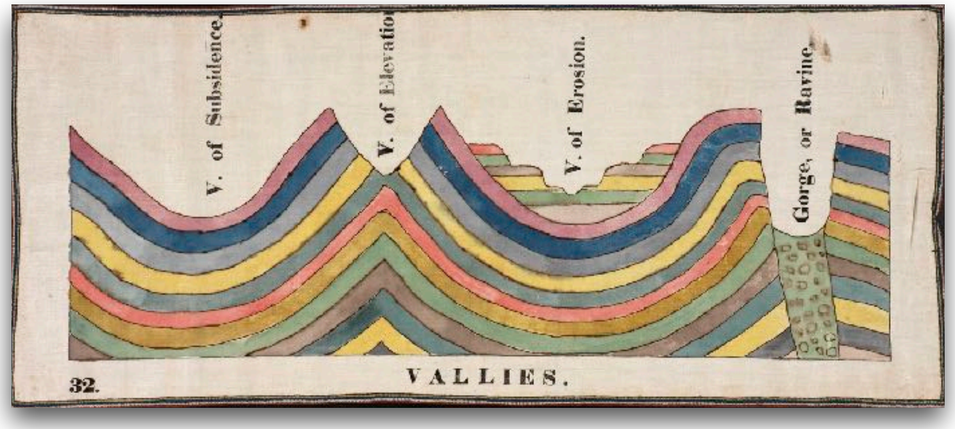
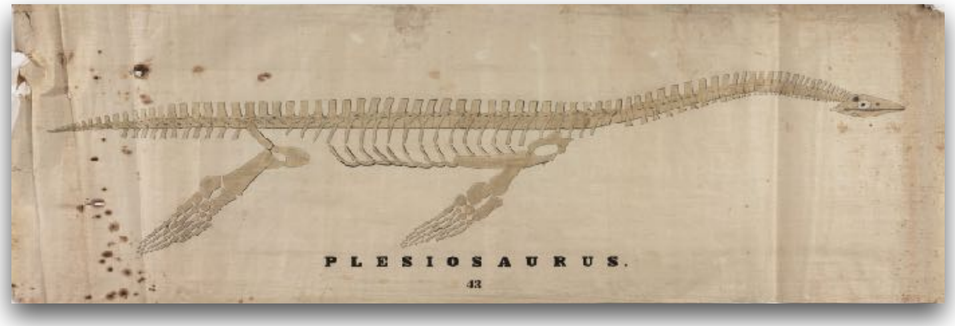


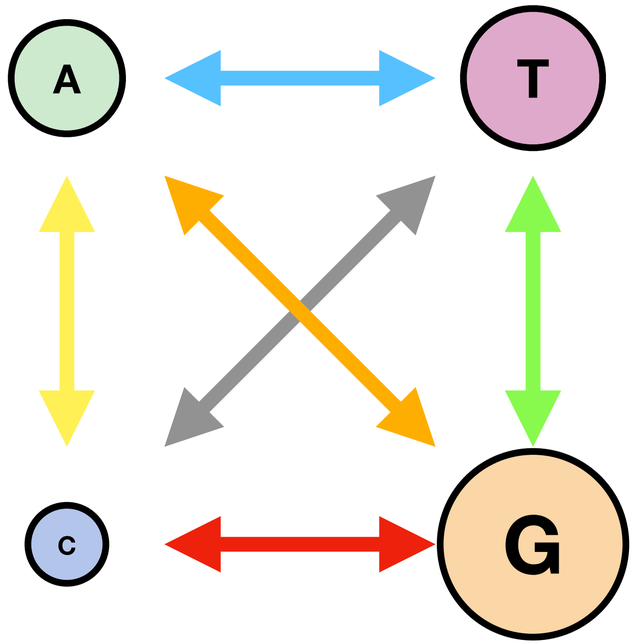


Maximum Likelihood (ML)

GTR (General Time Reversible):

It includes a full set of six substitution rate parameters (one for each pair of nucleotides) and allows any set of equilibrium base frequencies.

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



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

For example a normal distribution can be described by 2 parameters: the **mean** and the **variance**.

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

- rates of transitions / transversions / ... 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.

$$P(D|M)$$

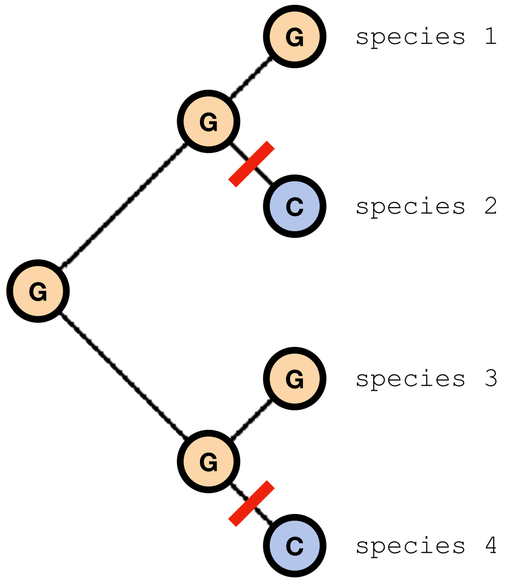
What is the probability of observed data given that a certain model is true?

P(D|M)

probability (biological 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
            *
```

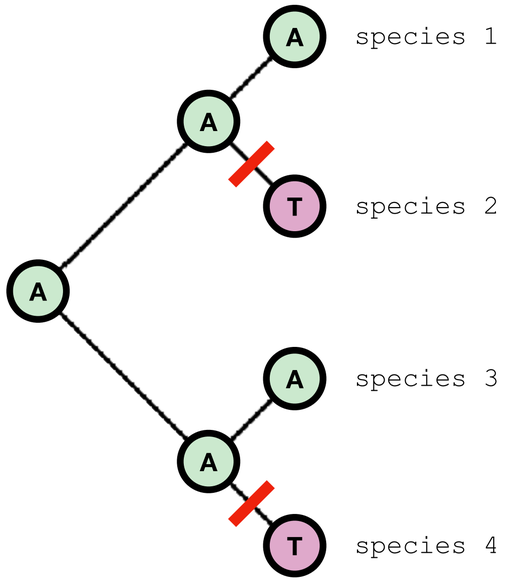


$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G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

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

$P_{site2} = P_{GC} \times P_{GC} \times P_G \times P_G \times P_G \times P_G \times P_{G(root)}$

```
                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
                        *
```

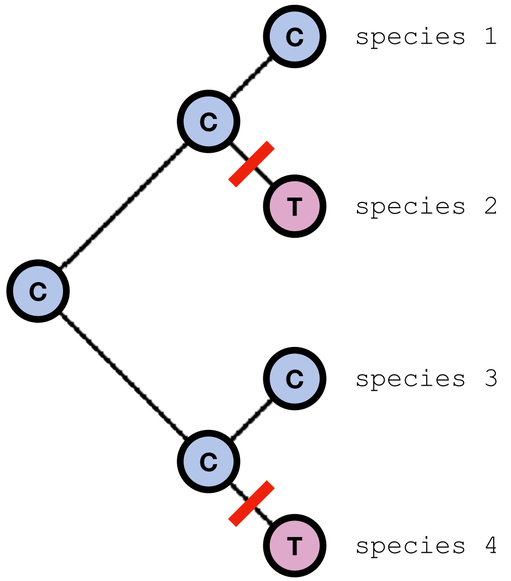


$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G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

$P_{AT} = r_{cT}\pi_T$

$P_{site4} = P_{AT} \times P_{AT} \times P_A \times P_A \times P_A \times P_A \times P_{A(root)}$

```
                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
                        *
```

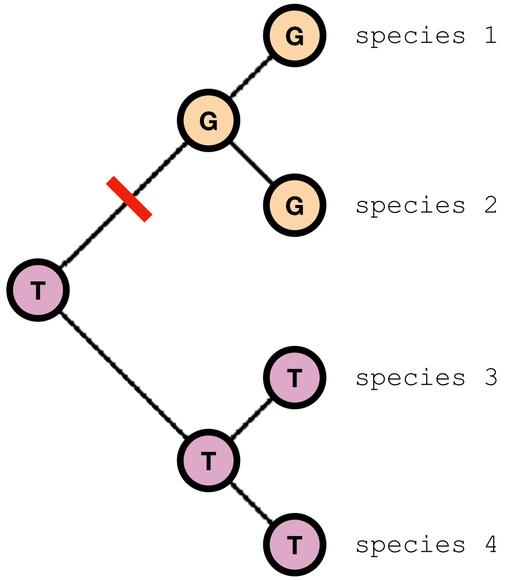


$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G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

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

$P_{site5} = P_{CT} \times P_{CT} \times P_C \times P_C \times P_C \times P_C \times P_{C(root)}$

```
                  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
                              *
```



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

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

$P_{site7} = P_{TG} \times P_G \times P_G \times P_T \times P_T \times P_T \times P_{T(root)}$

Note this is for one ancestral state assignment .. but what about ambiguous internal nodes?! 😱

We rely on **conditional likelihood …** the full likelihood sums over all possibilities!

- **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.

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).

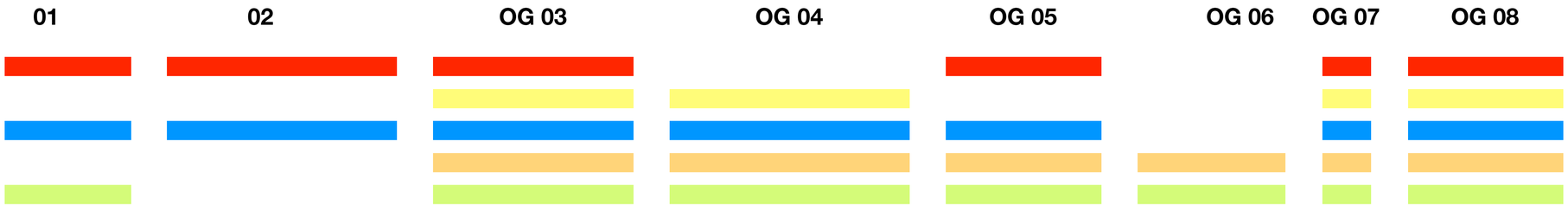
CONCATENATION

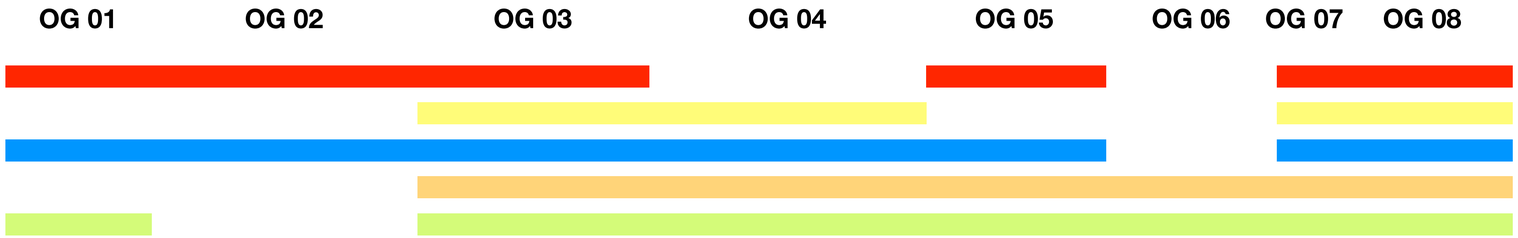
REMEMBER: it works for MP / ML / BI … but not coalescent methods!

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

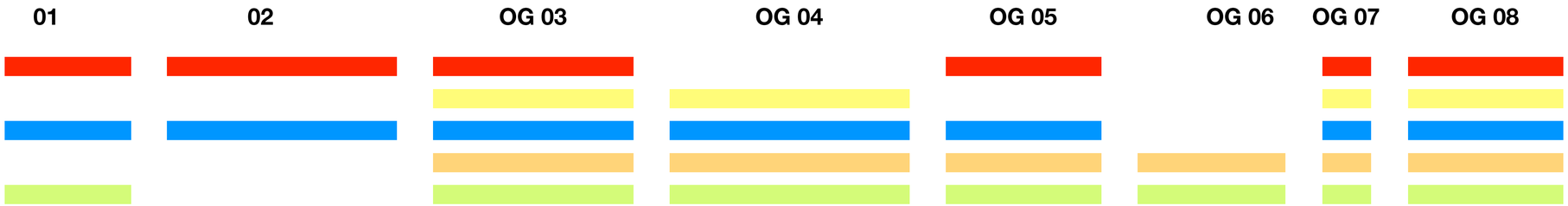
… orthology inference has generated a ton of orthogroups, we have aligned and filtered them …

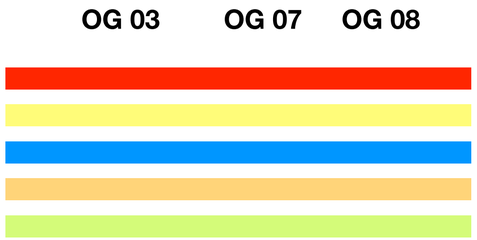
Concatenation - also known as the *supermatrix approach* - involves combining **multiple alignments** (gene or protein) into **a single large dataset** to infer the phylogeny.



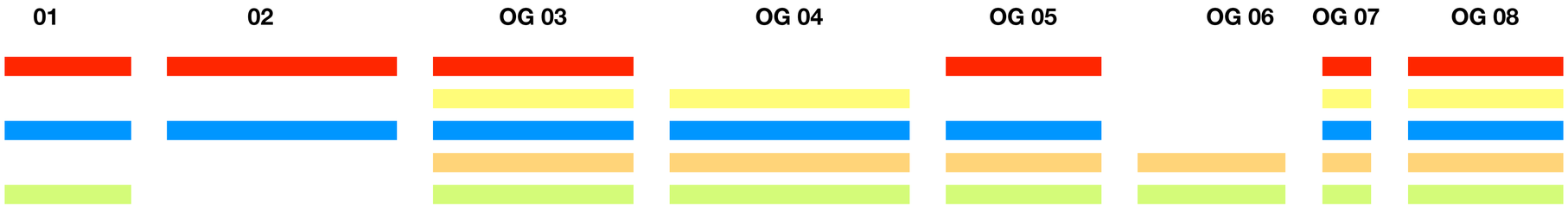


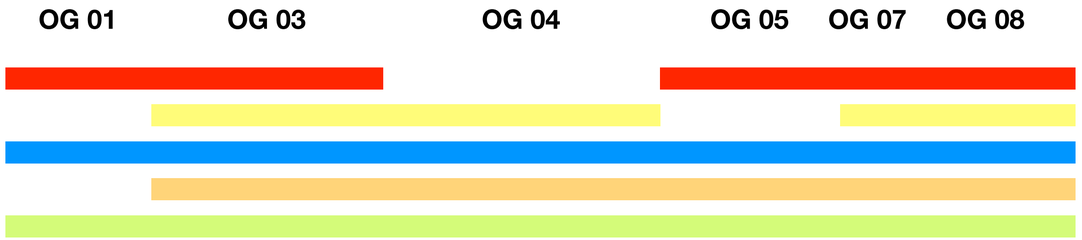
min. matrix occupancy = 00%





min. matrix occupancy = 100%





min. matrix occupancy = 50%

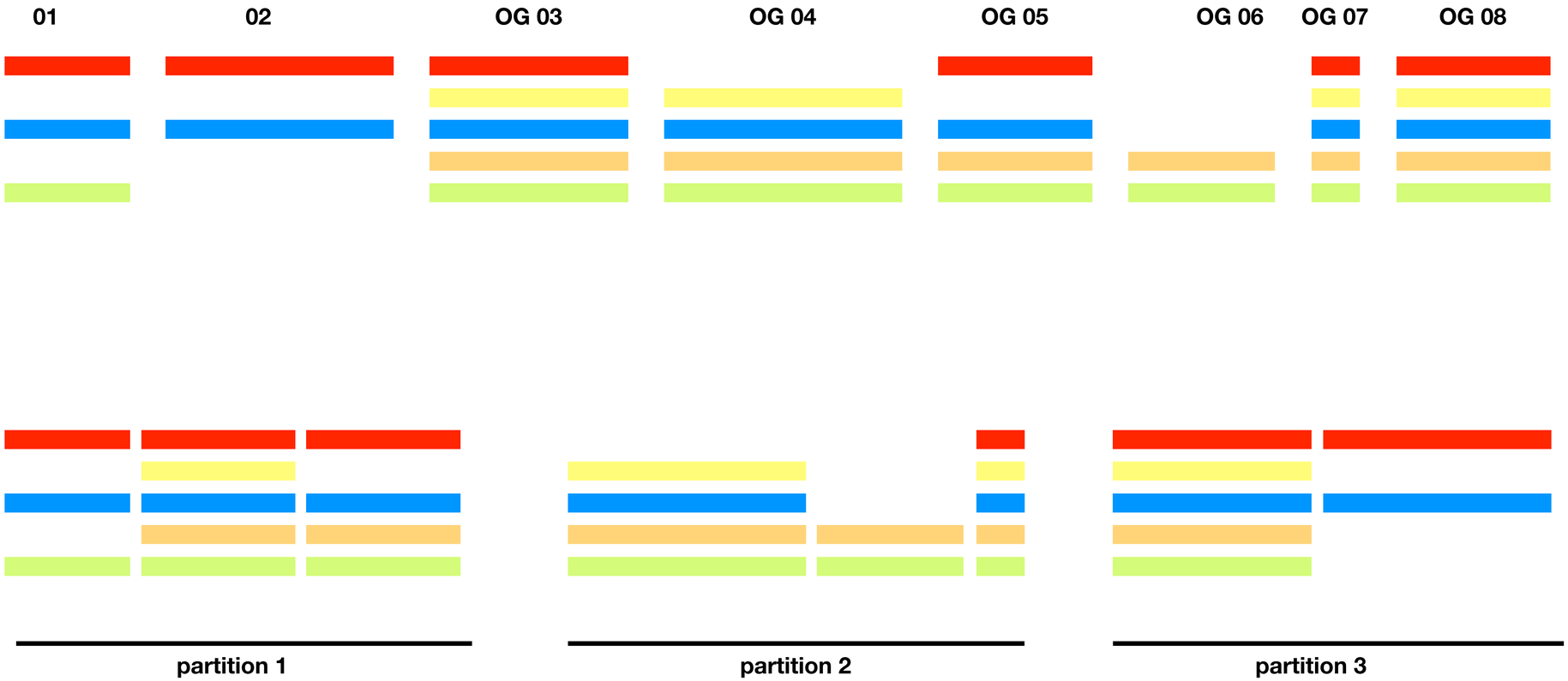
We **concatenate genes** to improve the resolution of phylogenomic analyses. Concatenation increases the strength of phylogenetic signal and reduces stochastic errors.

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-parameterization: assigning a unique model to every possible partition can lead to overfitting, where the model captures noise rather than true evolutionary signals.

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.

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

How are partitions merged?

Because adding a parameter (dimension) to a model will always ensure a maximum likelihood at least as large as without the parameter, some penalty must be imposed when parameters are added.

How large this penalty should be is not easy to define, which has led to many different possible criteria, e.g., the **AIC** (Akaike information criterion; Akaike 1974), **AICc** (second-order AIC; Sugiura 1978; Hurvich and Tsai 1989), and **BIC** (Bayesian information criterion; Schwarz 1978).

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)**.

FINISH