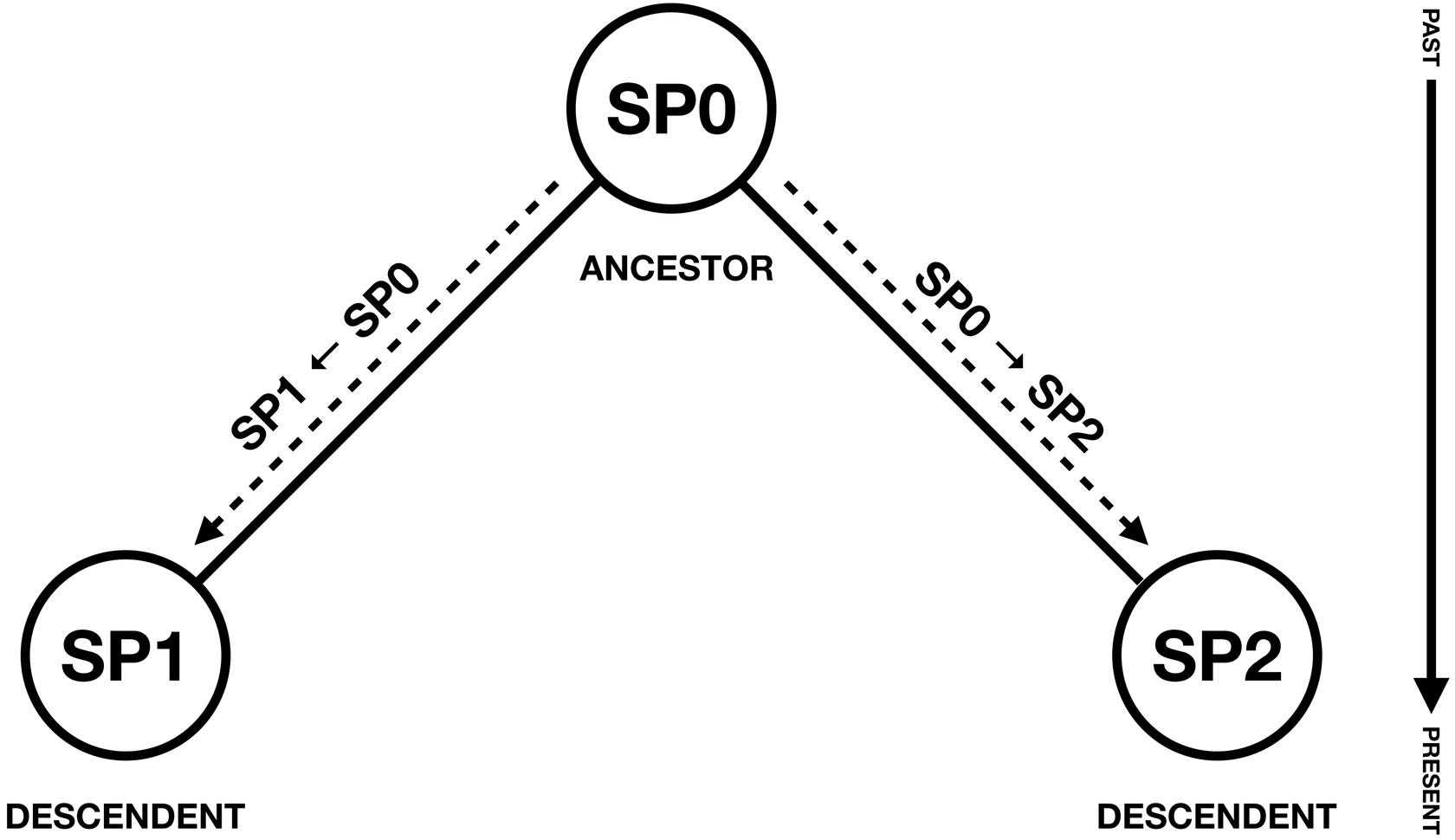
Today, most phylogenetics is done using molecular sequence data because:

- **Abundant and inexpehnsive:** sequences provide thousands to millions independent characters at a cheap price.
- **Objective:** molecular data is less subjective than morphological assessment.
- **Multiple resolutions:** different approcaches can resolve fine-scale and macroscopic differences relationships among closely related taxa.
- **Standardized:** computational tools and databases facilitate analyses and allow their reproducibility.

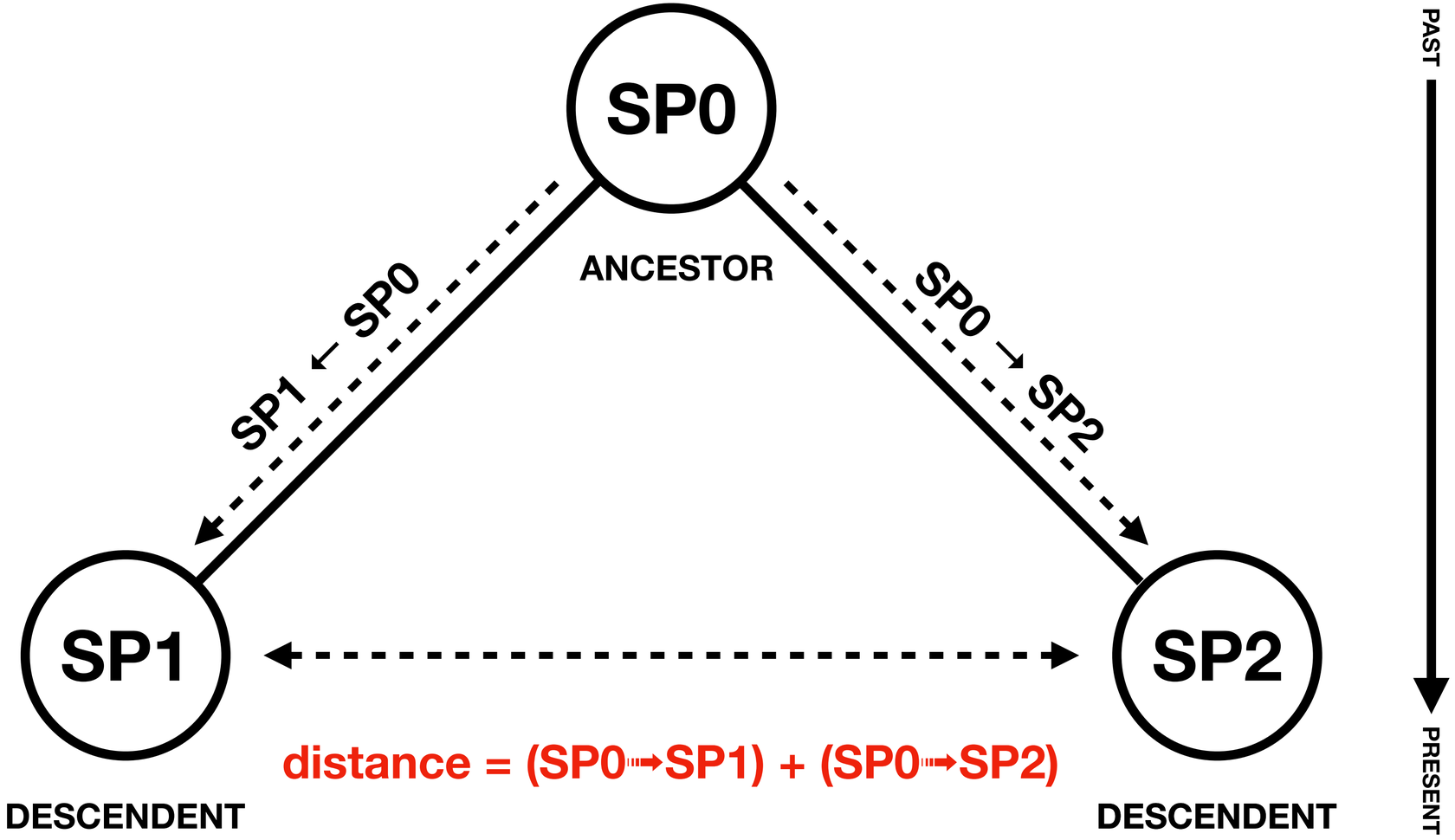
How to define the genetic distance between two taxa?



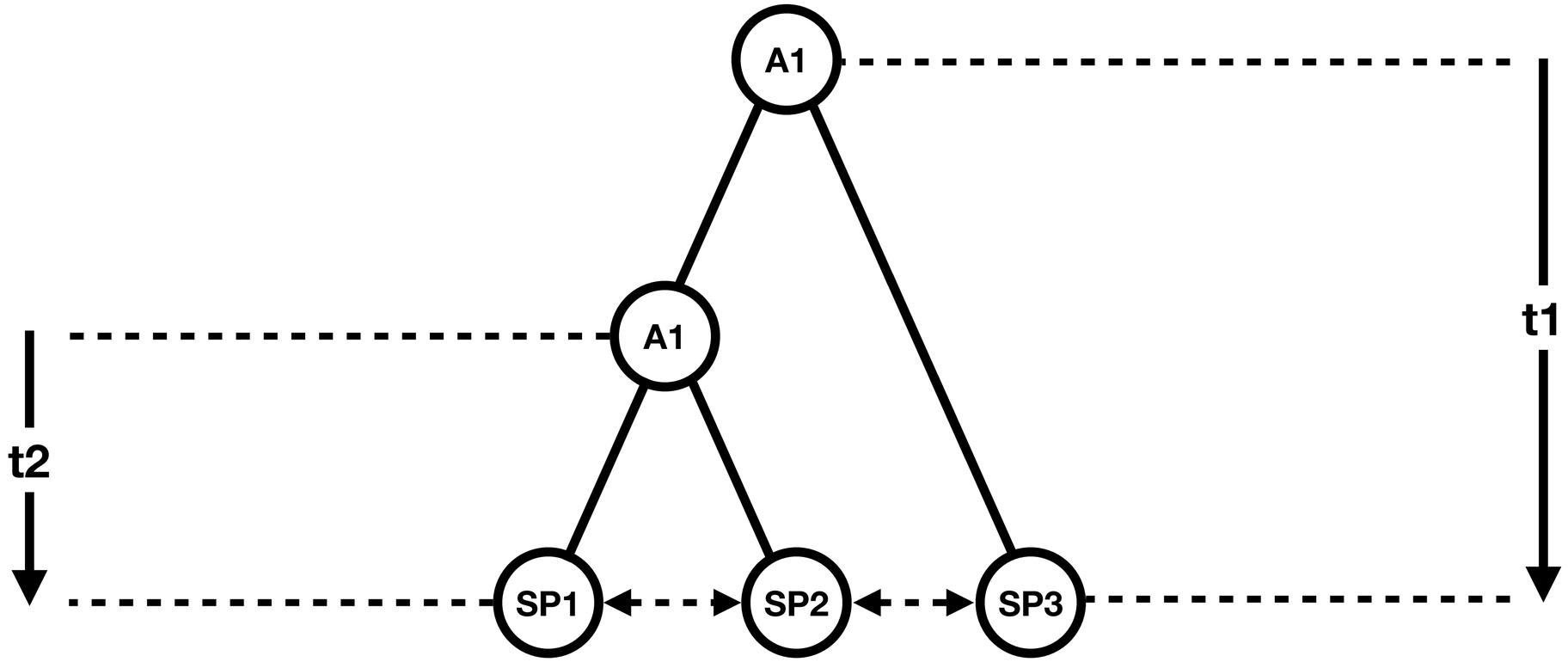
How to define the genetic distance between two taxa?



How to define the genetic distance between two taxa?



distance = (SP0 ➟ SP1) + (SP0 ➟ SP2)

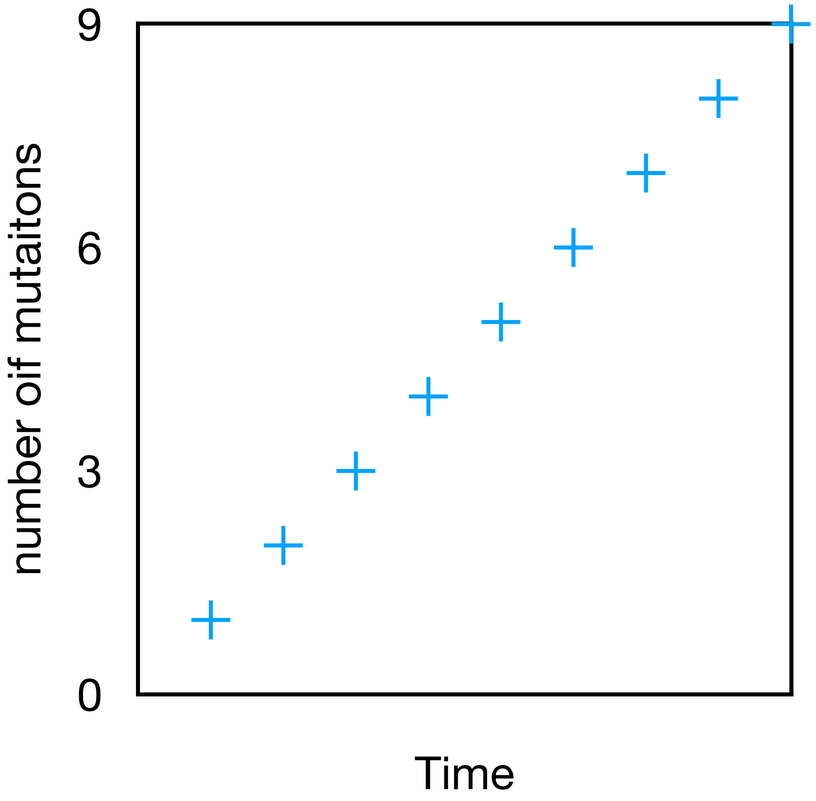


$t_1 > t_2$

$d_1 > d_2$

more ancient is the split
more time to accumulate changes
higher is the genetic distance

The molecular clock hypothesis suggests that **genetic changes** tend to accumulate at a relatively **constant rate** over long periods, which can help estimate divergence times between species.

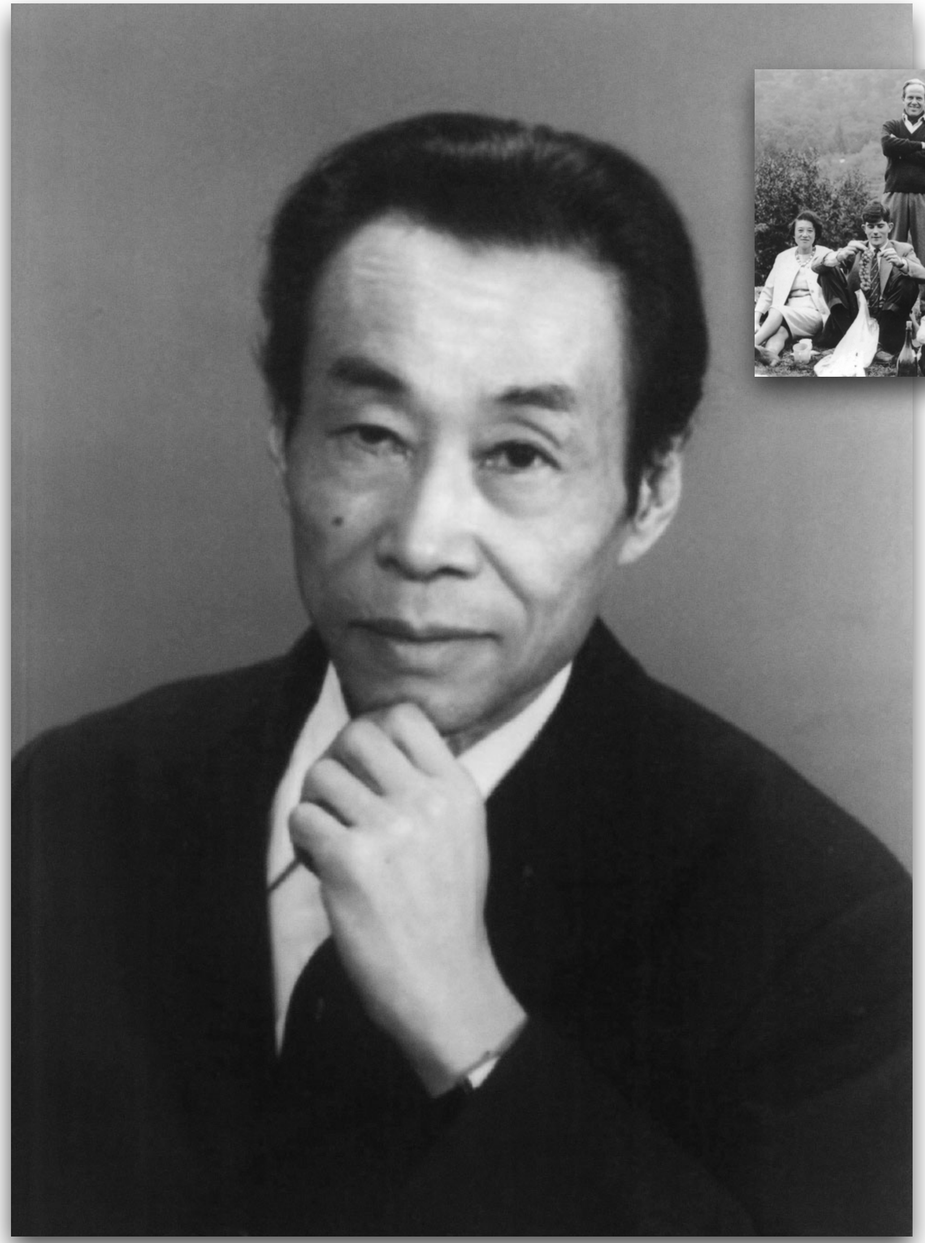


neutral theory & phylogenetics

Most molecular changes and within-species variability arise from random drift of neutral - or nearly neutral - mutant alleles, not from natural selection. Proposed by **Mooto Kimura**, this theory applies primarily to molecular evolution, while phenotypic traits are still shaped by natural selection (as *per* Darwin).

Implications for Phylogenetics:

- Since many molecular changes are neutral, they allow us to reconstruct evolutionary relationships.
- Since neutral mutations accumulate at a - relatively - constant rate over time, they allow us to estimate divergence times.



… however …

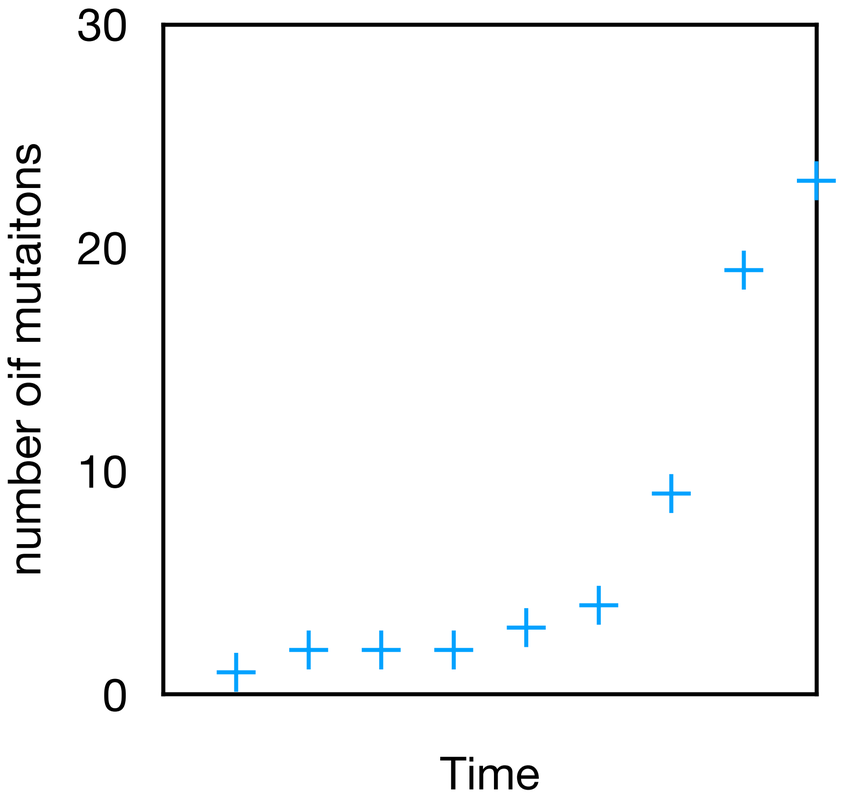
- **Mutation rates can vary:**
  Mutation rates can differ among genes, species, and even different parts of the genome due to factors like generation time, metabolic rate, and DNA repair mechanisms.

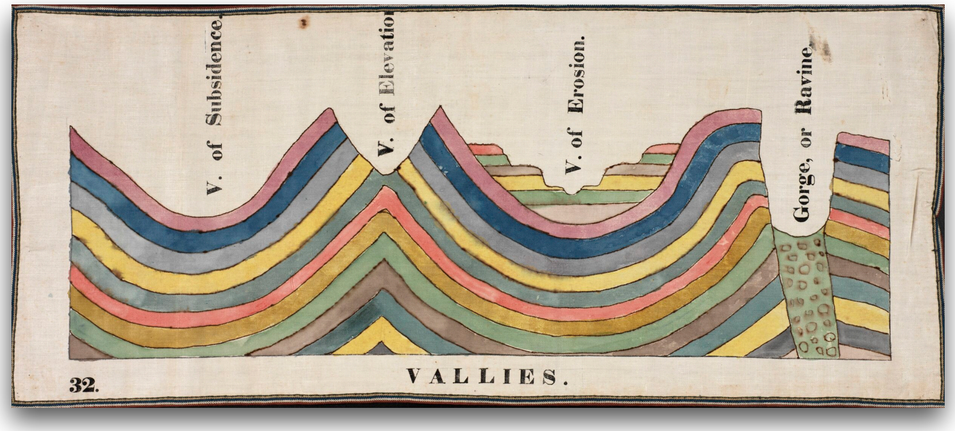
- **Selection can vary (across genes and genomes):**
  Some changes are subject to strong selective pressures and may be removed (purifying selection) or favored (positive selection), altering the rate of change.

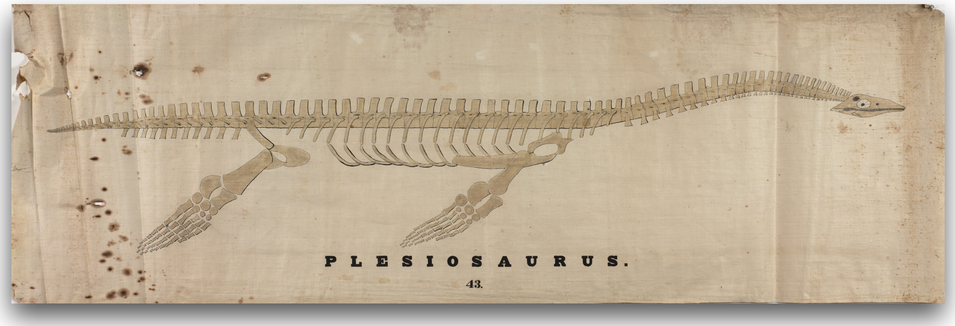
- **Environmental and Life-History Factors:**
  Organisms with shorter generation times or higher metabolic rates may accumulate mutations more quickly than those with longer generation times.

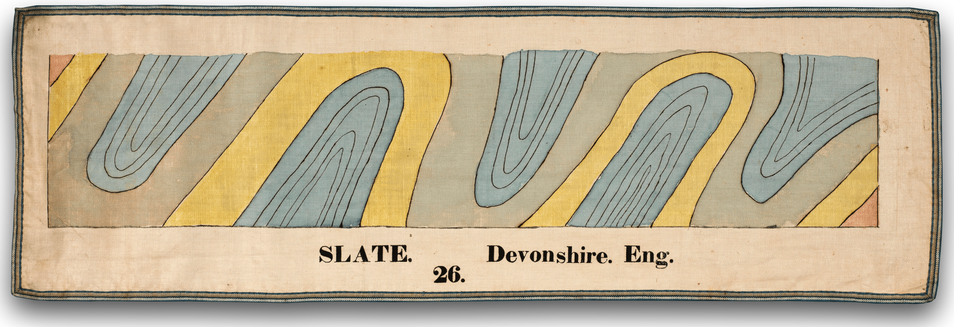
- **Practical Implications:**
  While the molecular clock provides a useful approximation for dating evolutionary events, calibrations using fossil records or known divergence times are essential to account for rate variability.

In summary, although genetic change can approximate a constant rate over evolutionary time, many factors cause deviations from a perfect molecular clock.


30
20
10
0
number oif mutaitons
Time

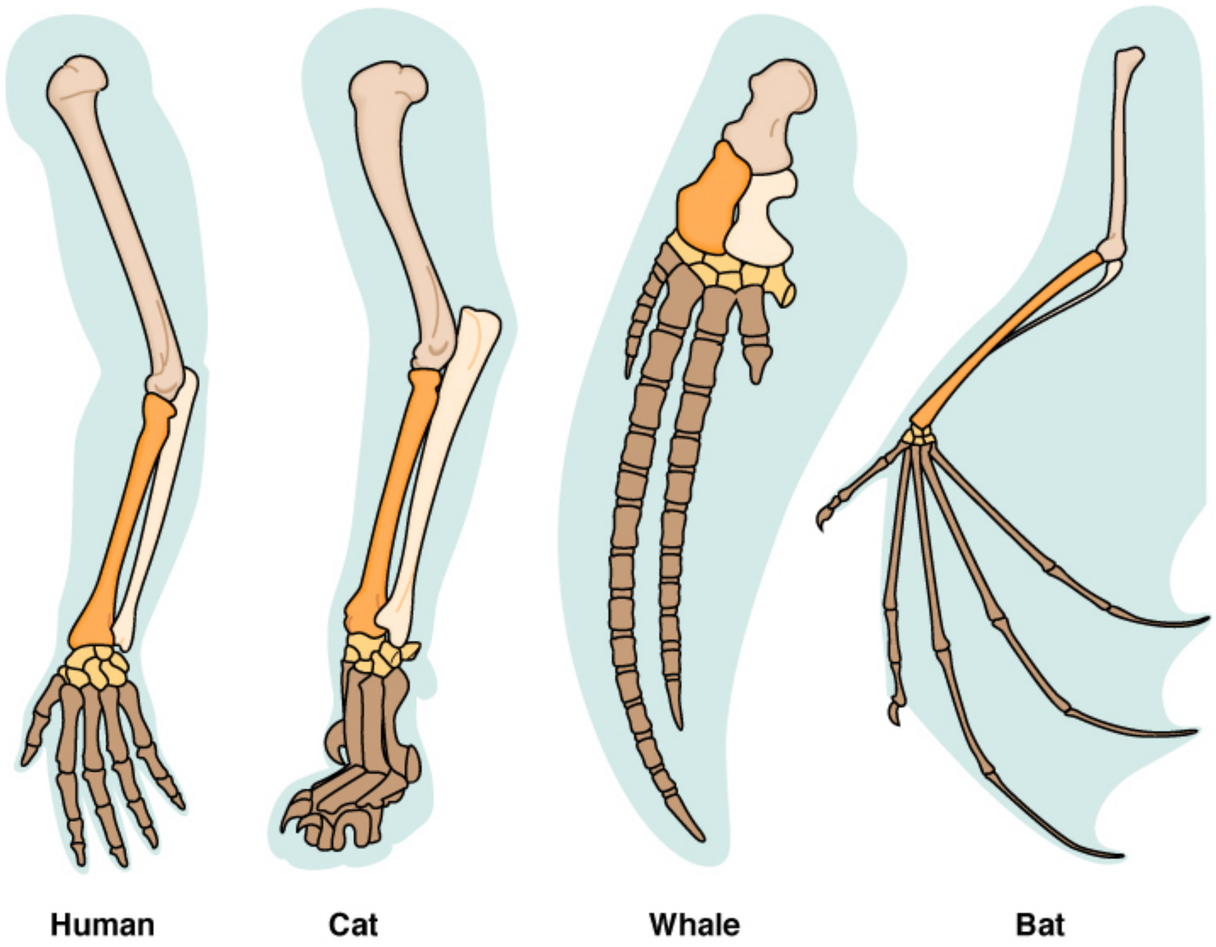






experimental design and orthology infererence

Phylogenetics is done using homologus characters! 😎



… but what does it mean when we think about genes? 😱

HOMOLOGY
characters sharing ancestry

ANALOGY
characters sharing function but not ancestry

😐

Compared characters **MUST** be homologous

… but within homology …

Orthology vs. Paralogy 🧬

- **Orthologs**

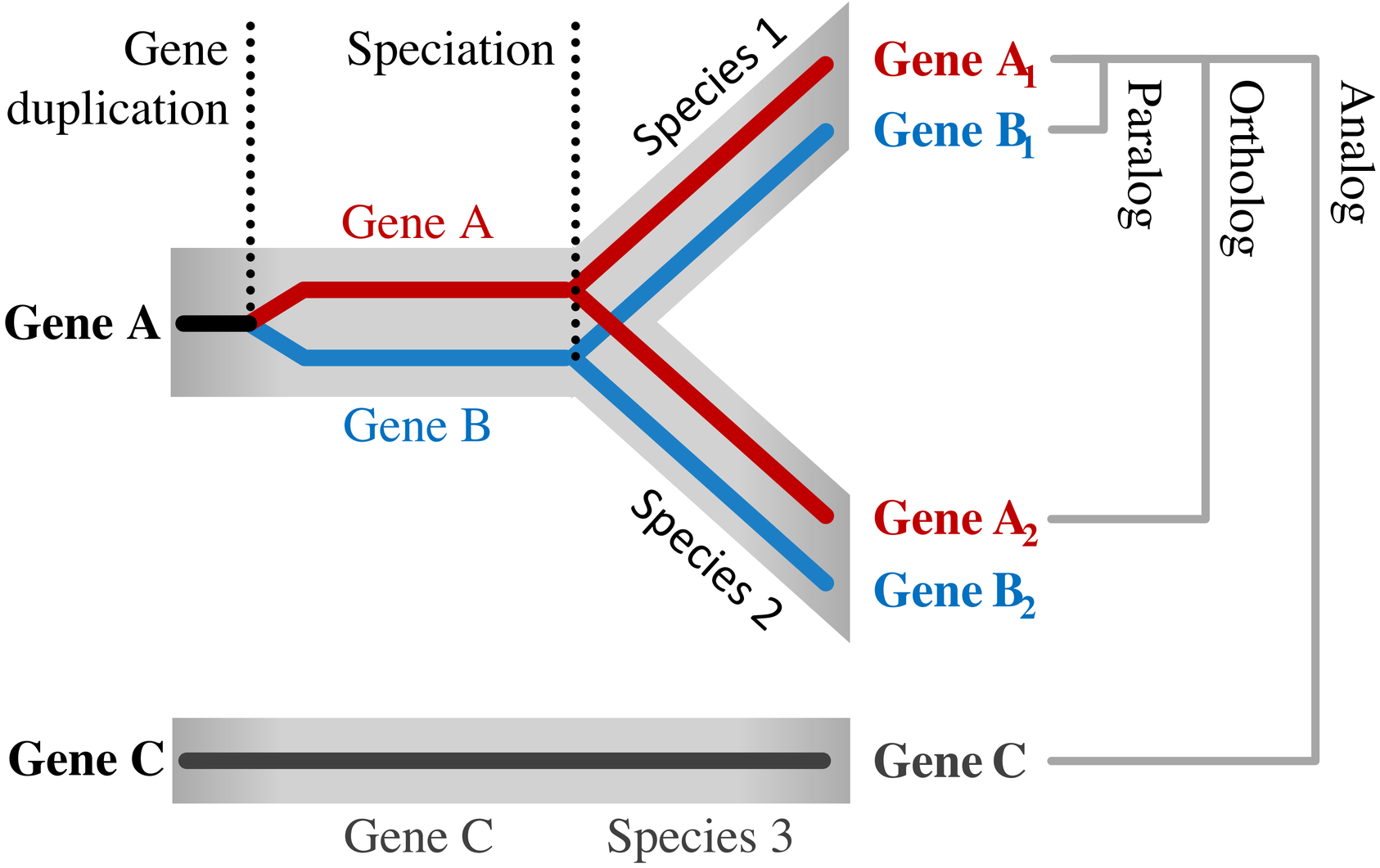
Genes that evolved from a common ancestral gene via **speciation** → **reflect species relationships** and may used for phylogenetic inference.

Two genes are orthologs if their MRCA is a **speciation** event!

- **Paralogs**:

Genes that originated from a common ancestral gene via **duplication** → may **confound phylogenies** as they do not reflect species relationships.

Two genes are paralogs if their MRCA is a **duplication** event!



wikipedia page on orthology where I took this figure 😛

Why in phylogenetics inference we are interested in strictly orhologs genes?

Since orthologs genes arise by speciation events,
they share the same evolutionary history of the underlying species.

the orthology conjecture

proposed in Nehrt et al. 2011

Orthologs genes are expected share the same **biological function**, while paralgos genes are belivied to differ in function.

However, as usual in biology, be aware of this latest corollary, recently ortholog conjecture has been **largely questioned** … see Stamboulian et al. 2020, Lynch and Conery 2000, and Gout and Lynch 2015 for for some interesting hints on fate of duplicated genes 😛

Btw, I think the orthology conjecture is mostly false … 😑

Classification of ortologs and paralogs!

Orthology is **always** defined by phylogenetics and unit of comparison.

One-to-One Orthologs: A single copy of the gene is present in both species, originating from a common ancestor and retained without duplication after a speciation event.

One-to-Many and Many-to-One Orthologs: After a speciation event, gene duplication occurs in one of the two species. This results in a single copy in one species and multiple copies in the other.

Many-to-Many Orthologs: After a speciation event, gene duplications occur in both species, leading to multiple orthologous copies in each lineage.

In-Paralogs: Paralogous genes that arise after a given speciation event. These duplicates are restricted to the lineage that experienced the duplication.

Out-Paralogs: Paralogous genes that originated before a given speciation event. These genes are related by a duplication event that occurred in the common ancestor of both species.

orthology inferences are inferred **pairwise**

when considering **multiple species**,
we should move to the concept of **orthogroup**

an orthogroup is a group of orthologs genes descending from the MRCA of a group of species
→ extending the concept of orthology to multiple species

an **orthogroup** is always defined by
a **reference speciation event**

Reciprocal Best BLAST Hit or **RBH** is an easy way to infer orthologs between 2 genomes.

- **Perform All-vs-All BLAST Searches**

Each gene in S1 is used as a query to search against the entire genome of S2.
The top hit by e-value or bit score is recorded.

Each gene in S2 is used as a query to search against the entire genome of S1.
The top hit by e-value or bit score is recorded.

- **Identify Reciprocal Best Hits**

Sp1 gene	Best Hit in Sp2	Sp2 gene	Best Hit in Sp1	RBH?
GeneA	GeneX	GeneX	GeneA	✅ Yes
GeneB	GeneY	GeneY	GeneC	❌ No
GeneC	GeneZ	GeneZ	GeneC	✅ Yes

Why is the Bi-Directional Best Hit Important?
Without a bi-directional comparison, a lost gene in one species could lead to incorrect inferences of orthology.

What is RBH biggest limitation?
Fails to Detect One-to-Many or Many-to-Many Orthologs. Cases where a single gene in Sp1 corresponds to multiple genes in Sp2 are ignored, and **RBH is strictly one-to-one**.

Beyond RBH:

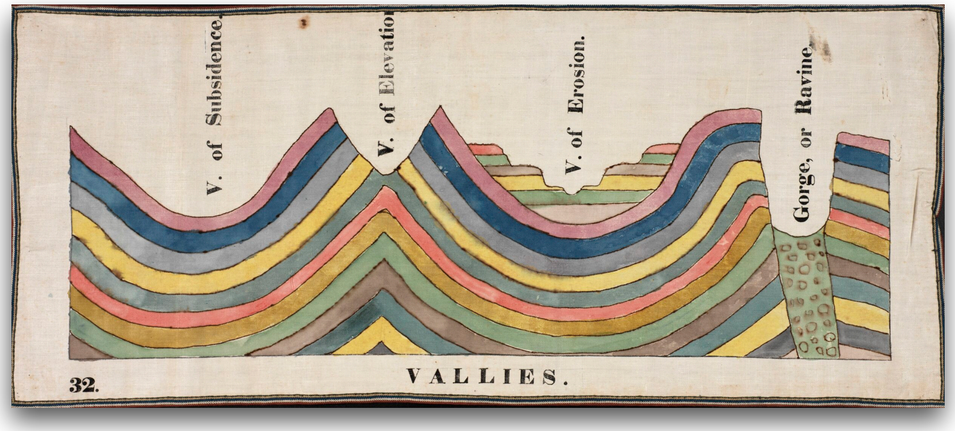
- **Clustering Orthologs into Orthogroups:** Groups multiple related genes across species, allowing many-to-many relationships.
- **Gene Tree Inference:** Uses phylogenetic reconstruction to distinguish true orthologs vs. paralogs, improving accuracy.

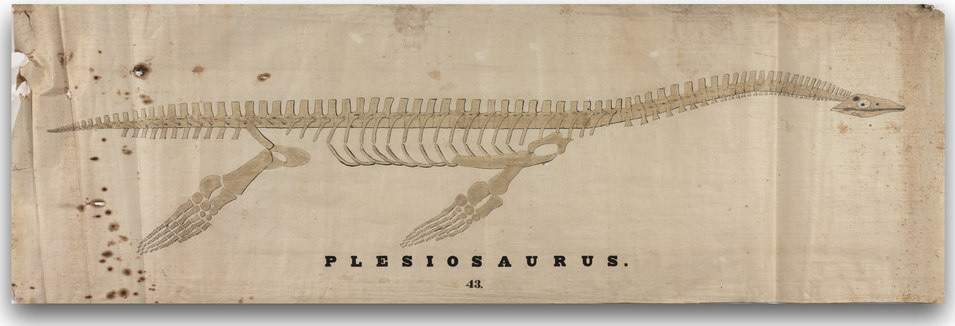
In **phylogenetics** we either want:

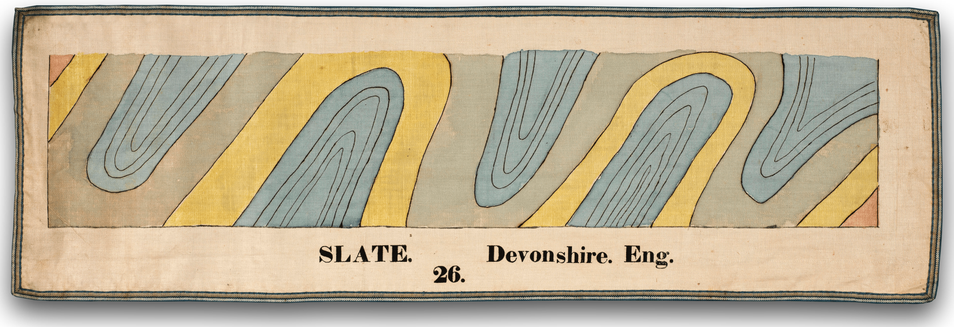
- **1-to-1** or **single-copy orthogroups** (i.e. with only one copy for specie)
- **trimmed orthogroups** (without genes derived from duplication events)

However the use of gene families (**paralogs + orthologs**) to infer species **phylogenetic relationship** is getting more and more attention lately!

.. for further infos see *Inferring Orthology and Paralogy* by Anisimova 2019.

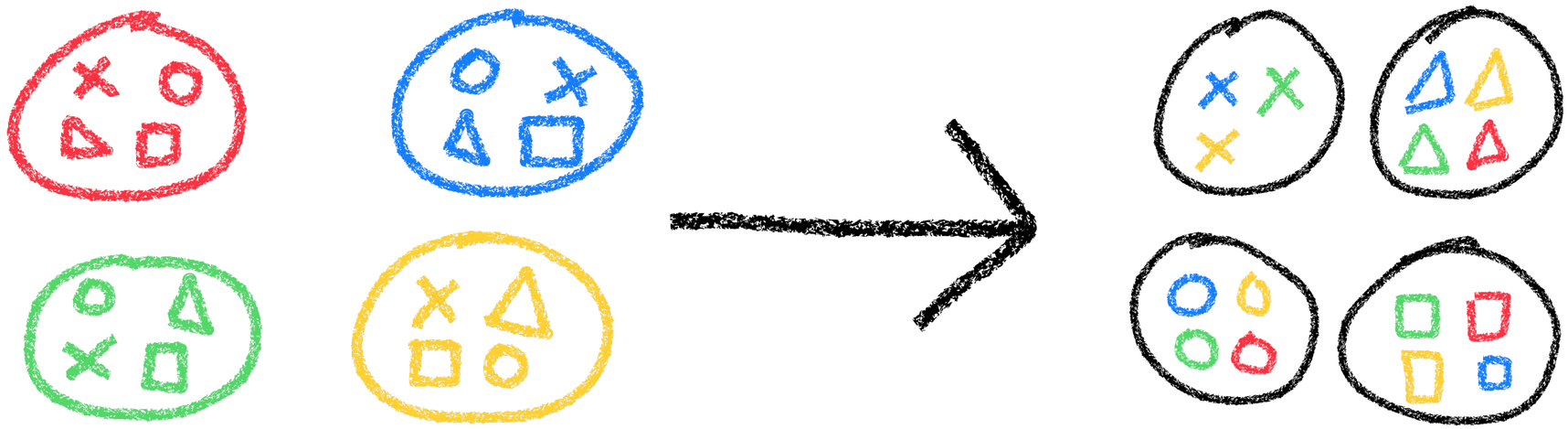






Sequence alignement and filtering

OK, I have my orthogroups … 😎



and now what?!

orthologous genes & orthologous sites!

```
species 1   AGGATCTGCAATTGCTCTTCTAATCTGTCTGATCAGGAT
species 2   AGG------AATTGCTCTTCTAATCTGTCT---CAGGAT
species 3   AGGATCTGCAATTGC---TCTAATCTGTCTGATCAGGAT
species 4   AGAATCTGCAATTGCTCTTCTGATCTGTCTGATCACGAT
species 5   AGGATCTGC---TGCTCTTCTGATCTGTCTGATCAGGAT
```

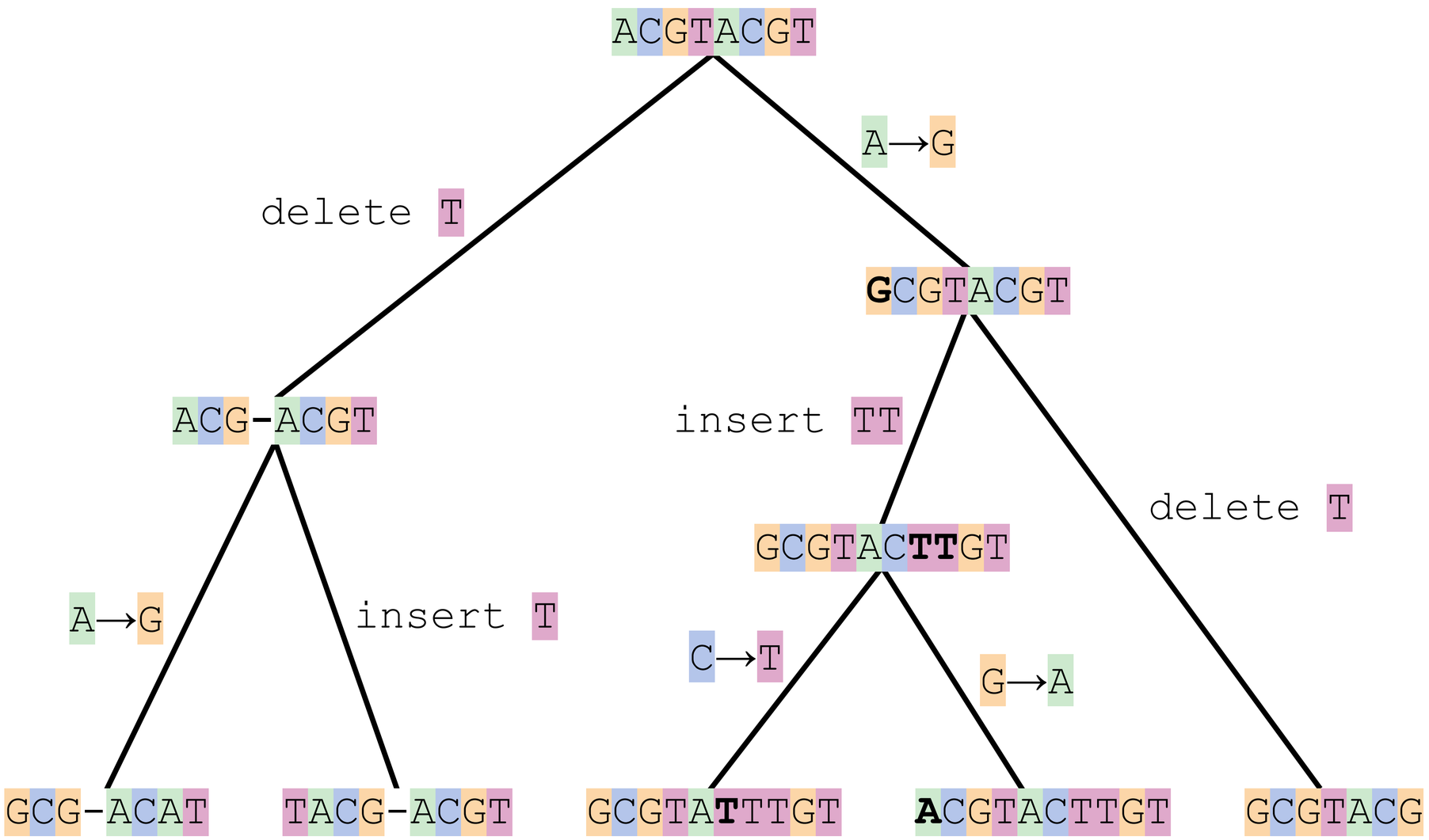
The goal of is to identify which **positions** are **orthologous**.

That is, their **evolutionary history is the same as that of the species**.

alignements (of nucleotides and proteins)

```
species 1   AGGATCTGCAATTGCTCTTCTAATCTGTCTGATCAGGAT
            ValArgSerCysSerCysValArgSerCysValValSer
species 2   AGG------AATTGCTCTTCTAATCTGTCT---CAGGAT
            Val------CysSerCysValArgSerCys---ValSer
species 3   AGGATCTGCAATTGC---TCTAATCTGTCTGATCAGGAT
            ValArgSerCysSer---ValArgSerCysValValSer
species 4   AGAATCTGCAATTGCTCTTCTGATCTGTCTGATCACGAT
            TrpArgSerCysSerCysValCysSerCysValArgSer
species 5   AGGATCTGC---TGCTCTTCTGATCTGTCTGATCAGGAT
            ValArgSer---SerCysValCysSerCysValValSer
```

sequences evolve on a tree


ACGTACGT
delete T
A→G
ACG-ACGT
GCGTACGT
insert TT
delete T
A→G
insert T
GCGTACTTGT
C→T
G→A
GCG-ACAT
TACG-ACGT
GCGTATTTGT
ACGTACTTGT
GCGTACG

These are the sequences sampled at present!

```
species 1   GCGACAT
species 2   TACGACGT
species 3   GCGTATTTGT
species 4   ACGTACTTGT
species 5   GCGTACG
```

This is the “true” alignement, which is the one where the nucleotides are arranged to be orthologous.

```
species 1   -GCG-AC-AT-
species 2   TACG-AC-GT-
species 3   -GCGTATTTGT
species 4   -ACGTACTTGT
species 5   -GCGTAC--G-
```

… but also this alignement does not look too bad 🙃!

```
species 1   -GCG-AC----A----T
species 2   ----TAC--G-AC--GT
species 3   -GCGTA-TT------TGT
species 4   A-CGTACTTGT-------
species 5   -GCGTAC--G--------
```

Which alignement is better?

This can be a bit of an art.

At its core you need to balance the number of gaps with number of mismatches

What price for a gap?

Gap penalties scoring:
There are two basic methods for assigning a cost ***c*** to a gap of length ***g*** in a sequence.

- Linear cost: $c=-dg$, where ***d*** is the gap open penalty
- Affine cost: $c=-d-(g-1)e$, where ***e*** is the gap extension penalty.

Typical values are d=10 and e=0.1.

Nucleotide substitution scoring:

- match (1)
- mismatch (-1)

alignment scores calculation:
Sum of match/mismatch scores
Minus gap penalties

There are two types of alignments: **pairwise alignments** and **multiple sequence alignments**. Each type of alignment has their own methods and algorithm, but they both attempt to maximize the similarity between sequences by inserting gaps when necessary to improve overall alignment.

Pairwise Alignment

- **Global alignment:** Needleman-Wunsch Algorithm

Aligns entire sequences from end to end. This algorithm is ideal when query sequences are of similar length and are expected to share homology across their full length.

- **Local Alignment:** Smith-Waterman Algorithm

Aligns specific regions of sequences rather than the full length. It is ideal when sequences are globally dissimilar but contain localized similarities, such as motifs or conserved domains.

Profile Alignment

1. Pairwise Alignments: Compute pairwise sequence distances.
2. Guide Tree Construction: Build a tree based on sequence similarity.
3. Progressive Alignment: Align the closest sequences first, then extend to profiles.
4. Final Refinement: Optimize alignment to improve accuracy.

The goal of the alignment process is to identify evolutionary events associated with homologies and ensure aligned sequences accurately reflect biological relationships.

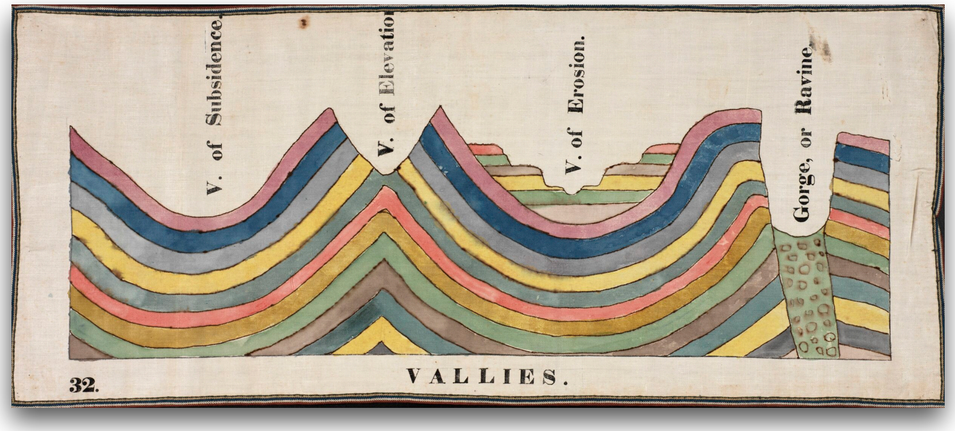
However:
- not all part of the sequences might be homologous
- for some part of the sequences homolgy might be not reconstructed correctly

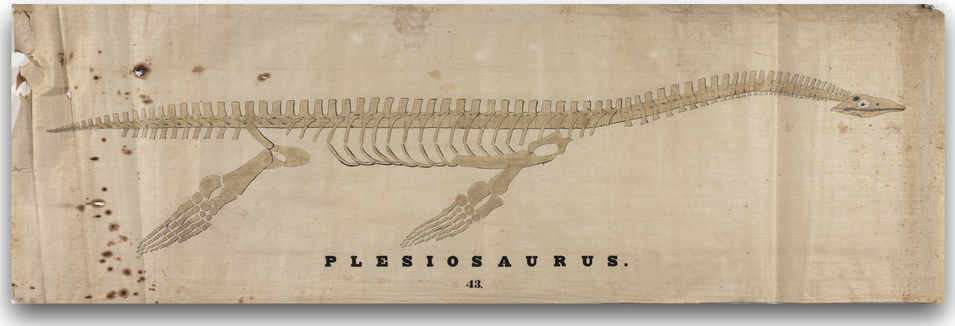
If the sequences are poorly aligned, you may want to consider trimming the poorly aligned areas.

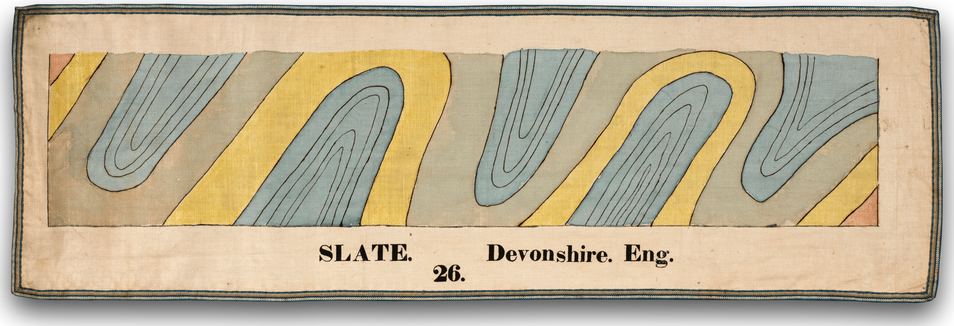
```
species 1  GCGTATTTGTCGCGAAAATCCCACGAA-GCG-AC-AT-C-AG--ATT-TATGAATCGACATG
species 2  GCGTATTTGTCGCGAAAATCCCACGAATA-G-AG-GT-G--TC-TG---ATGAATCCACATG
species 3  GCGTATTTGTCGCGAAAATCCCACGAA-GC---T-TGCC-AC--AG--CATGAATCGACATA
species 4  GCGTATTTGTCGCGAAAATCTCACGAA-ATGT-CTT-TC-AT--CT--CATGAATCGACATG
species 5  GCGTATTTGTCGCGAAAATCCCACGAA-GCTTAC--GC-AT--TAC--CATGAATCGACATG
```

Notable tools for the process are:

- **Gblocks** (Talavera & Castresana 2007)
- **Aliscore** (Kück et al. 2014)
- **BMGE** (Criscuolo & Gibaldo 2010)







distance-based
versus
charcter-based
algorythms

Distance-based methods

Distance-based phylogenetic trees are based on the **total number of evolutionary changes** between pairs of sequences.

Starting from the alignment, these methods look at all possible pairs of the aligned sequences and count how many characters are different at each position. These pairwise differences are represented in a **distance matrix**.

These methods are now best used for **exploratory analysis** of large datasets before conducting more intensive tree building using character-base methods.

Two distance-based methods are commonly used:

- **Unweighted Pair Group Method with Arithmetic Mean (UPGMA):** provides rooted and ultrametric trees. It assumes a constant molecular clock, meaning sequences evolve at a uniform rate over time.

- **Neighbor joining (NJ):** it does not assume a molecular clock, meaning it allows different lineages to evolve at different rates, and produce unrooted trees.

UPGMA

Step 1: Create a Distance Matrix

The first step is to calculate a pairwise distance matrix between sequences using a metric like p-distance, Jukes-Cantor or Kimura …

… more on these two in the following lessons.

Taxon	A	B	C	D
A	0	5	9	9
B	5	0	10	10
C	9	10	0	8
D	9	10	8	0

Step 2: Identify the Closest Pair

Find the smallest distance in the matrix. In this case, A & B have the shortest distance (5), so they are clustered first.

Taxon	A	B	C	D
A	0	5	9	9
B	5	0	10	10
C	9	10	0	8
D	9	10	8	0

Step 3: Compute a New Distance Matrix

The new cluster (A,B) must have its distance recalculated to all other taxa. The distance to each taxon is the average of the distances from A and B:

d (A,B) – C = d(A,C) + d(B,C) / 2

Taxon	(A,B)	C	D
(A,B)	0	9,5	9,5
C	9,5	0	8
D	9,5	8	0

Step 4: Repeat the Process of step 2 & 3

2. Find the next smallest distance and cluster pairs.
3. Recalculate new distances based on average linkage.

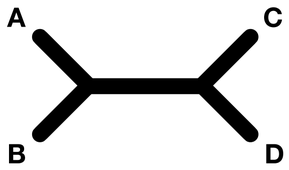
Taxon	(A,B)	(C,D)
(A,B)	0	9,5
(C,D)	9,5	0

Step 5: Final Merge

The remaining two clusters, (A,B) and (C,D), are joined, and the tree is complete.

((A,B),(C,D))

Does in remind you of anything?



PS: there is also WPGMA where the W stands for *weighted*. Basically it weights distances by cluster size, leading to more balanced trees.Recalculating Distances After Clustering.

d (F,G,H) – X = d(F,X) + (2 x d(G,H,X,) / 1+2

NJ

Step 1: Create a Distance Matrix

The same as with UPGMA!

They are both distance based methods in the end! 😜

Taxon	A	B	C	D	E	F	G	H
A	0	4	8	8	9	10	10	10
B	4	0	8	8	9	10	10	10
C	8	8	0	2	5	6	7	7
D	8	8	2	0	5	6	7	7
E	9	9	5	5	0	3	4	4
F	10	10	6	6	3	0	2	2
G	10	10	7	7	4	2	0	1
H	10	10	7	7	4	2	1	0

Step 2: Compute Net Divergence (Q-Matrix)

Instead of selecting directly the closest pair (as in UPGMA), NJ calculates a **Q-matrix**, based on net divergence of each taxon from all others (**R**).

Taxon	A	B	C	D	E	F	G	H
A	0	4	8	8	9	10	10	10
B	4	0	8	8	9	10	10	10
C	8	8	0	2	5	6	7	7
D	8	8	2	0	5	6	7	7
E	9	9	5	5	0	3	4	4
F	10	10	6	6	3	0	2	2
G	10	10	7	7	4	2	0	1
H	10	10	7	7	4	2	1	0

Step 2: Compute Net Divergence (Q-Matrix)

Step 2.1: Compute R(i) for Each Taxon For each taxon i, using: $R(i) = \sum_{k=1}^{n} d(i,k)$

R(A)=(0+4+8+8+9+10+10+10)=59
R(B)=(4+0+8+8+9+10+10+10)=59
…
R(G)=(10+10+7+7+4+2+1)=41
R(H)=(10+10+7+7+4+2+1)=41

Taxon	A	B	C	D	E	F	G	H
A	0	4	8	8	9	10	10	10
B	4	0	8	8	9	10	10	10
C	8	8	0	2	5	6	7	7
D	8	8	2	0	5	6	7	7
E	9	9	5	5	0	3	4	4
F	10	10	6	6	3	0	2	2
G	10	10	7	7	4	2	0	1
H	10	10	7	7	4	2	1	0

Step 2: Compute Net Divergence (Q-Matrix)

Step 2.2: the Q-value for each pair of taxa (i,j) is calculated as: $Q(i,j) = (n-2)d(i,j) - R(i) - R(j)$

- d(i,j) is the distance between taxa i and j.
- R(i) and R(j) are the net divergences for taxa i and j.
- n is the total number of taxa (initially 8).

Example: Q(G,H) = (8 – 2) × 1 – (41 + 41) = – 76

Taxon	A	B	C	D	E	F	G	H
A		-94.0	-54.0	-54.0	-44.0	-38.0	-40.0	-40.0
B	-94.0		-54.0	-54.0	-44.0	-38.0	-40.0	-40.0
C	-54.0	-54.0		-74.0	-52.0	-46.0	-42.0	-42.0
D	-54.0	-54.0	-74.0		-52.0	-46.0	-42.0	-42.0
E	-44.0	-44.0	-52.0	-52.0		-60.0	-56.0	-56.0
F	-38.0	-38.0	-46.0	-46.0	-60.0		-68.0	-68.0
G	-40.0	-40.0	-42.0	-42.0	-56.0	-68.0		-76.0
H	-40.0	-40.0	-42.0	-42.0	-56.0	-68.0	-76.0	

Step 3: Merge the Closest Pair and Compute New Distances

Step 3.1: Merge the closest pair.

The closest pair is made by G and H which have to be merged into a new cluster.

Taxon	A	B	C	D	E	F	G	H
A		-94.0	-54.0	-54.0	-44.0	-38.0	-40.0	-40.0
B	-94.0		-54.0	-54.0	-44.0	-38.0	-40.0	-40.0
C	-54.0	-54.0		-74.0	-52.0	-46.0	-42.0	-42.0
D	-54.0	-54.0	-74.0		-52.0	-46.0	-42.0	-42.0
E	-44.0	-44.0	-52.0	-52.0		-60.0	-56.0	-56.0
F	-38.0	-38.0	-46.0	-46.0	-60.0		-68.0	-68.0
G	-40.0	-40.0	-42.0	-42.0	-56.0	-68.0		-76.0
H	-40.0	-40.0	-42.0	-42.0	-56.0	-68.0	-76.0	

Step 3: Merge the Closest Pair and Compute New Distances

Step 3.2: compute new distances between the new cluster U and the remaining taxa:

To do so use the formula: $d(U,X) = \frac{1}{2}[d(G,X) + d(H,X) - d(G,H)]$

In this formula:

- d(G,H) is the original distance between G and H
- d(G,X) and d(H,X) are distances from G and H to any other taxon X

Taxon	A	B	C	D	E	F	(G,H)
A	0.0	4.0	8.0	8.0	9.0	10.0	9.5
B	4.0	0.0	8.0	8.0	9.0	10.0	9.5
C	8.0	8.0	0.0	2.0	5.0	6.0	6.5
D	8.0	8.0	2.0	0.0	5.0	6.0	6.5
E	9.0	9.0	5.0	5.0	0.0	3.0	3.5
F	10.0	10.0	6.0	6.0	3.0	0.0	1.5
(G,H)	9.5	9.5	6.5	6.5	3.5	1.5	0.0

Step 4: Compute the Branch Lengths of Merged Nodes

The branch lengths from the newly created node U to its child nodes G and H are computed as:

$$u_G = \frac{1}{2}\left[d(G,H) + \frac{R(G) - R(H)}{n-2}\right] \quad \text{and} \quad u_H = d(G,H) - u_G$$

Example for G and H:

$$u_G = \frac{1}{2}\left[1 + \frac{41-41}{8-2}\right] = \frac{1}{2}(1) = 0.5 \quad \text{and} \quad u_H = 1 - 0.5 = 0.5$$

Step 5: Repeat Until All Taxa Are Clustered

5.1 - Compute R for each taxa and a Q-matrix with the updated distances.

5.2 - Select next closest pair and merge, compute new distance matrix.

5.3 - calculate branchlengths

Continue until all taxa are merged into a NJ tree!

Key Properties of UPGMA

✅ Simple & Fast
✅ Produces a Rooted Tree
❌ Assumes a Constant Molecular Clock
❌ Sensitive to Unequal Evolutionary Rates

Key Properties of NJ

✅ Computationally efficient, scales well to large datasets
✅ Produces an unrooted tree
❌ Does not assume clock-like molecular evolution
❌ Less accurate than Maximum Likelihood (ML) or Bayesian Inference (BI)

Character-based methods

Character-based methods compare all sequences by **considering one character** (nucleotides or amino acids) in the alignment **at a time**.

As character-based methods incorporate **evolutionary models** making these methods more accurate than distance-based methods.

These methods construct multiple trees that are evaluated and ranked, resulting in selection of the best tree. Character-based methods include:

- **Maximum Parsimony (MP)**
- **Maximum Mikelihood (ML)**
- **Bayesian Inference (BI)**

PARSIMONY

Maximum Parsimony (MP) is a character-based phylogenetic method that seeks to reconstruct the tree with **the fewest evolutionary changes** (mutations) necessary to explain the observed data.

✅ Minimizes total character changes (Occam's Razor principle).

✅ Does not require an explicit evolutionary model, but posses an implicit one

❌ Computationally expensive for large datasets - the number of trees grows exponentially!

❌ Sensitive to homoplasy (convergent evolution or reversals).

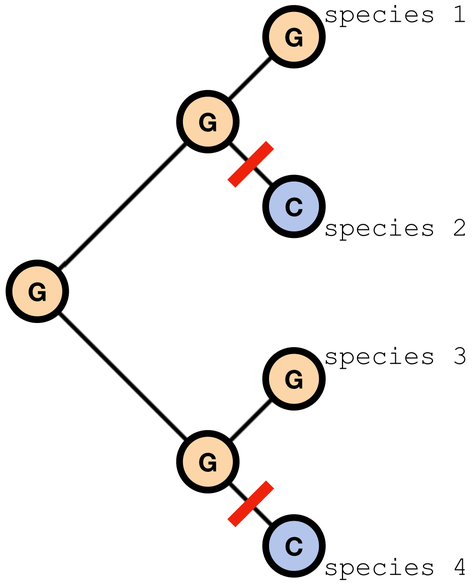
❌ Does not correct for multiple substitutions (it has no proper statistical model).

```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
species 5       A G G G C C G C
```

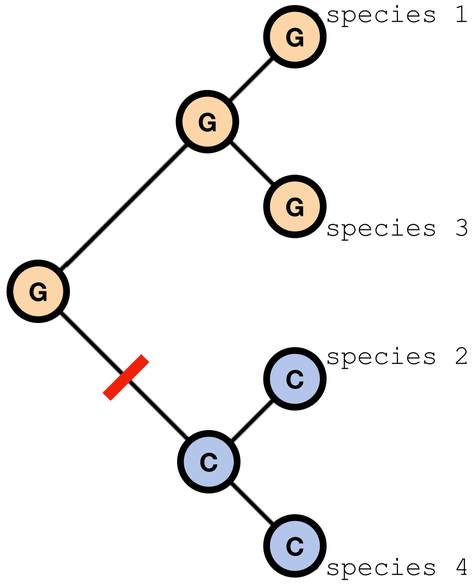
Parsimony uses only **informative sites**, which:
- contains at least two different character states
- each state appears in at least two taxa

1 - not informative, has only one character state
2 - **informative**, has two character states and each is present in more than one taxa
3 - not informative, has two character states but one is present in just one taxa (**singleton**)
4 - not informative, has three character states but one is present in just one taxa (**singleton**)
5 - ...
6 - ...
7 - ...
8 - ...

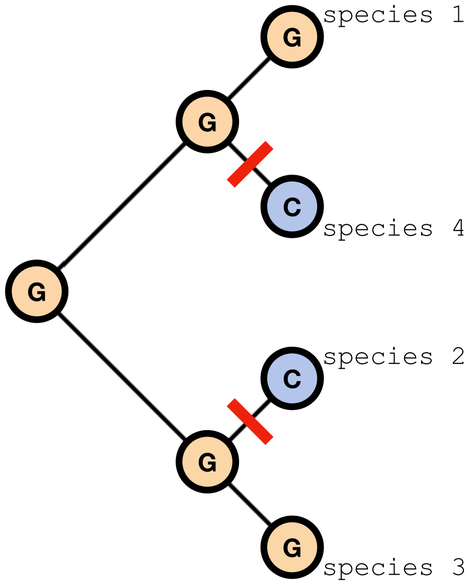
```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
                  *
```



SCORE = 2



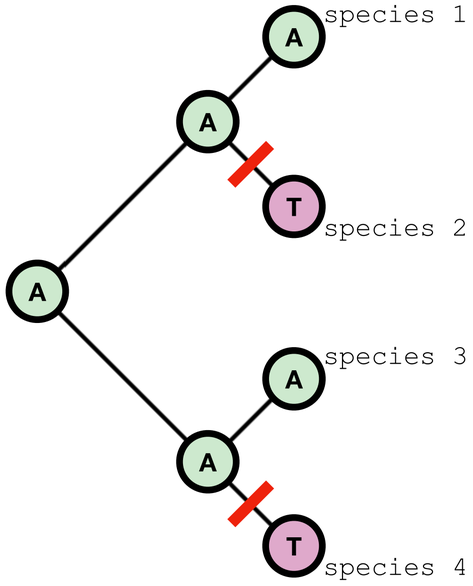
SCORE = 1



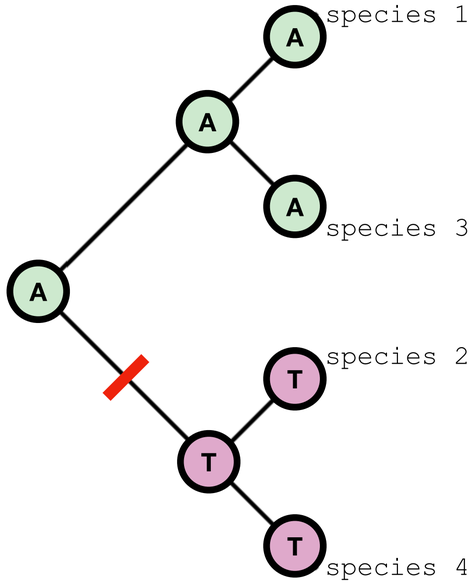
SCORE = 2

```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
                      *
```

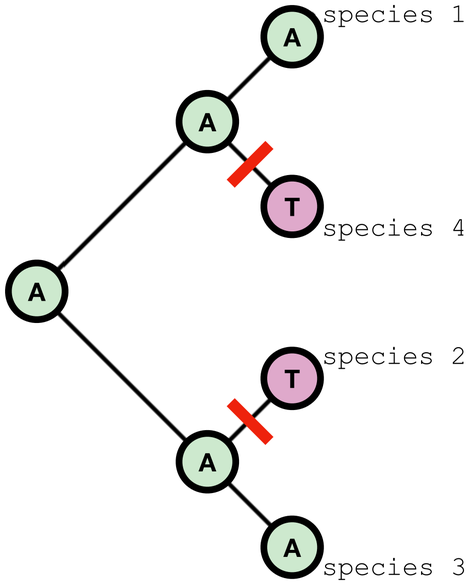
This site is congruent with site 2!



SCORE = 2



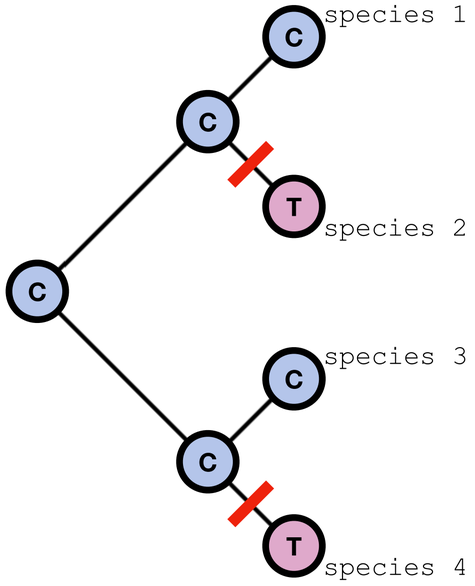
SCORE = 1



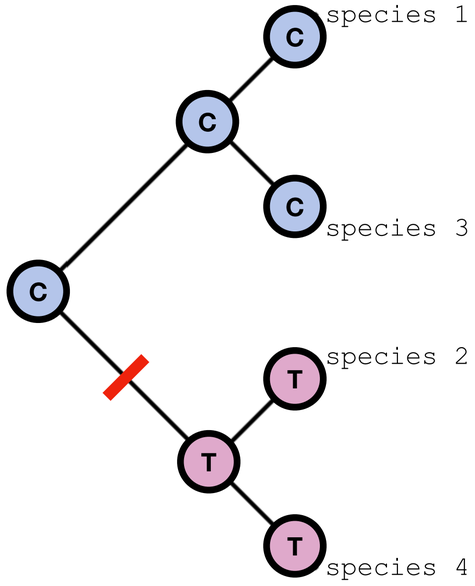
SCORE = 2

```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
                        *
```

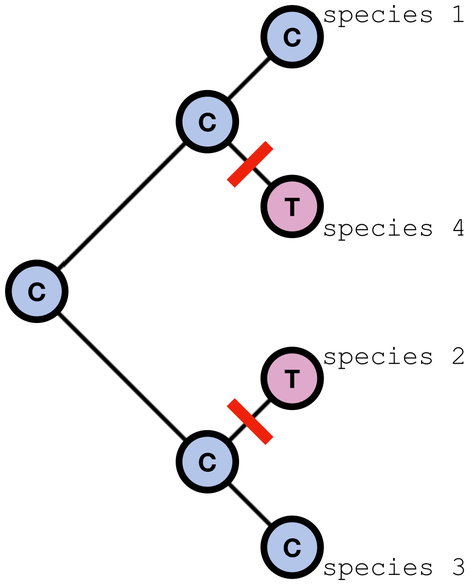
This site is congruent with site 2!



SCORE = 2



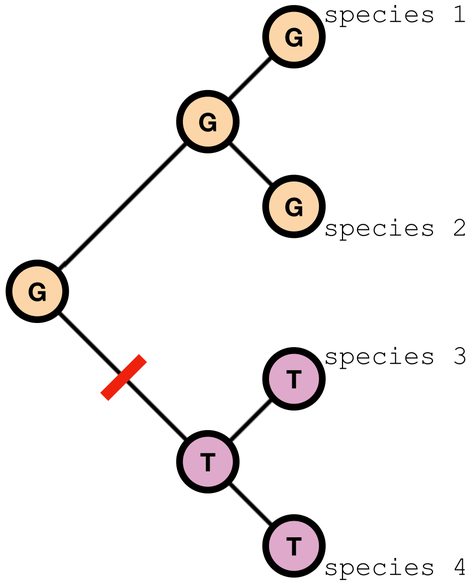
SCORE = 1



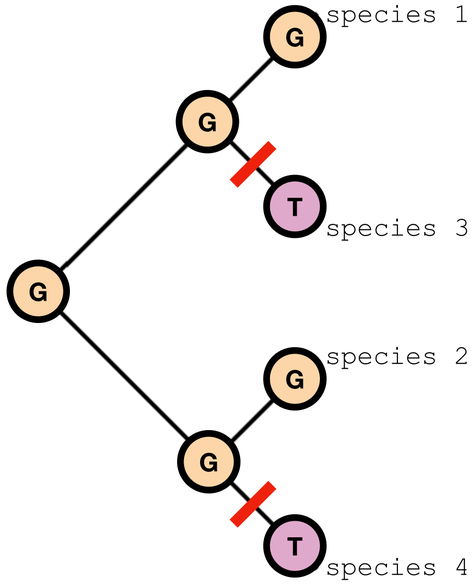
SCORE = 2

```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
                              *
```

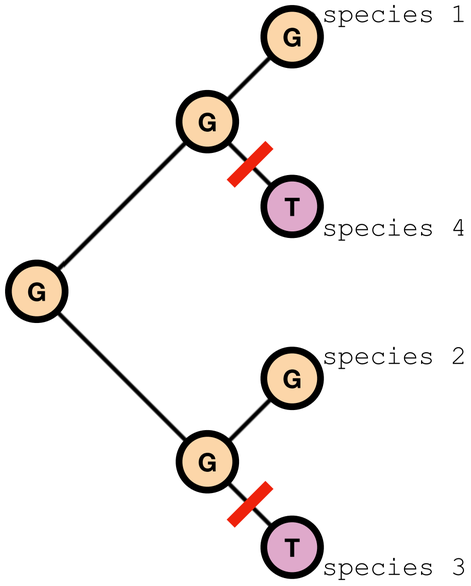
This site is incongruent with site 2!



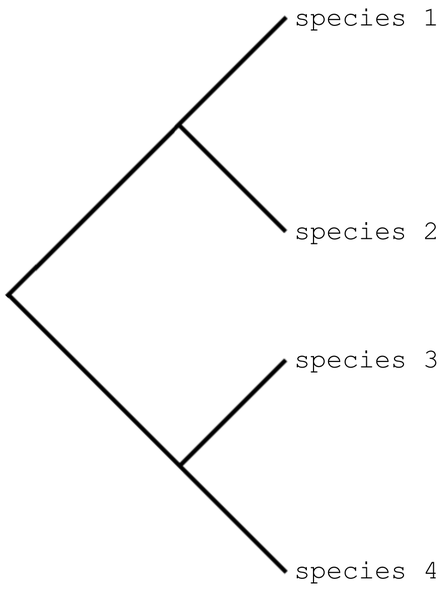
SCORE = 1

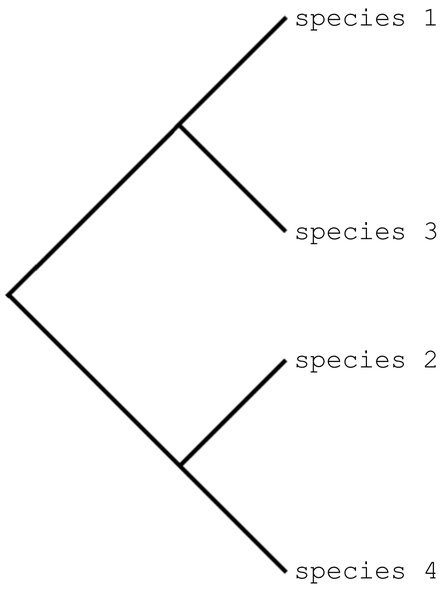


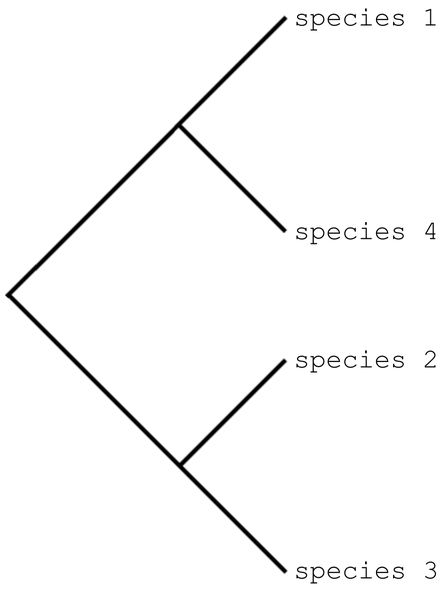
SCORE = 2



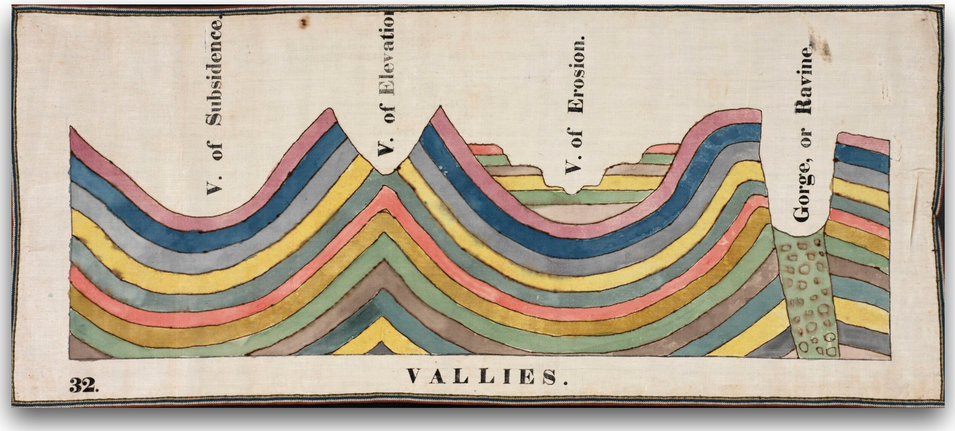
SCORE = 2

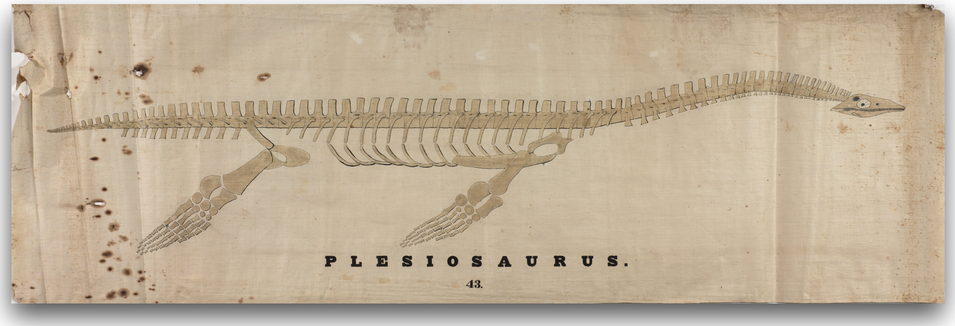


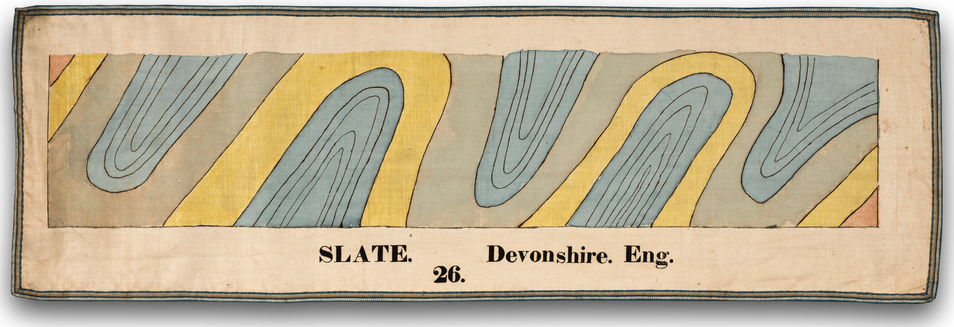




charachter (site)	topology 1	topology 2	topology 3
2	2	1	2
4	2	1	2
5	2	1	2
7	1	2	1
total	7	**5**	7







Models of sequence evolution

Nearest Neighbour Interchange (NNI) is a local search method that makes small, localized changes to the tree. It is the simplest and fastest tree rearrangement algorithm!

Subtree Pruning and Regrafting (SPR) is a more extensive tree rearrangement method than NNI. It moves entire subtrees rather than just swapping immediate neighbors.

Tree Bisection and Reconnection (TBR) is the most aggressive of these tree rearrangement methods and explores a much larger space of possible tree topologies. It has the lesser chance of local optima!

"*All models are wrong, but some are useful*"

Box, 1979

In biology, a **substitution model**, also called **models of sequence evolution**, are Markov models that describe changes over evolutionary time.

A substitution model is a mathematical framework that describes the probabilities of **changes** (substitutions) occurring in **sequence evolution over time**.

It accounts for different types of substitutions (e.g., nucleotide, amino acid, or codon changes) and is used to estimate evolutionary distances between sequences and to

by correcting for multiple substitutions that may have occurred at the same site.

In a **Markov model** is a type of stochastic model used to represent systems that transition between a finite or countable number of states in a chain-like manner.

The defining characteristic of a Markov model is the **Markov property**, which stipulates that the probability of transitioning to the next state depends solely on the current state and not on the sequence of events that preceded it.

Key components:

- **States:** The distinct configurations or conditions that the system can occupy.
- **Transition Probabilities:** The probabilities associated with moving from one state to another.

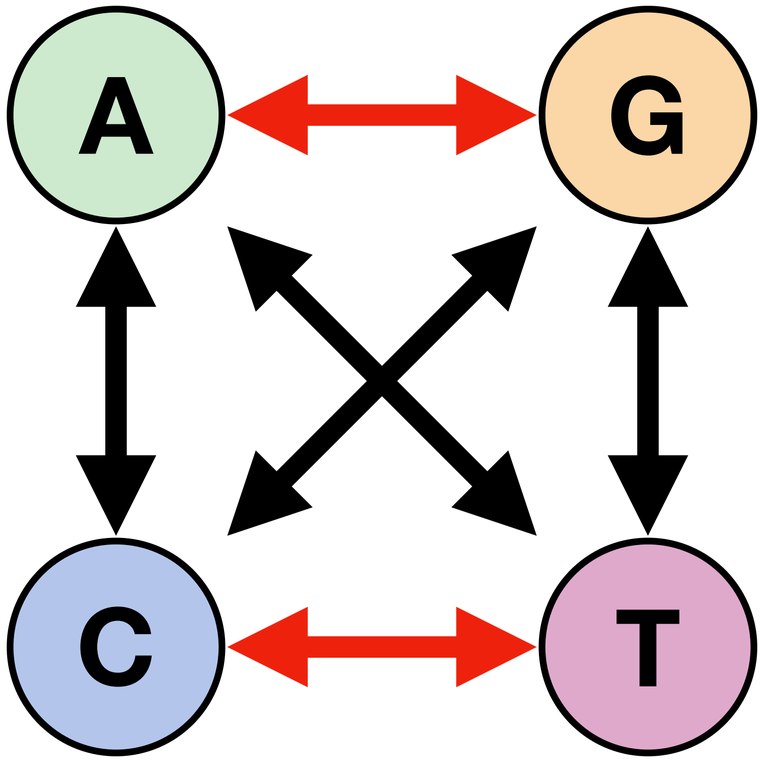
Mechanistic models: consider the biological process involved: mutational biases, natural selection… they have more interpretative power and particularly useful for studying the evolutionary forces and mechanisms.

Empirical models: describe the relative rates of substitution do not consider explicitly factors that influence the evolutionary process large quantities of sequence data.

Models of sequence substitution can be built for different evolutionary units:

- **nucleotides**: 4 states (A, C , G , T)
- **codons**: 61 states (AAA, AAC , AAT , …)
- **amino acids**: 20 states (Phe, Leu, Ile, …)

Nucleotide Substitution Models



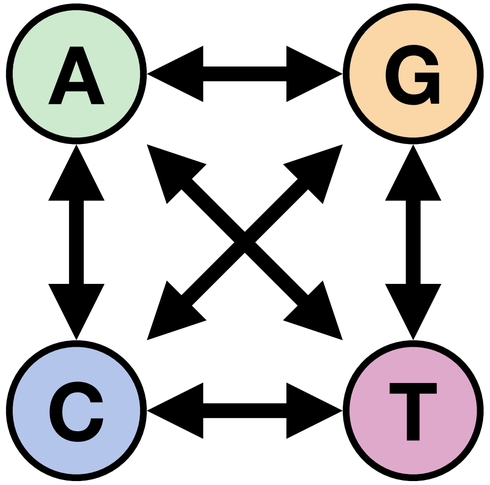
Transitions (Ts): Substitutions within the same nucleotide class A ↔ G (purines), C ↔ T (pyrimidines). More frequent than transversions due to structural similarity.

Transversions (Tv): Substitutions between different nucleotide classes (purine ↔ pyrimidine). Examples: A ↔ C, A ↔ T, G ↔ C, G ↔ T. Less common due to greater structural change.

Jukes–Cantor (JC69)

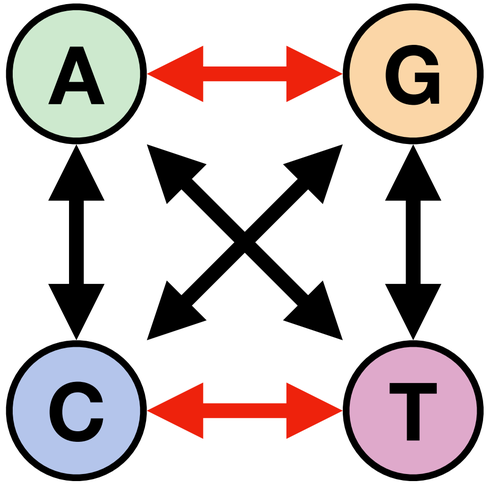
The simplest nucleotide model, JC69 assumes all nucleotides are equally frequent and all substitution rates are equal.

It has a single parameter (overall substitution rate) and corrects for multiple substitutions analytically. This model is often used as a baseline, though it may oversimplify real sequence evolution.



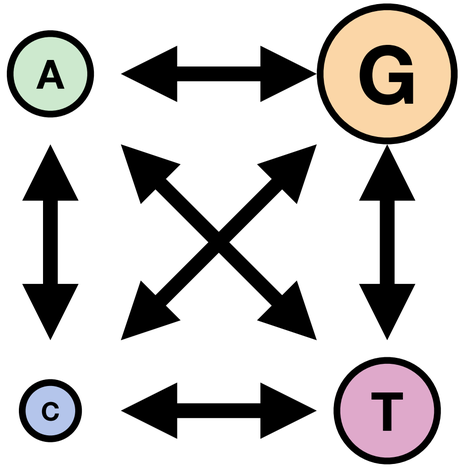
Kimura 2-Parameter (K80)

Kimura (1980) introduced a two-parameter DNA model that distinguishes between transitions and transversions. K80 still assumes equal base frequencies (25% each) but better fits data with the commonly observed transition bias. It uses a separate rate (alfa) for transitions (interchanges of A↔G or C↔T) and another (beta) for transversions.



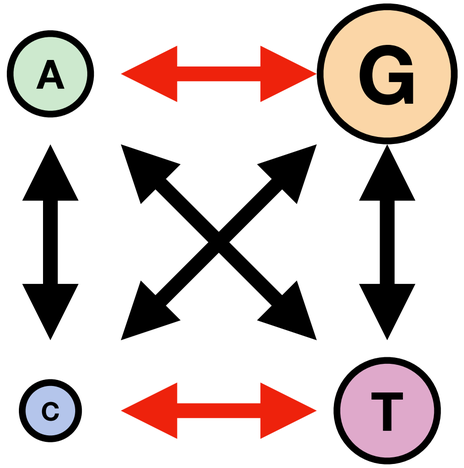
Felsenstein (F81)

Felsenstein's 1981 model extends JC69 by allowing unequal base frequencies while keeping all substitution rates equal . In F81, each nucleotide has its own equilibrium frequency … π_A, π_C, π_G, π_T … This accounts for compositional bias in sequences but assumes no transition/transversion rate difference.



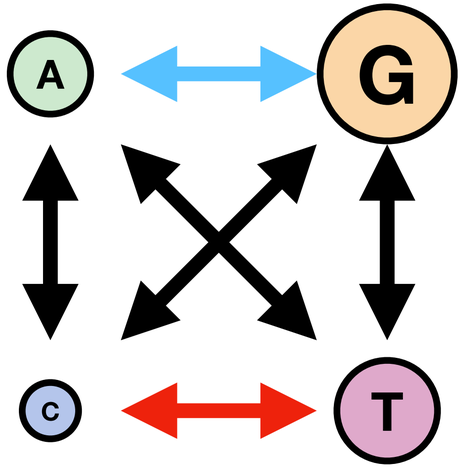
Felsenstein (F81)

Felsenstein's 1981 model extends JC69 by allowing unequal base frequencies while keeping all substitution rates equal . In F81, each nucleotide has its own equilibrium frequency … π_A, π_C, π_G, π_T … This accounts for compositional bias in sequences but assumes no transition/transversion rate difference.



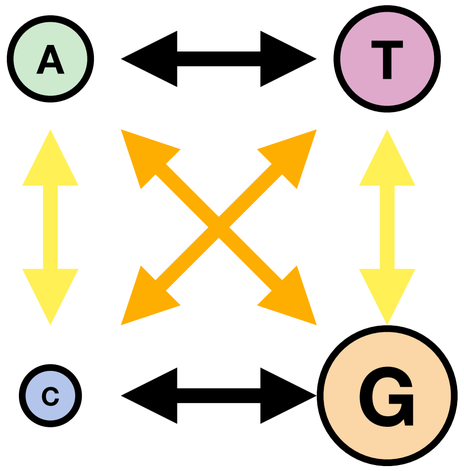
Tamura–Nei (TN93)

Tamura and Nei (1993) further refine transition rate parameters. TN93 allows two distinct transition rates while keeping a single rate for all transversions. It also permits unequal base frequencies.



Kimura 3-parameter (K3P)

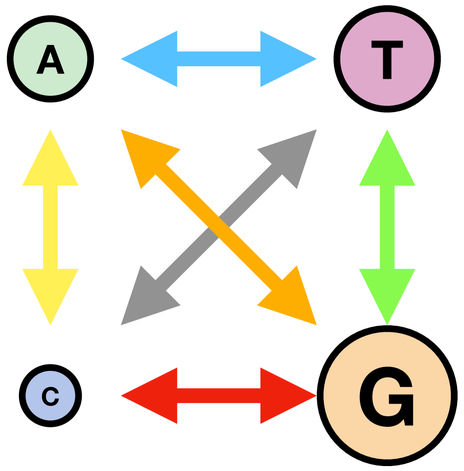
Kimura (1981) introduced the Kimura 3-parameter (K3P) model, which extends the Kimura 2-parameter (K2P) model by incorporating variable base frequencies. K3P assumes equal transition rates but allows two distinct transversion rates, making it more flexible than K2P while maintaining computational efficiency.



General Time Reversible (GTR):

GTR (Tavaré 1986) is the most general reversible model for nucleotides. It includes a full set of six substitution rate parameters (one for each pair of nucleotides) and allows any set of equilibrium base frequencies. GTR usually provides the best fit to data, at the cost of more parameters.

All simpler models (JC, HKY, TN93, etc.) are special cases of GTR with certain rates constrained equal.



Models are defined by the **Q-matrix**, specifying the instantaneous rates of substitution. It incorporates:

Equilibrium frequencies of nucleotides:

- π_A - equilibrium frequencies of A
- π_c - equilibrium frequencies of C
- π_G - equilibrium frequencies of G
- π_T - equilibrium frequencies of T

Substitution rate parameters between nucleotide pairs:

- **α** - substitution rate A to G
- **β** - substitution rate A to C
- **γ** - substitution rate A to T
- **δ** - substitution rate G to C
- **ϵ** - substitution rate G to T
- **η** - substitution rate C to T

Assuming equal frequencies, the off-diagonal elements of the Q matrix can be calculated as:
A→C: $q_{ac} = r_{ac} \times \pi_c = 1.6385 \times 0.25 = 0.4096$
A→G: $q_{ag} = r_{ag} \times \pi_g = 3.0090 \times 0.25 = 0.7523$
A→T: $q_{at} = r_{at} \times \pi_t = 1.6385 \times 0.25 = 0.4096$

The diagonal element q_{aa} is then calculated to ensure the row sums to zero: $q_{aa} = -(q_{ac} + q_{ag} + q_{at}) = -(0.4096 + 0.7523 + 0.4096) = -1.5715$. This process is repeated for all nucleotides!

General Time Reversible (GTR):

$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

Key Points:

The diagonal elements are set such that each row sums to zero, ensuring that the total probability remains conserved over time.

The model is time-reversible, meaning the process of substitution looks the same forward and backward in time when the system is at equilibrium.

Models ranked by complexity

Jukes-Cantor (JC): equal base frequencies, all substitutions equally likely

Felsenstein 1981 (F81): variable base frequencies, all substitutions equally likely

Kimura 2-parameter (K80): equal base frequencies, one transition rate and one transversion rate

Hasegawa-Kishino-Yano (HKY): variable base freqs, one transition and one transversion rate

Tamura-Nei (TrN): variable base frequencies, equal transversion rates, variable transition rates

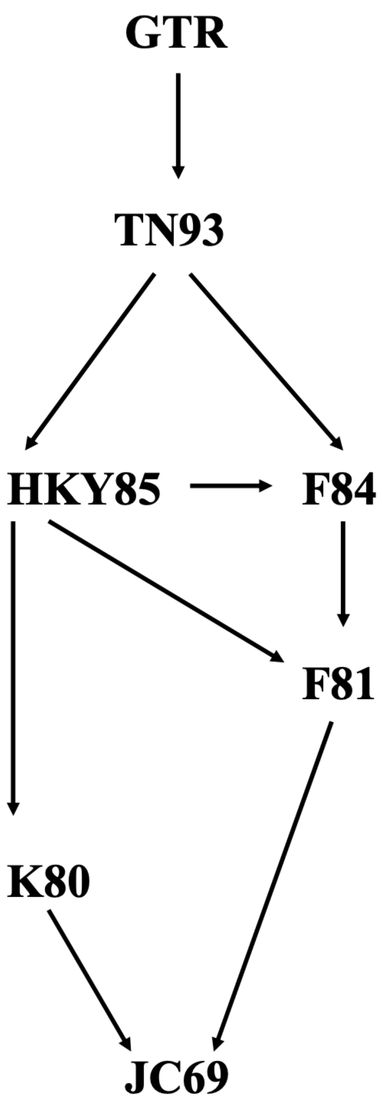
Kimura 3-parameter (K3P): variable base freqs, equal transition rates, two transversion rates

transition model (TIM): variable base freqs, variable transition rates, two transversion rates

transversion model (TVM): variable base freq, variable transversion rates, transition rates equal

symmetrical model (SYM): equal base freqs, symmetrical substitution matrix (A to T = T to A)

general time reversible (GTR): variable base frequencies, symmetrical substitution matrix



8

$a, b, c, d, e, \pi_A, \pi_C, \pi_G,$
$(a+b+c+d+e+f=1 \quad \pi_A,+\pi_C,+\pi_G,+\pi_T,=1)$

5

$a=c=d=f, b=e, \pi_A, \pi_C, \pi_G,$
(*Ti/Tv ratio, Y/R ratio and and three frequencies since* $\pi_A,+\pi_C,+\pi_G,+\pi_T,=1$)

4

$a=b=c=d=1, \quad b=e=k, \pi_A, \pi_C, \pi_G,$
(*Ti/Tv ratio and only thre e frequencies since* $\pi_A,+\pi_C,+\pi_G,+\pi_T,=1$)

3

$a=c=d=f=b=e=f, \pi_A, \pi_C, \pi_G,$
(*only three nt frequencies since* $\pi_A,+\pi_C,+\pi_G,+\pi_T,=$

1

$a=b=c=d=1, \quad b=e=k, \quad \pi_A=\pi_C=\pi_G=\pi_T$
(*Ti/Tv ratio*)

0

$a=b=c=d=e=f, \quad \pi_A=\pi_C=\pi_G=\pi_T$

Codon Substitutions Models

Translation involves the sequential recognition of triplets of adjacent nucleotides called **codons**.

Since there are four nucleotides, there are 4*4*4 = **64 possible codons**.

In the universal genetic code, 61 code for a specific amino acid, while the other three are **stop codons**. Given that **there are 61 codons but only 20 amino acids**, most amino acids are encoded by more than one codon.

The degeneracy of the genetic code allows **substitutions** to occur in DNA that do **not result in a change in the amino acid sequence**.

These substitutions are known as **synonymous substitutions** (i.e., those substitutions that do not result in an amino acid change), and are generally thought to be under relaxed selective pressure and more common than **non-synonymous substitutions** (i.e., those substitutions that do result in an amino acid change).

Muse and Gaut (MG94)

In 1994, Muse and Gaut introduced a codon substitution model that assigns different rates to synonymous vs. nonsynonymous substitutions.

In MG94, two parameters can be used:

- one for the rate of synonymous changes (**dS**)

- one for nonsynonymous changes (**dN**).

This allows for the estimation of the absolute rates of each type, reflecting selective constraints on protein-coding genes. Natural selection operates mainly at the protein level.

Goldman–Yang 1994 (GY94)

Goldman and Yang's codon model, also published in 1994, similarly distinguishes codon changes by their effect on the encoded amino acid. Rather than separate dN and dS, it is often parameterized in terms of their ratio **ω = dN/dS**.

$\omega < 1$ - purifying selection
$\omega = 1$ - neutral evolution
$\omega > 1$ - positive selection

The GY94 framework also typically incorporates a nucleotide mutation model (for the underlying base changes, e.g. a transition/transversion bias κ) and codon frequency parameters (e.g. using empirical codon frequencies or an F3×4 model of base frequencies at codon positions)

GY94 and MG94 laid the groundwork for most modern codon phylogenetic analyses by explicitly modeling selective pressure on proteins.

Building on GY94, Yang and Nielsen and others developed more complex codon models to detect selection.

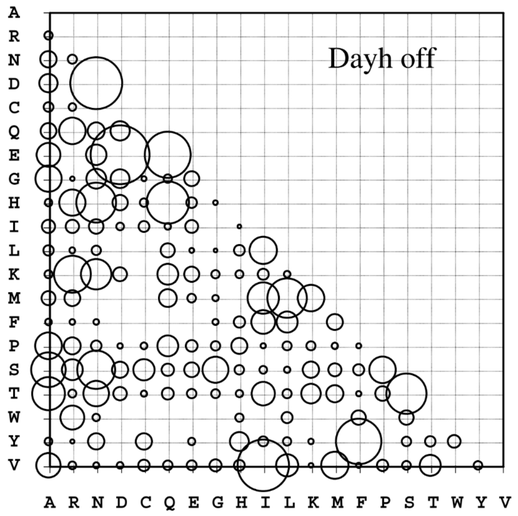
- **Site-specific models** in which the ω ratio varies across codon sites
- **branch-specific models** (Yang 1998) let ω differ among lineages to find episodes of adaptive evolution on particular branches.
- **Branch-site models** combine these ideas to detect positive selection affecting certain sites along specific lineages.

We will learn more on how to infer selection across sites and branches in **lesson 18**!

Amino Acid Substitution Models

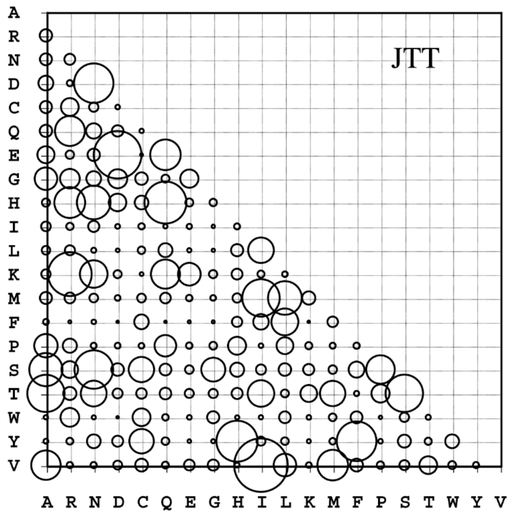
Dayhoff's PAM Model

Dayhoff et al. (1978) pioneered empirical protein evolution models! They compiled alignments of closely related proteins (>85% identity) to estimate amino acid replacement frequencies, resulting in the PAM (Percent Accepted Mutation) matrices.
For example, the PAM250 matrix corresponds to 250 accepted changes per 100 amino acids. The Dayhoff model (PAM) was foundational, though its limited dataset caused some rate biases (e.g. favoring globular protein substitutions).



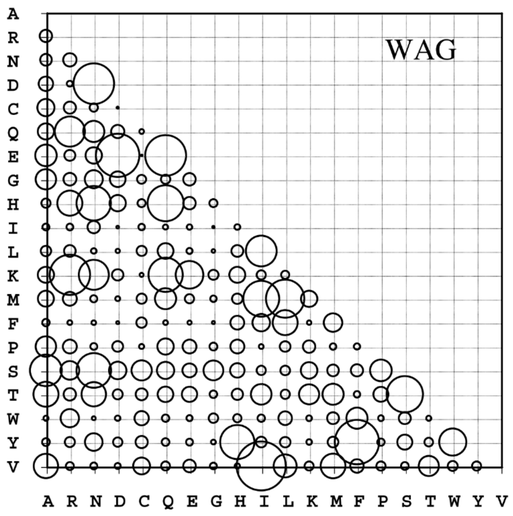
JTT (Jones–Taylor–Thornton):

JTT is an updated empirical model published in 1992 that built on Dayhoff's approach with a larger, more diverse protein database. Jones et al. derived a new 20×20 substitution matrix that better represents typical protein evolution, and it has been widely adopted for maximum-likelihood phylogenetic analyses.
They also produced a separate matrix for membrane proteins, noting that Dayhoff's matrix was biased toward soluble proteins!



WAG (Whelan and Goldman):

WAG is a 2001 amino acid model estimated using a maximum-likelihood approach on a broad set of protein families.
Unlike Dayhoff and JTT (which used observed counts and log-odds methods), WAG optimized the replacement rates to maximize likelihood, given a large training alignment dataset. This resulted in a more accurate general-purpose matrix that often fits protein data better than older matrices.



LG (Le and Gascuel): LG (2008) is a an improved empirical model derived from a vast alignment database using ML methods. Generally outperforming WAG/JTT, LG is a common choice for phylogenetics.

MtREV (Mitochondrial REV): To address the unique unique constraints of mitochondrial genomes, Adachi and Hasegawa (1996) developed the MtREV model from 20 vertebrate mitochondrial protein sequences. Similarly, specialized matrices like MtMam for mammalian mitochondria and HIVb for HIV proteins have been created.

BLOSUM (BLOcks SUbstitution Matrix): (Henikoff & Henikoff 1992) is a series of matrices (e.g. BLOSUM62) derived from conserved ungapped blocks of protein alignments. BLOSUM matrices are widely used for sequence alignment and database search, though they are not derived from an explicit evolutionary model.

+ possible to estimate a protein substitution rate matrix using **QMaker**!

1. Empirical and Mechanistic models:

Protein Substitution Models: Derived from extensive datasets, these models have pre-estimated parameters that remain fixed during analysis. While others, such as amino acid frequencies, can be estimated from the specific dataset.

Nucleotide Substitution Models: Mechanistic Models: Parameters, including substitution rates and base frequencies, are typically estimated directly from the dataset under analysis.

2. Distinguishing Model Parameters and Optimized Values:

Model Parameters: Refer to the structural components of the model, such as the number of substitution types or the inclusion of rate heterogeneity.

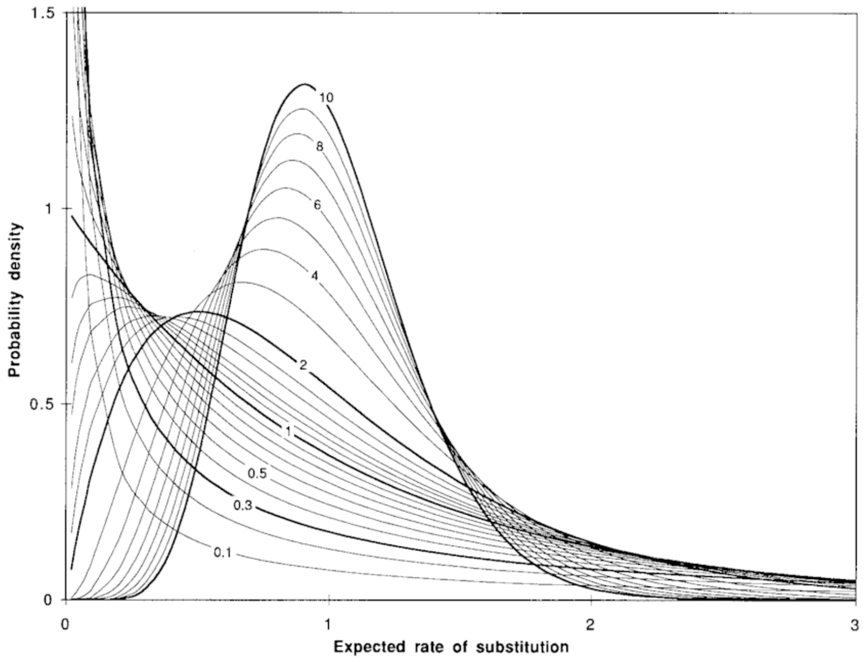
Optimized Numerical Values: These are the specific estimates of the parameters obtained by fitting the model to the data, reflecting the unique evolutionary characteristics of the dataset.

Key Assumptions:

The probability of a nucleotide changing from state ***i*** to state ***j*** depends solely on the current state ***i***, without influence from prior states. This reflects the Markov property, where the process lacks memory of past events.

The rates at which substitutions occur are constant over time. This implies that the evolutionary process is uniform throughout the timeline, without periods of accelerated or decelerated mutation rates.

The nucleotide frequencies remain constant over time, indicating that the system has reached a stationary distribution. At equilibrium, the proportion of each nucleotide does not change, reflecting a balance in the substitution process.



Rate Variation: Real sequence data rarely evolve at a uniform rate across all sites.

- **Gamma-distributed rate heterogeneity (+Γ)** assumes sites have variable rates drawn from a gamma distribution (with shape parameter α), allowing sites to evolve faster or slower than average

- **Invariant sites (+I)** models allow a proportion of sites to be evolutionarily invariable (zero rate)

These modifiers can be added to any base model (e.g. HKY+Γ, GTR+Γ+I) to better capture among-site rate variation, significantly improving likelihood fit in many cases!

What about distances?

How to define the genetic distance between two taxa?

Observed distance or p-distance

the proportion of nucleotide sites at which the two sequences differ.

```
species 1    AGAATCTGCAATTGCTCTGCTGATCTGTCTGATCACGATAT
species 2    AGGATCTGCAGTTGCTCTTCCGATCTGTCTGATCAGGATAT
               *         *         *  *                *
```

Expected distance or corrected distance

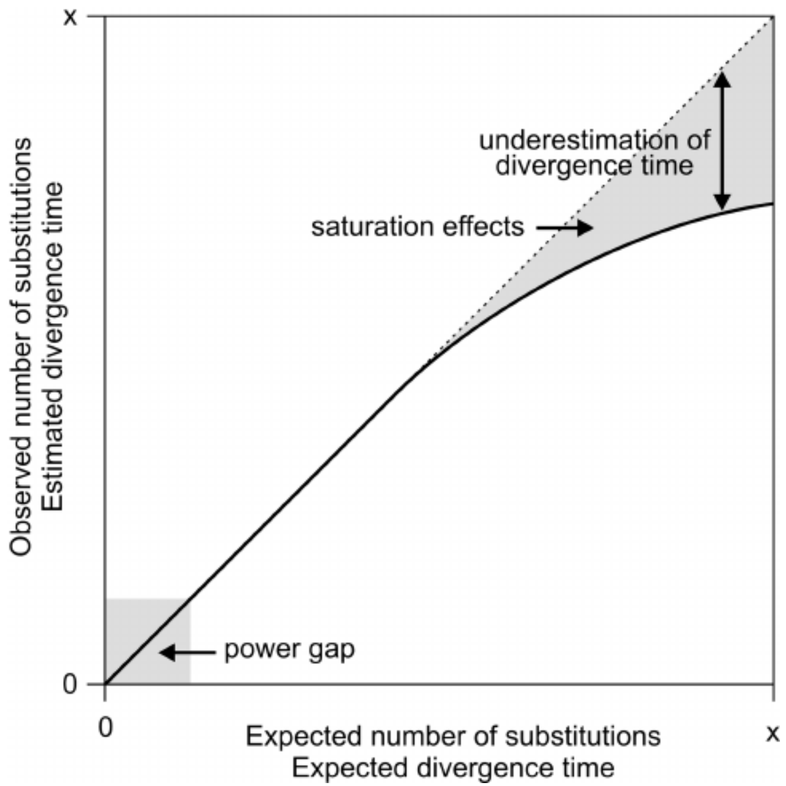
an estimate of the actual number of substitutions that have occurred, using substitution models to correct observed distances by considering the likelihood of multiple substitutions over time.

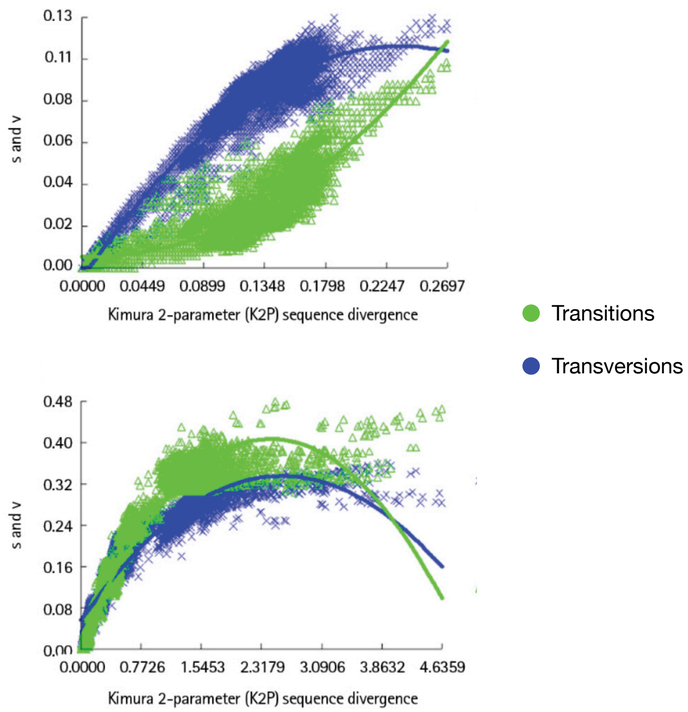
```
species 1    AGAATCTGCAATTGCTCTGCTGATCTGTCTGATCACGATAT
              T          |         |  A                |
              A          |         C  G                A
              ↓          ↓         ↓  ↓                ↓
species 2    AGGATCTGCAGTTGCTCTTCCGATCTGTCTGATCAGGATAT
```

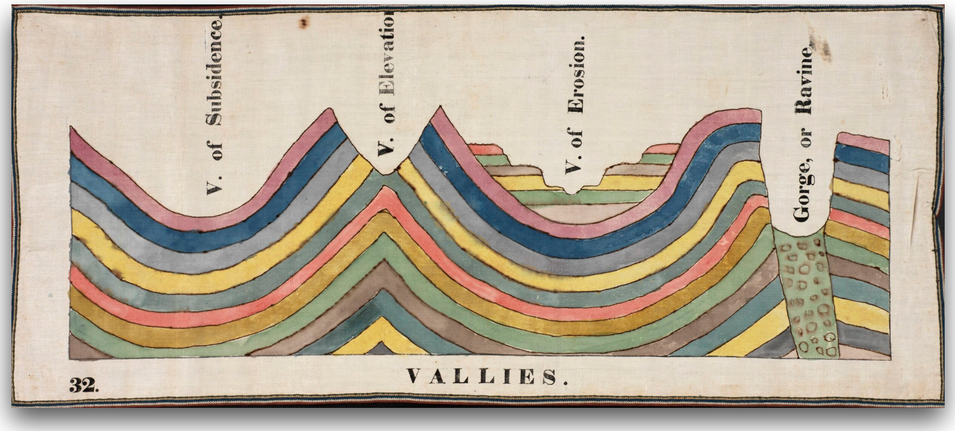
Substitution Saturation: when multiple nucleotide or amino acid substitutions have happened at the same site over time, leading to an underestimation of the true evolutionary divergence between sequences. Leads to:

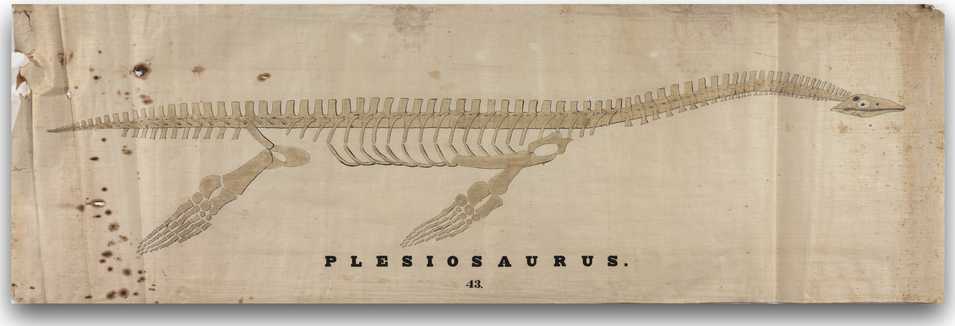
Loss of Phylogenetic Signal: Saturation diminishes the informative value of sequences.

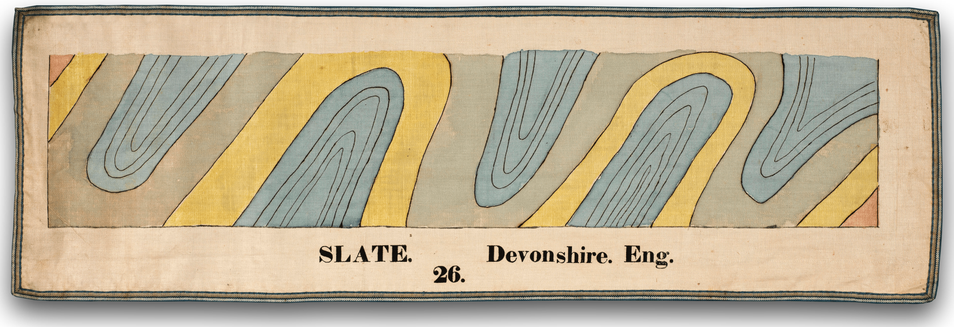
Long Branch Attraction: Highly divergent lineages may appear erroneously similar due to saturation.











Maximum Likelihood (ML)

Maximum Likelihood is a **statistical method for estimating parameters of a probability model**.

For example a normal distribution can be described by two parameters: the mean and variance.

Instead, in molecular phylogenetics there are a wide plethora of **parameters**, which include:

- rates of transitions / trasversions / ... between bases
- base composition
- descriptors of rate heterogeneity across sites
- branchlengths
- the tree itself 😵‍💫

Likelihood is defined as a quantity proportional to the probability of observing data given the model.

$$P(D|M)$$

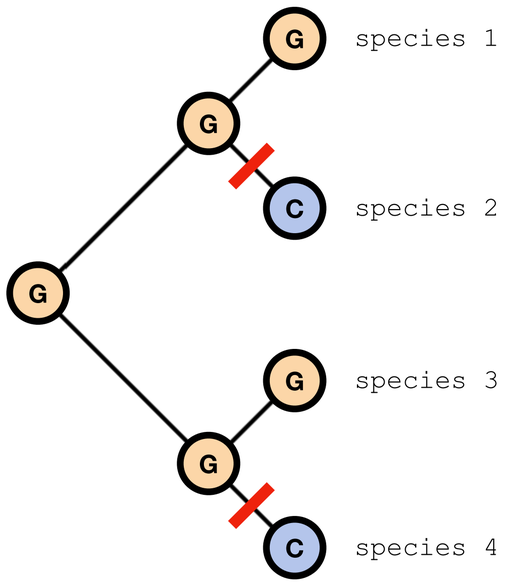
What is the probability of observed data given that a certain model is true?

P(D|M)

probability (data | model of evolution)

probability (sequences | substitution model & tree)

```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
                  *
```

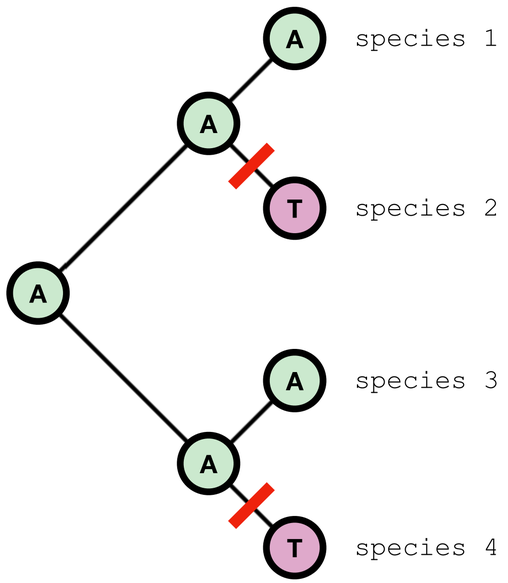


$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

$P_{GC} = r_{GC}\pi_G$

$P_{site2} = P_{GC} \times P_{GC} \times P_G \times P_G \times P_G \times P_G$

```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
                      *
```

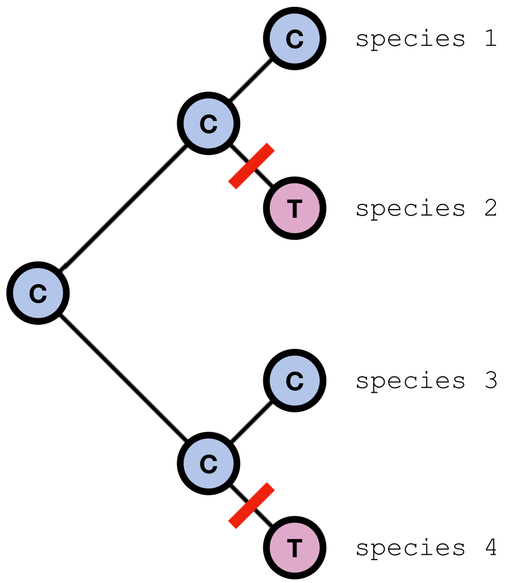


$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

$P_{AT} = r_{AT}\pi_A$

$P_{site4} = P_{AT} \times P_{AT} \times P_A \times P_A \times P_A \times P_A$

```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
                        *
```

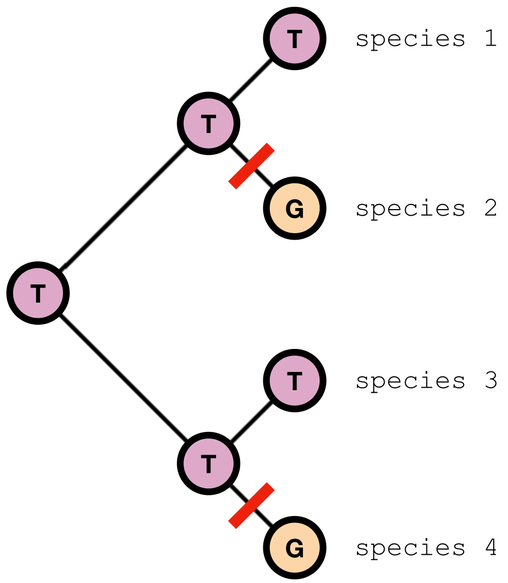


$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

$P_{CT} = r_{CT}\pi_C$

$P_{site4} = P_{CT} \times P_{CT} \times P_C \times P_C \times P_C \times P_C$

```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
                              *
```



$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

$P_{TG} = r_{TG}\pi_T$

$P_{site7} = P_{TG} \times P_{TG} \times P_T \times P_T \times P_T \times P_T$

But what about about ambiguous internal nodes?! 😱

We rely on **conditional likelihood**!

- **At terminal nodes:** For observed data at the tips, the algorithm assigns a conditional likelihood of 1 to the observed state and 0 to all others.
- **At internal nodes:** The algorithm computes the conditional likelihood for each possible state by integrating the likelihoods from descendant nodes, weighted by transition probabilities.

By summing over all possible states at each node, the algorithm incorporates the uncertainty of ancestral states into the overall likelihood calculation. This integration ensures that ambiguities are naturally integrated into the model, providing a comprehensive likelihood that reflects all possible evolutionary scenarios.

Calculating these probabilities directly for all possible ancestral states is computationally intensive, especially for large trees. To address this, algorithms like **Felsenstein's tree-pruning algorithm** are employed to efficiently compute the likelihood by systematically summing over ancestral states in a recursive manner. If you want to know more read **here**!

Likelihood of a Phylogenetic Tree = $\prod_{j=1}^{n} P(\text{site}_j \mid T)$

Here, $P(\text{site}_j \mid T)$ represents the probability of the observed data at site_j given the tree T, and the product is taken over all n sites in the dataset.

The calculation of the total likelihood as a product of individual site probabilities is based on the assumption that each site in the sequence alignment evolves independently.

Character-based phylogenetic inference is really about tree-scoring, not tree-finding .. 🏕️

In phylogenetics, the **log-likelihood** is preferred over the raw **likelihood** for several key reasons:

- **Numerical Stability**: Likelihood values are often extremely small. Taking the logarithm transforms these small products into manageable sums, reducing computational errors.

- **Computational Efficiency**: Logarithms convert products into sums. Sums are easier and faster to compute, especially when differentiating or integrating.

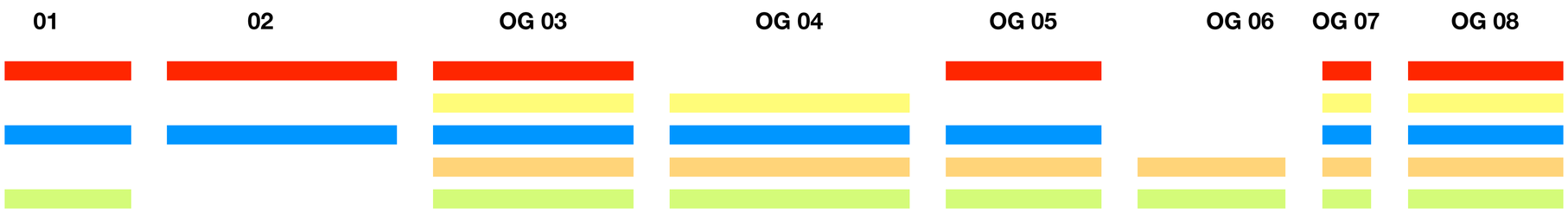
- **Model Comparison**:Log-likelihood values facilitate the use of statistical tests, such as the Likelihood Ratio Test (LRT).

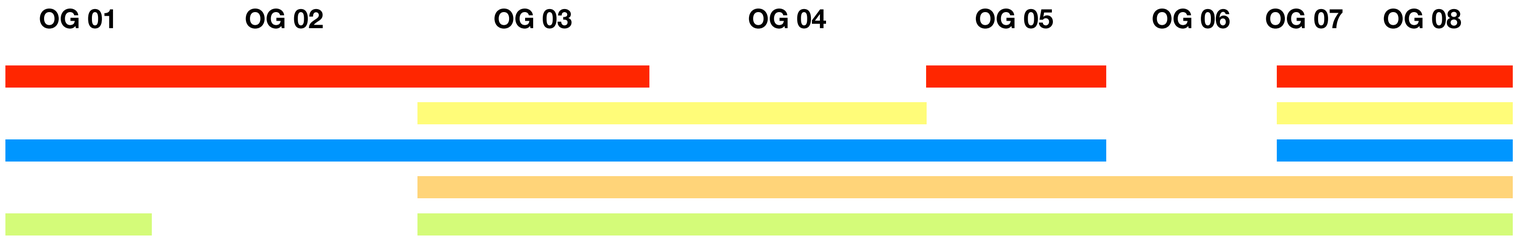
CONCATENATION

What we have seen until now is shown in single alignements …
… but what about the real life scenario?

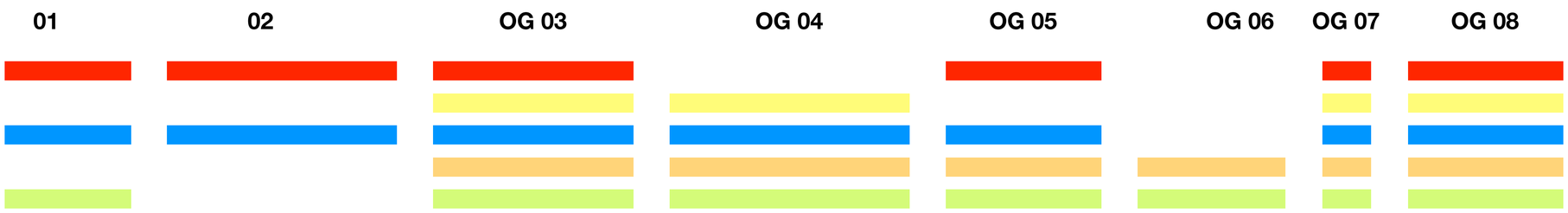
… orthology inference has generated a ton of orthogroups, we have aligned and filtered them …

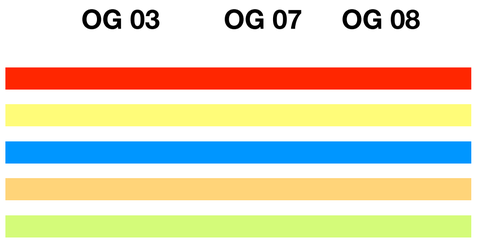
Concatenation - also known as the *supermatrix approach* - involves combining **multiple alignments** (gene or protein) into **a single large dataset** to infer the phylogeny.



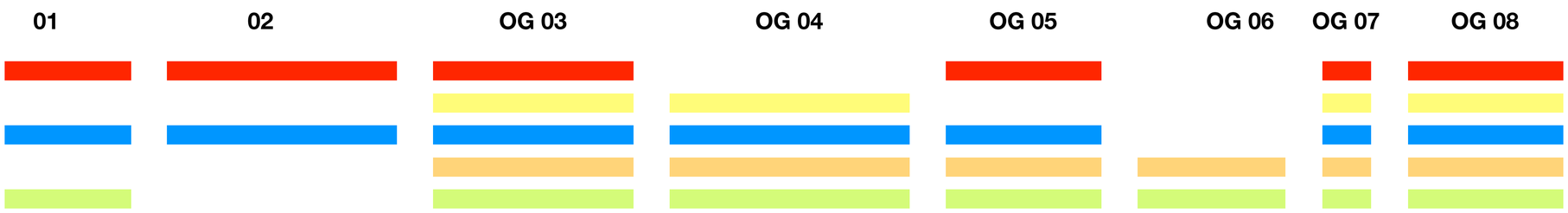


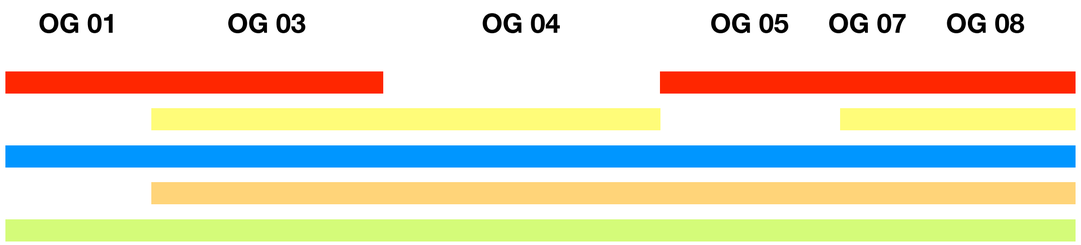
min. matrix occupancy = 00%





min. matrix occupancy = 100%





min. matrix occupancy = 50%

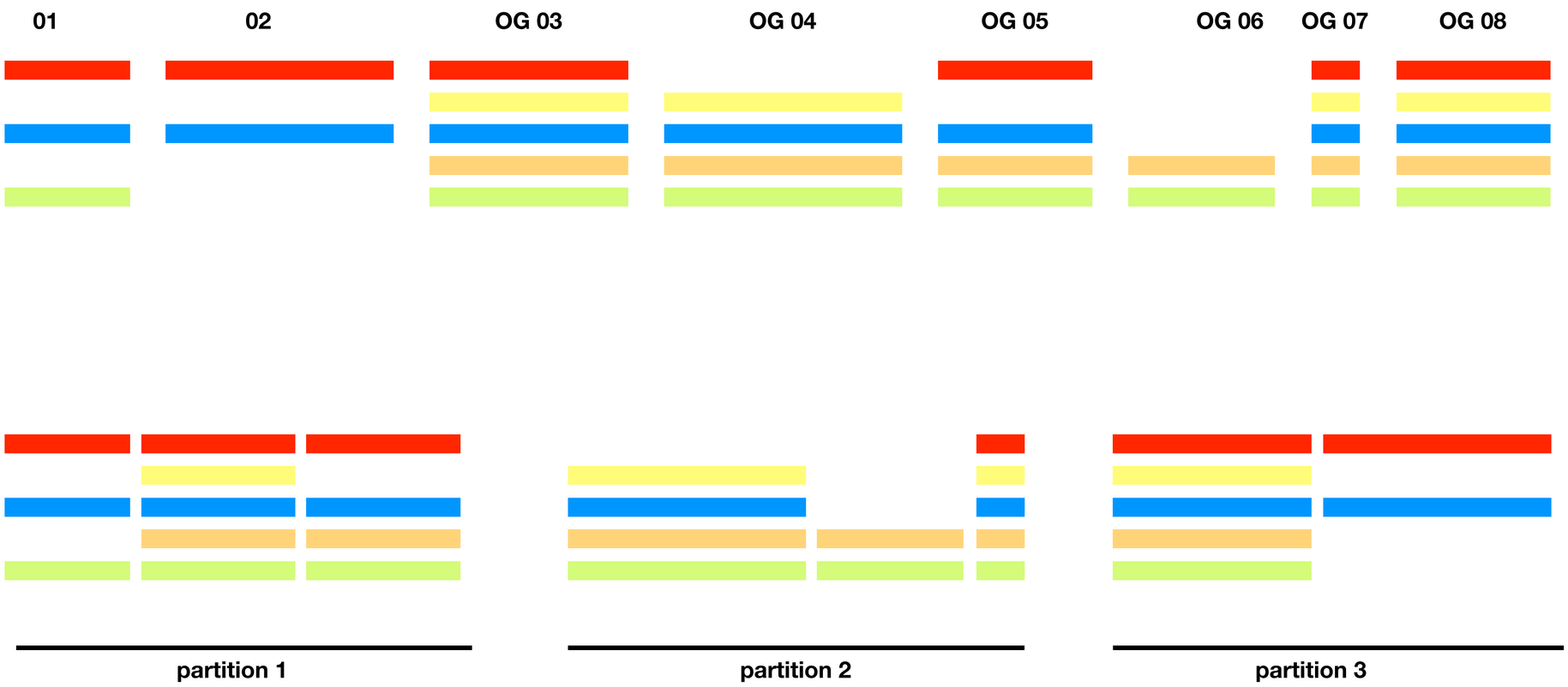
We **concatenate genes** to improve the resolution of phylogenomic analyses. Concatenation increases the strength of phylogenetic signal and reduce stochastic errors.

That is, inaccuracies caused by random sampling effects due to limited sequence data (or/and insufficient genetic variation) leading to misleading topologies and weakly supported relationships.

Challenges of concatenation:

⚠️ **Heterotachy & model violations:** evolutionary rates can vary across genes and taxa.
⚠️ **Incongruence between gene trees:** gene tree conflicts due to incomplete lineage sorting or ..
⚠️ **Partitioning strategy:** decided *a priori*

Alternative approach: coalescent-based methods! We will talk about that in **lesson 13**!


01
02
OG 03
OG 04
OG 05
OG 06
OG 07
OG 08
partition 1
partition 2
partition 3

Why do we merge gene into partitions?

Avoiding over-oarameterization: Assigning a unique model to every possible partition can lead to overfitting, where the model captures noise rather than true evolutionary signals.

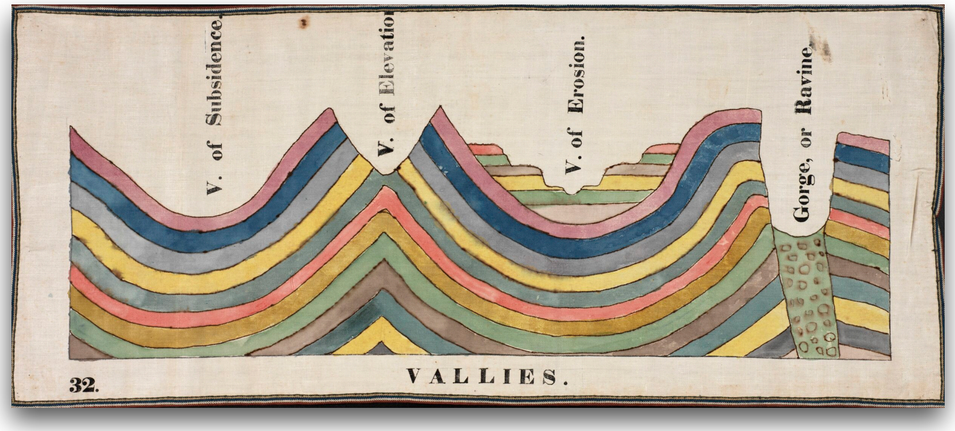
Improving computational efficiency: Fewer partitions reduce the number of parameters to estimate, leading to faster computations and more efficient analyses.

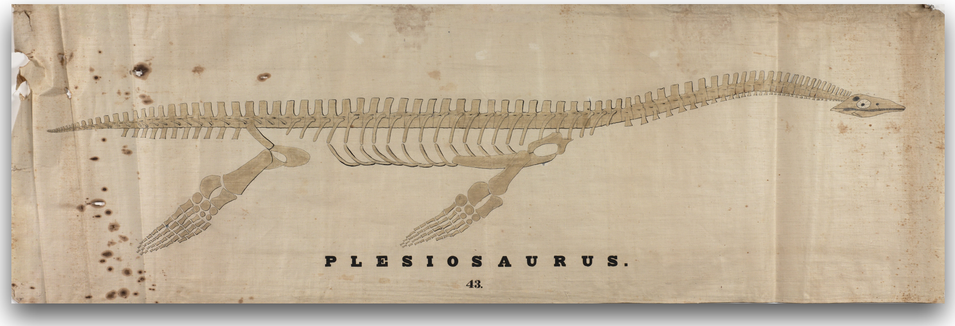
Enhancing statistical power: Combining partitions with similar evolutionary patterns increases the amount of data informing each parameter estimate, resulting in more robust and reliable inferences.

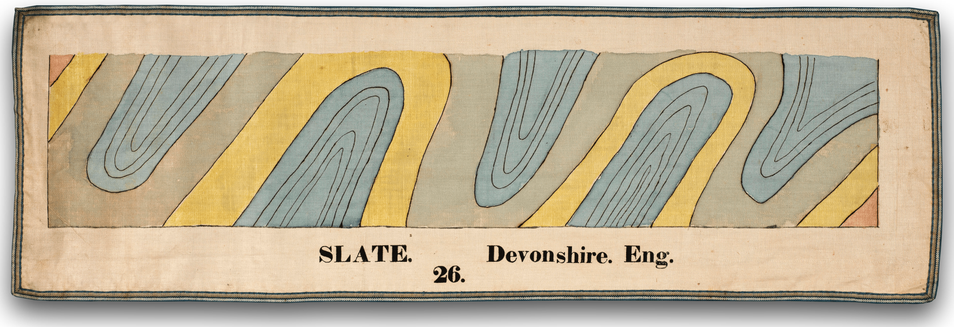
How are partitions merged?

Metrics like the **Akaike Information Criterion** (AIC) or **Bayesian Information Criterion** (BIC) evaluate the trade-off between model fit and complexity. More on that in the practicals …

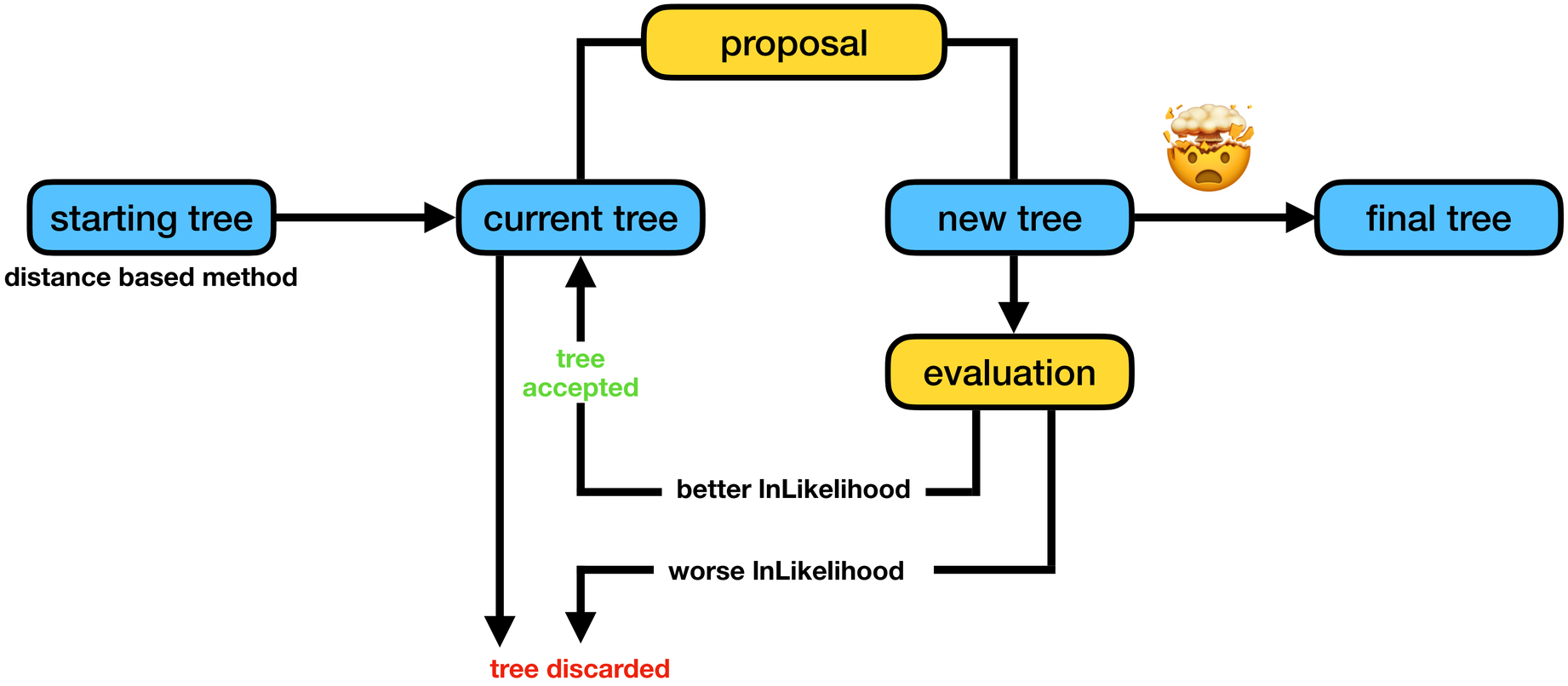
Inadequate partitioning can lead to biased phylogenetic estimates if too few partitions are used (**under-partitioning**) or can overfit the data if too many partitions are used (**over-partitioning)**.







Bayesian Inference (BI)


proposal
starting tree
distance based method
current tree
new tree
final tree
tree accepted
evaluation
better lnLikelihood
worse lnLikelihood
tree discarded

proposal

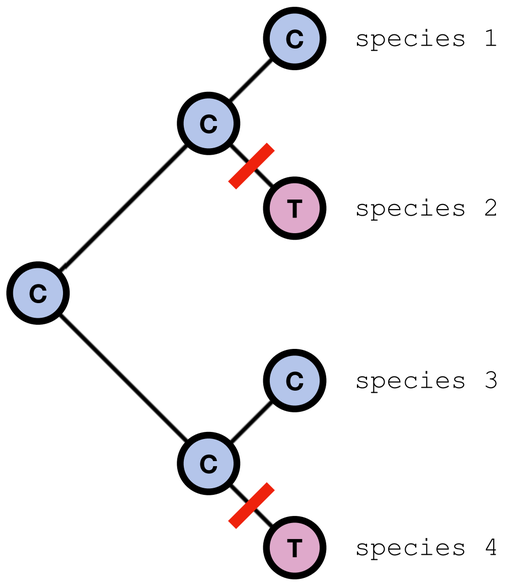
Nearest Neighbour Interchange (NNI) is a local search method that makes small, localized changes to the tree. It is the simplest and fastest tree rearrangement algorithm!

Subtree Pruning and Regrafting (SPR) is a more extensive tree rearrangement method than NNI. It moves entire subtrees rather than just swapping immediate neighbors.

Tree Bisection and Reconnection (TBR) is the most aggressive of these tree rearrangement methods and explores a much larger space of possible tree topologies. It has the lesser chance of local optima!

evaluation

```
                1 2 3 4 5 6 7 8
species 1       A G G A C C G A
species 2       A C T T T C G G
species 3       A G G A C C T T
species 4       A C G T T C T T
                        *
```



$$Q = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{G} & \mathbf{C} & \mathbf{T} \end{array} \\ \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix} \end{array} \begin{array}{c} \\ \mathbf{A} \\ \mathbf{G} \\ \mathbf{C} \\ \mathbf{T} \end{array}$$

$P_{CT} = r_{CT}\pi_C$

$P_{site4} = P_{CT} \times P_{CT} \times P_C \times P_C \times P_C \times P_C$

Maximum Likelihood P(D|M)
probability of observed data given a model

Bayesian Inference P(M|D)
probability the model is correct given the data

… but why Bayesian Inference ?!

We refer to Bayesian Inference due to the **Bayes Theorem**.

It describes how to **update our beliefs based on new evidence** …

… and it provides a mathematical framework for conditional probability!

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

- $P(A|B)$ is how often A happens *given that B happens*
- $P(B|A)$ is how often B happens *given that A happens*
- $P(A)$ is how likely is A to happen on its own
- $P(B)$ is how likely is B to happen on its own

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

- $P(A|B)$ ***- the posterior probability***
- $P(B|A)$ ***- the likelihood***
- $P(A)$ ***- the prior probability***
- $P(B)$ ***- the marginal likelihood***

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example:

- **fires are rare: P (Fire) = 0.01**
- **smoke is fairly common: P (Smoke) = 0.1**
- **dangerous fires make smoke: P (Smoke I Fire) = 0.9**

What's the probability of fire when I see smoke?! **P (Fire I Smoke) = ?!**

… how are these concepts releted to phylogeny?!

$$P(Tree|Data) = \frac{P(Data|Tree)P(Tree)}{P(Data)}$$

Where:

- $P(Tree|Data)$ probability of a given tree given the data **- *posterior probability***
- $P(Data|Tree)$ probability of the data given that the tree is correct **- *likelihood***
- $P(Tree)$ the prior belief about which trees are more likely **- *prior probability***
- $P(Data)$ ***the marginal likelihood …*** we are not getting deep into that 🥲

MCMC is a class of algorithms used to approximate complex posterior distributions where direct sampling is challenging by generating a sequence of samples. Composed by:

- **Markov Chain:** a sequence of possible events where the probability of each event depends only on the state attained in the previous event.

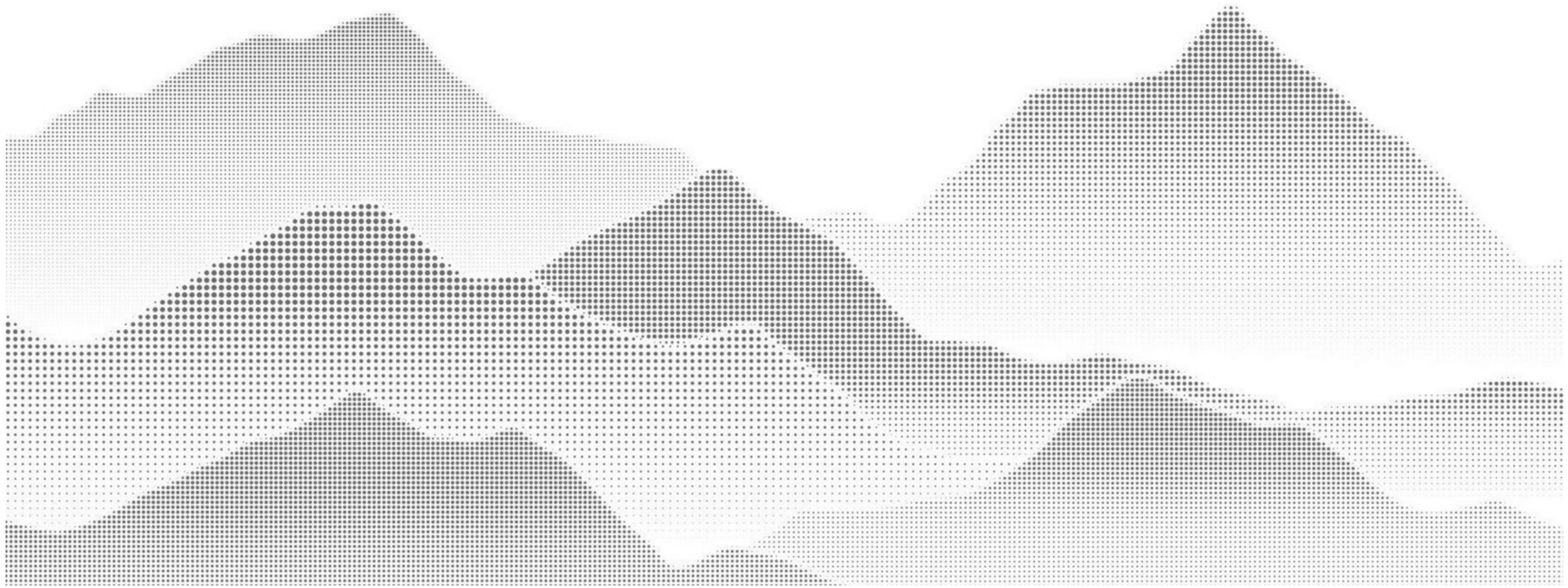
- **Monte Carlo:** random sampling to obtain numerical results, used to estimate functions and mimic the operations of complex systems.

How MCMC works:

- **Initialize:** start with an arbitrary point in the parameter space.
- **Propose:** generate a new point based on a proposal distribution.
- **Evaluate:** decide whether to accept or reject the new point
- **Iterate:** repeat proposal and acceptance to build a chain of samples

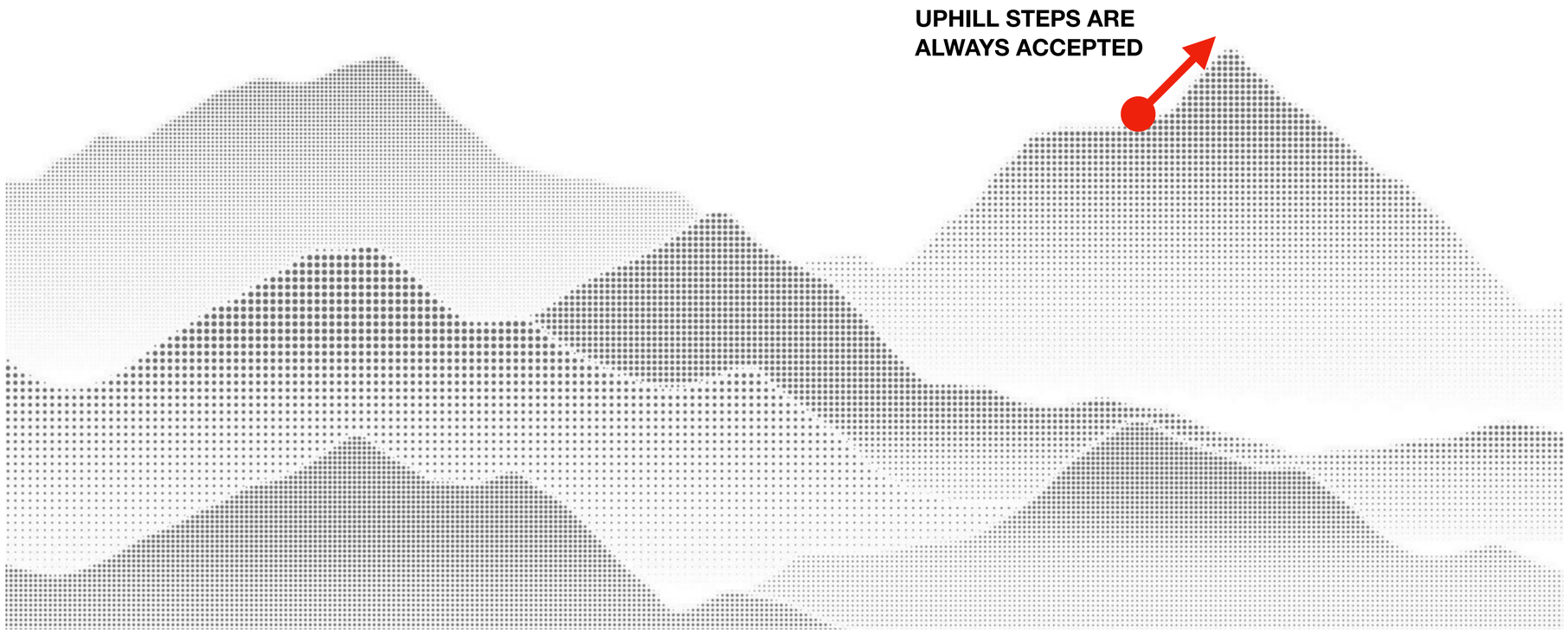
Metropolis-Hastings *criterion* takes a step if a random num. between 0 and 1 is less or equal to R.

R = proposed state / current state



Metropolis-Hastings *criterion* takes a step if a random num. between 0 and 1 is less or equal to R.

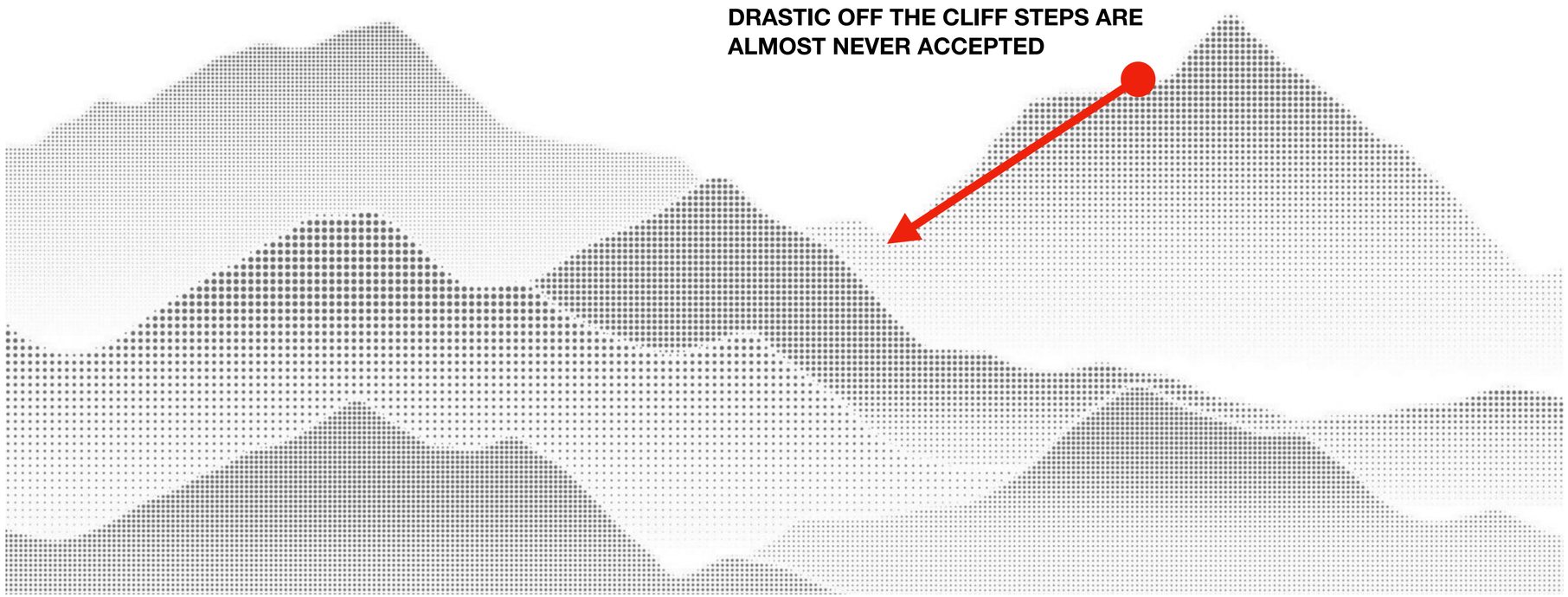
- **Moves to higher-probability states always accepted!**
  current state = 100; proposed state = 120; R = 120 / 100 = 1.20



Metropolis-Hastings *criterion* takes a step if a random num. between 0 and 1 is less or equal to R.

- **Moves to higher-probability states always accepted!**
  current state = 100; proposed state = 120; R = 120 / 100 = 1.20

- **Moves to very low probability states nearly always rejected!**
  current state = 115; proposed state = 20; R = 20 / 115 = 0.17

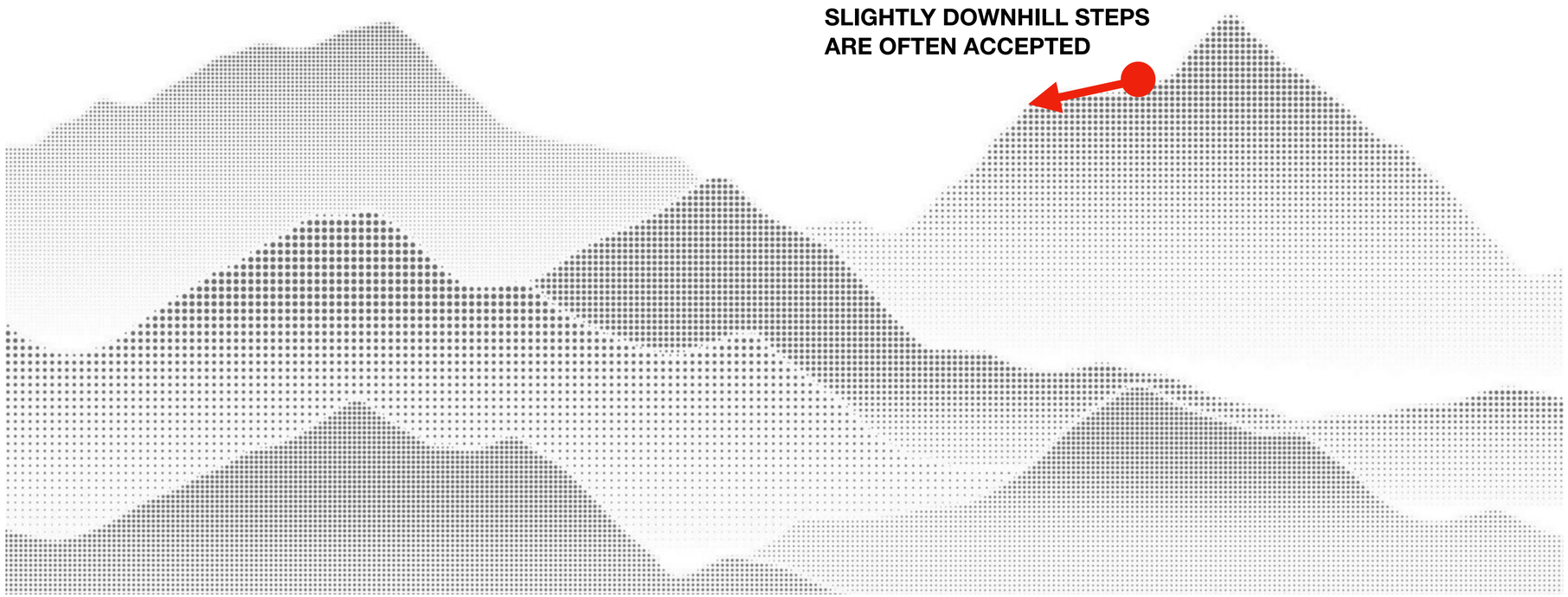


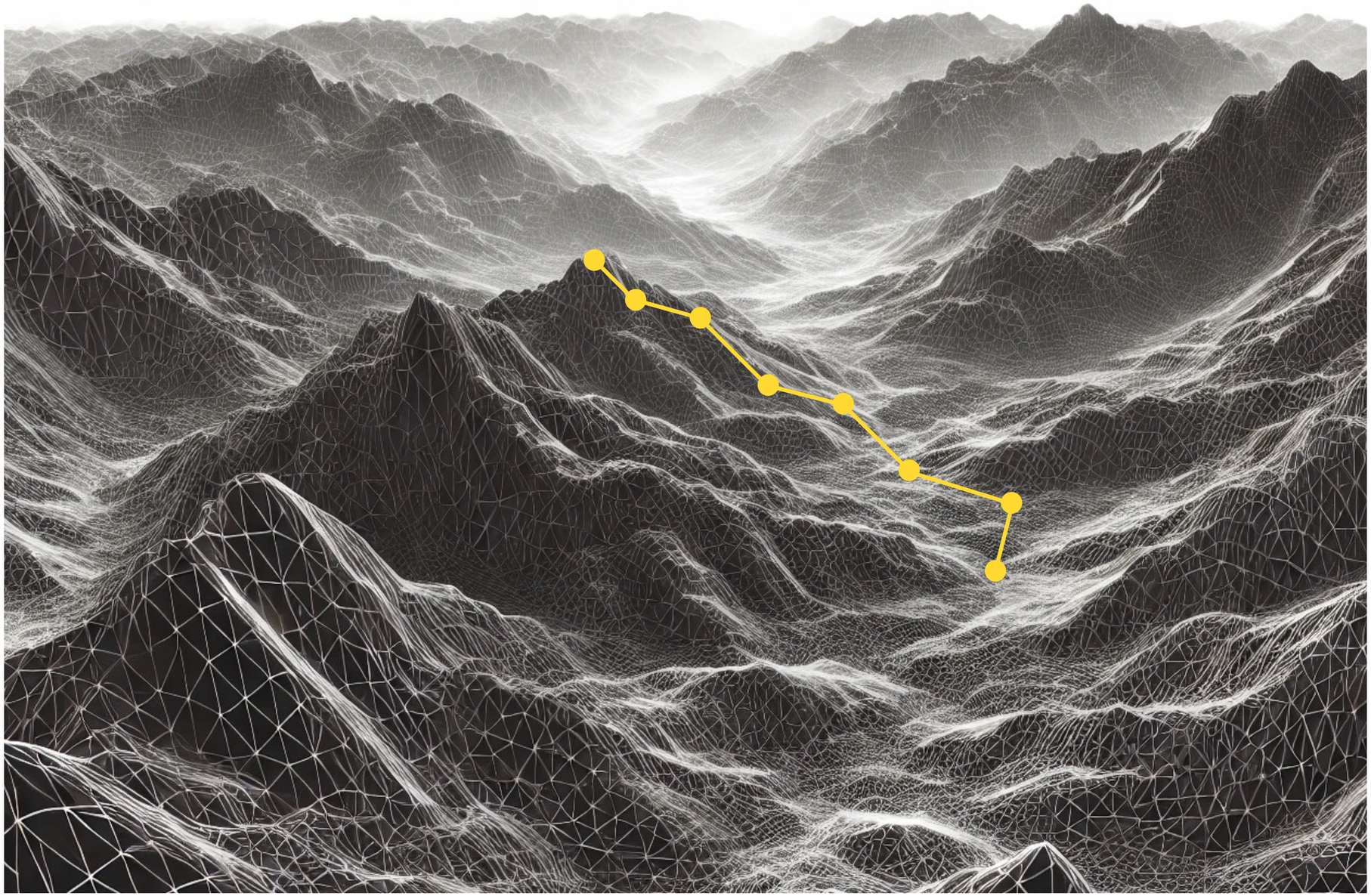
Metropolis-Hastings *criterion* takes a step if a random num. between 0 and 1 is less or equal to R.

- **Moves to higher-probability states always accepted!**
  current state = 100; proposed state = 120; R = 120 / 100 = 1.20

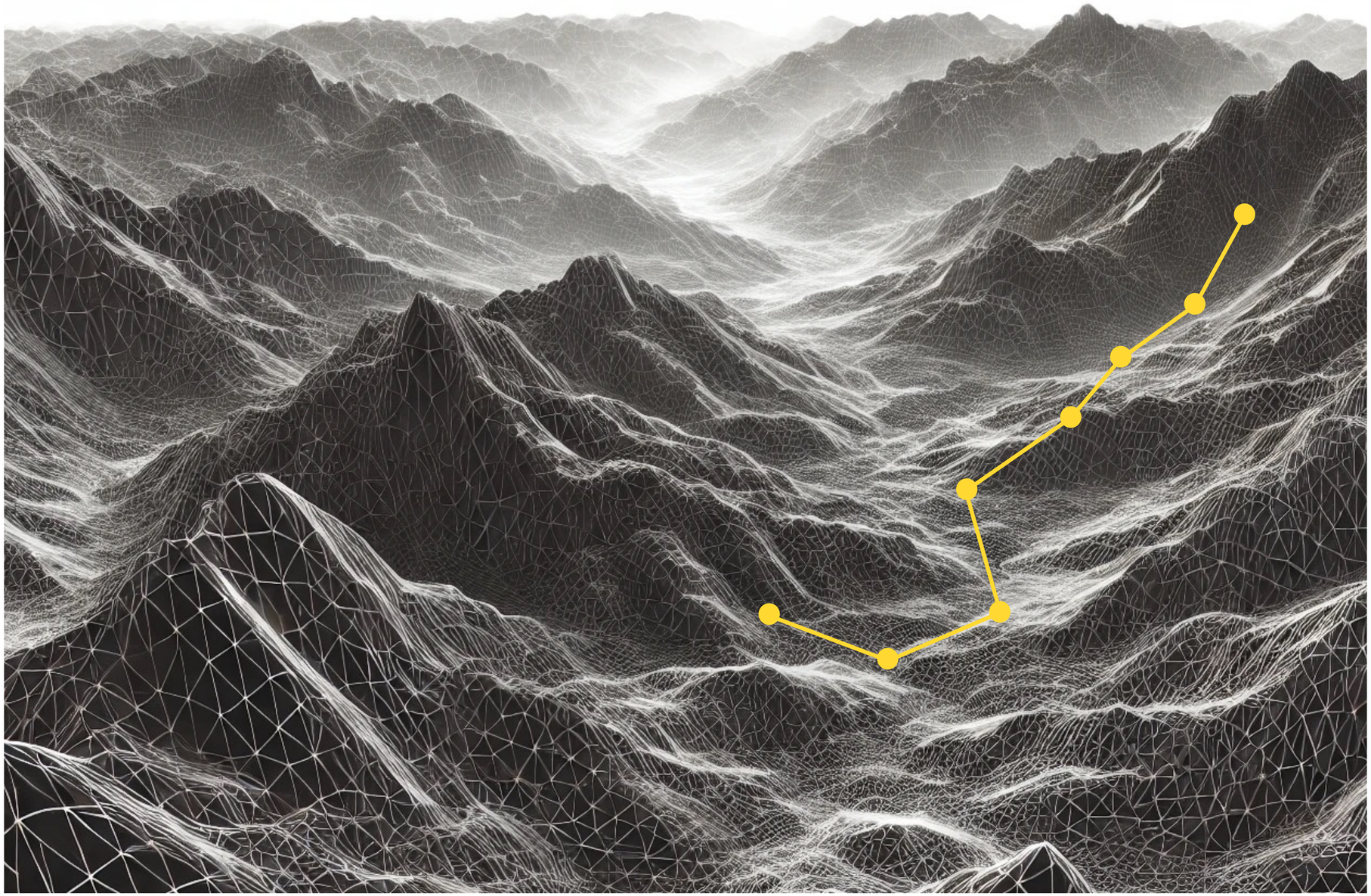
- **Moves to very low probability states nearly always rejected!**
  current state = 115; proposed state = 20; R = 20 / 115 = 0.17

- **Moves to slightly low probability states nearly always accepted!**
  current state = 102; proposed state = 93; R = 93 / 102 = 0,92

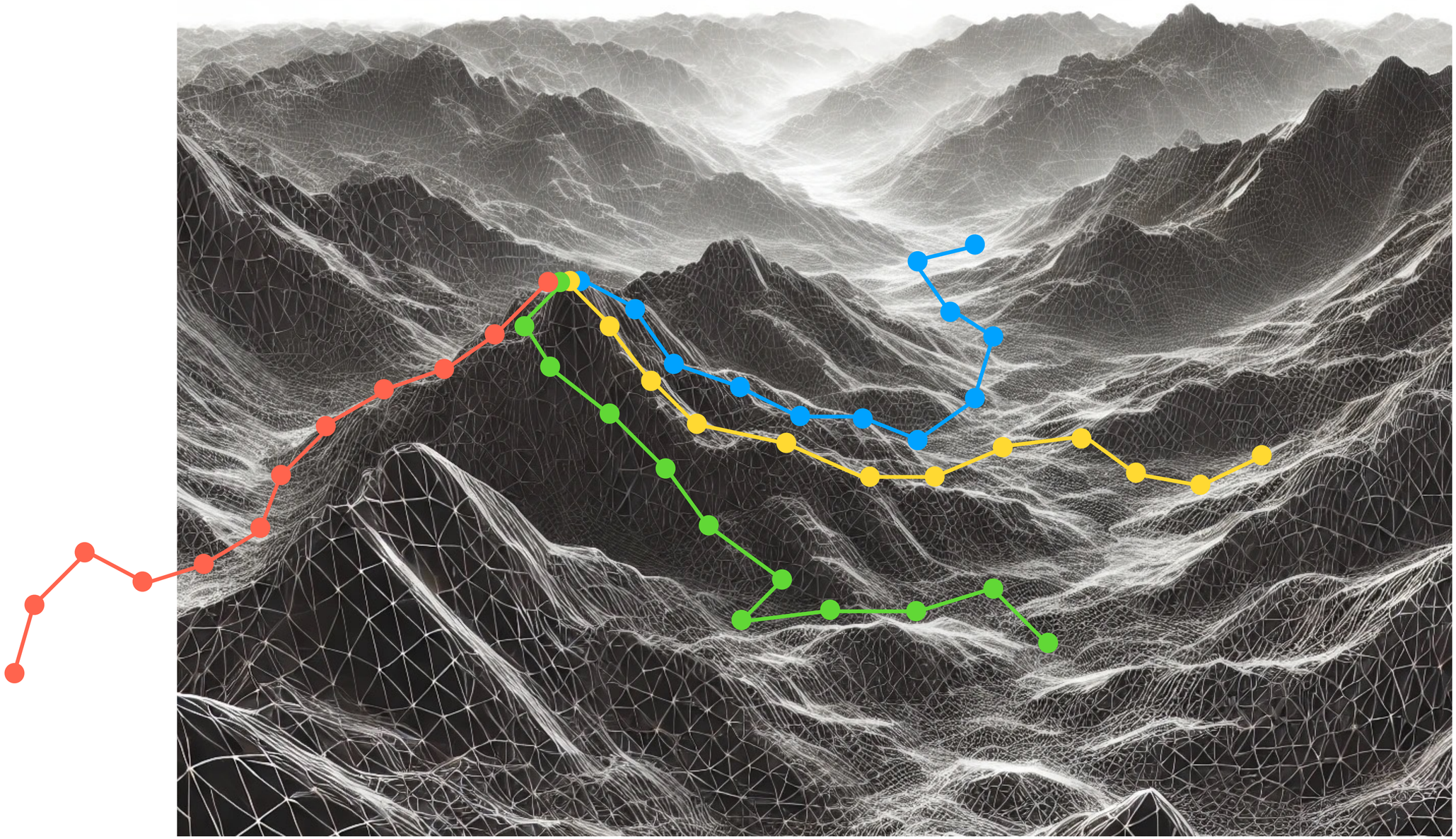




… are we on the highest peak?!



… are we still ascending?!



Reproducibility of the results across independent runs started from different initial conditions.

What to expect during MCMC?

Log-Likelihood:

- Starts increasing rapidly as the chain explores better-fitting trees.
- Eventually stabilizes, indicating a *plateau*.

Tree Topology:

- Initially fluctuates as the chain explores different possibilities.
- Over time, a consistent topology emerges.

Model Parameters:

- Parameters fluctuate widely at first.
- As sampling continues, they settle within a narrower range.

What is at the peak?!

stationary / target distribution

The goal of MCMC is to construct a chain …
whose **stationary distribution** matches the desired **target distribution**!

By running the chain for a sufficient number of iterations,
the samples generated will approximate this target distribution,
allowing for **accurate estimation of statistical properties**.

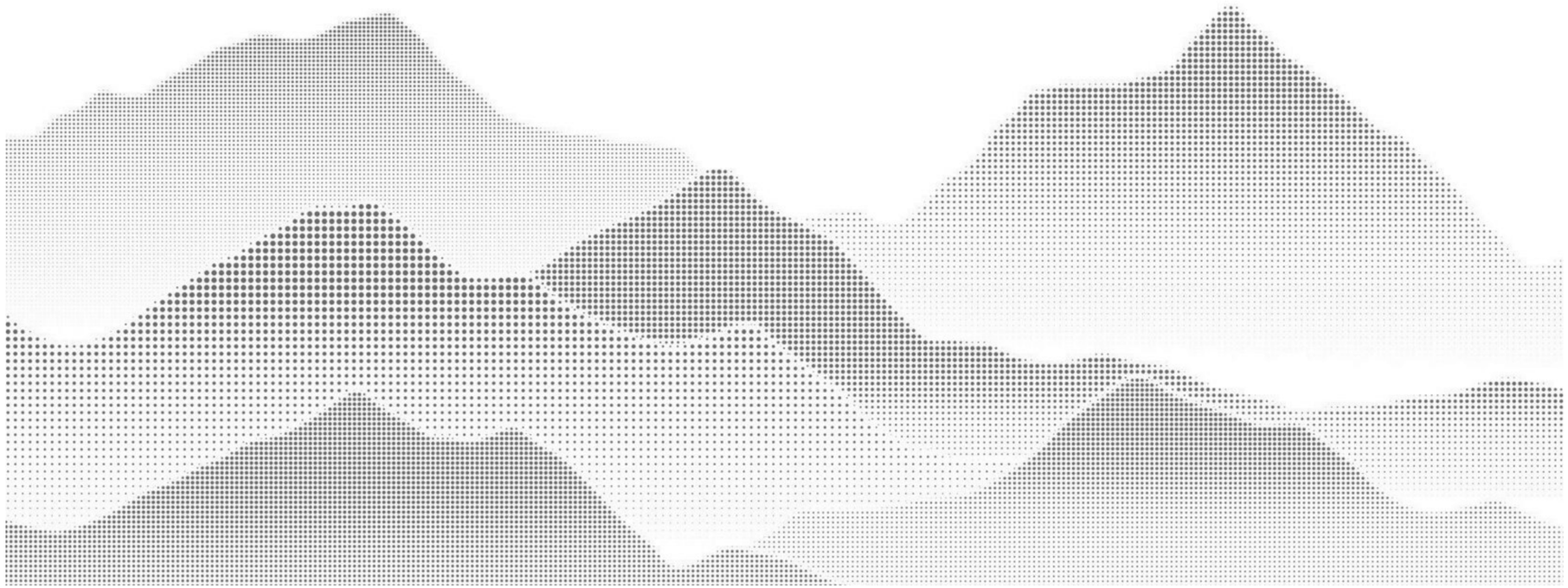
MCMC can get stuck in **local optima**, especially in complex phylogenetic landscapes. MCMC actually involves running **multiple chains** at the same time:

Cold Chain: represents the standard MCMC chain.

Heated Chains: explore a flattened posterior for broader exploration of the parameter space.

Hot escape local optima by exploring unlikely regions, **cold** refine high-probability trees.

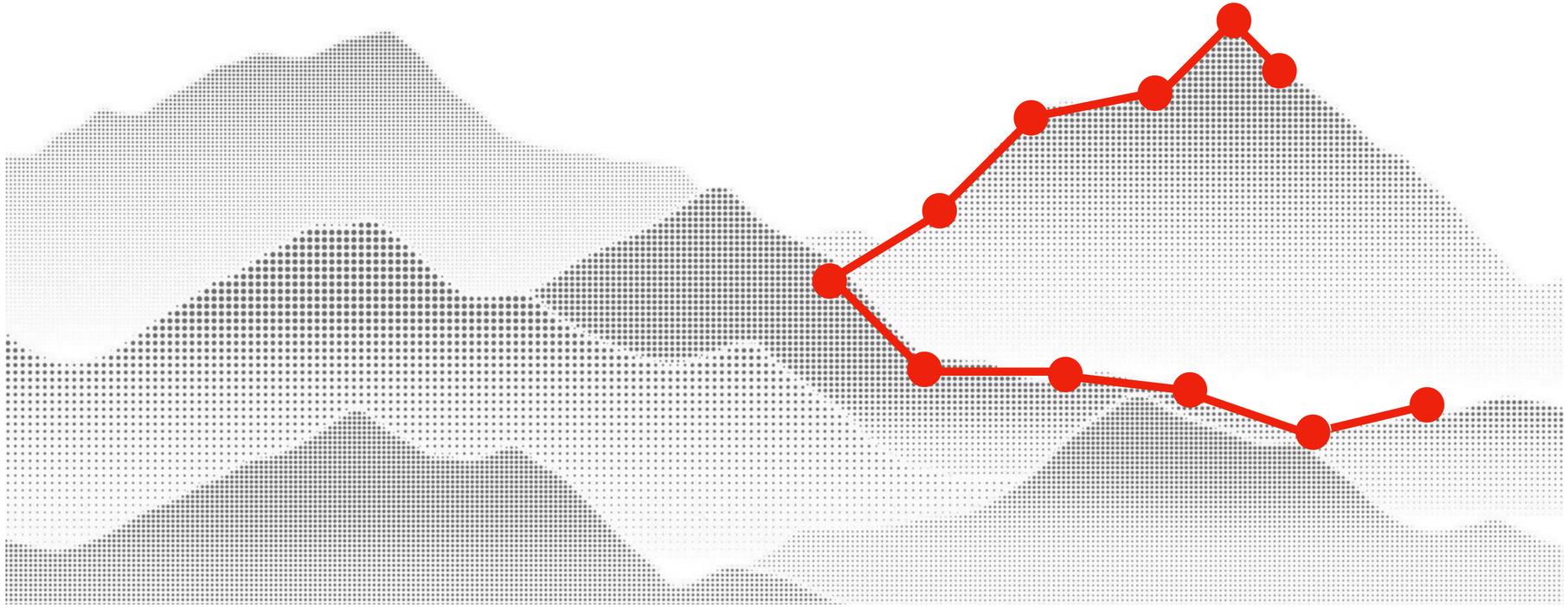
The likelihood function is raised to the power of 1/T, where T is the **temperature**. When T>1 the likelihood values become less extreme, allowing more movement between different trees.



Burn-in refers to the initial set of iterations where the chain may not have yet reached the target distribution.

These early samples do not represent the target / stationary distribution, which providing accurate estimates of the parameters of interest.

Typically, 10-25% of the MCMC samples are removed, but the exact percentage depends on convergence **diagnostics**.



Among the post-inference diagnostics **Effective Sample Size (ESS)** are the more important.

ESS estimates the **number of independent samples** in an MCMC chain.

Due to **autocorrelation**, MCMC samples are not fully independent.

Autocorrelation means that **consecutive samples are similar**.

Basically, we can think to it as a chain that is still going uphill.

If the ESS of a parameter is small, then the **estimate of its posterior distribution will be poor**.

What size ESS is adequate? Basically, the larger the better. ESSs < 100 are considered weak and having ESSs > 200 would be better. On the other hand, chasing ESSs > 10000 may simply be a waste of computational resources.

How can I make ESS better? Improved proposal strategies can enhance the exploration of the parameter space, reducing autocorrelation. Increased MCMC run length can help achieve a sufficient number of independent samples, mitigating the effects of autocorrelation.

Should all parameters have high ESS? Not necessarily. While low ESS values can indicate poor mixing, if parameters of interest have adequate ESS, occasional low ESS in other parameters might be acceptable. However, key metrics like likelihoods should maintain high ESS for robust analysis.

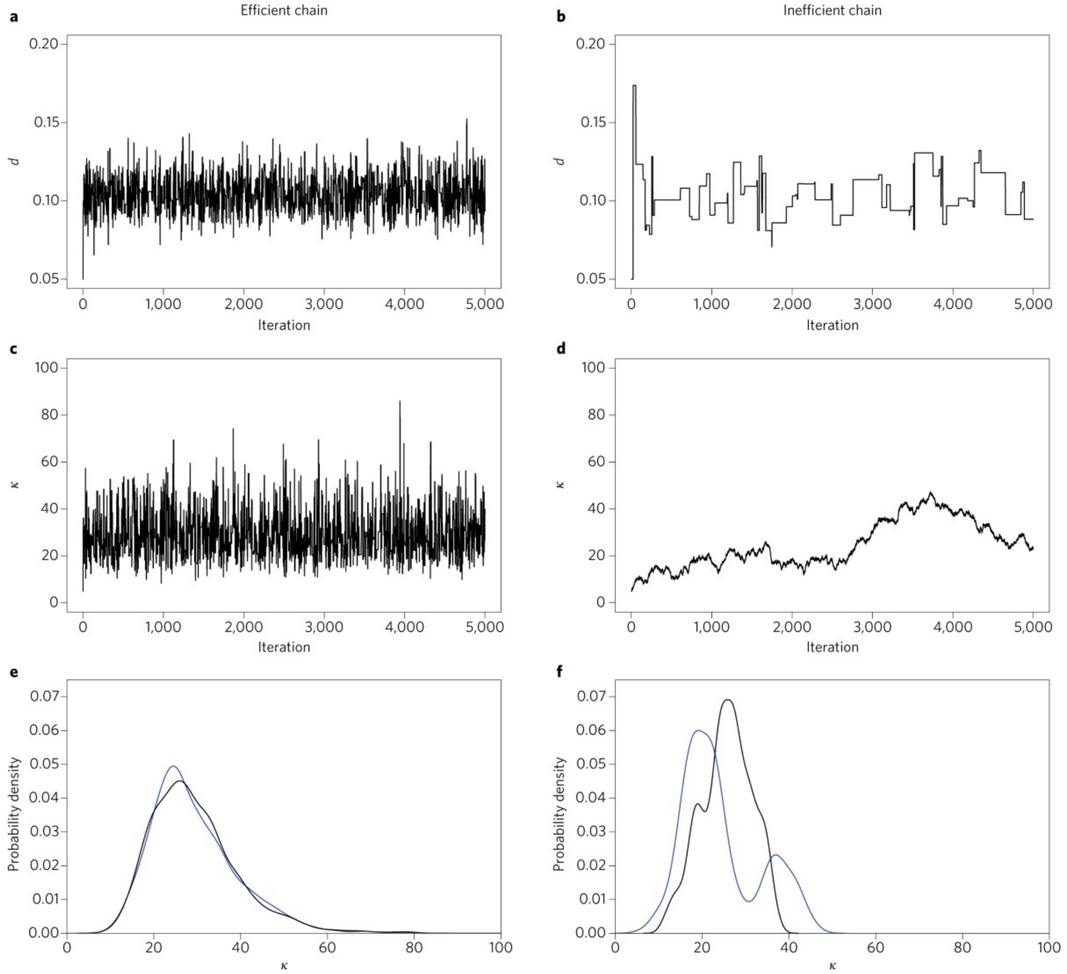


Fig. 2 | Trace plots and histograms for *d* and *κ* from sampling a posterior distribution using efficient and inefficient MCMC chains. The posterior distribution used is shown in Fig. 1c. **a**,**c**, Trace plots of *d* (**a**) and *κ* (**c**) for an efficient chain with good mixing. The window sizes are $w_d = 0.12$ and $w_\kappa = 180$, with acceptance proportions $P_{jump} = 30.4\%$ for *d* and 29.8% for *κ*, achieving Eff = 23% for *d* and 20% for *κ*. **b**,**d**, The corresponding trace plots for an inefficient chain with poor mixing, with $w_d = 5$ and $w_\kappa = 1$. In **b** the window for *d* is too wide, and most proposals are rejected ($P_{jump} = 1.5\%$), so the chain is often stuck at the same value for many iterations, leading to poor mixing with Eff = 1.79%. In **d** the window for *κ* is too small, so most of the proposals are accepted (with $P_{jump} = 98.6\%$), but the chain makes small baby steps and is very slow in traversing the posterior parameter space, with Eff = 1.28%. **e**,**f**, Histograms of *κ* for two runs (shown by the black and blue lines) of the efficient (**e**) and inefficient (**f**) chains (sample size $n = 10{,}000$). The posterior mean (and standard deviation) calculated using a very long run of the efficient chain is 0.104 (0.0114) for *d*, and 29.2 (10.0) for *κ*.

… we end up with a ton of trees! 🌳🌴🌲

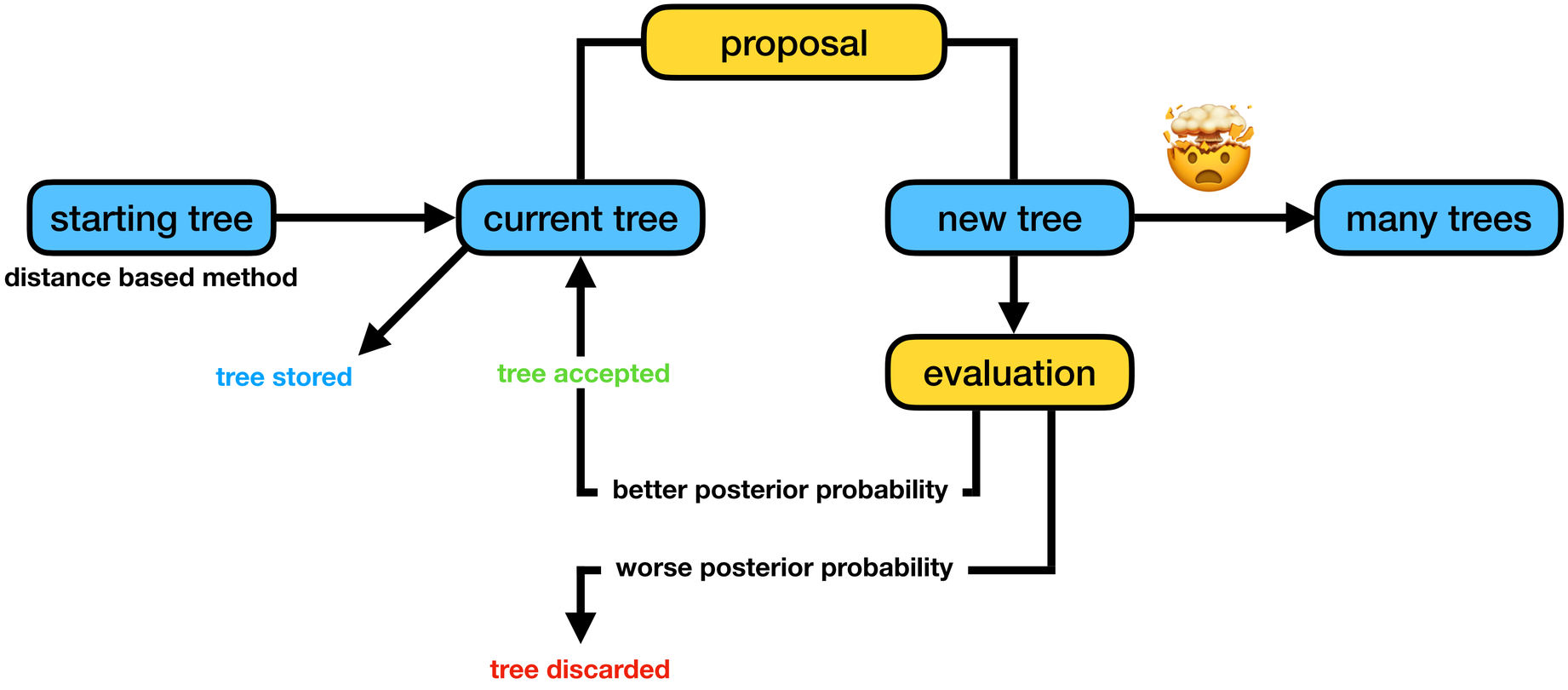
… and then we need to summarize the posterior distribution of tree.

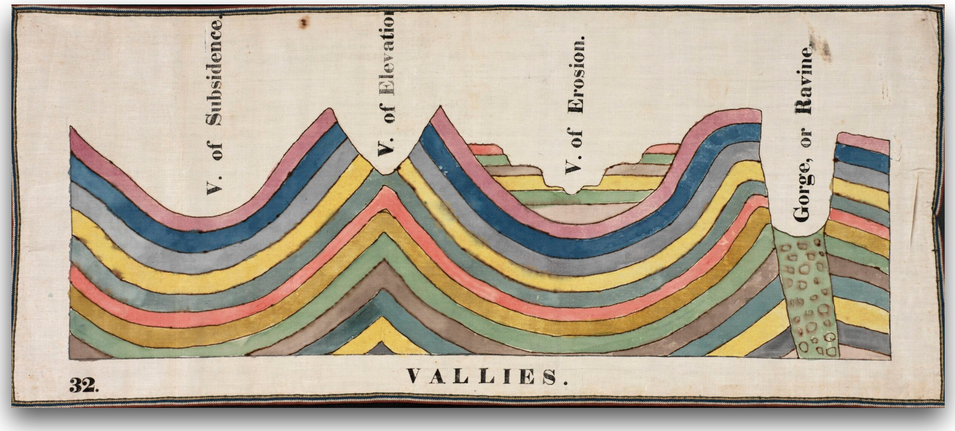
Majority-Rule Consensus Tree: this method identifies clades that appear in more than 50% of the posterior trees. It provides a summary of well-supported clades but may result in a tree topology that was not explicitly sampled during the analysis.

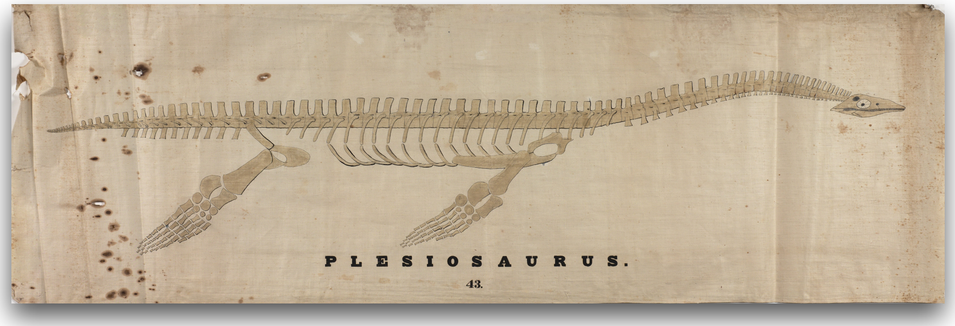
Maximum Clade Credibility (MCC) Tree: the single tree from the posterior distribution that has the highest product of posterior probabilities across all its clades and where the combined support for all its clades is maximized. A poorly supported clade can significantly reduce the tree credibility.

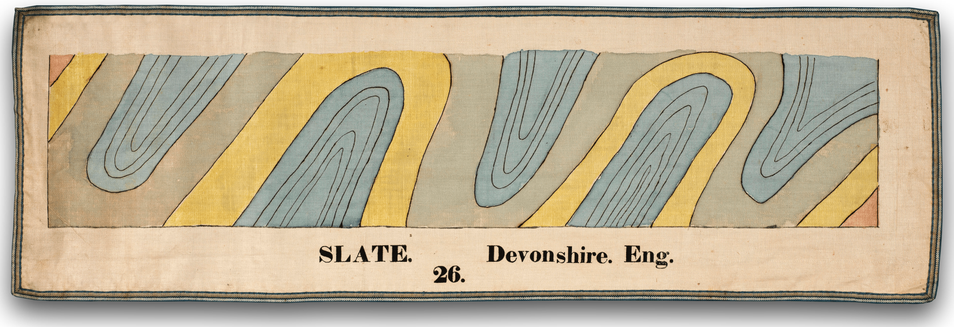
Maximum Sum of Clade Credibilities (MSCC) Tree: the tree with the highest sum of posterior probabilities for its clades. Less sensitive to individual low-probability clades than MCC. However, it favor trees with more clades, as each additional clade contributes positively to the total sum.

Median Clade Credibility Tree: aims to minimize the variance in branch lengths across the posterior distribution and thus provides a summary that reflects the central tendency of branch lengths, which can be particularly useful when branch length estimates are of primary interest.


proposal
starting tree
current tree
new tree
many trees
distance based method
tree stored
tree accepted
evaluation
better posterior probability
worse posterior probability
tree discarded







Support Metrics

Phylogenetic trees are just …

Phylogenetic trees are just … hypotheses!

It is important to know how much we can be **confident** in our findings…

Teese are the metrics we are going to explore:

- nonparametric & parametric bootstrap (**BPs**)
- Tranfer Bootstrap Expectation (**TBE**)
- parametric & nonparametric jackknife
- SH-like approximate Likelihood Ratio Test (**SHaLRT**)
- genes & sites Concordance Factors (**gCFs** & **sCFs**)
- Posterior Probabilities (**PP**)
- Quartet Support (**QS**) … will see that in **lesson 13**!

non-parametric bootstrapping

Is a general statistical method for estimating the parameters of an unknown probability distribution by resampling from a sample that was drawn from this distribution.

Transferred to phylogenetics by **Felsenstein** in this paper from 1985.

This is the process step-by-step:

- start with an alignment of length **n**
- generate a bootstrapped alignment:
  - randomly sample **n** columns with replacement
  - arrange the new columns into a bootstrapped alignment of length **n**
- build a tree from the bootstrapped alignment with the same method as for the original alignment
- repeat the process many times and store all resulting bootstrap trees
- the **bootstrap proportion (BPs)** is the percentage of trees containing a given bipartition

It reflects the consistence with the sequence data support that branch.

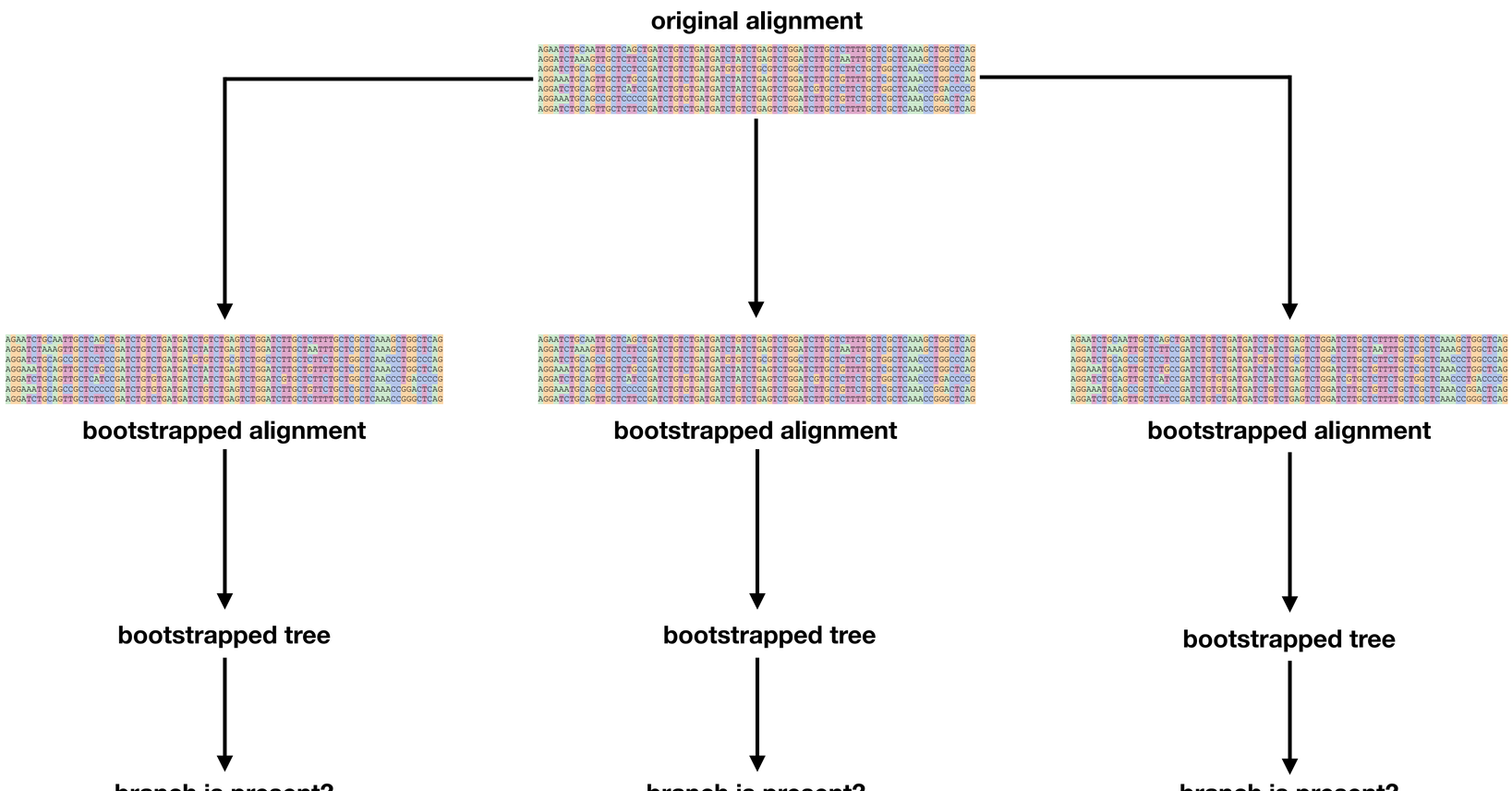
BPs are typically shown on the "*best tree*" or the majority-rule consensus of bootstrap trees.

```
species 1   AGAATCTGCAATTGCTCAGCTGATCTGTCTGATGATCTGTCTGAGTCTGGATCTTGCTCTTTTGCTCGCTCAAAGCTGGCTCAG
species 2   AGGATCTAAAGTTGCTCTTCCGATCTGTCTGATGATCTATCTGAGTCTGGATCTTGCTAATTTGCTCGCTCAAAGCTGGCTCAG
species 3   AGGATCTGCAGCCGCTCCTCCGATCTGTCTGATGATGTGTCTGCGTCTGGCTCTTGCTCTTCTGCTGGCTCAACCCTGGCCCAG
species 4   cGGAAATGCAGTTGCTCTGCCGATCTGTCTGATGATCTATCTGAGTCTGGATCTTGCTGTTTTGCTCGCTCAAACCTGGCTCAG
species 5   AGGATCTGCAGTTGCTCATCCGATCTGTGTGATGATCTATCTGAGTCTGGATCGTGCTCTTCTGCTGGCTCAACCCTGACCCCG
species 6   AGGAAATGCAGCCGCTCCCCCGATCTGTGTGATGATCTGTCTGAGTCTGGATCTTGCTGTTCTGCTCGCTCAAACCGGACTCAG
species 7   AGGATCTGCAGTTGCTCTTCCGATCTGTCTGATGATCTGTCTGAGTCTGGATCTTGCTCTTTTGCTCGCTCAAACCGGGCTCAG
```

```
species 1   GAACAATGGTA…                                                    …AGTAAA
species 2   GTACTATGTTT…                                                    …TGTTAT
species 3   GCACCCTGTTC…                                                    …CGTCAC
species 4   GTACTATGGTT…                                                    …TGTTAT
species 5   GAAGACTGTGA…                                                    …AGTAAA
species 6   GCAGCAGGCTC…                                                    …CGGCAC
species 7   GTACTAGGTTT…                                                    …TGGTAT
```

Sampling with replacement!

Is the same information present in this bootstrapped alignment?



node	branch is present?	branch is present?	branch is present?	BPs
1	Y	Y	Y	100%
2	Y	N	N	33%
3	Y	Y	Y	100%
4	N	N	N	0%
5	N	Y	Y	66%
..	…	…	…	…
last	Y	Y	Y	100%

parametric bootstrapping

A **model-based approach** to estimate support using simulated data.

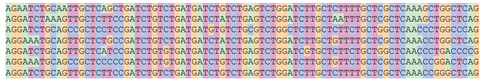
This is the process step-by-step:

- estimate from the original alignment **X**:
  - model of substitution
  - a phylogenetic tree **T**
- use the model and tree T to simulate replicate alignments $\mathbf{X}_i$
- phylogenetic trees $\mathbf{T}_i$ are inferred for each of the alignments $\mathbf{X}_i$
- branch support values are derived - as in non-parametric bootstrapping - by computing the percentage with which a certain branch occurs in the set of generated trees $\mathbf{T}_i$

⚠️ Important:

- parametric bootstrap values **depend heavily on the model** estimated from the original data
- if the model is misspecified, support values may be biased or misleading

original alignment



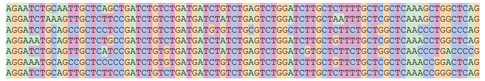
↓

estimate model and tree

↓

simulation of alignments

↓

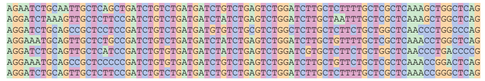


simulated alignment

↓

bootstrapped tree

↓

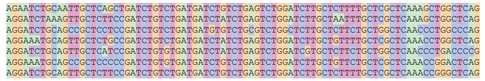


simulated alignment

↓

bootstrapped tree

↓



simulated alignment

↓

bootstrapped tree

↓

node	branch is present?	branch is present?	branch is present?	BPs
1	Y	Y	Y	100%
2	Y	N	N	33%
3	Y	Y	Y	100%
4	N	N	N	0%
5	N	Y	Y	66%
..	…	…	…	…
last	Y	Y	Y	100%

transfer bootstrap expectation (TBE)

The bootstrap is used extensively 🤩 but scaling to thousands of tips introduces new challenges 🤯. Felsenstein BPs yield **low support on deep branches** and is **not robust to taxon sampling** noise.

This is due to its **binary scoring** system: a branch is either **present** or **absent** in a replicate tree.

Lemoine et al. 2018 introduced TBE which uses a **transfer index based on gradual distances** of a branch in the replicate tree to the reference branch, capturing partial support in bootstap trees.

In detail:

- for a reference clade **R**, split the tips into two complementary groups.
- for each bipartition **B** in the bootstrap tree, compute the transfer distance: that is the minimum number of tips that need to be moved from one side to the other so that **B** matches **R**
- select the branch with the smallest transfer distance - i.e. the best-matching clade
- compute the transfer index from this match: $\text{Transfer Index} = 1 - \frac{\text{Transfer Distance}}{\min(|R|, |\neg R|)}$
- averaging the transfer indices across all replicates gives the TBE value for the clade **R**

jacknife

A resampling method related to BPs but uses **deletion rather than resampling** with replacement.

This is the process step-by-step:

- generate a jacknifed alignment from the starting alignment
  - randomly delete a fixed proportion of alignment columns
  - typically, 50% of sites are removed per replicate as im Felsenstein 1985
  - equivalent to drawing n/2 sites without replacement from an alignment of length n
- build a tree from the jacknifed alignment with the same method as for the original alignment
- repeat the process many times and store all resulting jacknife trees
- the **jacknife** is the percentage of trees containing a given bipartition.

It reflects the consistence with the sequence data support that branch.

Jacknife is typically shown on the "*best tree*" or the majority-rule consensus of bootstrap trees.

… an parametric version of the jacknife is also possible, but not shown!

```
species 1   AGAATCTGCAATTGCTCAGCTGATCTGTCTGATGATCTGTCTGAGTCTGGATCTTGCTCTTTTGCTCGCTCAAAGCTGGCTCAG
species 2   AGGATCTAAAGTTGCTCTTCCGATCTGTCTGATGATCTATCTGAGTCTGGATCTTGCTAATTTGCTCGCTCAAAGCTGGCTCAG
species 3   AGGATCTGCAGCCGCTCCTCCGATCTGTCTGATGATGTGTCTGCGTCTGGCTCTTGCTCTTCTGCTGGCTCAACCCTGGCCCAG
species 4   cGGAAATGCAGTTGCTCTGCCGATCTGTCTGATGATCTATCTGAGTCTGGATCTTGCTGTTTTGCTCGCTCAAACCTGGCTCAG
species 5   AGGATCTGCAGTTGCTCATCCGATCTGTGTGATGATCTATCTGAGTCTGGATCGTGCTCTTCTGCTGGCTCAACCCTGACCCCG
species 6   AGGAAATGCAGCCGCTCCCCCGATCTGTGTGATGATCTGTCTGAGTCTGGATCTTGCTGTTCTGCTCGCTCAAACCGGACTCAG
species 7   AGGATCTGCAGTTGCTCTTCCGATCTGTCTGATGATCTGTCTGAGTCTGGATCTTGCTCTTTTGCTCGCTCAAACCGGGCTCAG
```

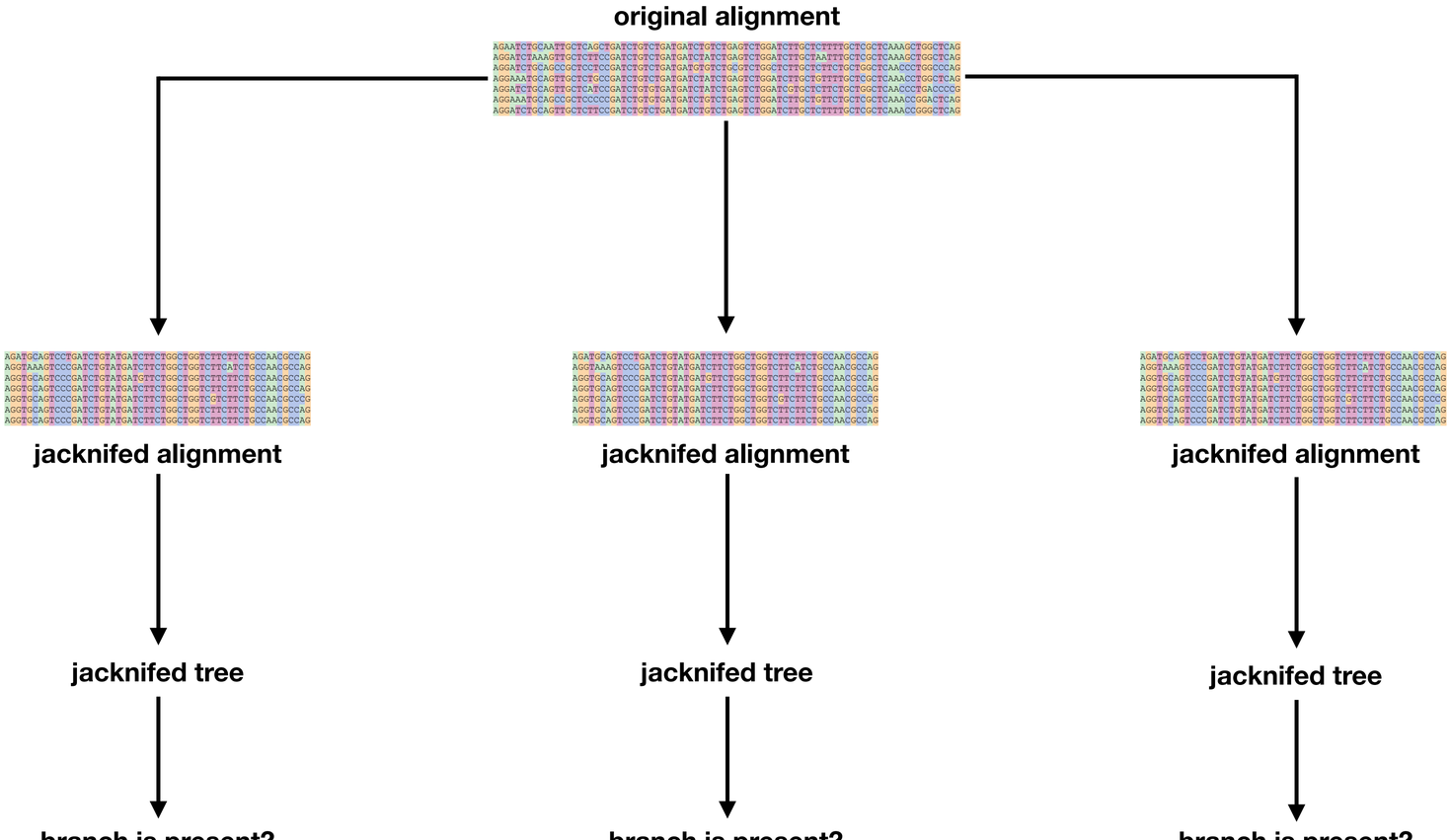
```
species 1   AGA   TGCA   G TC  CTGATCTGT   ATGATCT TCTG G CTGG TCTT C  TT   CT GC CAA  C G C CAG
species 2   AGG   TAAA   G TC  CCGATCTGT   ATGATCT TCTG G CTGG TCTT C  AT   CT GC CAA  C G C CAG
species 3   AGG   TGCA   G TC  CCGATCTGT   ATGATGT TCTG G CTGG TCTT C  TT   CT GC CAA  C G C CAG
species 4   AGG   TGCA   G TC  CCGATCTGT   ATGATCT TCTG G CTGG TCTT C  TT   CT GC CAA  C G C CAG
species 5   AGG   TGCA   G TC  CCGATCTGT   ATGATCT TCTG G CTGG TCGT C  TT   CT GC CAA  C G C CCG
species 6   AGG   TGCA   G TC  CCGATCTGT   ATGATCT TCTG G CTGG TCTT C  TT   CT GC CAA  C G C CAG
species 7   AGG   TGCA   G TC  CCGATCTGT   ATGATCT TCTG G CTGG TCTT C  TT   CT GC CAA  C G C CAG
```

```
species 1   AGAATCTGCAATTGCTCAGCTGATCTGTCTGATGATCTGTCTGAGTCTGGATCTTGCTCTTTTGCTCGCTCAAAGCTGGCTCAG
species 2   AGGATCTAAAGTTGCTCTTCCGATCTGTCTGATGATCTATCTGAGTCTGGATCTTGCTAATTTGCTCGCTCAAAGCTGGCTCAG
species 3   AGGATCTGCAGCCGCTCCTCCGATCTGTCTGATGATGTGTCTGCGTCTGGCTCTTGCTCTTCTGCTGGCTCAACCCTGGCCCAG
species 4   cGGAAATGCAGTTGCTCTGCCGATCTGTCTGATGATCTATCTGAGTCTGGATCTTGCTGTTTTGCTCGCTCAAACCTGGCTCAG
species 5   AGGATCTGCAGTTGCTCATCCGATCTGTGTGATGATCTATCTGAGTCTGGATCGTGCTCTTCTGCTGGCTCAACCCTGACCCCG
species 6   AGGAAATGCAGCCGCTCCCCCGATCTGTGTGATGATCTGTCTGAGTCTGGATCTTGCTGTTCTGCTCGCTCAAACCGGACTCAG
species 7   AGGATCTGCAGTTGCTCTTCCGATCTGTCTGATGATCTGTCTGAGTCTGGATCTTGCTCTTTTGCTCGCTCAAACCGGGCTCAG
```

```
species 1   AGATGCAGTCCTGATCTGTATGATCTTCTGGCTGGTCTTCTTCTGCCAACGCCAG
species 2   AGGTAAAGTCCCGATCTGTATGATCTTCTGGCTGGTCTTCATCTGCCAACGCCAG
species 3   AGGTGCAGTCCCGATCTGTATGATGTTCTGGCTGGTCTTCTTCTGCCAACGCCAG
species 4   AGGTGCAGTCCCGATCTGTATGATCTTCTGGCTGGTCTTCTTCTGCCAACGCCAG
species 5   AGGTGCAGTCCCGATCTGTATGATCTTCTGGCTGGTCGTCTTCTGCCAACGCCCG
species 6   AGGTGCAGTCCCGATCTGTATGATCTTCTGGCTGGTCTTCTTCTGCCAACGCCAG
species 7   AGGTGCAGTCCCGATCTGTATGATCTTCTGGCTGGTCTTCTTCTGCCAACGCCAG
```

Sampling without replacement!

Is the same information present in the jacknifed alignment?



node	branch is present?	branch is present?	branch is present?	JF
1	Y	Y	Y	100%
2	Y	N	N	33%
3	Y	Y	Y	100%
4	N	N	N	0%
5	N	Y	Y	66%
..	…	…	…	…
last	Y	Y	Y	100%

approximate Likelihood Ratio Test (aLRT)

aLRT is a fast, branch-specific metric, related to the **conventional Likelihood Ratio Test (LRT)**.

The standard LRT uses the test statistics $2(L_1 - L_0)$:

L_1 - **alternative hypothesis**: is the log-likelihood of the current tree
L_0 - **null hypothesis**: is the log-likelihood of the same tree, but with the branch of interest collapsed

aLRT approximation: L_2 is the second best NNI configuration around the branch of interest

Why use L_2 in aLRT?

- **much faster** than evaluating a collapsed model or full likelihood surface
  L_2 is computed by optimising only over the branch of interest and the four adjacent branches
  other parameters are fixed at their optimal values corresponding to the best ML tree

- allows efficient estimation of **branch-specific support**, especially in large trees
  focuses on local rearrangements around a branch - just 4 adjacent branches!

⚠️ Local rearrangements approach may miss topologies distant to current one but with higher lnL!

Posterior Probabilities (PPs)

Bayesian Inference is associated to **Posterior Probabilities** - also called clade credibility values.

- frequency that a bipartition appears in sampled trees ate the stationary phase of MCMC chain
- ranges from 0 to 1 - a PP of 0.95 means that the clade appeared in 95% of sampled trees

⚠️ important:

- PPs tend to be higher than non-parametric BPs
- sensitive to model choice and priors

Concordance Factors (CFs)

Quantify how many genes or sites support each branch in a reference tree.

Two types of CFs:

- gene Concordance Factor **(gCF)**:
  proportion of **gene trees** that contain a specific branch (or split).

- site Concordance Factor **(sCF)**:
  proportion of **informative sites** in the alignment that support a branch

Interpreting CFs values:

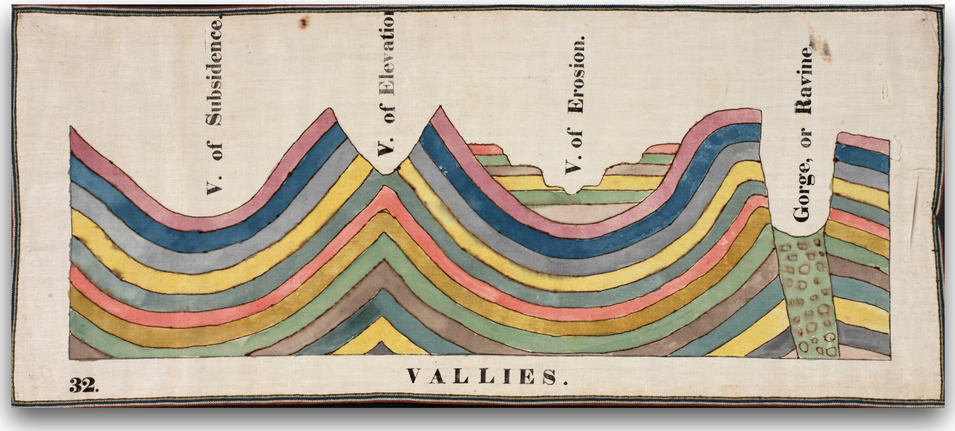
- High (e.g., >70%) - strong signal and high agreement across loci/sites
- Moderate (40–70%) - moderate support; possible discordance
- Low (<40%) - weak support or substantial gene tree conflict

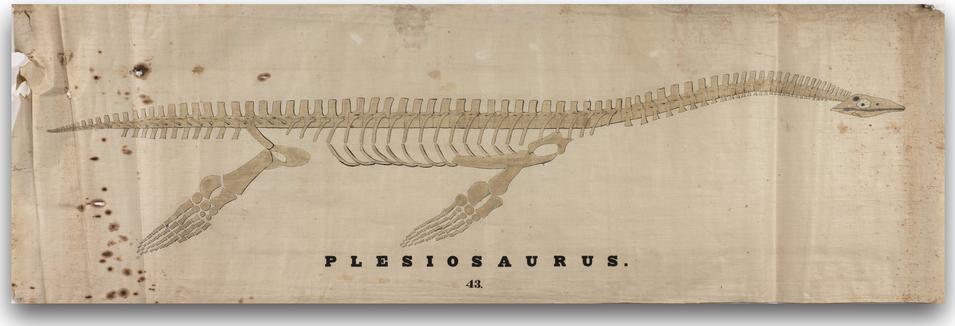
⚠️ Different support **metrics** have different **pitfalls**:

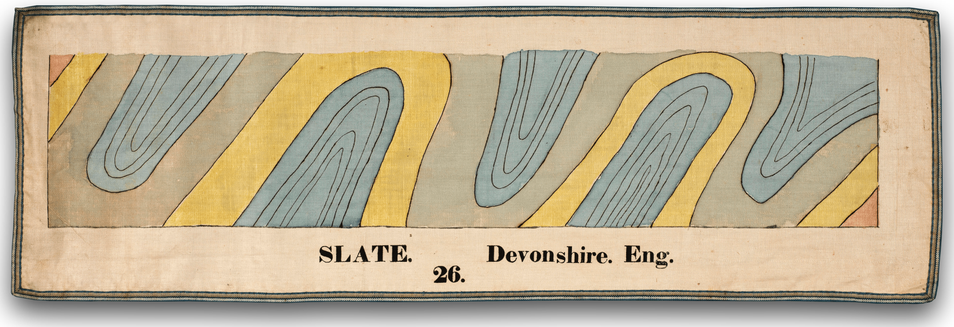
- parametric supports assume the evolutionary model is the correct one.
- standard metrics - PPs or BPs - can support conflicting topologies.
- …

🙃 Good practices:

- different support metrics must be compared
- observe consistent outcomes from different approaches







Complex models

Sequence data evolve under **heterotachy**!

That is **rate variation across sites** - and lineages of course!

Heterotachy is known to **mislead** phylogenetic inference …

… and we need **complex models 🤯 to deal with heterotachy!**

Partition Models

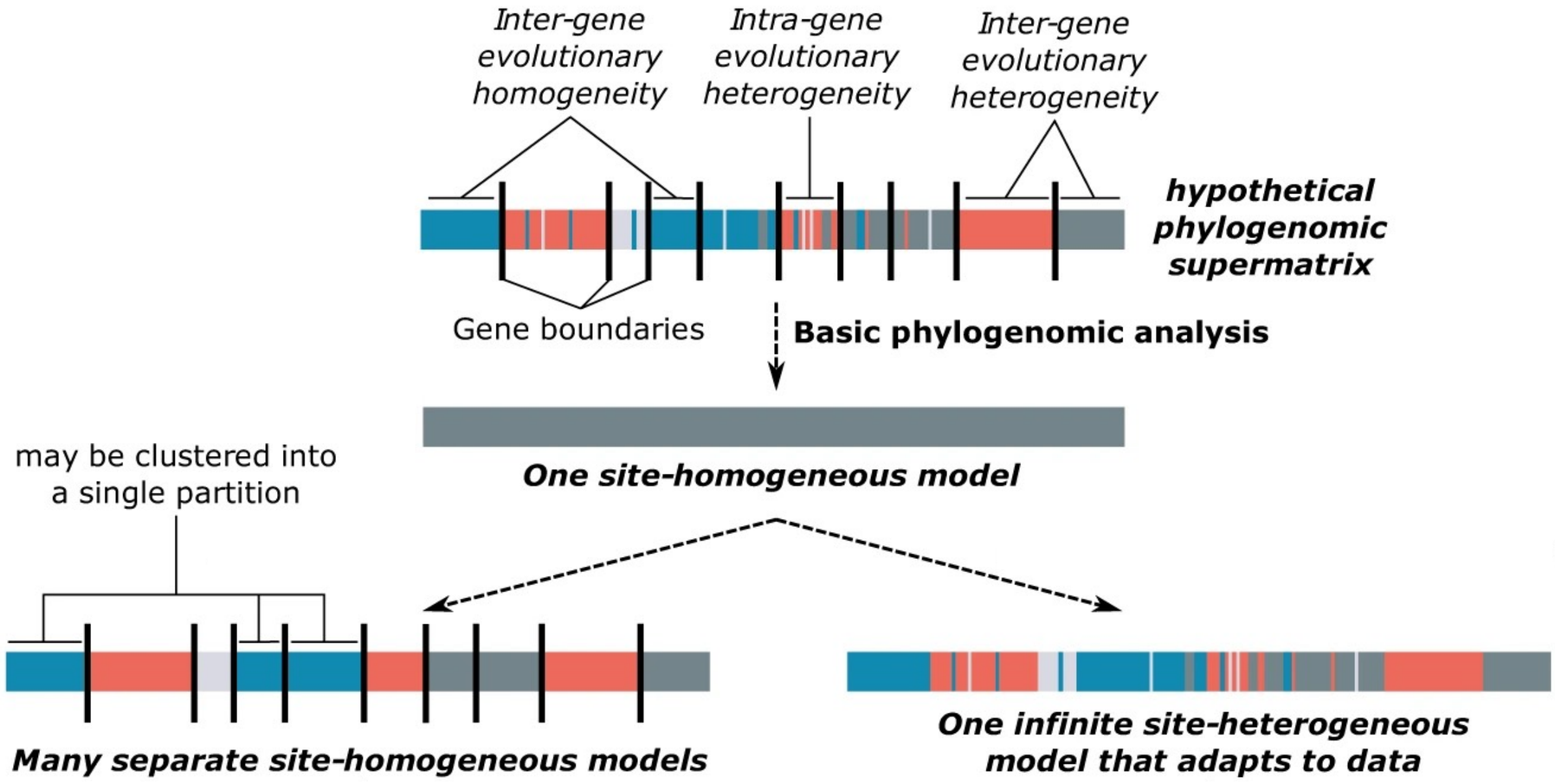
- Divide a multiple sequence alignment into distinct subsets (e.g. genes or codon positions).
- Each subset (partition) evolves under its own substitution model and parameters.
- Useful when prior knowledge about data structure is available.

Mixture Models

- Do not assign sites to specific subsets beforehand.
- Each site has a probability of belonging to multiple model components.
- Model heterogeneity within sites by combining multiple substitution processes.

Key Difference

- Partition models use predefined structure.
- Mixture models infer structure probabilistically.

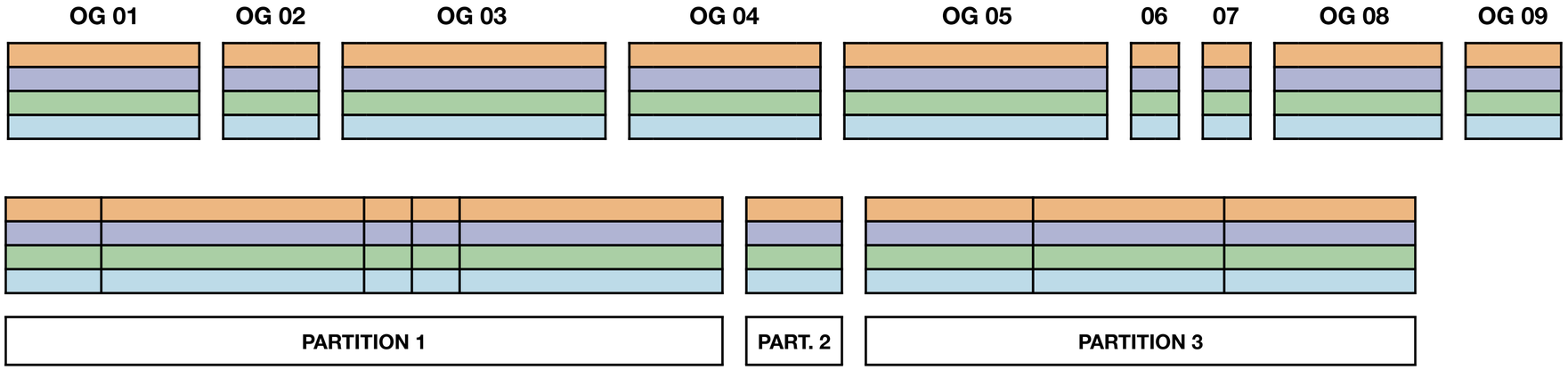

Inter-gene evolutionary homogeneity
Intra-gene evolutionary heterogeneity
Inter-gene evolutionary heterogeneity
hypothetical phylogenomic supermatrix
Gene boundaries
Basic phylogenomic analysis
One site-homogeneous model
may be clustered into a single partition
Many separate site-homogeneous models
One infinite site-heterogeneous model that adapts to data

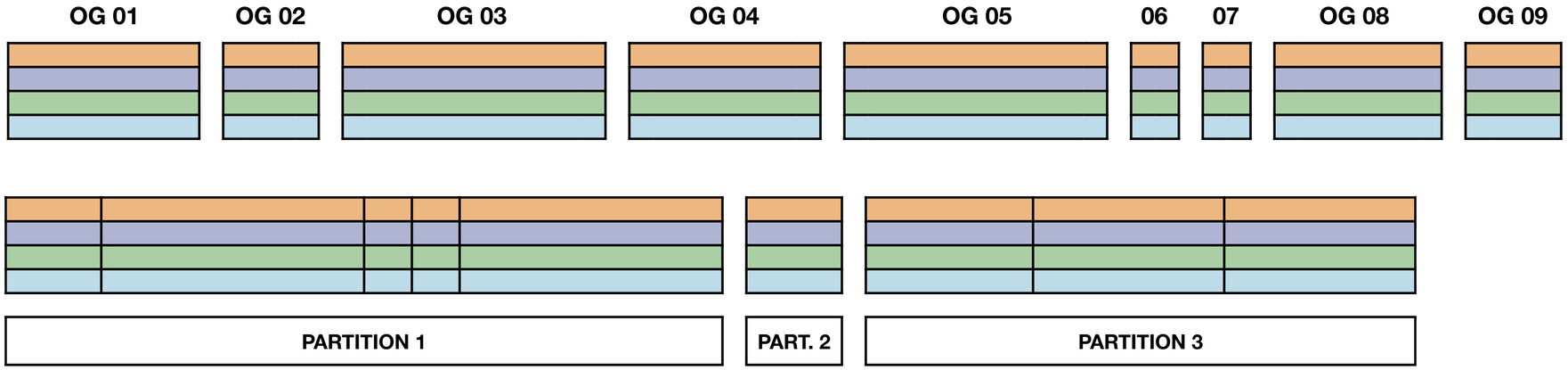
The types of **partition models** are:

- **Edge-linked partition model with equal branch lengths**
  All partitions share the same set of branch lengths.

- **Edge-linked partition model with proportional branch lengths**
  Each partition has branch lengths that are proportional to a shared set.

- **Edge-unlinked partition model**
  Each partition has a different set of branch lengths.


OG 01
OG 02
OG 03
OG 04
OG 05
06
07
OG 08
OG 09
PARTITION 1
PART. 2
PARTITION 3



EDGE - LINKED PARTITION MODEL

3X MODELS

A	P_1
C	P_1
T	P_1
G	P_1

Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1

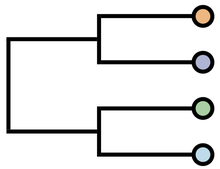
A	P_2
C	P_2
T	P_2
G	P_2

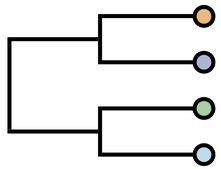
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2

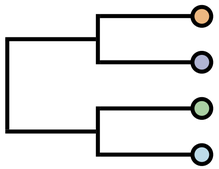
A	P_3
C	P_3
T	P_3
G	P_3

Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_1	Q_3
Q_3	Q_3	Q_1	Q_3
Q_3	Q_3	Q_3	Q_3

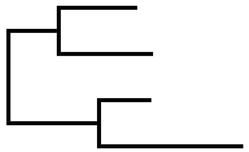
1X TOPOLOGY

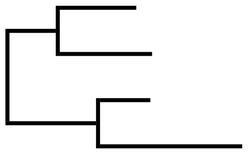


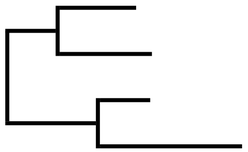


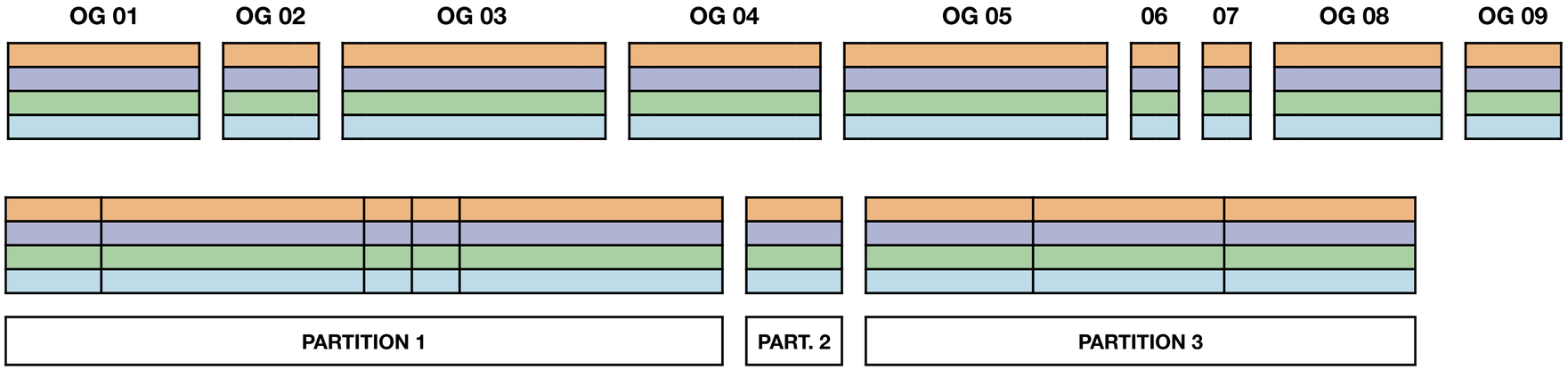


1X BRANCHLENGTHS

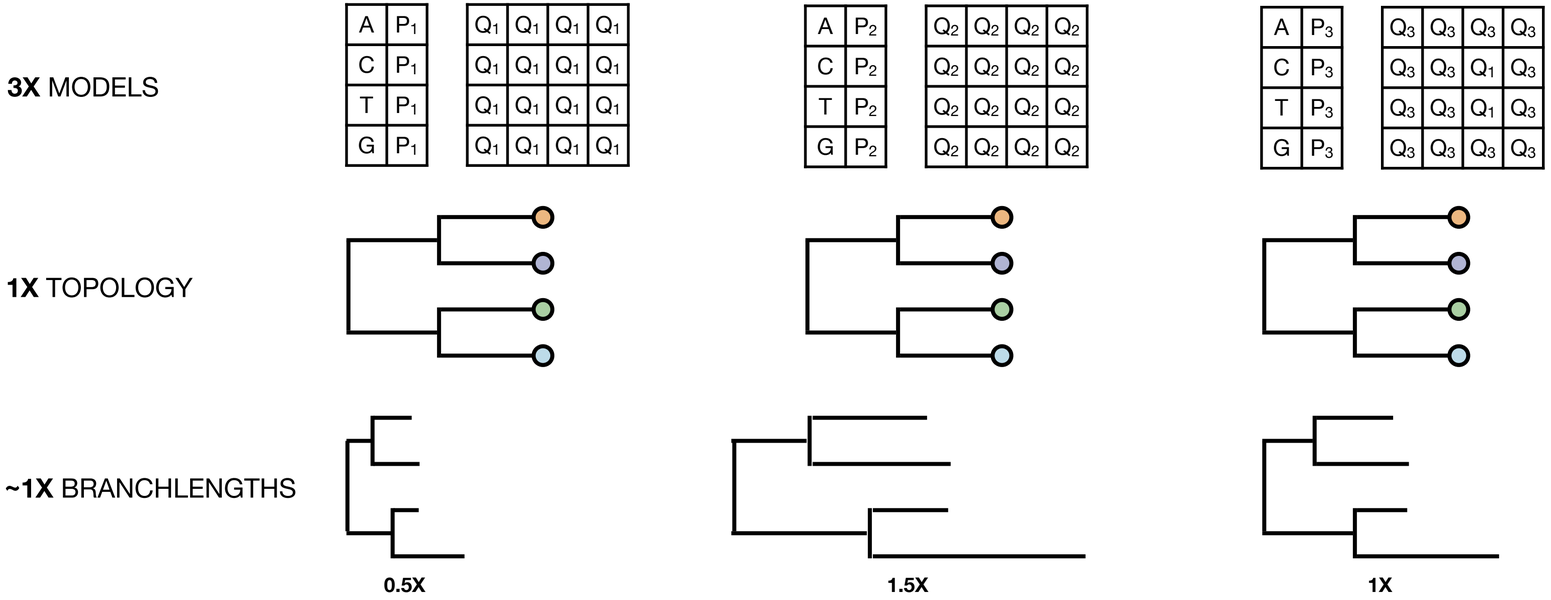


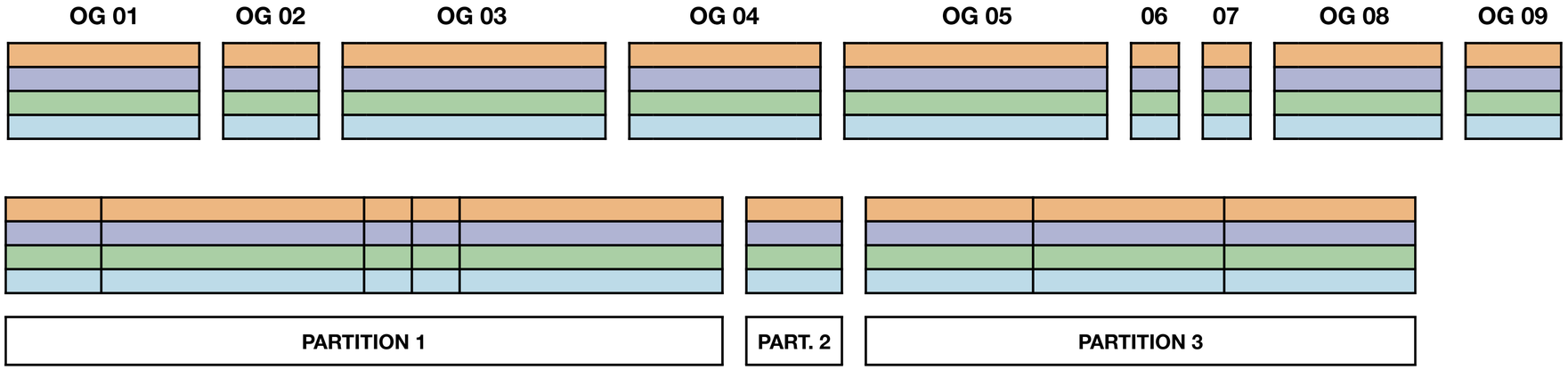






EDGE - PROPORTIONAL PARTITION MODEL





EDGE - UNLINKED PARTITION MODEL

3X MODELS

A	P_1
C	P_1
T	P_1
G	P_1

Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1

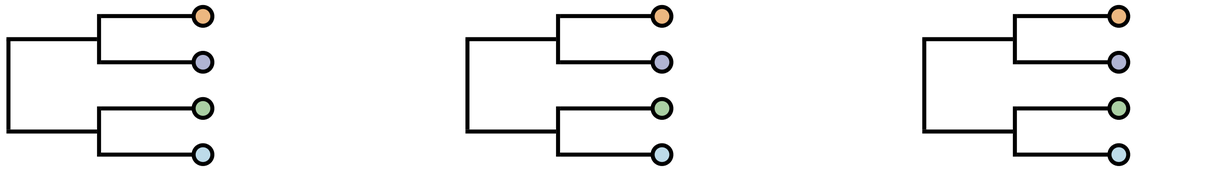
A	P_2
C	P_2
T	P_2
G	P_2

Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2

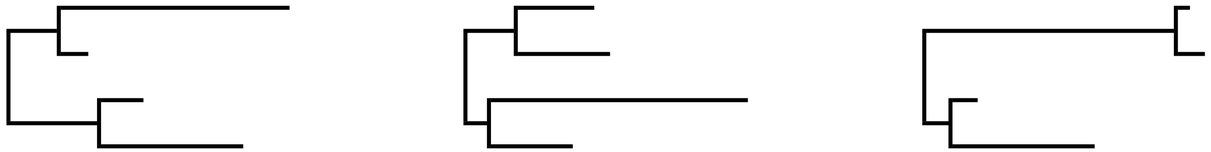
A	P_3
C	P_3
T	P_3
G	P_3

Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_1	Q_3
Q_3	Q_3	Q_1	Q_3
Q_3	Q_3	Q_3	Q_3

1X TOPOLOGY



3X BRANCHLENGTHS



Which partition model?

- edge-equal model is typically unrealistic
  Every part of the alignment is treated as if it evolved at the same speed along the same branches. However we know that different genes, codon positions, or functional regions often evolve at different rates.

- edge-unlinked model can overfit if there are many short partitions
  Partitions contain few informative sites and not enough data in each partition to reliably estimate its own branch lengths … the model begins to fit noise in the data rather than the true evolutionary signal.

- edge-proportional model is recommended as it represents a good tradeoff between oversimplification and overparametrization 🧐

RATE MIXTURE MODEL

Site evolves under the same substitution model, but with a different rate multiplier - e.g. +G or +K.

PROFILE MIXTURE MODEL

Model variation in frequencies (aa or nt) across sites, but they share the same substitution matrix.

Empirical profile models use fixed frequency profiles derived from large datasets - e.g. LG+C60.
Bayesian profile models infer the number and composition of profiles from the data - e.g. CAT.

FULL MIXTURE MODEL

Each site may evolve under a different substitution model, not just a different profile or rate.

RATE MIXTURE MODEL

A	C	G	C	T	G	C	A	A	T	A	C	T
A	C	C	C	T	C	C	A	T	G	C	C	T
A	C	C	C	A	G	C	A	G	C	T	C	C
A	C	C	C	A	A	C	A	A	C	A	G	C

K_1
K_2
K_3

3X GAMMA CATEGORIES

PROFILE MIXTURE MODEL

A	C	G	C	T	G	C	A	A	T	A	C	T
A	C	C	C	T	C	C	A	T	G	C	C	T
A	C	C	C	A	G	C	A	G	C	T	C	C
A	C	C	C	A	A	C	A	A	C	A	G	C
0.1	0.1	0.2	0.8	0.3	0.2	0.1	0.2	0.7	0.1	0.3	0.1	0.8
0.7	0.7	0.5	0.1	0.4	0.7	0.1	0.5	0.2	0.1	0.4	0.8	0.1
0.2	0.2	0.3	0.1	0.3	0.1	0.8	0.3	0.1	0.8	0.3	0.1	0.1

A	P_1
C	P_1
T	P_1
G	P_1

A	P_2
C	P_2
T	P_2
G	P_2

A	P_3
C	P_3
T	P_3
G	P_3

3X FREQUENCIES

FULL MIXTURE MODEL

A	C	G	C	T	G	C	A	A	T	A	C	T
A	C	C	C	T	C	C	A	T	G	C	C	T
A	C	C	C	A	G	C	A	G	C	T	C	C
A	C	C	C	A	A	C	A	A	C	A	G	C
0.1	0.1	0.2	0.8	0.3	0.2	0.1	0.2	0.7	0.1	0.3	0.1	0.8
0.7	0.7	0.5	0.1	0.4	0.7	0.1	0.5	0.2	0.1	0.4	0.8	0.1
0.2	0.2	0.3	0.1	0.3	0.1	0.8	0.3	0.1	0.8	0.3	0.1	0.1

A	P_1	Q_1	Q_1	Q_1	Q_1
C	P_1	Q_1	Q_1	Q_1	Q_1
T	P_1	Q_1	Q_1	Q_1	Q_1
G	P_1	Q_1	Q_1	Q_1	Q_1

A	P_2	Q_2	Q_2	Q_2	Q_2
C	P_2	Q_2	Q_2	Q_2	Q_2
T	P_2	Q_2	Q_2	Q_2	Q_2
G	P_2	Q_2	Q_2	Q_2	Q_2

A	P_3	Q_3	Q_3	Q_3	Q_3
C	P_3	Q_3	Q_3	Q_3	Q_3
T	P_3	Q_3	Q_3	Q_3	Q_3
G	P_3	Q_3	Q_3	Q_3	Q_3

3X SUBSTITUTION MODELS

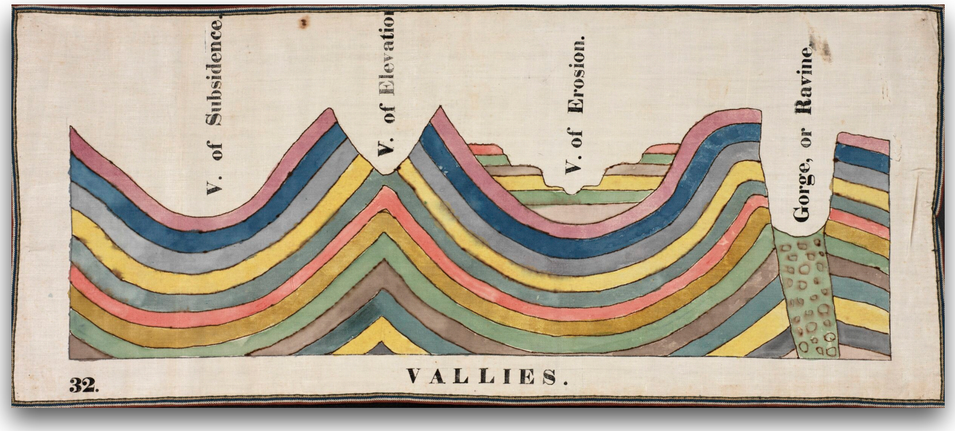
Both partition and mixture models allow more than one substitution model along the sequences and accommodate heterogeneity in evolutionary processes. **However**:

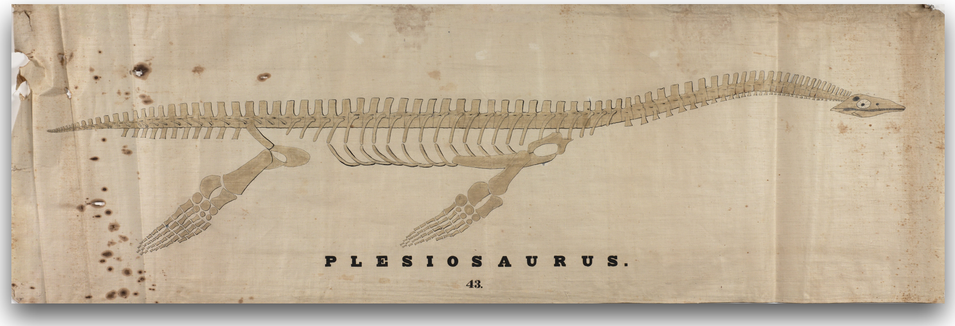
Partition Models:

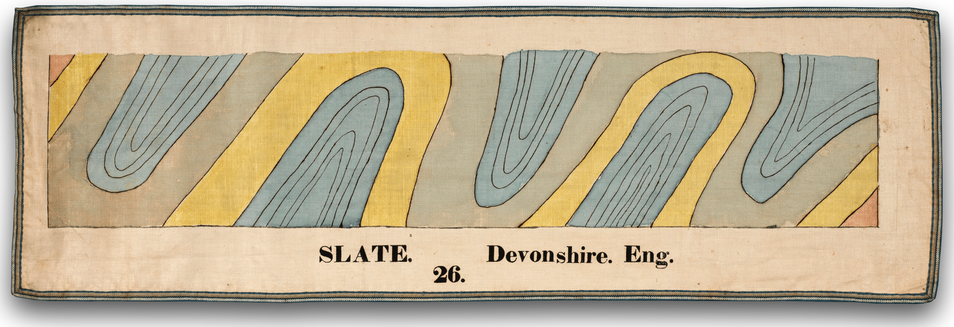
- **Sites assignment**: each site is assigned to a specific partition based on predefined *criteria*.
- **Model application**: distinct substitution models are applied to each partition.
- **Assumption**: the assignment of sites to partitions is known and fixed prior to analysis.

Mixture Models:

- **Sites assignment:** no assignment of sites to a class, probability of belonging to multiple classes.
- **Model application**: a weighted average of substitution models is used.
- **Assumption**: the site-to-class assignment is unknown.







Hypothesis testing

Phylogenetic trees are just …

Phylogenetic trees are just … hypotheses!

Phylogenetic trees are just … hypotheses!

… and we can test them!

Tree topology tests

Statistical tests used to compare alternative phylogenetic trees and assess whether differences in their fit to the data are significant.

Why use them?

- to test specific evolutionary hypotheses e.g. monophyly of a group
- to assess the impact of constraints or model choices on tree inference
- to reduce search space when exhaustive searches are impossible

Tree topology tests and **constrained inference** serve different purposes, however they are closely related conceptually.

We enforce certain relationships during tree search, often based on well-established knowledge. For example, Vertebrata is monophyletic.

Narrows tree space and focuses on biologically plausible topologies.

- **hard constraints**: enforce strict relationships
  A and B must be sisters.

- **soft constraints**: more flexibile within broader groupings
  A, B, and C must be monophyletic, but their branching order is not fixed

Common tree topology tests:

- **bp-RELL test**
  fast but less robust than formal statistical tests.

- **Kishino–Hasegawa (KH) test**
  pairwise test - assumes trees were pre-specified, not data-driven

- **Shimodaira–Hasegawa (SH) test**
  Adjusts for multiple comparisons between topologies

- **Approximately Unbiased (AU) test**
  More accurate, especially for many competing trees

- **Expected Likelihood Weights (ELW) test**
  Computes relative support for each tree based on expected likelihoods.

They all rely on the **RELL approximation**!

The **Resampling Estimated Log-Likelihoods (RELL) approximation** (Kishino et al. 1990) is a cornerstone method used in tree topology tests.

In topology tests we want to estimate how **tree lnL vary across replicates**.

Full parametric bootstrapping would require **re-estimating parameters and tree likelihoods** for each replicate - computationally expensive.

It is a **bootstrap-based method** used to **approximate the sampling distribution** of differences in tree likelihoods across sites - without the need to optimize again parameters for each replicate.

RELL simplifies this by:
- keeping likelihoods fixed per site and tree
- resampling sites only, not full trees or parameters
- making bootstrap replicates of site-wise likelihoods

Compute site-wise log-likelihoods for each tree topology using a matrix with:
- rows = **sites**
- columns = **trees**
- values = **lnL**

	site 1	site 2	site 3	site 4	site 5	site 6	site 7	site 8	site 9
tree 1	-104124.0580	-1043324.0580	-104344.0350	-204124.0580	-1045324.0580	-104344.0350	-204124.0580	-1945324.0580	-1047724.0580
tree 2	-272924.0580	-104124.0580	-1043324.0580	-272924.0580	-104344.0350	-204124.0580	-1045324.0580	-104344.0350	-204124.0580
tree 3	-1043324.0580	-104344.0350	-272924.0580	-104124.0580	-1043324.0580	-204124.0580	-204124.0580	-1043324.0580	-104344.0350
tree 4	-104124.0580	-1043324.0580	-104344.0350	-204124.0580	-104124.0580	-1043324.0580	-204124.0580	-104344.0350	-204124.0580

bp-RELL test

In brief:

1. calculate Site-wise Log-Likelihoods for each input tree
2. use the RELL method to bootstrap alignment sites and generate many pseudo-replicates

   then for each bootstrap pseudo-replicate:

   3. sum the resampled site log-likelihoods to get total likelihoods
   4. identify the best-scoring tree in that replicate
5. the bootstrap proportion (bp) for a tree is the fraction of replicates where it had the highest lnL

Interpreting results:

- bp values range from 0 to 1. Higher values represent more frequent best performance
- bp-RELL is fast and easy to compute, but it is not a formal statistical test

Kishino-Hasegawa (KH) test

Start with *two trees* that you want to compare. Assumes *trees are hypothesis-driven*.

1. calculate Site-wise Log-Likelihoods of the 2 trees - how much each site supports each tree
2. for each site, calculate the difference in lnL between Tree $_A$ and Tree $_B$

   $\Delta_{lnL} = lnL_{Tree\ A} - lnL_{Tree\ B}$

   this creates a list of *per site* differences
3. paired t-test to assess if the average log-likelihood difference is significantly different from zero
4. returns a p value indicating whether the worse-scoring tree is statistically rejected

Interpreting results:

- $p < 0.05$: the worse-scoring tree is significantly worse - reject it!
- $p \geq 0.05$: no significant difference between the two trees - cannot reject either tree 🙃

Shimodaira-Hasegawa (SH) test

Designed to compare multiple phylogenetic trees using the same alignment data. It asks: *"Are any of these trees significantly worse than the best one?"*. In brief:

1. calculate Site-wise Log-Likelihoods of multiple input trees

2. identify the "*best tree*" by summing site log-likelihoods of each tree

3. bootstrap resampling of sties using the RELL method to create many replicates of dataset

4. for each replicate:
   - sum the resampled site-wise log-likelihoods for each tree
   - compute log-likelihood differences between each tree and the "*best tree*"

5. this produces a bootstrap distribution of likelihood differences for each tree

6. p values are calculated as described previously and adjusted for multiple test comparisons

Both KH and SH give back a p value, the latter for each tree relative to the "*best tree*". However:

- KH test compares two or few trees, has no correction for multiple testing
- SH test compares many trees, conservative as it adjusts for testing multiple topologies

Approximately Unbiased (AU) test

In brief:

1. calculate Site-wise Log-Likelihoods of multiple input trees

2. identify the "*best tree*" by summing site log-likelihoods of each tree

3. multiscale bootstrap resampling of sites using the RELL method, but at different block sizes (e.g. 0.5X 1X and 2X the dataset size) to better approximate the true distribution

4. for each replicate:
   - sum the resampled site-wise log-likelihoods for each tree
   - compute log-likelihood differences between each tree and the "*best tree*"

5. calculate an accurate distribution of likelihood differences, accounting for selection bias

6. p values are then calculated from this improved distribution

AU provide p values for each tree relative to the best, and is recommended over KH and SH tests:

- KH test is too liberal if used with data-derived trees, leading to inflated Type I errors
- SH test becomes too conservative and rejects few trees when testing many trees

Expected Likelihood Weights (ELW) test

ELW test provides a way to estimate the relative support of each candidate tree, based on *how well it explains the data*. Rather than returning p values, it assigns weights that sum to 1 across all trees.

1. calculate Site-wise Log-Likelihoods of multiple input trees

2. identify the "*best tree*" by summing site log-likelihoods of each tree

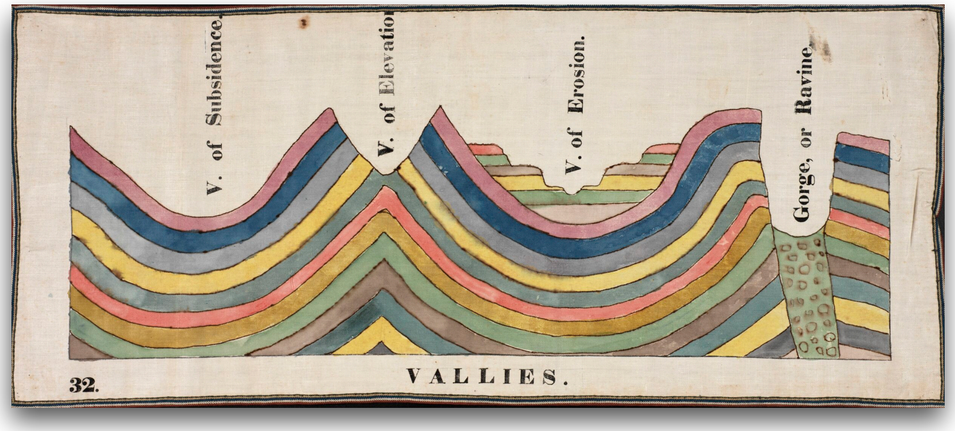
3. bootstrap resampling of sties using the RELL method to create many replicates of dataset

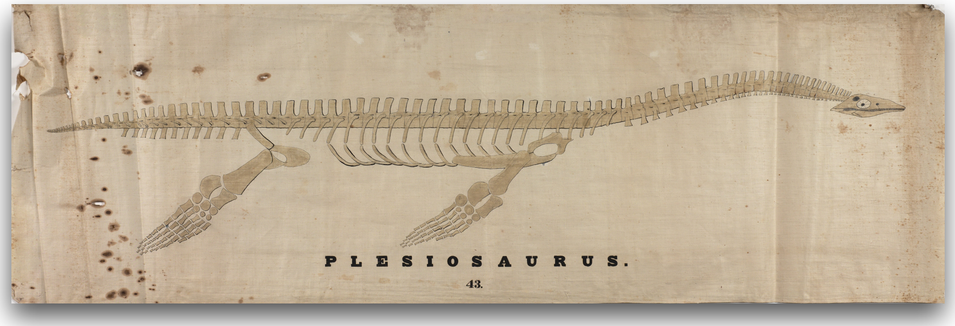
4. for each replicate:
   - sum the resampled site-wise log-likelihoods for each tree
   - compute log-likelihood differences between each tree and the "*best tree*"
   - compute relative likelihood of each tree, based on how its total likelihood compares to others.

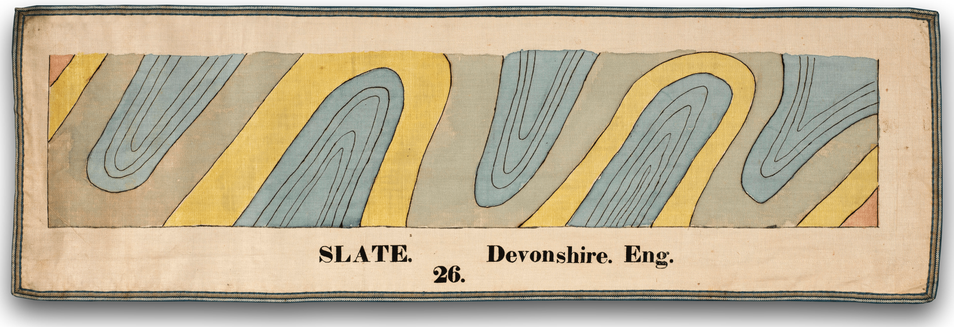
5. calculate ELW for each tree as the proportion of replicates in which it performs well, weighted by its relative likelihood

Interpreting results:

ELW values range from 0 to 1 and sum to 1 across all trees A tree with a high ELW (e.g., >0.95) has strong support Lower values indicate less support, but not necessarily rejection







discordance, ILS & the coalescent

Species tree
Represents the evolutionary history of species - cladogenesis or the sequence of speciation events.

Gene tree(s)
Traces the evolutionary history of a single locus and may differ biologically from the species tree.

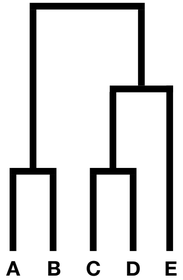
A common cause of gene tree discordance is **Incomplete Lineage Sorting (ILS)**,
which is especially likely when:

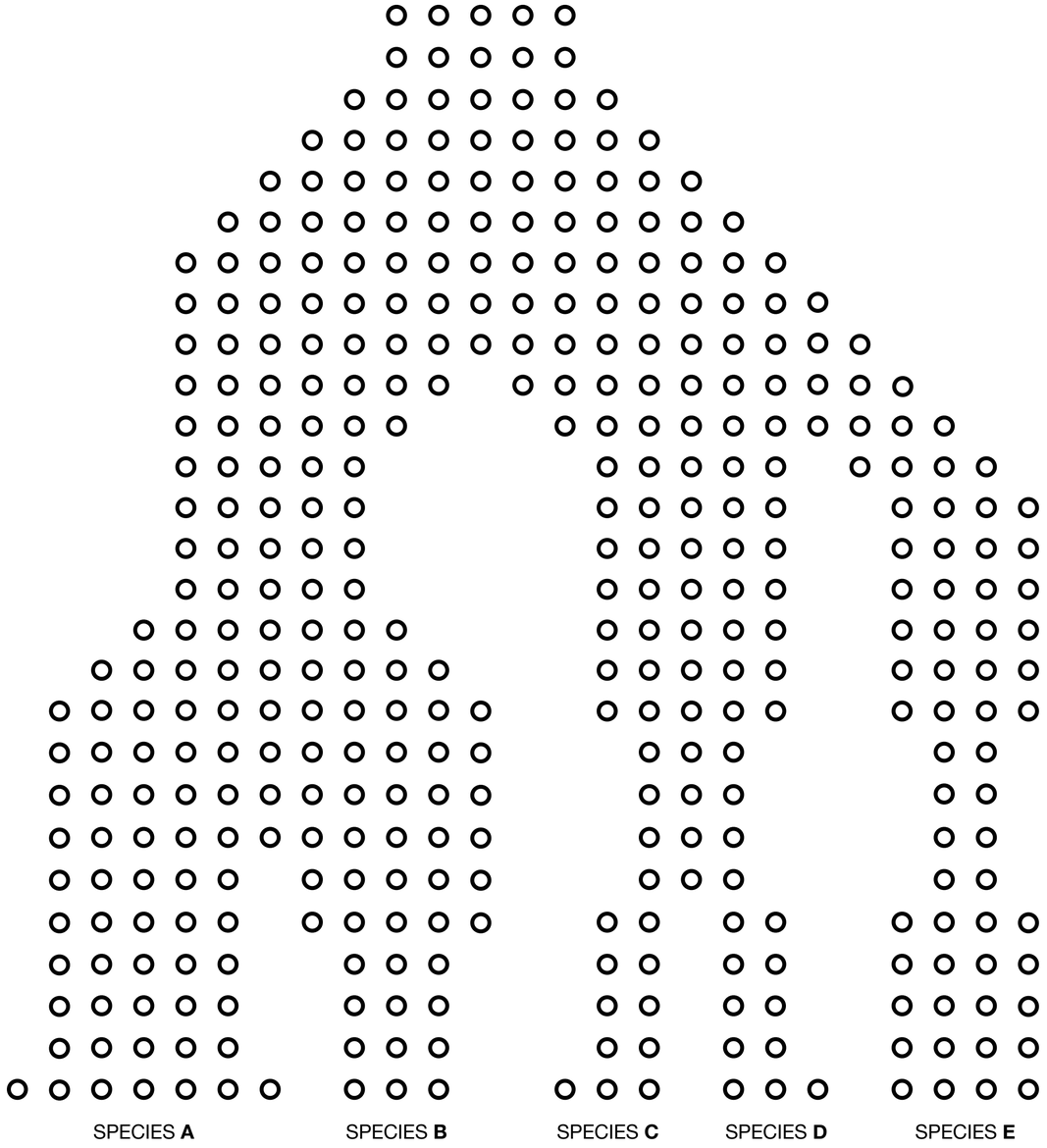
- effective population sizes are large
- speciation events are closely spaced in time

You can **think of ILS as**:

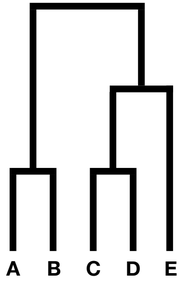
- ancestral polymorphisms that persist across speciation events,
- alleles sorting into descendant lineages *erratically*, not reflecting the species splits.

SPECIES TREE

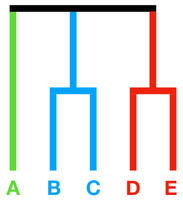


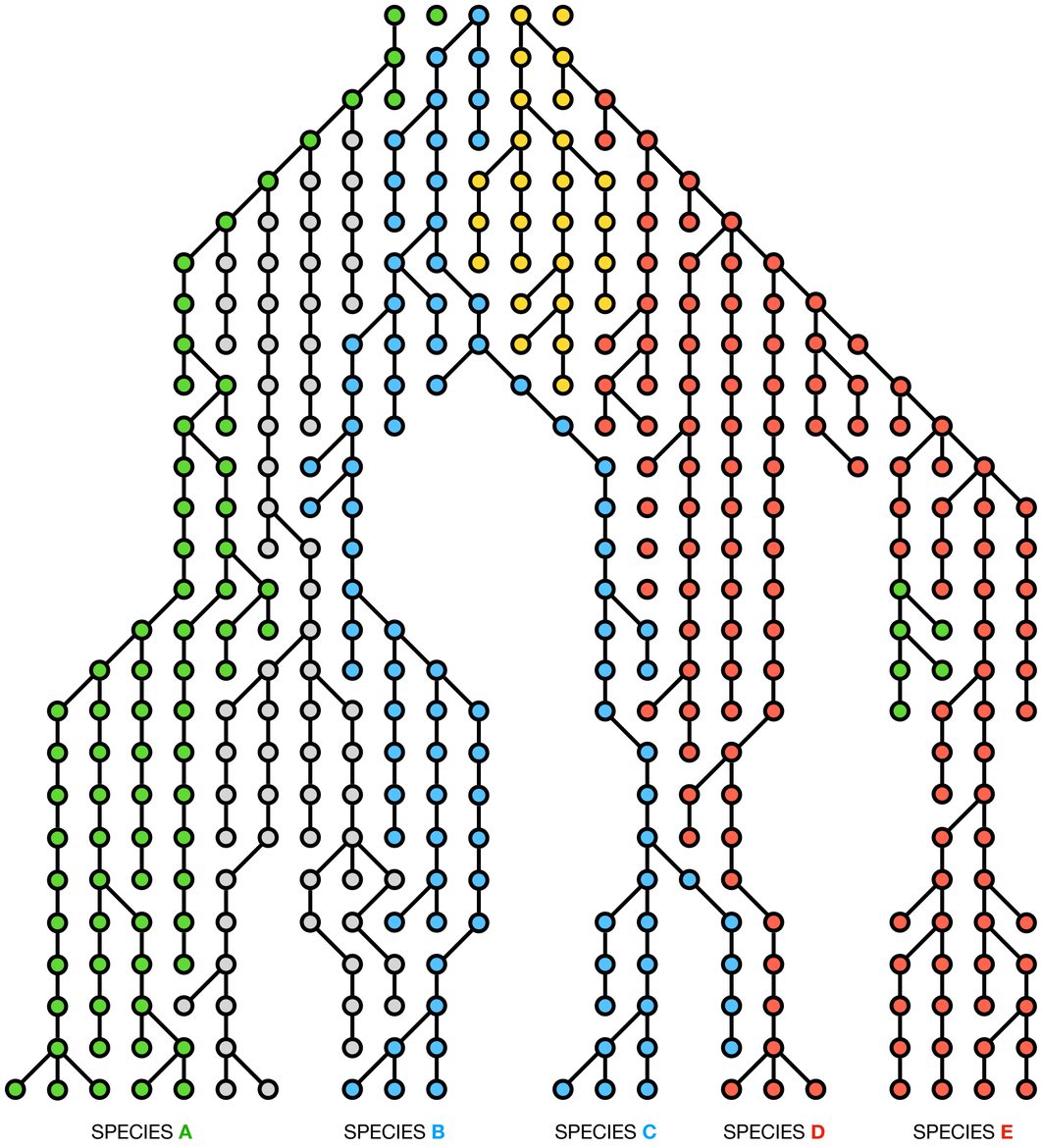


SPECIES TREE



GENE TREE





While ILS is a major cause of gene tree conflict …

… also other biological processes can result in gene tree discordance:

- **Horizontal Gene Transfer (HGT)**
  Transfer of genetic material between unrelated lineages, common in microbes.

- **Hybrid Introgression (HI)**
  Transfer of genetic material between species through hybridization followed by backcrossing.

These processes can result in **true biological conflict** between gene trees and the species tree!

In Lesson 14, we'll explore the **technical artifacts and sources of error** that can create the **illusion of conflict** between gene trees and the species tree — even when no true biological discordance exists.

By the way … **Robinson-Foulds (RF) distance** …

… are a commonly used metric to **compare two phylogenetic trees** with the same set of taxa.

How it works:

It measures the symmetric difference between two trees. Specifically, it counts the number of bipartitions (splits) that are present in Tree A but not in Tree B and present in Tree B but not in Tree A.

- Higher RF distance = greater disagreement between the trees.
- Lower RF distance = smalle disagreement between the trees.
- Often normalized to range from 0 (identical) to 1 (completely different)

While RF distance captures topological differences, it treats all differences equally … regardless of their biological relevance.

Other more nuanced metrics aim to assess the phylogenetic informativeness of the differences.

Concatenation-based species tree inference

- a single data matrix containing aligned sequences (species × genes)
- assumes all genes share the same evolutionary history = one tree

Coalescent-based species tree inference

- a set of unrooted, binary gene trees (each with leaves labeled by species) Models gene tree
- infers a species tree from multiple gene histories discordance due to incomplete lineage sorting

The **coalescent** is a framework that describes how genes sampled from an extant population trace back to a **common ancestor** over time.

The **coalescent theory** sits between phylo- and pop- genetics.

It's a **retrospective model**: the coalescent looks **backward in time** to understand the **genealogical relationships** between individuals or alleles.

The coalescent is based on the mathematical notion that mutations within genes - **new alleles** - can be **traced backwards in time**, to the point where the mutation initially occurred.

Given that this is a retrospective, instead of describing these mutation moments as **divergence** events, these appear as moments where mutations come back together as **coalescence** events.

There are a number of applications of coalescent theory, and it is particularly fitting process for understanding the **neutral demographichistory of populations and species**.

The initial coalescent model was described in the 1980s, built upon by a number of different ecologists, geneticists and mathematicians. However, **John Kingman** is often attributed with the fomalization of the original coalescent model.

Key **assumptions** of the coalescent:

- **constant population size** - effective population size (N_e) does not change over time
- **random mating** - all genes have equal chance of being parent of any gene in the next gen
- **no selection** - all alleles are neutral and drift is the only force acting on allele frequencies
- **no migration** - no gene flow from other populations and no spatial or geographic structure
- **non-overlapping generations**
- **no recombination within loci**
- **infinite sites or infinite alleles model**
    - **infinite sites**: mutations occur at unique sites, no back or parallel mutations
    - **infinite alleles**: every mutation creates a completely new allele

What's the chance two gene copies coalesce?

Imagine two haploid gene copies sampled from a population. In each generation:

- each gene copy chooses a parent randomly from the previous generation
- in a population of effective size N_e, there are $2N_e$ possible parental gene copies.

So…

- the probability they coalesce in the previous generation: $P(\text{coalescence}) = \frac{1}{2N_e}$
- the probability they don't coalesce in the previous generation: $P(\text{no coalescence}) = 1 - \frac{1}{2N_e}$

What's the chance two gene copies share a common ancestor *t* generations ago?

- the probability they coalesce a generation t: $P(\text{coalescence at generation } t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \cdot \frac{1}{2N_e}$

This forms the basis of the exponential distribution of coalescent times.

PAST

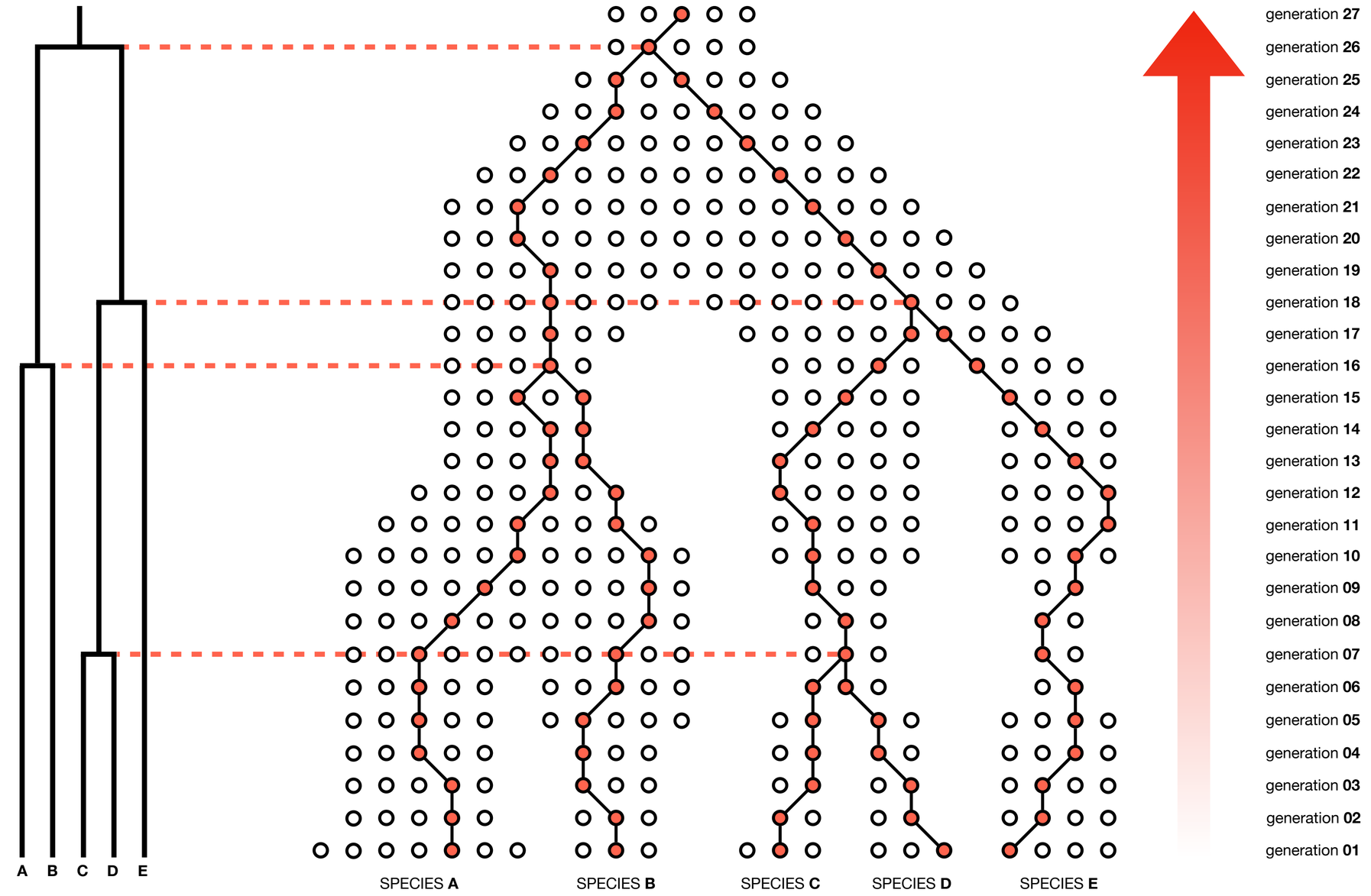
DIVERGENCE

PRESENT

PAST

COALESCENCE

PRESENT


generation 27
generation 26
generation 25
generation 24
generation 23
generation 22
generation 21
generation 20
generation 19
generation 18
generation 17
generation 16
generation 15
generation 14
generation 13
generation 12
generation 11
generation 10
generation 09
generation 08
generation 07
generation 06
generation 05
generation 04
generation 03
generation 02
generation 01
A B C D E
SPECIES A
SPECIES B
SPECIES C
SPECIES D
SPECIES E

This is theoretically complicated … 🤯 … and we won't dive into the full mathematical depths of coalescent theory.

However, also consider that the coalescent gives us a **null model** - a baseline expectation - of how gene lineages and mutations would **coalesce back in time assuming ideal conditions**.

Comparing observed data to predictions of the coalescent, we can:

- detect population size changes
- infer population structure or migration
- identify signals of selection or recombination

In coalescent-based inference there are different:

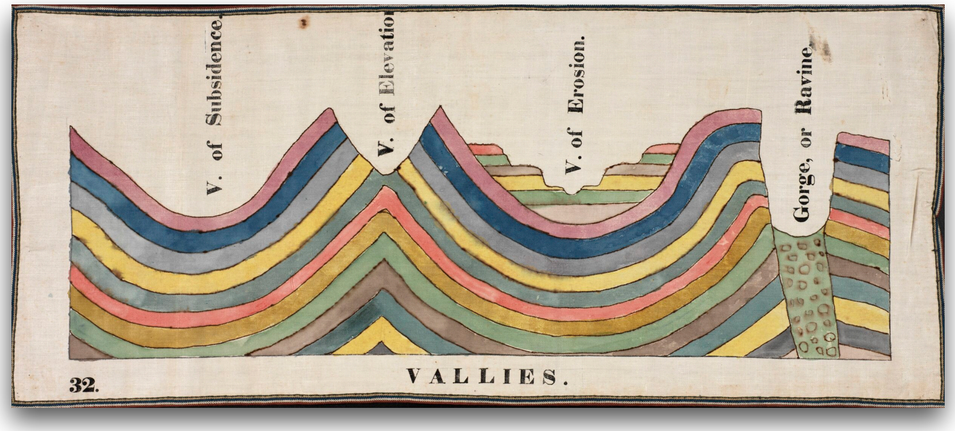
branch support metrics

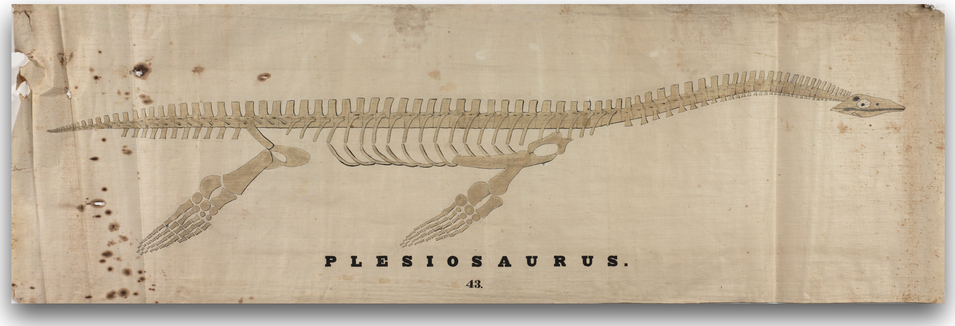
- **Local Posterior Probability (LPP)**
  Probability a branch is correct based on quartet frequencies.
  Uses the Dirichlet prior - a pseudocount of 1 per topology: $\mathrm{LPP}_{T1} = \frac{(n_1 + 1)}{(n_1 + n_2 + n_3 + 3)}$
  Ranges from 0 to 1.

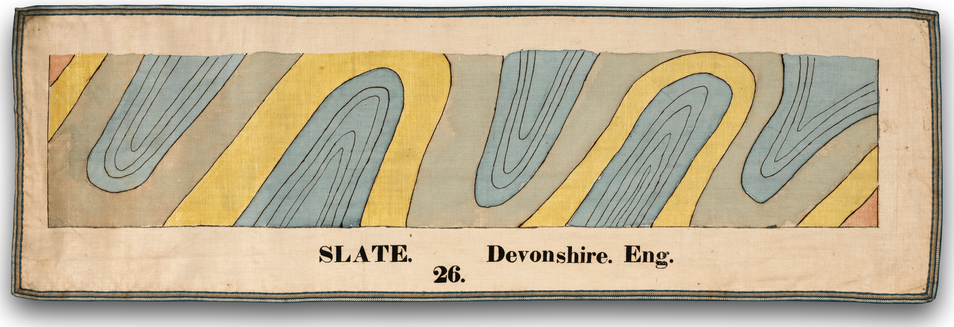
- **Multi-locus Bootstrap (MLBS)**
  Branch stability across bootstrapped gene trees.
  Ranges from 0 to 100%.

branch length units

- **Coalescent Units**
  The expected number of coalescent events per unit time, scaled by population size.
  So, 1 coalescent unit = $2N_e$ generations. This scaling allows comparisons of coalescent processes across populations of different sizes.







biases in phylogenetics

stochastic bias

How stochastic bias happens:

- Random noise introduced by limited data can outweigh the true phylogenetic signal.
- This can happen even when the model is correct.

Common causes:

- insufficient sequence length or number of genes
- low phylogenetic signal (few informative sites)
- high variance in substitution processes

Mitigating stochastic bias:

- Use longer alignments and more loci to average out noise - phylogenomics!
- Focus on genes with high phylogenetic signal - we will see that in the practicals!
- Filter hyper variable / poorly aligned and conserved / uninformative portions of alignement.
- Use branch support metrics to assess confidence in inferred relationships.
- Adding more taxa to break up long branches can reduce stochastic noise.

some guidelines / ballpark numbers from literature

- **dozens of genes** can resolve shallow divergences - recent speciation events - but deeper or more complex trees require hundreds of loci.

- **100–500 genes** significantly mitigate stochastic errors in resolving phylogenies at intermediate evolutionary depths.

- **thousands of genes** are require for genome-scale studies and are recommend to recover accurate phylogenetic signals across a range of divergence times.

systematic bias

Model assumptions and violations in phylogenetics

Most models of sequence evolution assume that the process is:

- **Stationary** – base or amino acid frequencies remain constant over time
- **Reversible** – the process looks the same forward and backward in time
- **Homogeneous** – the same model applies across all branches of the tree

These are collectively known as **SRH assumptions**.

SRH assumptions simplify inference, but may not reflect biological reality:

- Violations of SRH assumptions are common in real datasets.
- Such violations can lead to systematic bias and incorrect topologies.

Mitigating model violation:

- Test for model violations prior to tree reconstruction.
- Do phylogenetic subsampling and exclude problematic partitions - will see in the practicals!
- Apply non-SRH models when appropriate such as non-reversible models.
- Use partitioned analyses or profile mxiture models to better fit real data.

- unmodelled **heterogeneity in rates** - also known as Long Branch Attraction or LBA.

  Can mislead tree reconstruction methods - especially parsimony - so that fast-evolving lineages may be incorrectly inferred as closely related.

- unmodelled **heterogeneity in composition**

  Differences in base or amino acid composition across taxa can cause unrelated lineages with similar compositions to cluster together.

How LBA happens:

- In long branches, the probability of multiple substitutions at the same site increases.
- Excess of subs. leads to some parallel ones that falsely inflate similarity of unrelated taxa.

Common causes:

- of course uneven rates of evolution across taxa
- poor taxon sampling - missing intermediate branches
- deep divergence times

Mitigating LBA:

- Improved Taxon Sampling and add intermediate taxa to break long branches Reduces the chance of convergent substitutions being interpreted as shared ancestr

- Use Better-Fitting Models Employ e.g., site-heterogeneous models like CAT, GTR+Γ) These better account for rate variation and reduce LBA effects

- Avoid or Supplement Parsimony Use likelihood- or Bayesian-based methods with model testing Parsimony is more sensitive to LBA, especially with high divergence

- Remove Problematic Taxa (with caution!) If LBA is suspected and unsolvable, consider excluding the long-branched taxa.

How heterogenity in compositions happens:

- Phylogenetics assume that nt and aa composition is homogeneous across lineages.
- Taxa that evolve with distinct frequencies cluster based on composition rather than ancestry.

Common causes:

- of course uneven rates of evolution across taxa
- poor taxon sampling - missing intermediate branches
- deep divergence times

Mitigating heterogenity in compositions:

- Add intermediate taxa to break long branches Reduces the chance of convergent substitutions being interpreted as shared ancestr

- Use better-fitting models - more specifically profile mixture models - these better account for rate variation and reduce LBA effects

- Use amino acid recoding (e.g., Dayhoff6, SR4) or RY-coding for nucleotides to reduce noise from compositional differences.

- If certain taxa strongly deviate in composition and distort topology, consider excluding them after sensitivity analyses.

+ unmodelled saturation

- Sequences accumulate multiple substitutions at the same sites over time.
- Divergence underestimation: multiple substitutions along a branch interpreted as one or few.
- Homoplasy increase: identical states arise independently in different lineages.
- Leads to a loss of phylogenetic signal, especially at deeper divergences.
- Results in underestimation of true evolutionary distances and produces misleading topologies

Common causes:

- High substitution rates relative to divergence time
- Use of highly variable or fast-evolving markers
- Long evolutionary timescales without model correction
- Relying on nucleotide data for deep splits

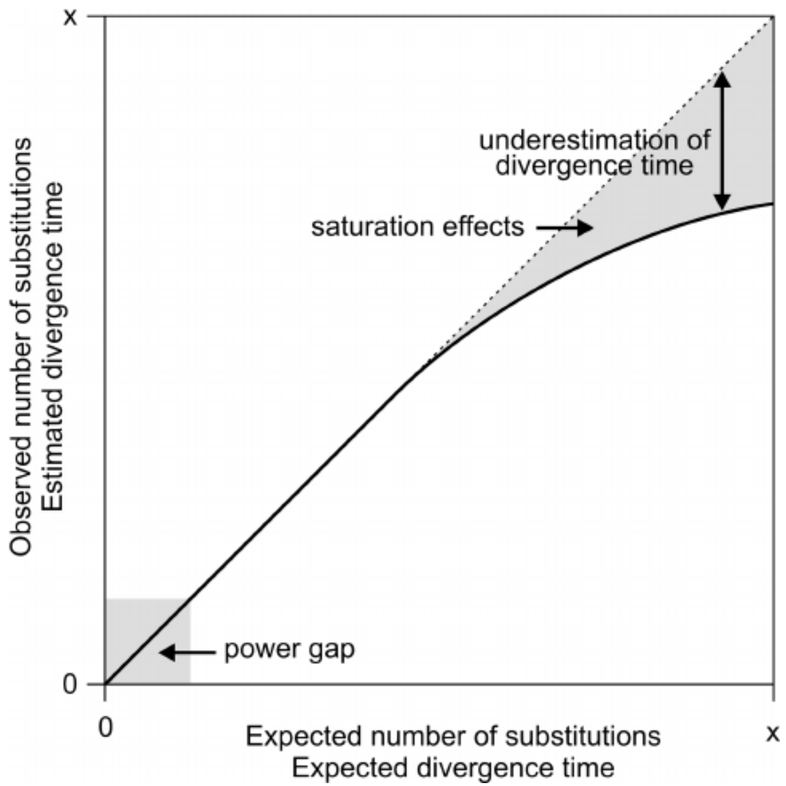
Mitigating Saturation:

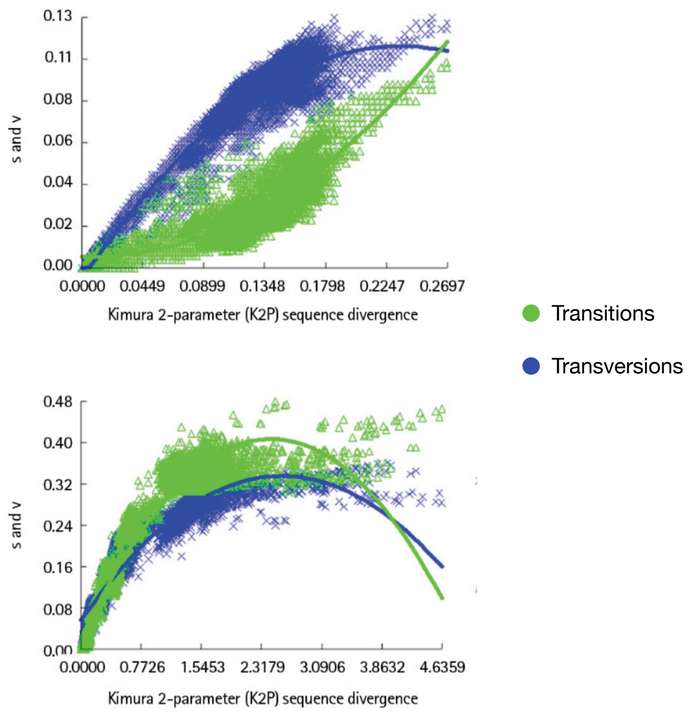
- Choose less saturated and slowly evolving genes or regions
- Use amino acid alignments over nucleotides for deep phylogenies.
- Apply models that account for multiple hits and rate heterogeneity (gamma).
- Recoding strategies as RY coding (nucleotides) or Dayhoff-type recoding (proteins).

Interpreting a saturation plot:

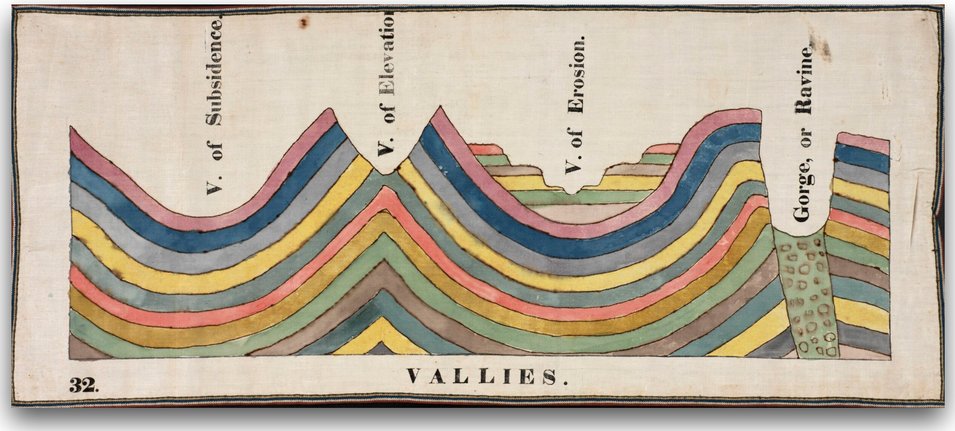
- Early divergence (left side): substitutions increase linearly; good phylogenetic signal.
- Later divergence (right side): substitution rate slows and plateaus—this indicates saturation.

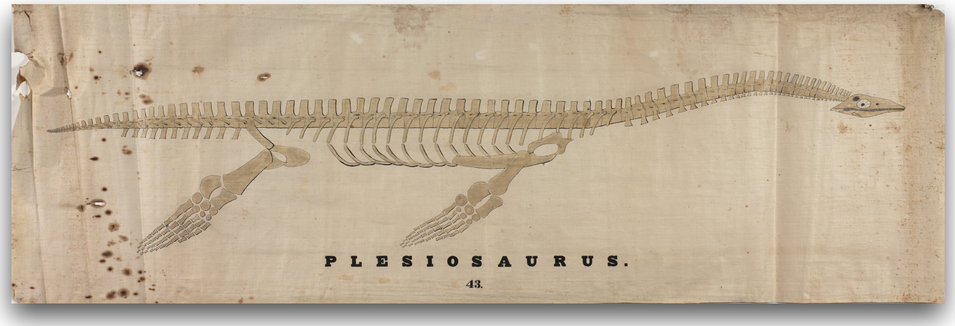
Typically, **transitions** saturate faster than **transversions** because they're biochemically more frequent (purine ↔ purine, pyrimidine ↔ pyrimidine).

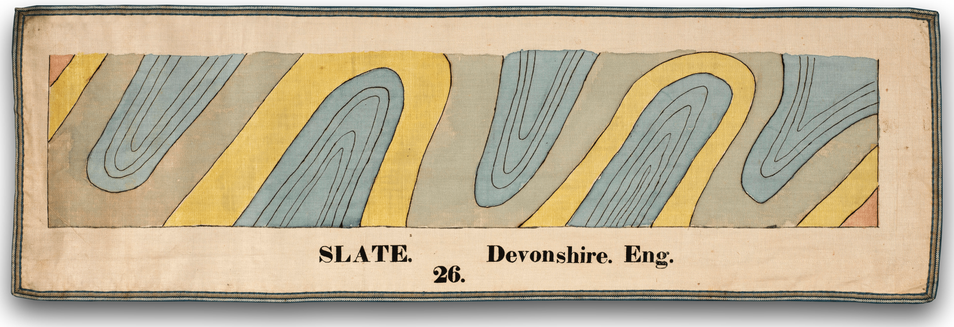




bias	stochastic	systematic
cause	random error due to limited data	incorrect model and assumptions
consistency	random; decreases with more data	persistent; remains despite more data
mitigation	increase data quantity (genes, taxa)	appropriate models and assumptions

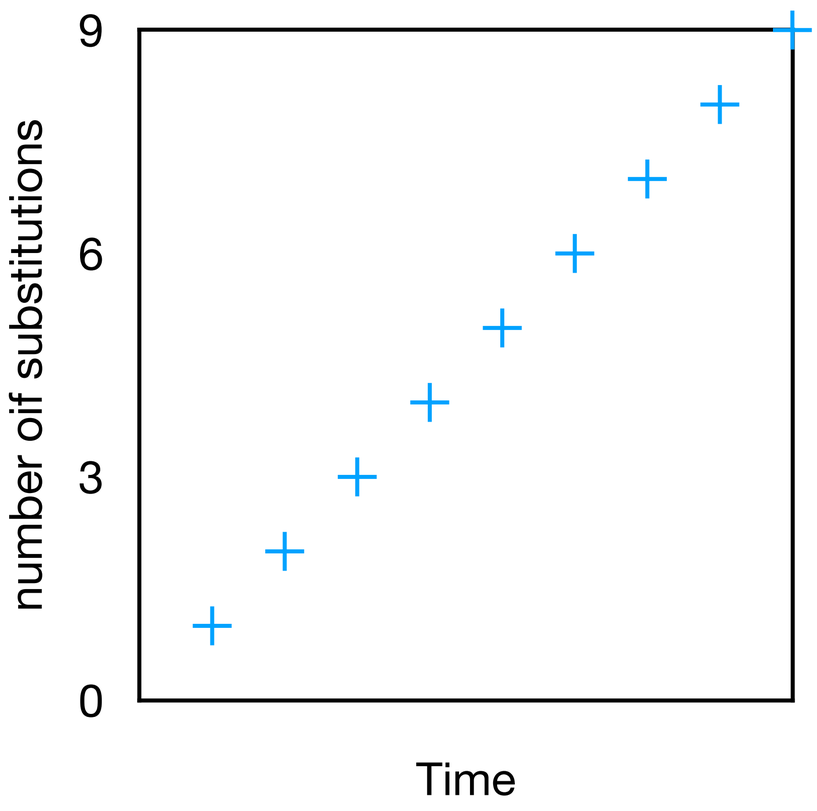




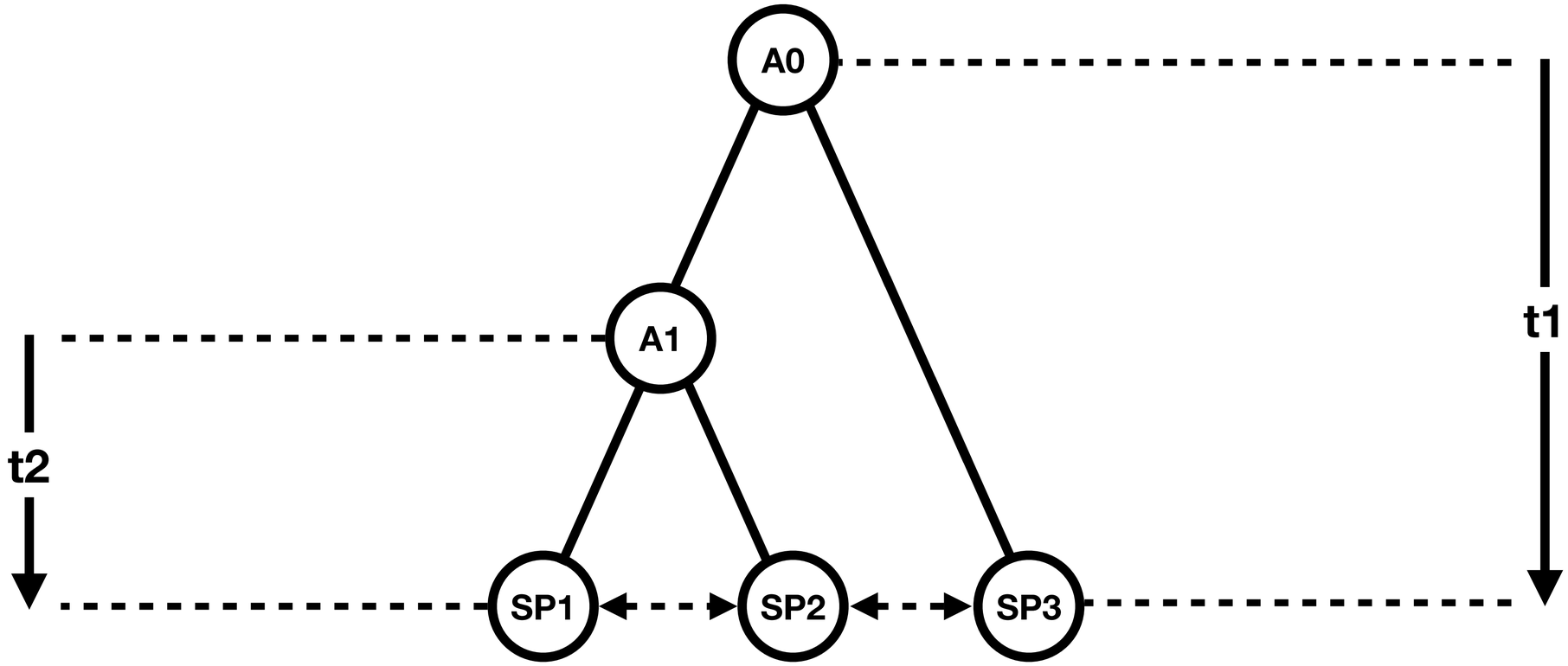


divergence time analyses

The molecular clock hypothesis suggests that **genetic changes** tend to accumulate at a relatively **constant rate** over long periods 🕰️

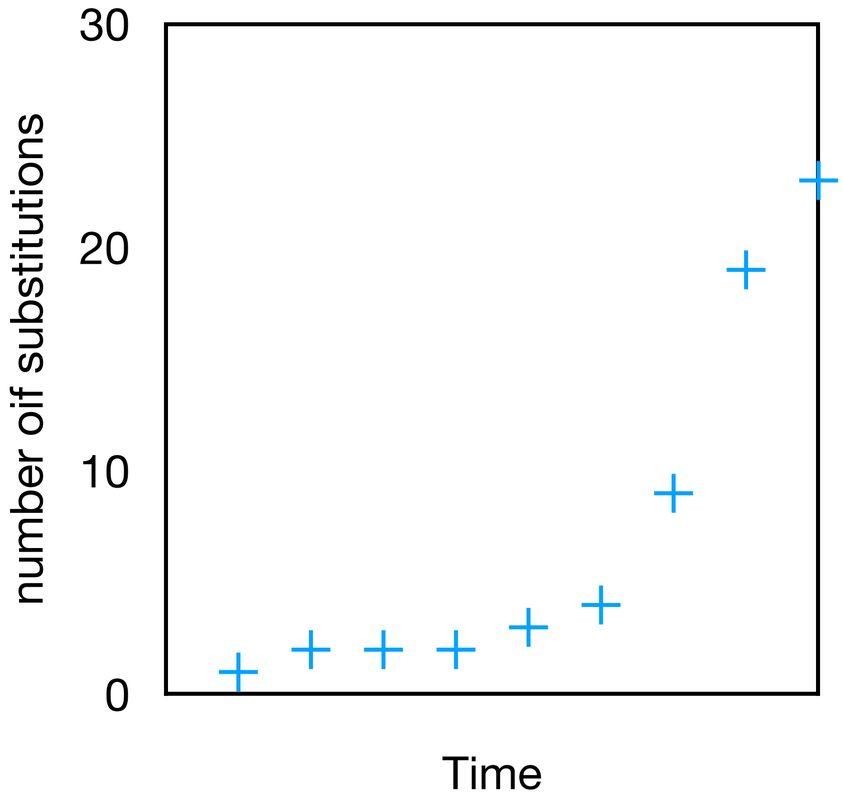


That would make estimating divergence times between species a breeze 😌



$$\text{Divergence Time } (sp_1\text{–}sp_3) = \frac{\text{Molecular Distance } (sp_1\text{–}sp_3)}{2}$$

… however, **the molecular clock assumption is often violated** …



… here, the plot suggest a non-linear pattern, suggesting accelerated rates at some point of the evolutionary history of a lineage …

🕰️ Strict Molecular Clock

assumes **constant rate of evolution** across all lineages
often **unrealistic** for large or heterogeneous datasets
can produce **biased age estimates** when rate variation exists

From the **relaxation** of the strict molecular clock **assumption** we get …

🕰️ Local Molecular Clock

different **substitution rates** apply to different **monophyletic groups** within a tree
each **local clock group** evolves at its own constant rate

🕰️ Relaxed Molecular Clock

allows **rate variation among lineages**
2 main types:

- **Uncorrelated clocks**: rates vary independently (e.g., lognormal or exponential priors)
- **Correlated clocks**: rates in neighboring branches are related

🕰️ Autocorrelated Relaxed Clock

- Substitution rates change gradually along the tree
- Daughter branches tend to inherit their parent's rate
- Models evolutionary "clock drift" (e.g., lognormal distribution)

🕰️ Uncorrelated (General) Relaxed Clock

- Substitution rates are drawn independently for each branch
- No assumption of rate similarity between parent and daughter
- Common distributions: lognormal, exponential, gamma

Least Squares Dating (LSD)

- fits divergence times to branch lengths by **minimizing squared differences**
- uses a **strict or relaxed clock**, depending on rate smoothing
- requires a fixed topology and branch lengths, extremely fast and ideal for large datasets
- less sensitive to model violations, but lacks full probabilistic framework
- software: IQ-TREE

Penalized Maximum Likelihood (PML)

- semi-**parametric** approach balancing fit and rate smoothness
- allows rate heterogeneity while discouraging sudden shifts
- requires a **fixed topology and branch lengths**
- computationally efficient for large-scale phylogenies
- software: r8s, treePL

Bayesian Inference

- fully **probabilistic framework**: models uncertainty in topology, rates, and dates
- can implement the **Fossilized Birth-Death (FBD) model** including speciation, extinction, etc
- avoids arbitrary calibration bounds by integrating fossil distributions
- can implement **total-evidence dating** combining morphological + molecular data
- software: BEAST, RevBayes

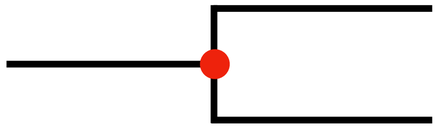
Tip Dating

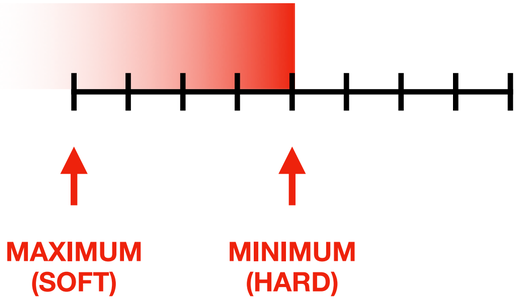
- dates are assigned directly to the tips of the tree
- uses dated samples (e.g. ancient DNA or fossils as terminals)
- integrates “extinct” information as part of the phylogenetic inference
- allows estimation of divergence times without node calibrations
- more sensitive to calibrations
- often used in total-evidence dating approaches

Node Dating

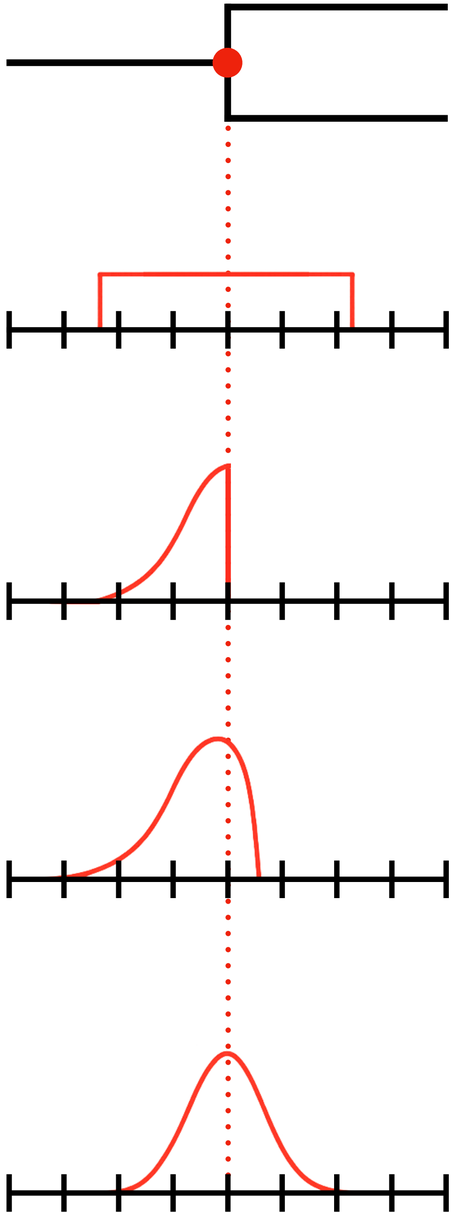
- dates are assigned to internal nodes of a phylogeny
- Relies on fossil calibrations placed at ancestral nodes
- Fossils provide minimum age constraints
- Assumes correct placement of fossils and accurate tree topology
- Common in analyses using only extant taxa

fossils
...
usually give only the minimum age constraint





more common foossil prior shapes



Uniform

- equal probability between minimum and maximum ages
- dimple and non-informative, often unrealistic
- use only when maximum bound is well-justified

Lognormal

- probability concentrated just above the minimum
- heavy tail toward older ages
- divergence likely predated the fossil, but not arbitrarily far back

Exponential

- peaks at the minimum bound, then declines rapidly
- suitable when fossil is likely close to the actual divergence
- simple and interpretable with one parameter (mean)

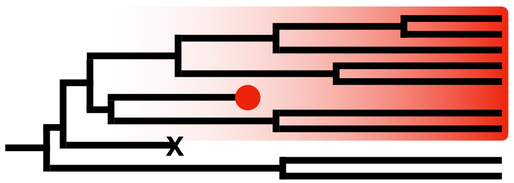
Normal

- bell-shaped, centered at a hypothesized mean
- requires strong justification for both center and spread
- less common for fossils, more for geological calibrations

fossils placement: Crown *versus* Stem

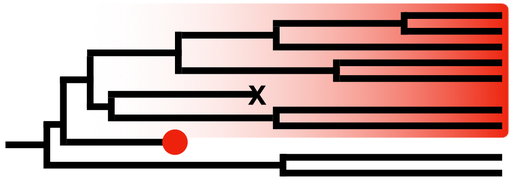
Crown Group

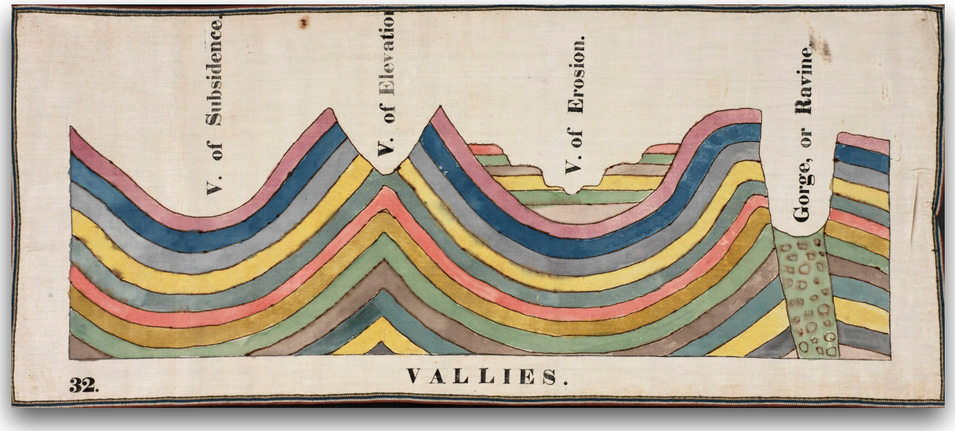
- includes the last common ancestor of all living members of a clade and all of its descendants.
- fossils placed here are within the clade, sharing all its defining synapomorphies.
- Exp: a fossil assigned to crown birds must postdate the divergence from non-avian dinosaurs.

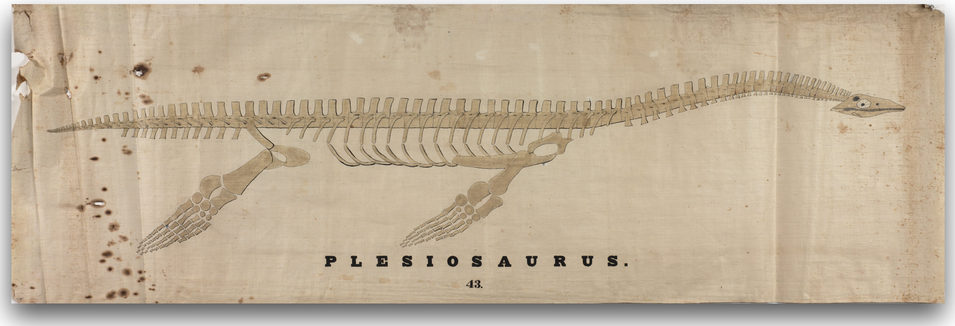


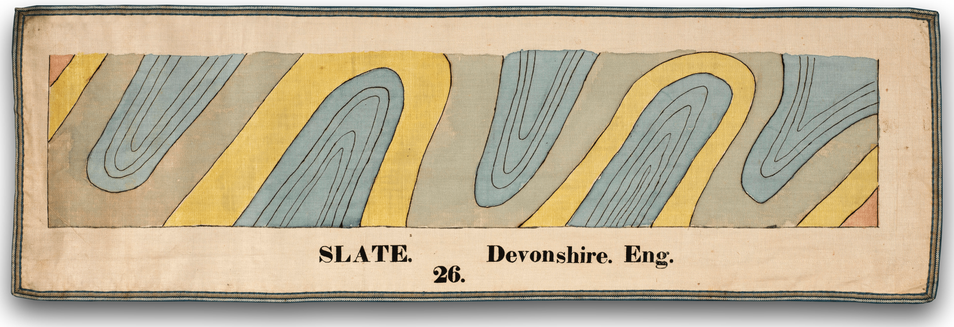
Stem Group

- includes extinct lineages closer to a crown group than to any other, but outside the crown itself.
- stem fossils help bracket the minimum age of the crown group.
- Example: feathered theropod fossils - e.g. Archaeopteryx - are stem birds.









Inferring selection

Phylogenetics aims to reconstruct patterns of **shared ancestry** among different organisms …

… while phylogenetic trees are themselves the result of these inferences, they can also be powerful **tools for studying evolutionary processes**.

Translation involves the sequential recognition of triplets of adjacent nucleotides called **codons**.

Since there are four nucleotides, there are 4*4*4 = **64 possible codons**.

Given that in the universal genetic code **there are 61 codons for a specific amino acid**, and three are stop codons, **but only 20 amino acids**.

Most amino acids are encoded by more than one codon.

The degeneracy of the genetic code allows two types of substitutions:

Nonsynonymous substitutions:

- Do **change** the amino acid sequence of a protein.
- Their rate (*dN*) is expected to reflect **selective pressure** on protein.

Synonymous substitutions:

- Do **not change** the amino acid sequence of a protein.
- Their rate (*dS*) is assumed to be **neutral** with respect to selection.

These assumptions - especially about *dS* neutrality - are **widely debated**!

Nonetheless, **codon models** provide insight into selective forces.

Muse and Gaut (MG94)

In 1994, Muse and Gaut introduced a codon substitution model that has **different rates for synonymous versus nonsynonymous substitutions**.

In codon models, two parameters are used:

- ratio of the number of syn. substitutions per syn. site (**dS**).
- ratio of the number of nonsyn. substitutions per nonsyn. site (**dN**).

This allows for the estimation of the absolute rates of each type, reflecting selective constraints on protein-coding genes.

Goldman – Yang 1994 (GY94)

Goldman and Yang's codon model, also published in 1994, similarly distinguishes codon changes by their effect on the encoded amino acid. Rather than separate dN and dS, it is often parameterized in terms of their ratio **ω = dN/dS**.

$\omega < 1$ - purifying selection
$\omega = 1$ - neutral evolution
$\omega > 1$ - positive selection

The GY94 framework also typically incorporates a **nucleotide mutation model** - it includes a transition/transversion bias κ - and **codon frequency parameters**- using empirical frequencies or an F3×4 model of base frequencies at codon positions.

GY94 and MG94 laid the groundwork for most modern codon phylogenetic analyses by explicitly modeling selective pressure on proteins.

More complex codon models to detect selection followed shortly after:

- **site-specific models** let dNdS varies across codons
- **branch-specific models** let dNdS vary across lineages
- **branch-site models** let dNdS vary across lineages and codons

- **site-specific models** let dNdS vary across codons … dNdS values are averaged throughout all branches of our phylogeny … but we know that selection can differ between different lineages!

- **branch-specific models** let dNdS vary across lineages … dNdS values are averaged across all codons of our alignment … but we know that selection can differ widely between different gene!

- **branch-site models** relax at the same time the previous assumptions to detect positive selection on specific sites and lineages …

…some **jargon** in codon models ,,,

Foreground: different class(es) of branches of the phylogeny that are allowed to have a different dN/dS - whether higher or lower - than the rest.

Background: all other branches not under explicit testing and that provide the baseline for comparison with the foreground.

Purifying Selection ($\omega < 1$) also known as negative selection, acts to eliminate harmful mutations that alter protein function. It preserves the integrity of important genes by discouraging changes at the amino acid level. This is the most common form of selection in coding regions, especially in genes essential for basic cellular processes.

Neutral Evolution ($\omega = 1$) occurs when mutations neither benefit nor harm the organism—they simply drift through populations over time. In these cases, synonymous and nonsynonymous changes accumulate at similar rates. This process is expected at sites where function is not tightly constrained, such as in pseudogenes or certain non-coding regions.

Positive Selection ($\omega > 1$) promotes mutations that offer an advantage to the organism. These changes are often fixed rapidly in populations and can lead to new or enhanced functions. It's a hallmark of evolutionary innovation, commonly observed in genes involved in immune defense, sensory perception, or reproductive function. An elevated dN/dS ratio is a strong indicator that positive selection may be at work.

Relaxed Selection ($\Delta\omega > 0$ between lineages) isn't defined by an absolute ω value, but by a comparison between two evolutionary contexts. If a gene or lineage previously under strong purifying selection (low ω) experiences reduced functional constraint, its ω may increase — not necessarily above 1, but higher than before. It reflects weaker purifying selection, not active adaptation and os detected by comparing ω across branches and gens.

A **dNdS > 1** is considered **unrealistic** when inferred for whole gene.

Such a value would suggest that all codons are accumulating **twice as many nonsynonymous mutations** as synonymous ones.

This signal is rarely linked to genuine positive selection, even when associated with specific branches.

These extreme values point to issues such as **pseudogenization** or **errors** in alignment and orthology, which can impact **downstream analyses**.

This aspect can be leveraged for **phylogenetic subsampling** and dNdS can be useful for finding problematic regions, sequences or whole genes.

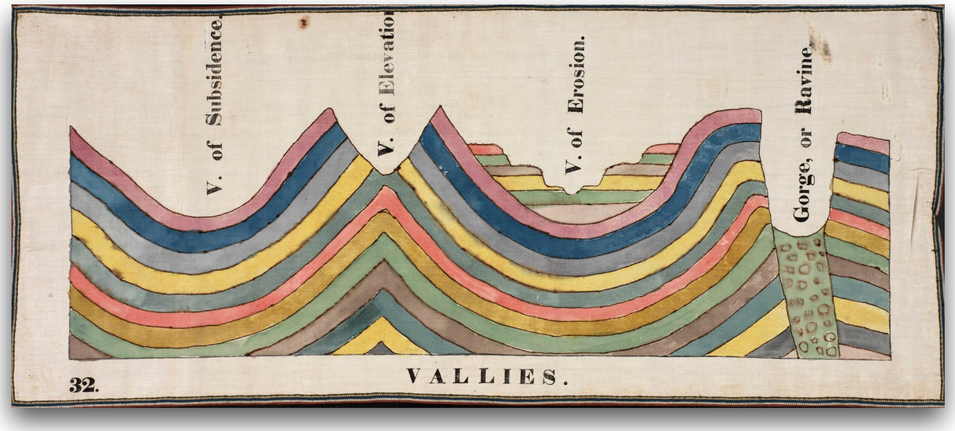
codon-based alignments

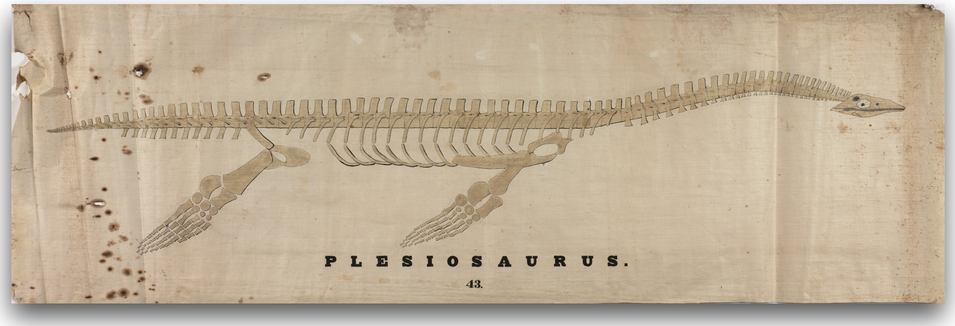
A type of multiple sequence alignment where **nucleotide sequences are aligned in sets of three nucleotides** (codons).

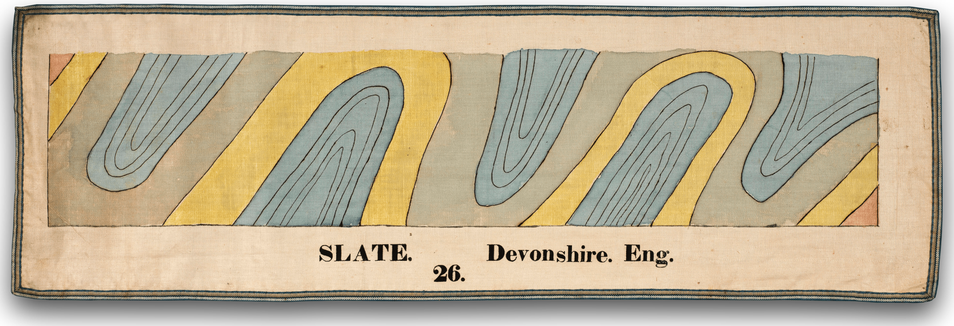
Insertions or deletions - **indels** - must involve entire codons - **multiples of three nucleotides** - to avoid disrupting the reading frame.

Ensures that the reading frame is preserved across all sequences, **preventing frame-shift mutations** that are not biologically plausible.

… fundamental for codon models!







modelling trait evolution on phylogenies: discrete traits

What can you do with phylogenetic comparative methods?

- model fitting - testing evolutionary hypotheses on how trait evolve.
- ancestral state(s) reconstructions - **will do in the practical!**
- diversification - speciation and extinction rates.
- biogeography 🌍.
- …

Character Types:

- **Discrete Characters:** traits with a finite number of distinct states. *Examples:* presence or absence of wings, flower color categories.

- **Continuous Characters:** traits that vary along a measurable scale. *Examples:* body length, enzyme activity levels, beak depth.

Analytical Approaches:

- **Discrete Characters:** analyzed using models like the Mk model, which describes the evolution of discrete traits as a Markov process.

- **Continuous Characters:** often examined with models as Brownian Motion (BM) and Ornstein-Uhlenbeck (OU) processes.

Discrete characters

The most commonly used model for discrete character evolution on trees is a model called the **Mk model**.

- **M** for the process modelled follows the Markov property
- **k** as the model is generalized to include an arbitrary number (k) states

The Q matrix is the core of the modeland gives the instantaneous transition rates between states… **does it remember you anything ?!**

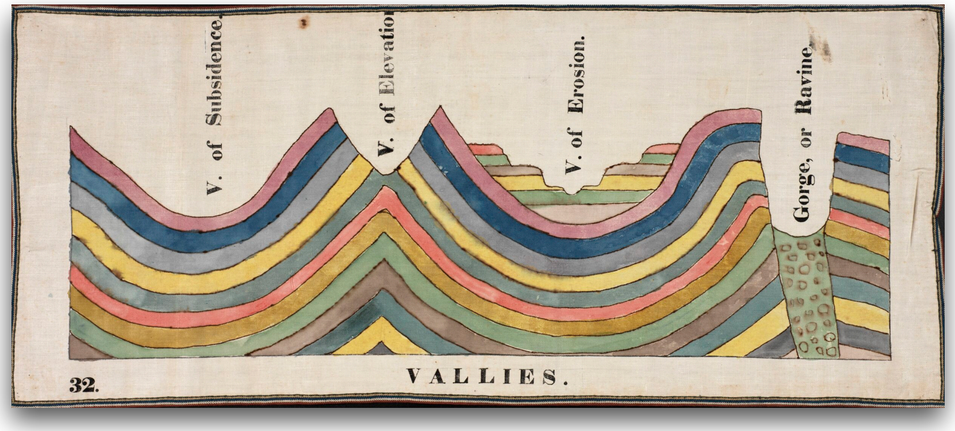
Depending on the assumptions we make about the transition rates, the Mk model can be parameterised in several different ways:

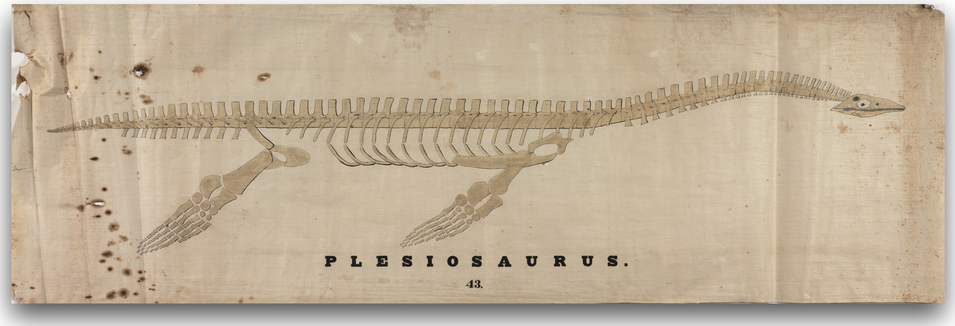
- **ER** - Equal Rates: all transitions have the same rate.
- **SYM** - Symmetrical: different rates but forward and reverse trans. equal.
- **ARD** - All Rates Different: each transition has its own rate.

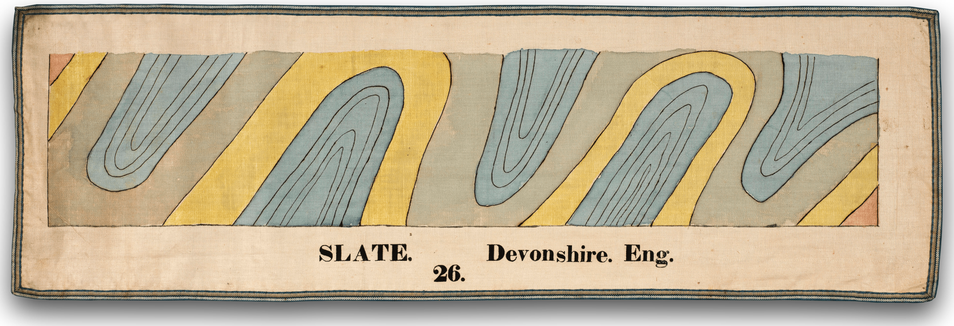
Joint *versus* marginal ancestral reconstruction

- Joint reconstruction is finding the set of character states at all nodes that (jointly) maximize the joint probability (likelihood) of the entire tree.

- marginal reconstruction is finding the state at the current node that maximizes the likelihood integrating over all other states at all nodes.







modelling trait evolution on phylogenies: continuous traits

Character Types:

- **Discrete Characters:** traits with a finite number of distinct states.
  Examples: presence or absence of wings, flower color categories.

- **Continuous Characters:** traits that vary along a measurable scale.
  Examples: body length, enzyme activity levels, beak depth.

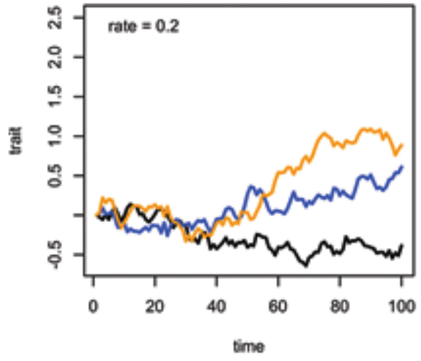
Analytical Approaches:

- **Discrete Characters: Mk model**

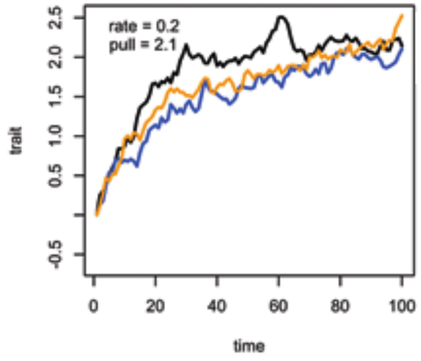
- **Continuous Characters: ??!**

Models of continuous trait evolution

- **Brownian Motion (BM)**:
  trait evolve randomly and variance increases with time



- **Ornstein-Uhlenbeck (OU)**:
  add stabilizing selection anf traits evolve toward an optimum



Models of continuous trait evolution

What is Brownian Motion (BM)?

- a random walk model of trait evolution.
- at each step, trait changes by a small, normally distributed random amount.
- no preferred direction or optimum - purely stochastic.

Key Features

- trait variance increases linearly with time since divergence.
- closely related species tend to have more similar traits.
- trait evolution is neutral: no selection toward an optimum.

BM is s a **null model** for trait evolution.

It is a good fit for traits evolving neutrally or with drift.

Models of continuous trait evolution

What is the OU Model?

- An extension of Brownian Motion that includes stabilizing selection.
- Traits evolve with a tendency toward an optimal value θ.

Key Features

- Variance increases with time but levels off as values are “pulled” toward the optimum.
- Includes three parameters:
  - **θ**: trait optimum
  - **α**: strength of selection toward the optimum
  - **σ2**: evolutionary rate - as in BM!

When to Use OU?

- When traits show selective constraints or adaptive peaks.
- For testing **adaptive hypotheses** or detecting **convergent evolution**.

What Is Phylogenetic Signal?

- The tendency for related species to resemble each other more than expected under a model of random trait distribution.
- Reflects the non-independence of species due to shared ancestry.

Why It Matters

- Strong signal: trait variation is structured by phylogeny.
- Weak signal: trait may evolve rapidly or independently of phylogeny.
- Helps choose the right model and correct for autocorrelation.

Pagel's λ (Lambda)

- A branch-length scaling parameter that quantifies how much the trait variation is explained by the phylogeny.
- Modifies the internal branches of the tree to best match the observed trait covariance.

How It Works

The internal branches of a phylogenetic tree represent the shared evolutionary history between species.

Pagel's λ rescales these internal branches to reflect how much shared ancestry influences trait similarity.

λ = 1 - full phylogenetic covariance - trait evolves under Brownian motion.
λ = 0 - internal branches collapsed - trait evolve independently of phylo.

Blomberg's K

- A variance-based metric that quantifies how much trait similarity among species is explained by their shared evolutionary history.

- Specifically, it compares observed trait variance across a phylogeny to the variance expected under Brownian motion (BM) evolution.

How It Works

- The core concept: $K = \frac{\text{Observed phylogenetic signal}}{\text{Expected signal under Brownian motion}}$

- $K \approx 1$ Trait matches BM expectations (moderate signal)
- $K > 1$ Trait is more conserved than expected
- $K < 1$ Trait is less conserved (weak signal or convergence)

a phylogenetic problem: non-independence in data

Species are not independent data points! 🐍🦎🧸🦧🦍

They share evolutionary history! 🐟🐡🐠🦀🪲

… and this creates statistical non-independence.

What happens if we disregard phylogeny?

- Violates the assumption of independent distributed residuals.
- inflated Type I error rates - false positives!

Use PIC and PGLS!