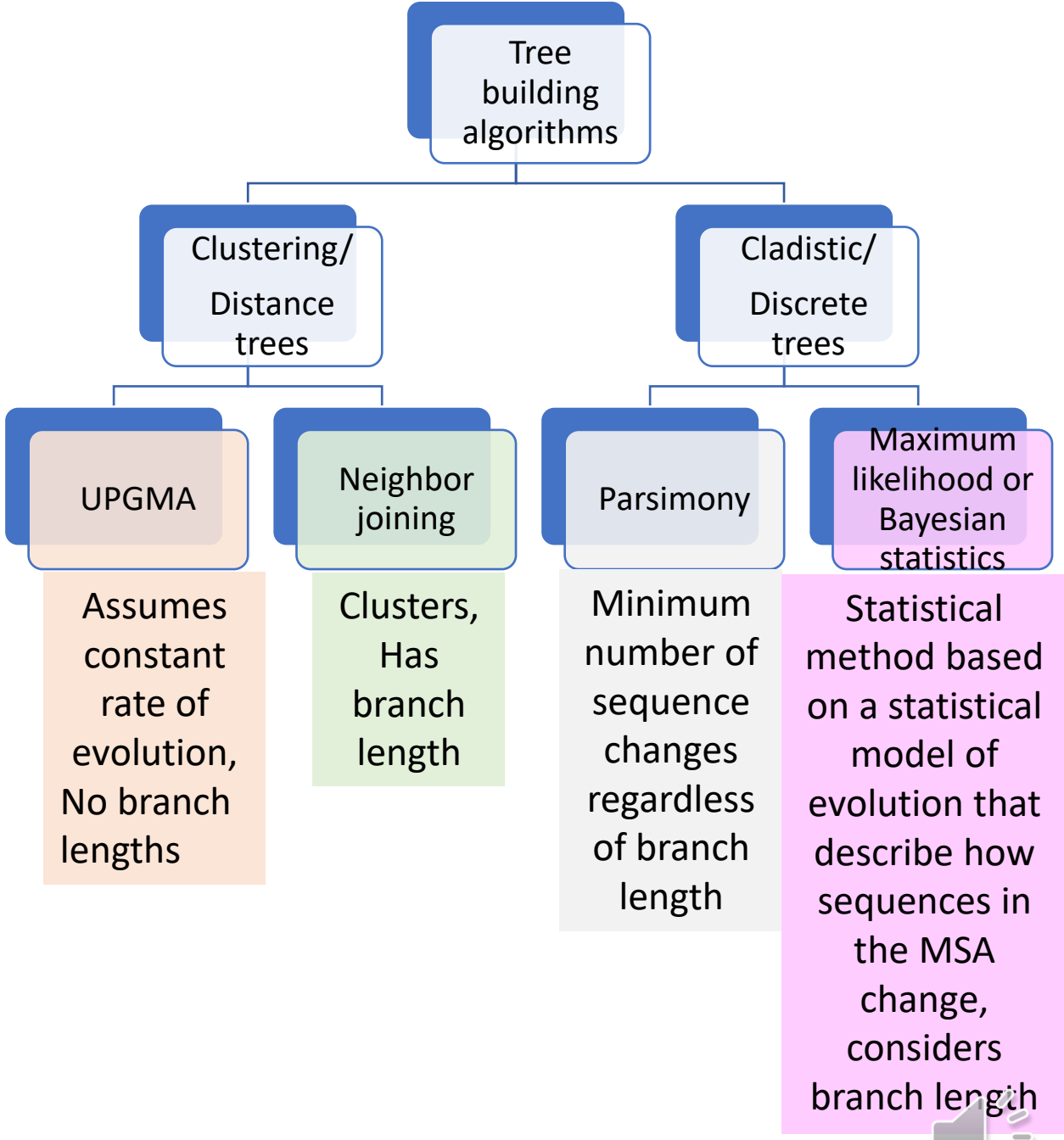
How to build a tree?

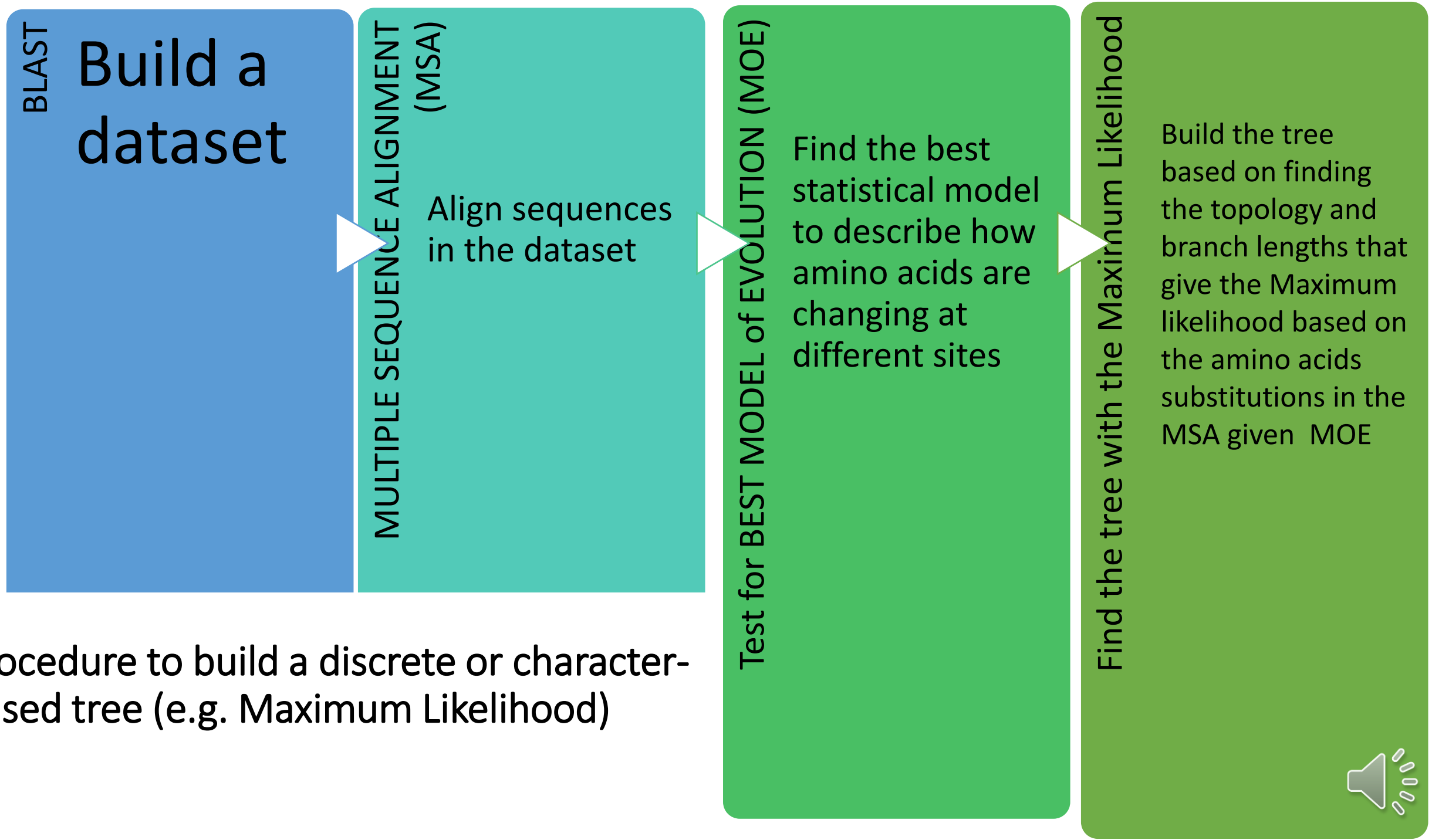
How to build a tree?

1. Build a dataset for your protein or gene
   - BLAST – What species? What database? What substitution matrix? What e-value?
   - Informative sequence names
2. Build a multiple sequence alignment.
3. Choose a tree building algorithm



In short, distance trees are based on pairwise sequence identities

- Different operations are done to reach the final tree, but it is only based on distance from pairwise sequence identities (or other metric such as morphological characters).



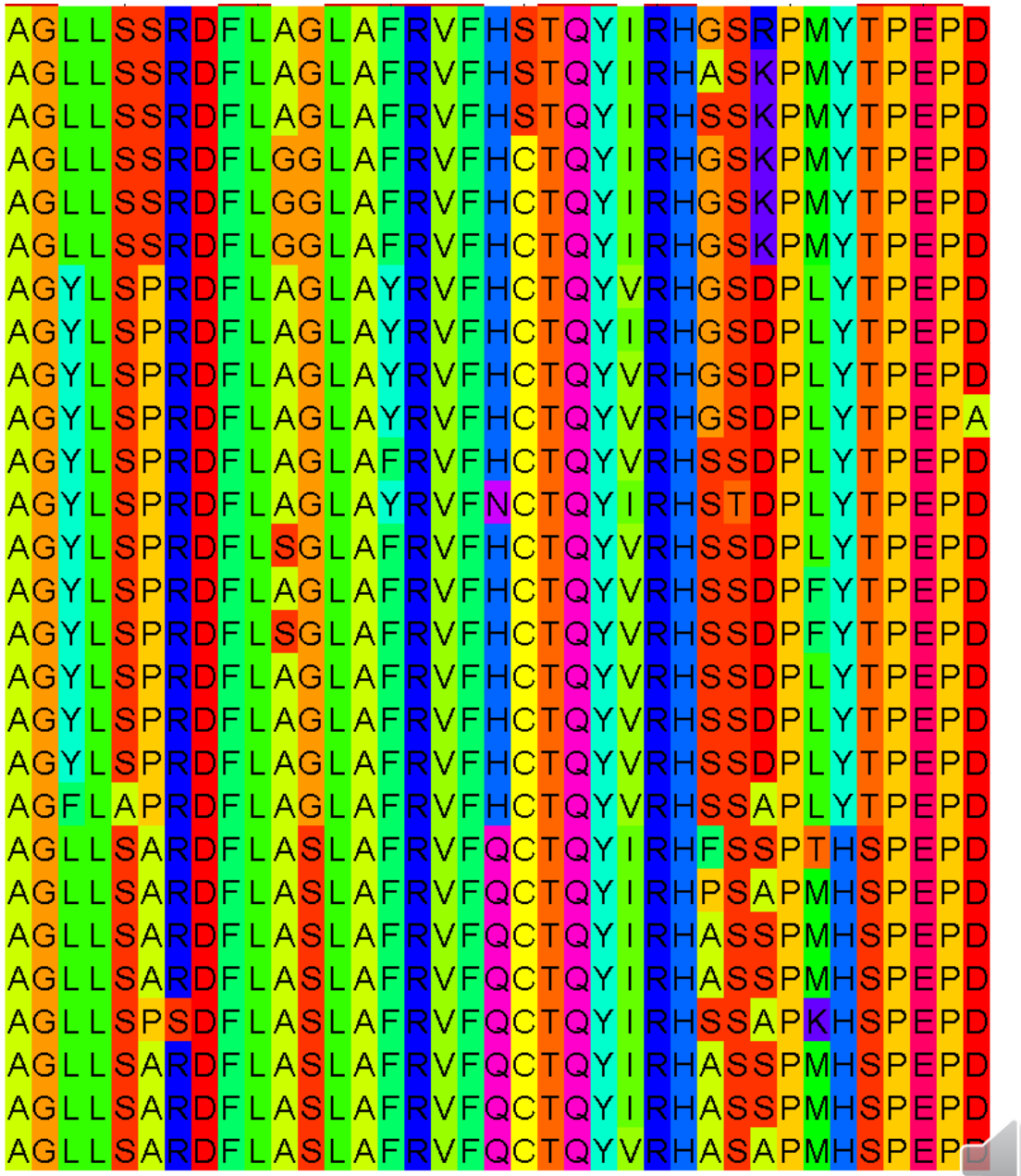
Procedure to build a discrete or character-based tree (e.g. Maximum Likelihood)

Discrete/character-based methods

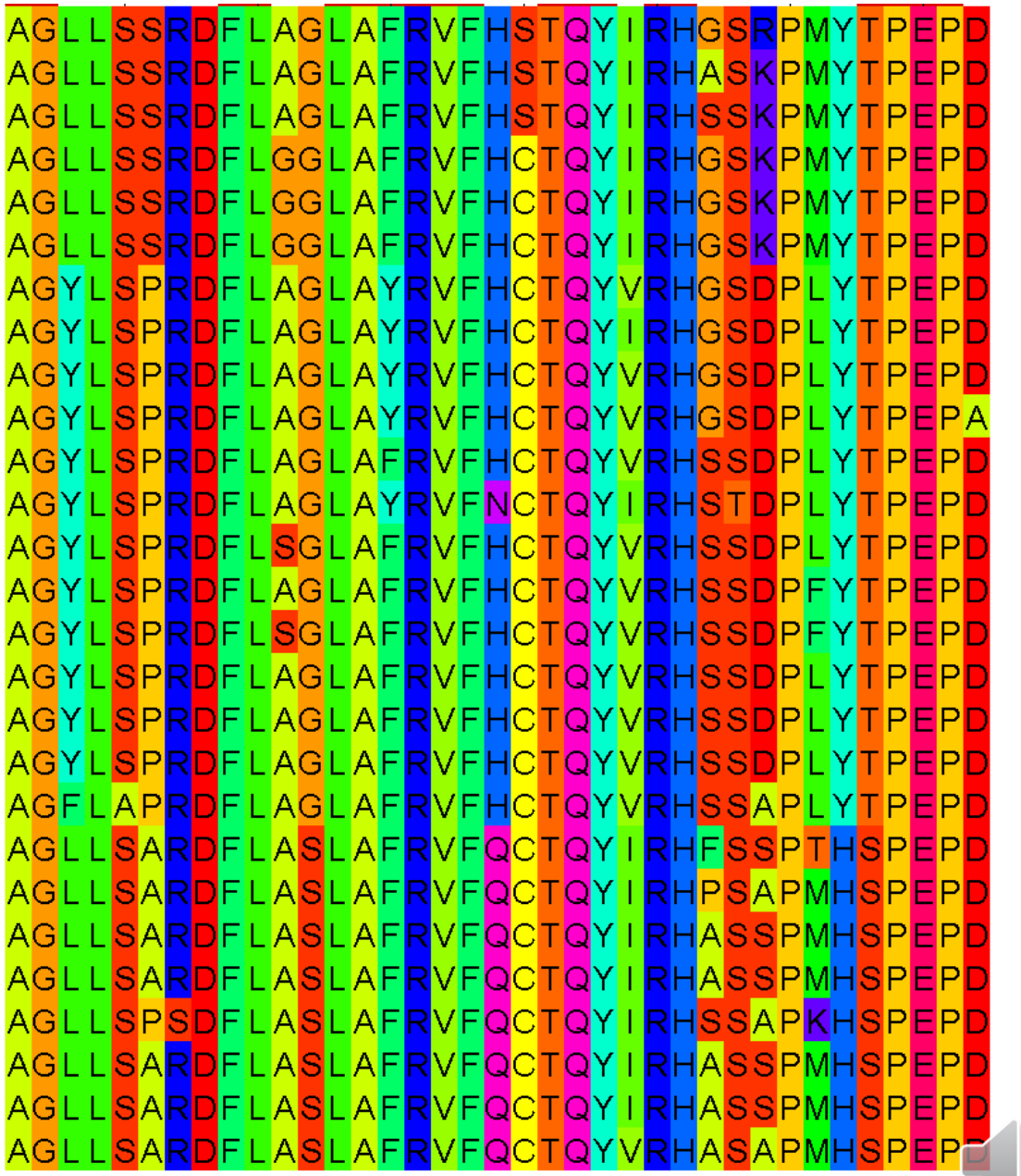
These methods are character based: The tree is built based on how the sequences in the MSA changing per site.

- Maximum likelihood: e.g. RAXML, PhyML
- Bayesian statistics: e.g. MrBayes

(protein or nucleotide sequences)



1. Do sequences in different alignments evolve at the same rate?



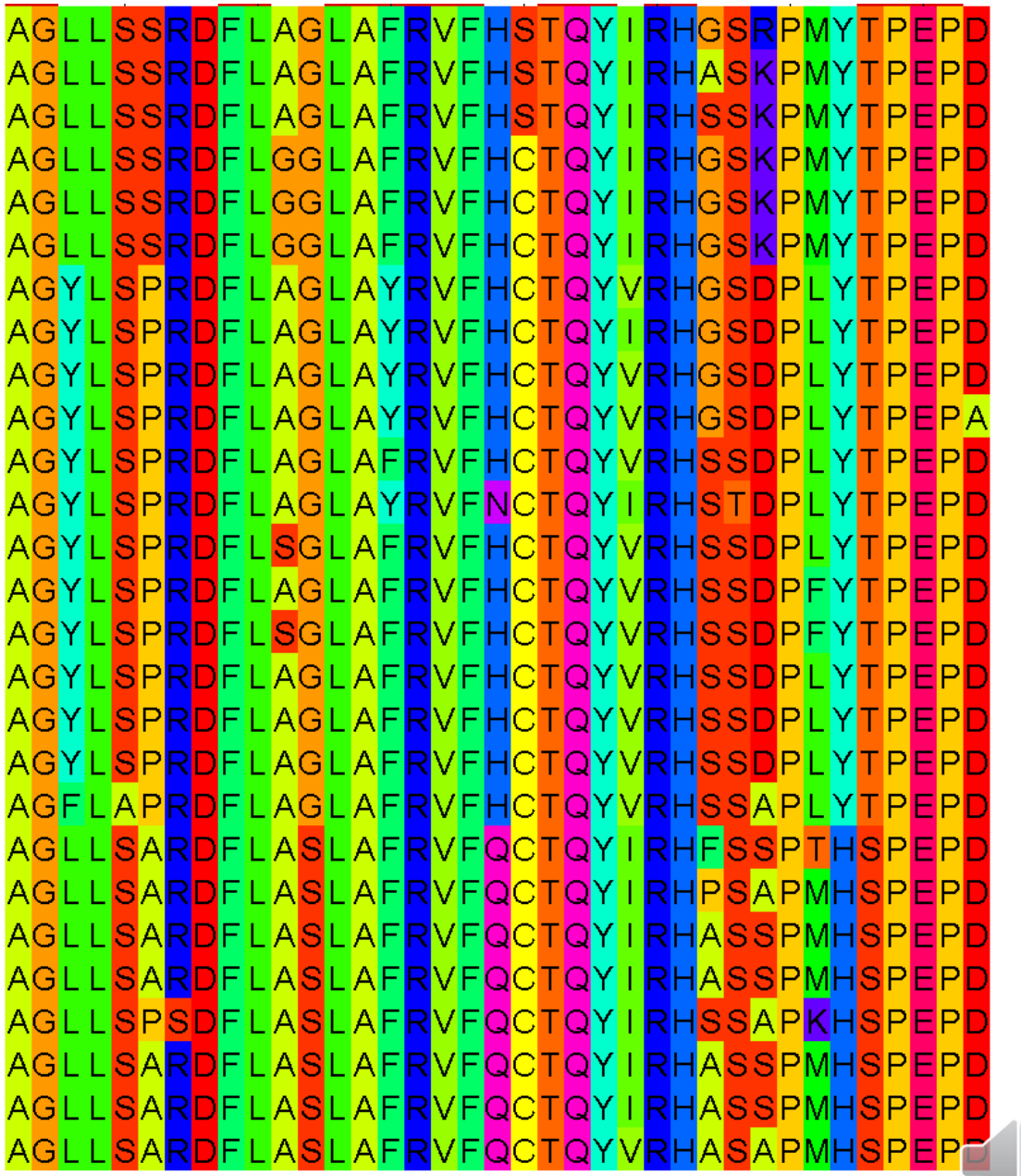
1. Do sequences in different alignments evolve at the same rate?

No, rates vary!
Different alignments (and different parts of alignments) tend to have different rate distributions.

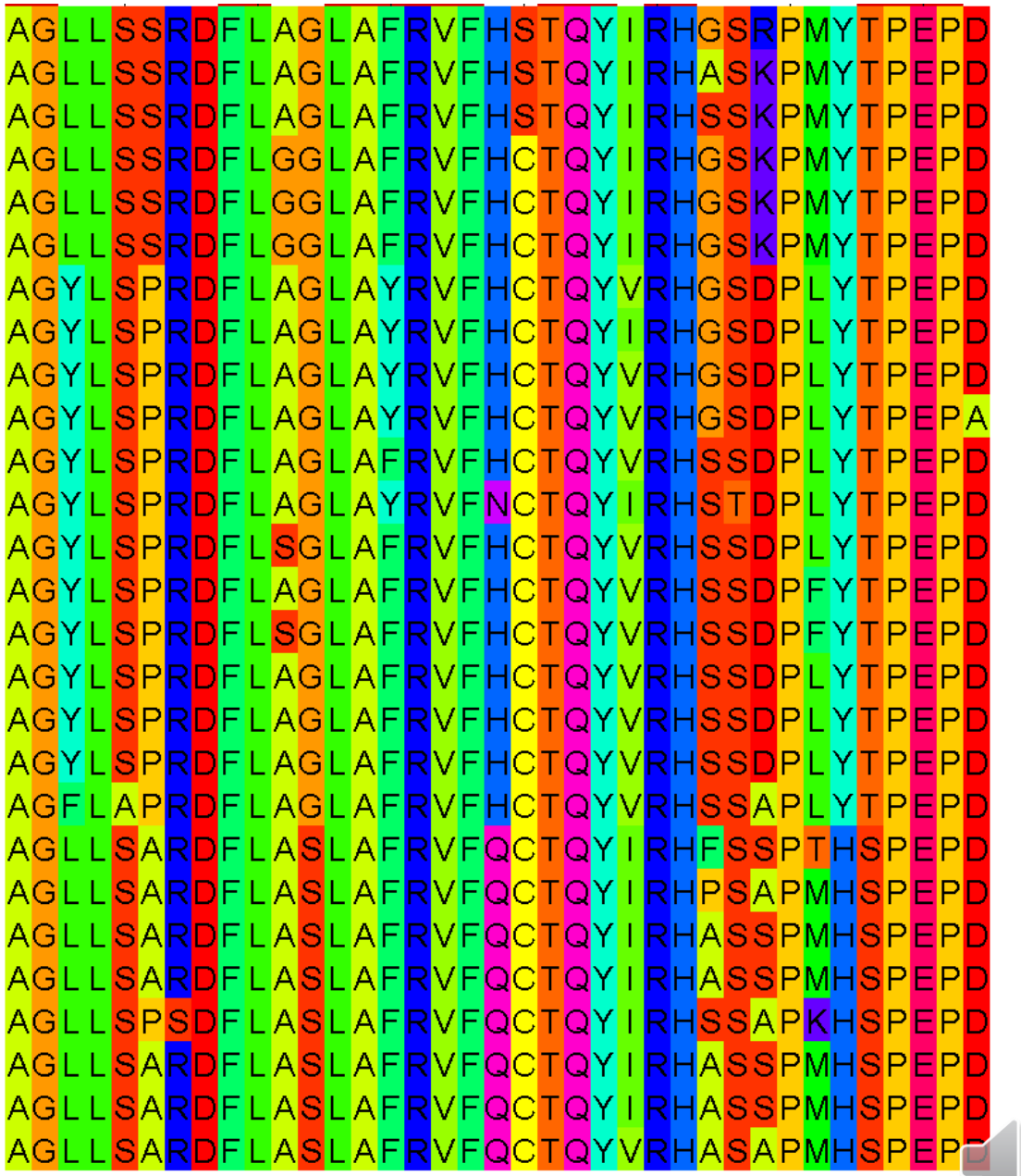
Remember different BLOSUM and different PAM matrices?

JTT, LG, HIVb, mtMam, etc…
Lots of different substitution matrices.

We need to know which substitution matrix is the better fit for our data!



2. Do all sites in a protein evolve at the same rate?

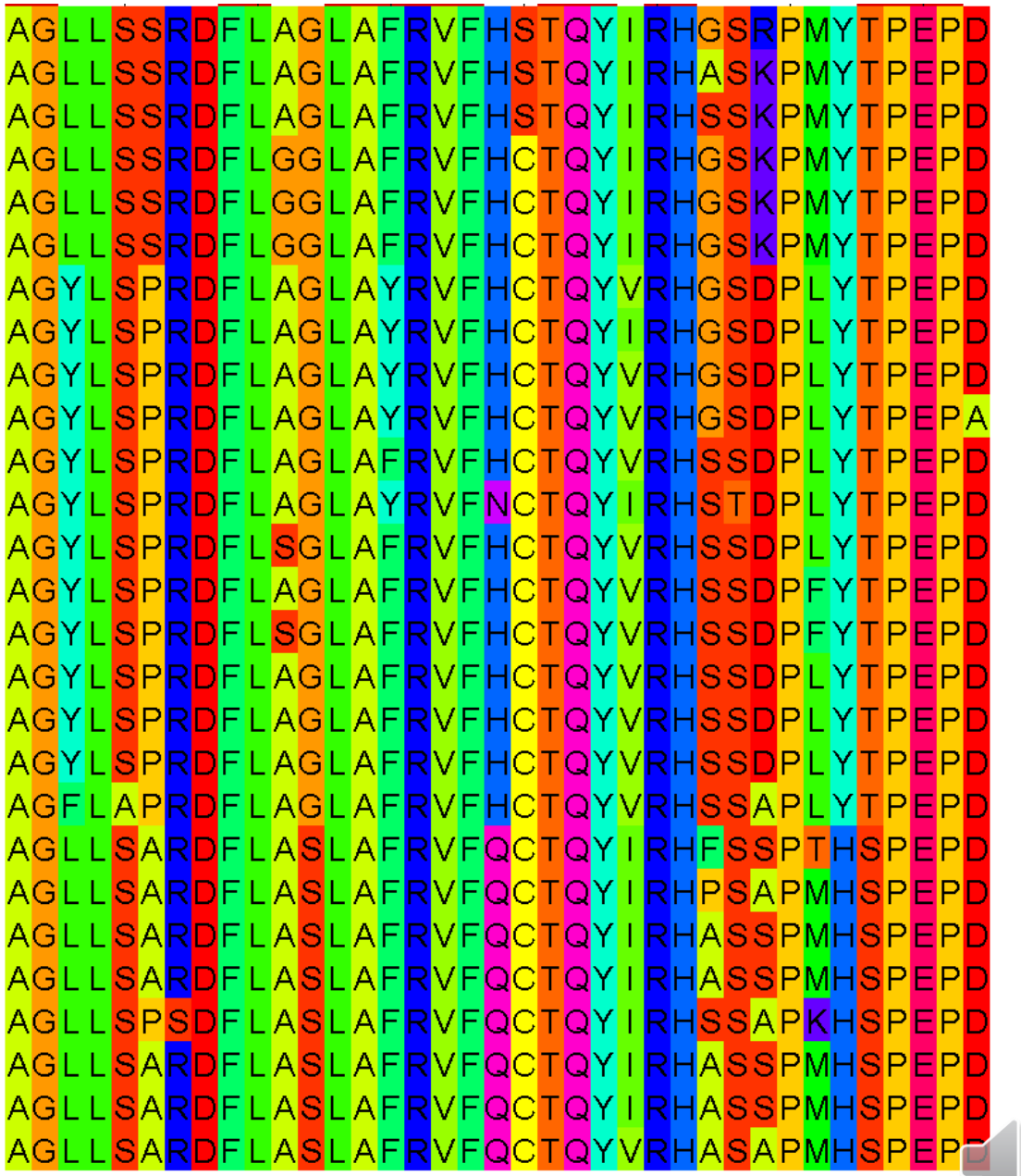


2. Do all sites in a protein evolve at the same rate?

No, usually not.

4-6 rate categories are commonly used to describe different rates at different sites.

The distribution of rates in an alignment is modelled by a GAMMA DISTRIBUTION.



Gamma distribution (G)

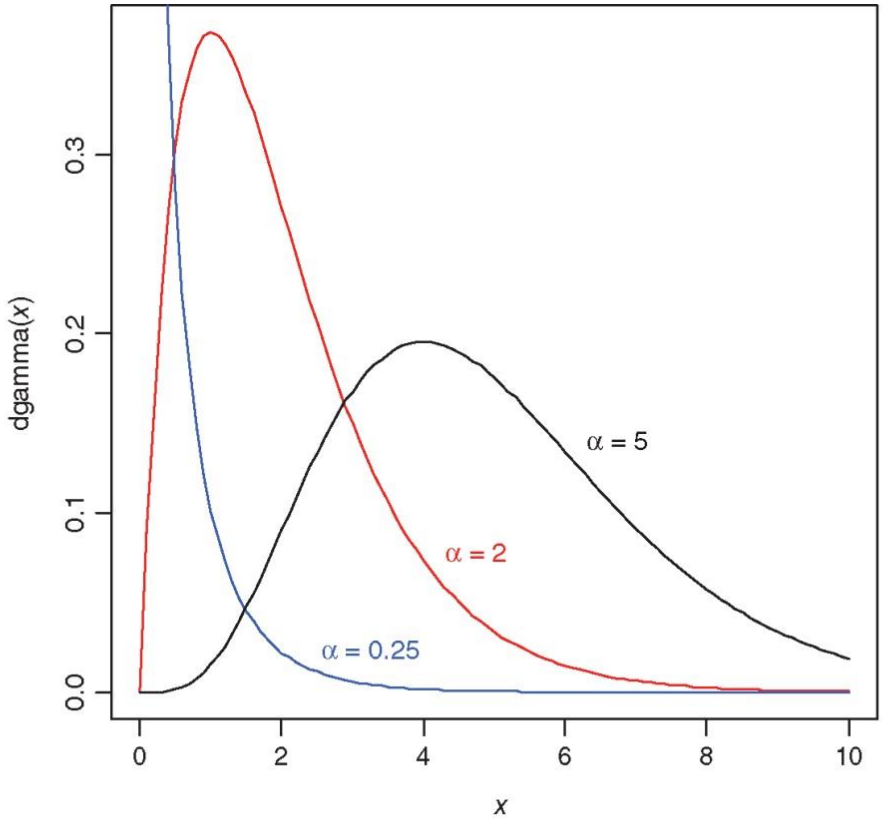


FIGURE 7.21 *Bioinformatics and Functional Genomics*, Third Edition, Jonathan Pevsner.

Companion Website: www.wiley.com/go/pevsnerbioinformatics

The gamma distribution shows the rate distribution for a multiple sequence alignment.

4-6 rate categories are commonly used.

The shape parameter for the gamma distribution is called α.

If α is small = large variation
Some sites are evolving very rapidly and some very slowly.

If α is large = little variation
Most sites evolve at the same rate.

We need to know if our data warrants Gamma in our model of evolution.

Amino acid frequencies (F)

Do the amino acid frequencies in our dataset deviate from those of the dataset used to build the substitution matrix?

****

Observed number of invariant sites: 54

Observed aminoacid frequencies:

A: 0.064 C: 0.022 D: 0.057 E: 0.085 F: 0.057

G: 0.050 H: 0.029 I: 0.044 K: 0.066 L: 0.100

M: 0.013 N: 0.032 P: 0.051 Q: 0.040 R: 0.059

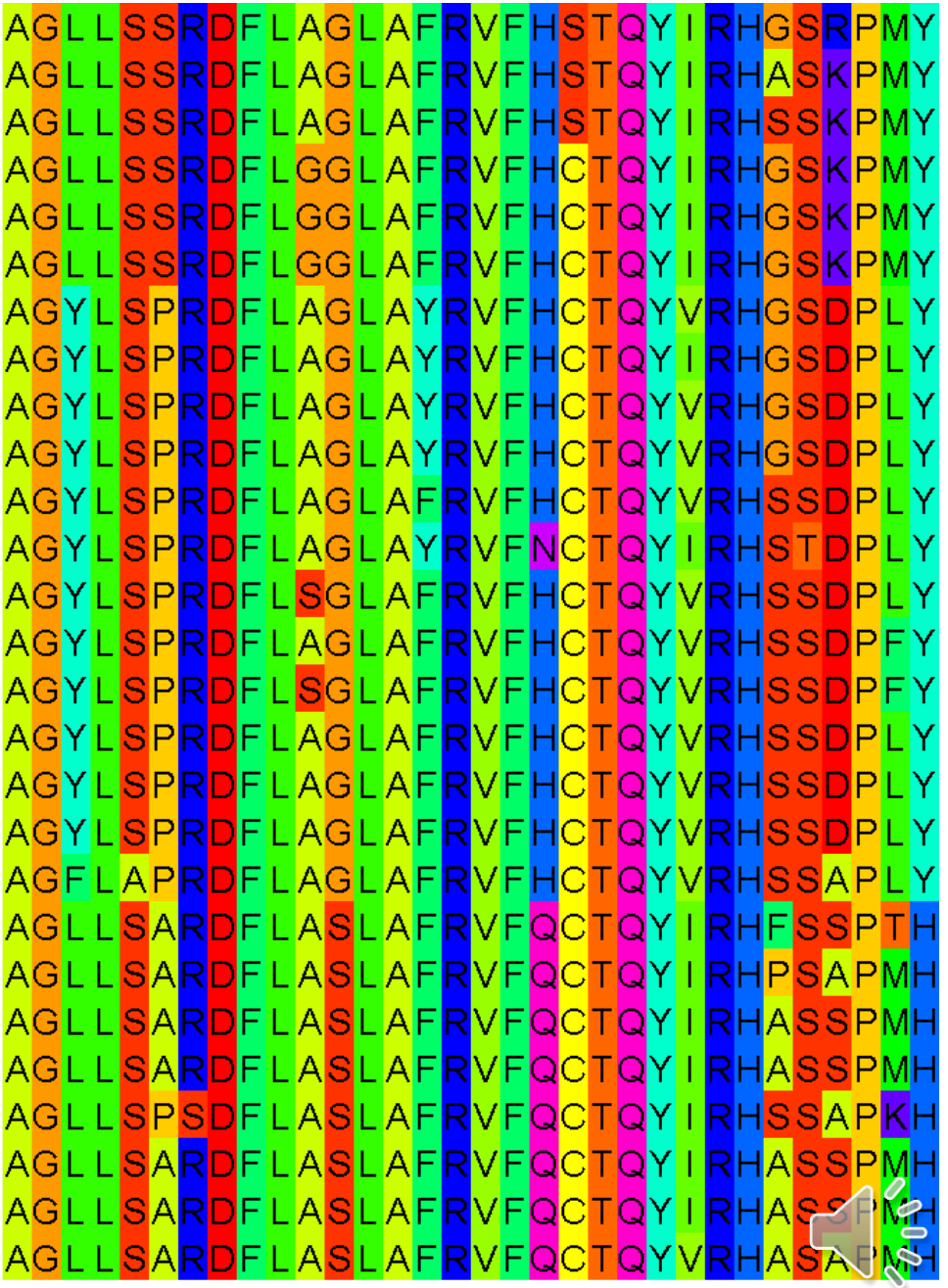
S: 0.074 T: 0.050 V: 0.058 W: 0.007 Y: 0.041

****

Invariant sites (I)

Do the invariant sites warrant them being considered as their own group?

An invariant site is conserved. That means that the amino acid is the same across all sequences.



Statistical methods such as Likelihood and Bayesian methods need a MODEL of EVOLUTION

The MODEL of EVOLUTION describes how the sequences are changing in the MSA.

The Model of evolution has 4 components:

1. Substitution matrix
2. Gamma distribution (G)
3. Amino acid frequencies (F)
4. Invariant sites (I)

This is the likelihood function

$$L_D = f(X|t, v, \theta)$$

L_D likelihood
X data (sequence alignment)
t tree
v branch lengths
θ model of evolution

Maximum Likelihood

The model of evolution includes:

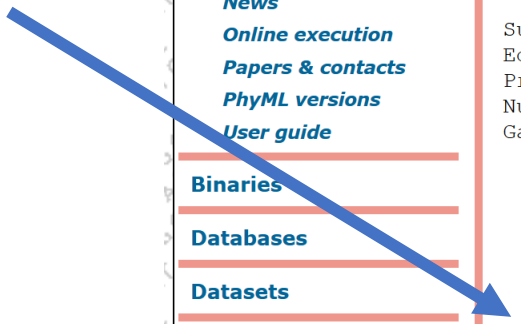
Substitution matrix + Amino acid frequencies (F) + Invariant sites (I) + Gamma (G)

Akaike Information Criterion AIC = $-2\ln L_D + 2K$ (K = number of parameters)

AIC does a trade-off between the goodness of fit and the number of parameters. Too many parameters can make the model slow.

Test for the best fitting model of evolution for each alignment before your build a tree.

Substitution matrix



Home
Organization
Citations & Statistics
Partners
Online programs
PhyML
Benchmarks
Datasets
Downloads
FAQ
News
Online execution
Papers & contacts
PhyML versions
User guide
Binaries
Databases
Datasets
NGS

PhyML-SMS:

Please cite:
"SMS: Smart Model Selection in PhyML."
Vincent Lefort, Jean-Emmanuel Longueville, Olivier Gascuel.
Molecular Biology and Evolution, msx149, 2017.

Analysis name : likelihood

PhyML results :

- **Download (zip format)**
- **Tree Visualisation**

Best model: JTT +G+I+F

```
Substitution model                          : JTT
Equilibrium frequencies                     : Empirical
Proportion of invariable sites              : estimated (0.276)
Number of substitution rate categories      : 4
Gamma shape parameter                       : estimated (2.810)
```

Model	Decoration	K	Llk	AIC	BIC
JTT	+G+I+F	28	-2837,58667	5731,17334	5842,93434
JTT	+G+F	27	-2839,67822	5733,35644	5841,12598
JTT	+G+I	9	-2866,95898	5751,91796	5787,84114
JTT	+G	8	-2869,30766	5754,61532	5786,54703
HIVb	+G+I+F	28	-2849,55424	5755,10848	5866,86948
Flu	+G+I+F	28	-2854,26236	5764,52472	5876,28572
WAG	+G+I+F	28	-2861,10674	5778,21348	5889,97448
MtREV	+G+I+F	28	-2862,96321	5781,92642	5893,68742
DCMut	+G+I+F	28	-2863,46768	5782,93536	5894,69636
Dayhoff	+G+I+F	28	-2863,50695	5783,01390	5894,77490
VT	+G+I+F	28	-2864,38341	5784,76682	5896,52782
CpREV	+G+I+F	28	-2868,63533	5793,27066	5905,03166

`Equilibrium frequencies : Empirical` means that F, the amino acid frequencies are estimated from the multiple sequence alignment used to build the tree. This is +F.

If it said
`Equilibrium frequencies : Model`
The amino acid frequencies from the dataset used to make the substitution matrix is used. No F.

Maximum Likelihood

$$L_D = f(X|t, v, \theta)$$

L_D likelihood
X data (sequence alignment)
t tree
v branch lengths
θ model of evolution

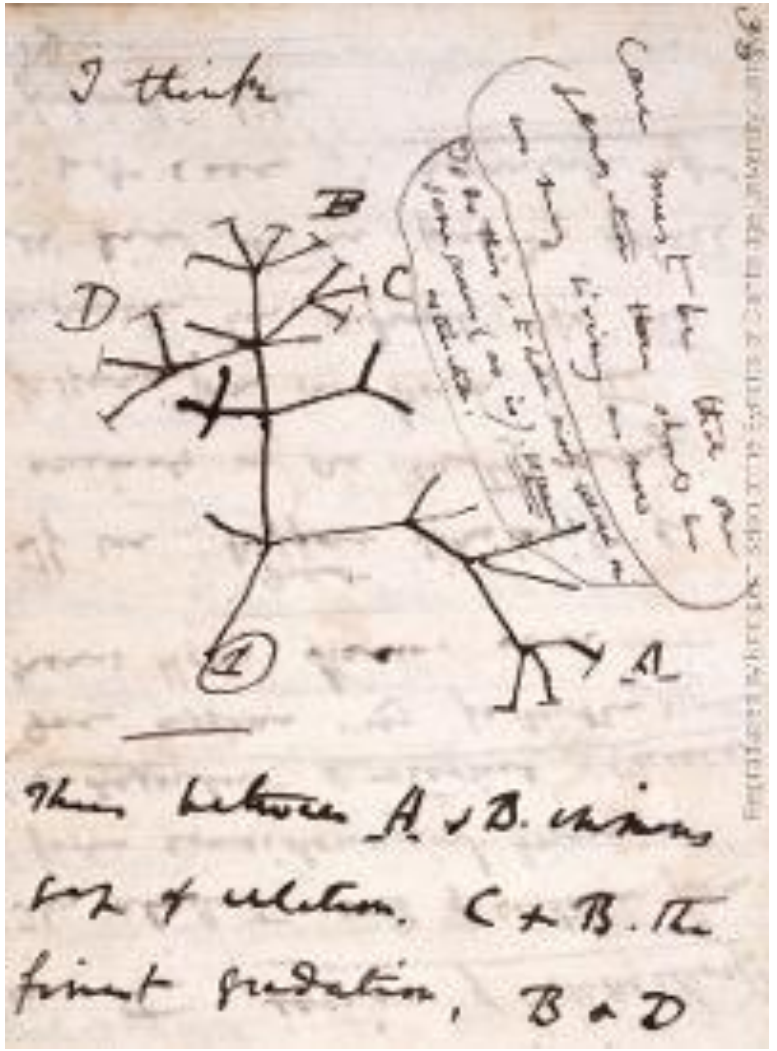
- Where do the trees & branch lengths come from?
  - Starting tree – a distance tree
  - Generate a set of similar trees
  - Calculate L_D
  - Continue with the tree that has the maximum L_D
  - Generate a set of similar trees and repeat until no better tree is found.

How confident can we be that the final tree represents the true history?

Bootstrapping

- Random sampling of the data => Many trees
- In this tree distribution
- How often is a node/branch reoccurring?

Posterior probability or likelihood supports at nodes



http://www.nhm.ac.uk/galleries/galleries-home/treasures/specimens/darwin-origins/

Guinea P

Mouse

Cat

Hamster

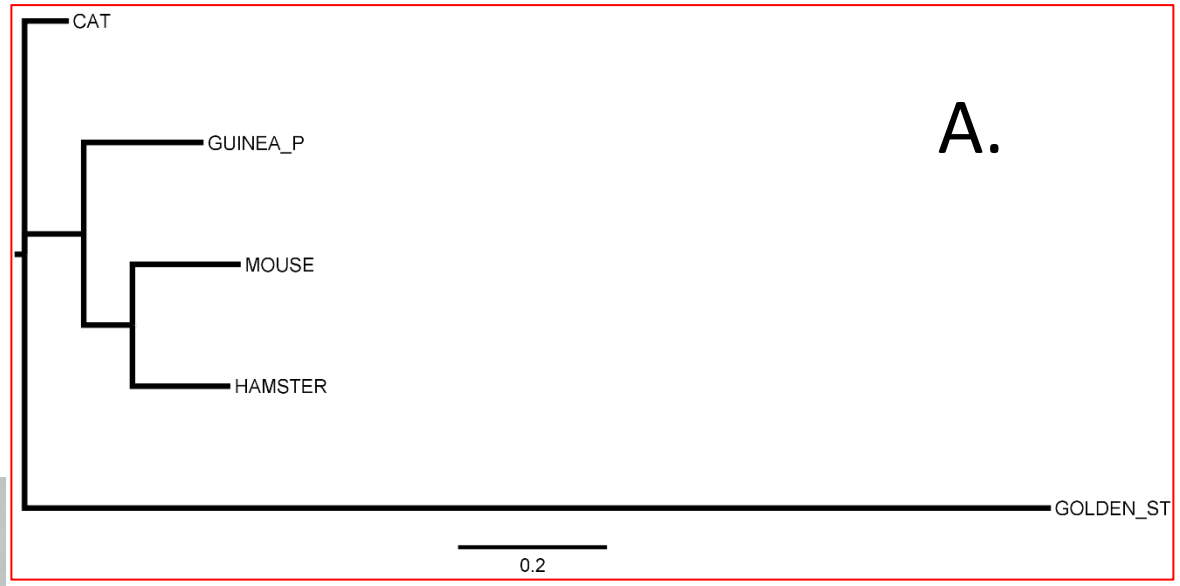
Golden St

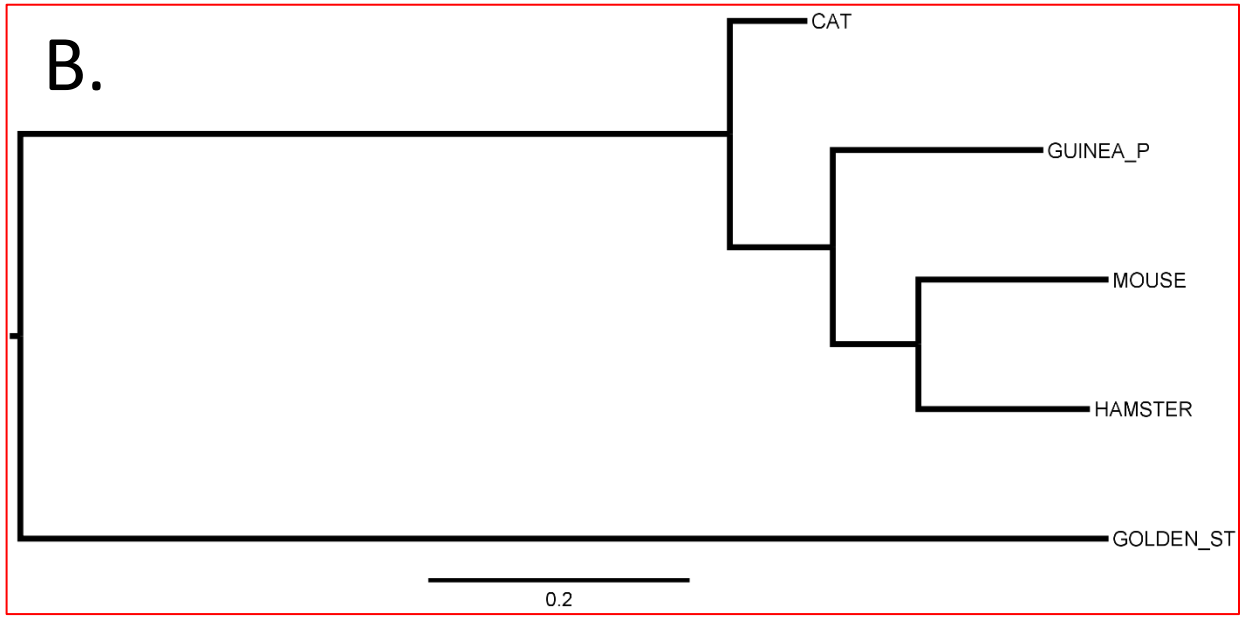
These panels show the same tree.

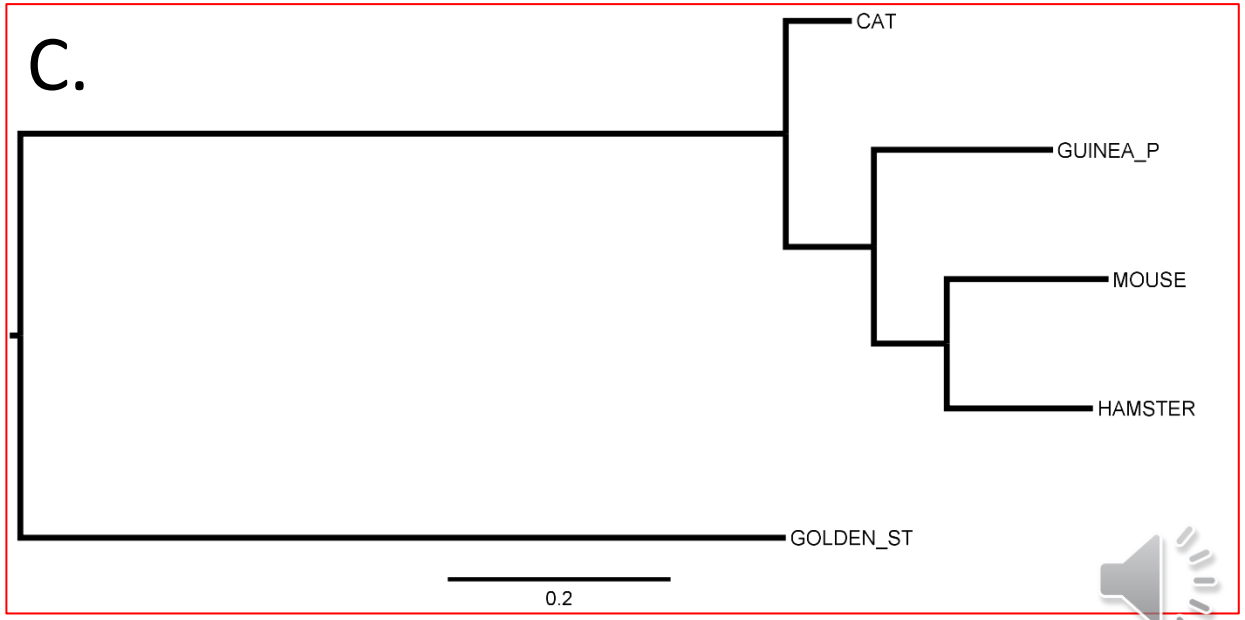
A. Unrooted (as is from the treebuilder)
B. Rooted at midpoint
C. Rooted with outgroup



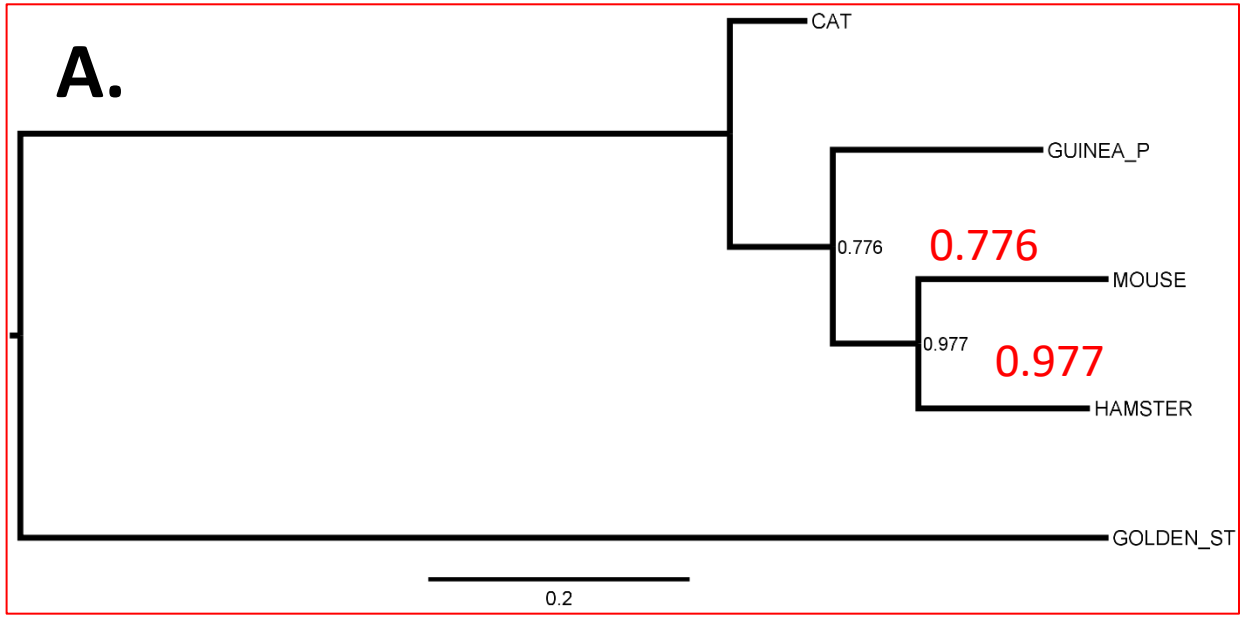


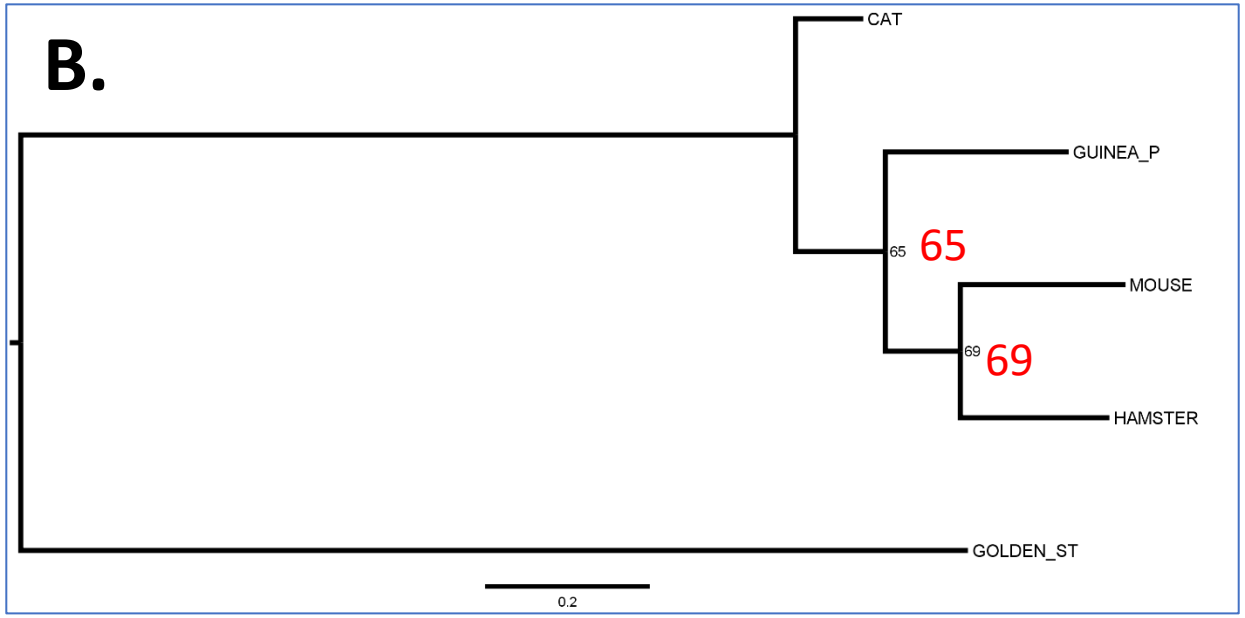






These panels show trees built with the same data (that is the same multiple sequence alignment) but with
(A) SH-like supports and
(B) 100 bootstraps.
Both trees are rooted at midpoint.





SH-like supports are fast to generate, but often less accurate than bootstrap values.
For SH-like supports the max value is 1.0 for 100 bootstraps, the max value is 100.
In the bootstrapped tree, 65 means that this branch is found in 65 out of 100 trees.