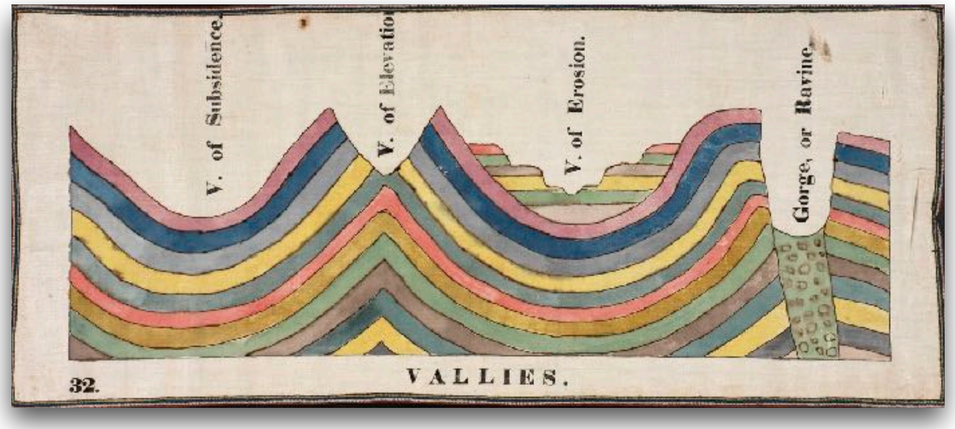
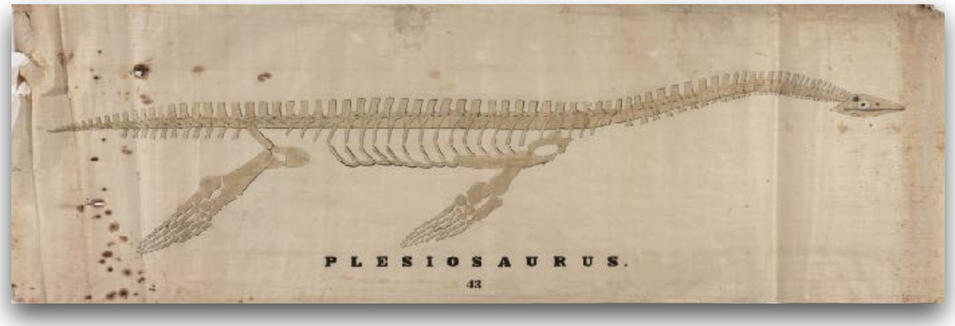




Complex models

Simple substitution models assume that all sites in an alignment evolve under the same substitution process.

However real sequence data often violate these assumptions …

Heterogenity can affect:

- rates across sites / across lineages / across sites & lineages
- composition across sites
- …

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

PARTITION MODELS

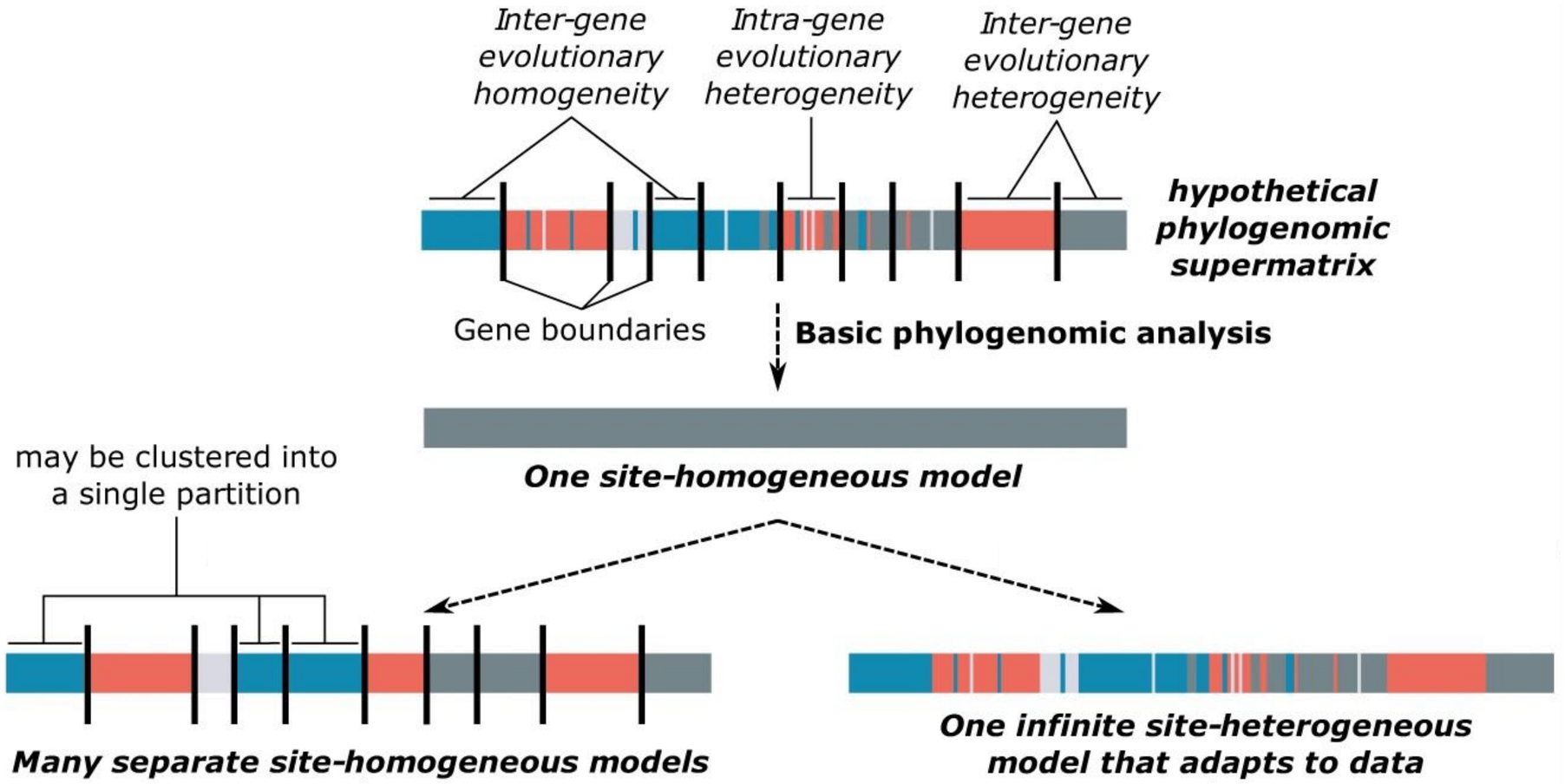
- Divide a multiple sequence alignment into distinct subsets (e.g. genes or codon positions).
- Each subset (partition) evolves under its own substitution model and parameters.
- Useful when prior knowledge about data structure is available.

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

Three types of **partition models**:

Three types of **partition models**:

EDGE-LINKED WITH **EQUAL BRANCH LENGTHS**

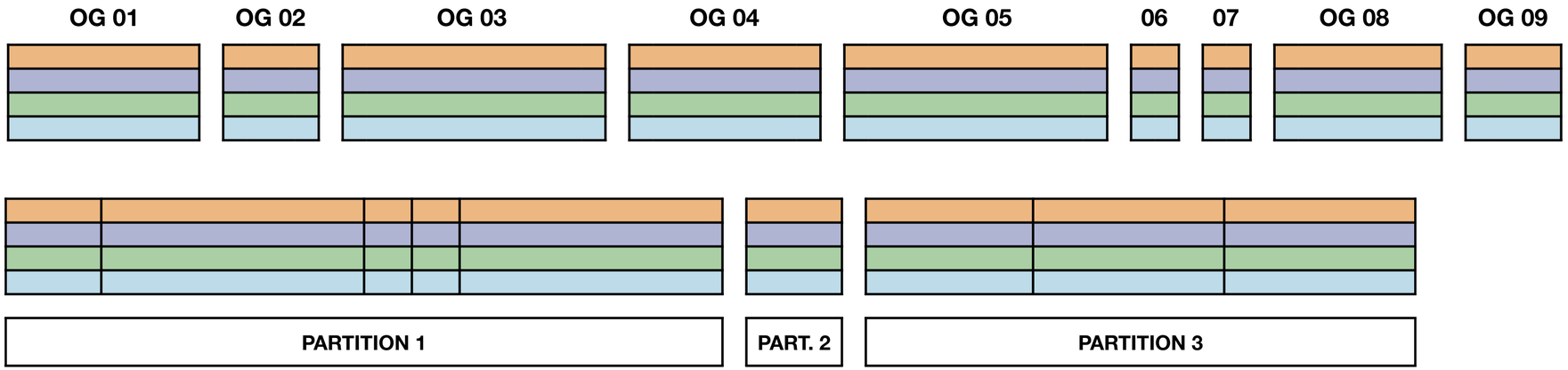
All partitions share an identical set of branch lengths.

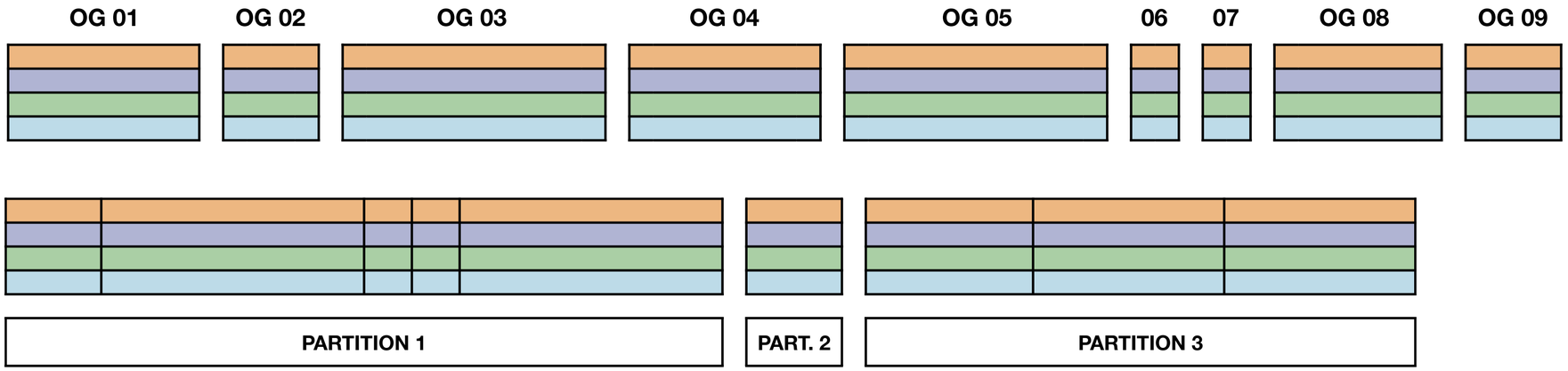
EDGE-LINKED WITH **PROPORTIONAL BRANCH LENGTHS**

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

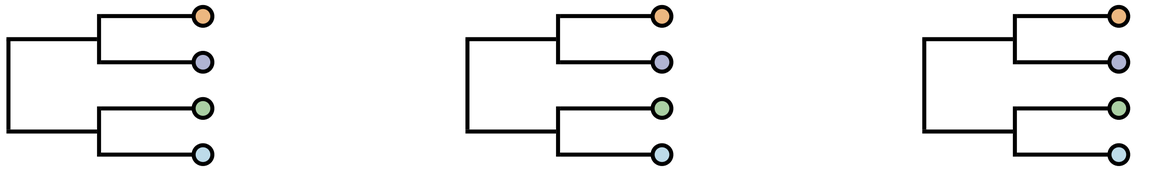
A	P_2
C	P_2
T	P_2
G	P_2

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

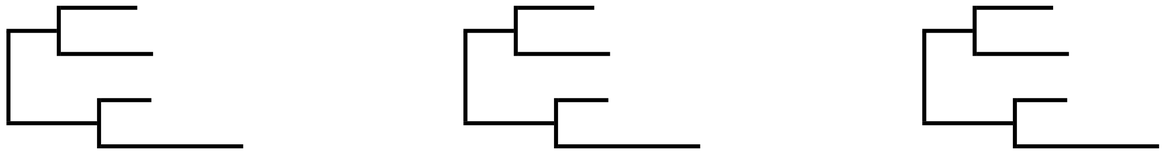
A	P_3
C	P_3
T	P_3
G	P_3

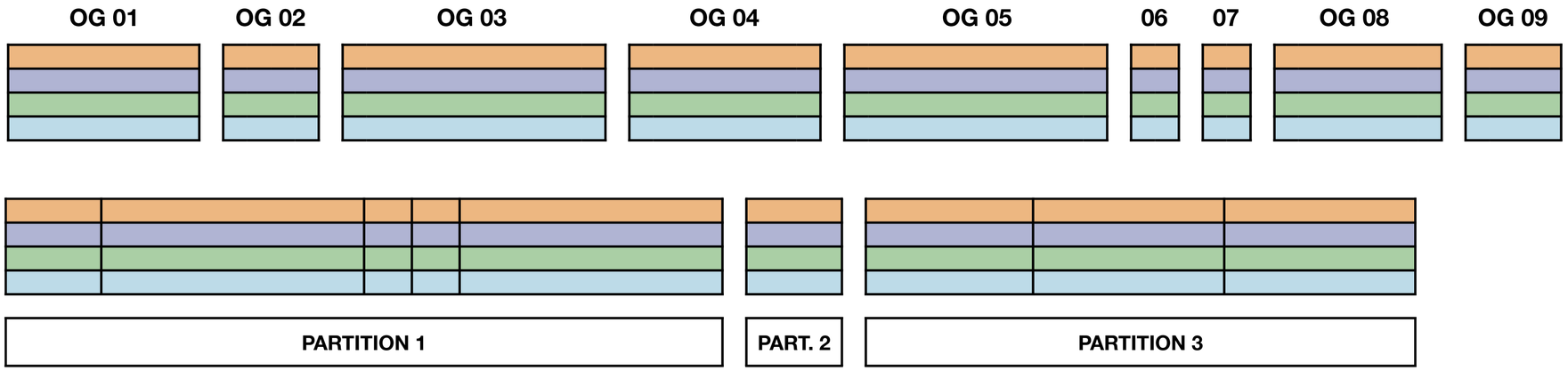
Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_3	Q_3

1X TOPOLOGY



1X BRANCHLENGTHS





EDGE - PROPORTIONAL 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

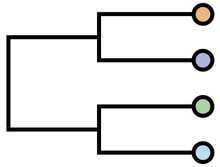
A	P_2
C	P_2
T	P_2
G	P_2

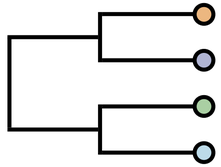
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2

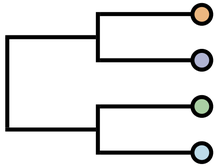
A	P_3
C	P_3
T	P_3
G	P_3

Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_3	Q_3

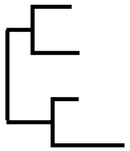
1X TOPOLOGY



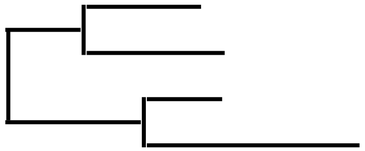




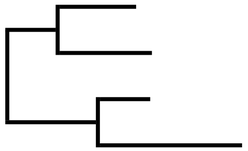
~1X BRANCHLENGTHS



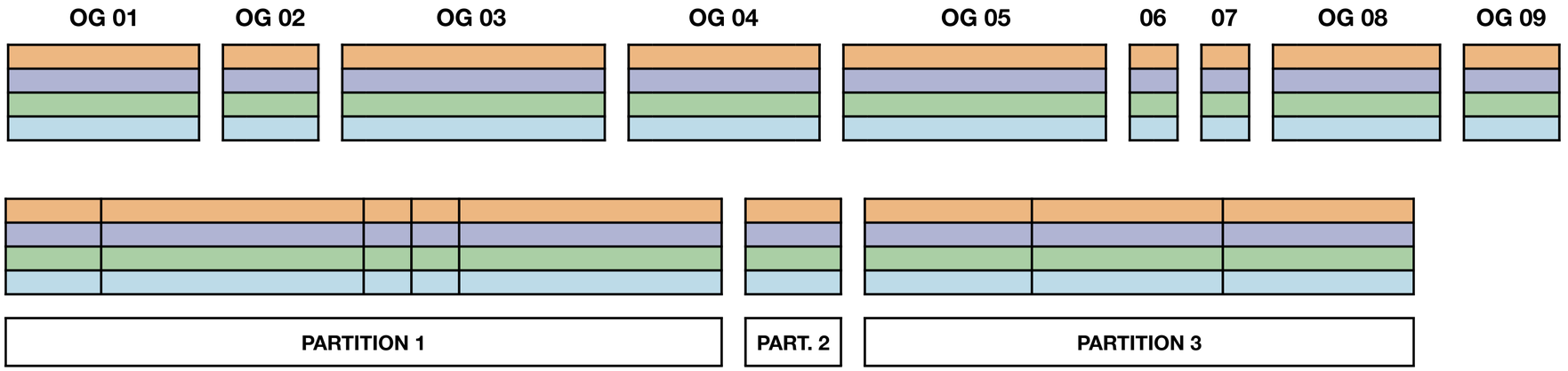
0.5X



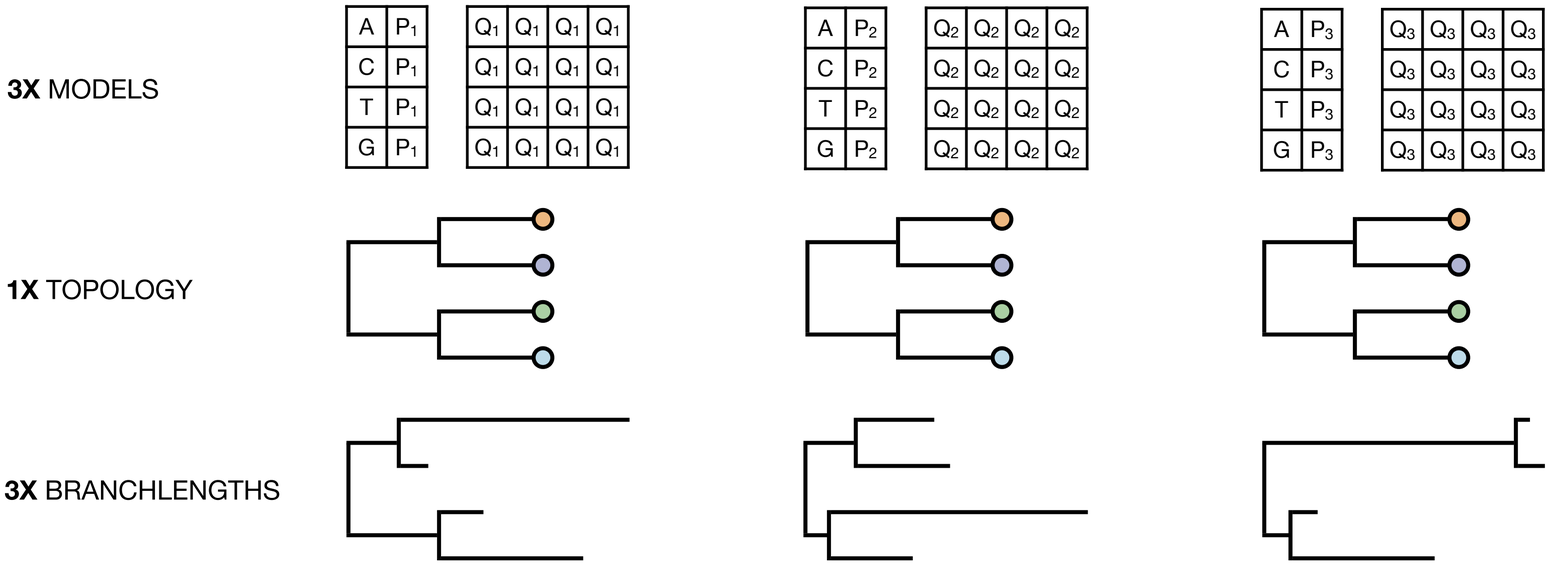
1.5X



1X



EDGE - UNLINKED PARTITION MODEL



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 🧐

Three types of **mixture models**:

RATE MIXTURE MODEL

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

PROFILE MIXTURE MODEL

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

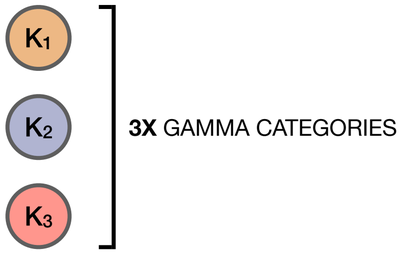
Empirical profile models use fixed frequency profiles derived from large datasets - e.g. LG+C60.
Bayesian profile models infer the number and composition of profiles from the data - e.g. CAT.

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
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

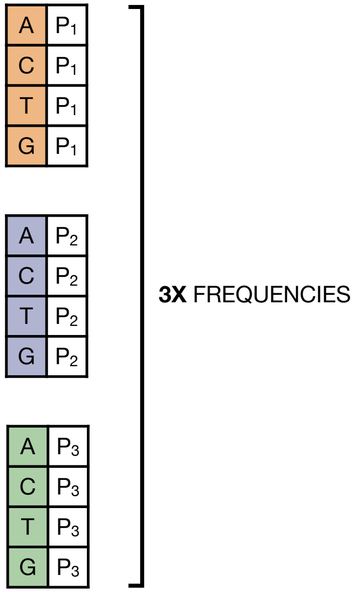


A	P_1
C	P_1
T	P_1
G	P_1

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

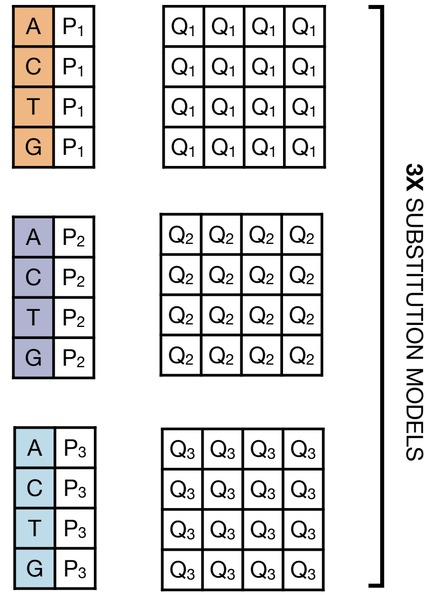
PROFILE MIXTURE MODEL

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



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



HETEROTACHY

the same site evolves at different rates on different branches.

… +Γ assumes one set of branch lengths for all sites.

… but the **GHOST** (General Heterogeneous evolution On a Single Topology) model specifically addresses heterotachy.

Sites are assigned probabilistically to classes (like a mixture model) each with its own branch lengths and Q matrix.

Similar to the edge-unlinked partition model, but no a priori partitioning needed …

PS: all previous models relax assumptions about rates and frequencies, but **MAST** is the first to relax the topology itself. This accounts for ILS, introgression, and recombination within a single concatenated ML framework. We will lear better on how to account for topological discordance in lesson 12! 👀

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**: the likelihood is a weighted sum of per-component likelihoods.
- **Assumption**: the site-to-class assignment is *a priori* unknown.

FINISH