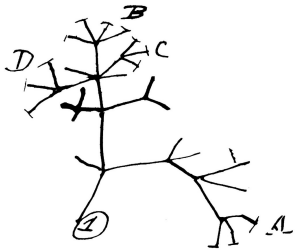


IPS-164 INTRODUCTION TO PHYLOGENETICS 2022

Lecture 9
Reconstructing phylogenetic trees. Part II

Sergei Tarasov

Beetle curator & Docent
Finnish Museum of Natural History, University of Helsinki

- @tarasov_sergio
- sergei.tarasov@helsinki.fi
- https://www.tarasovlab.com

PLAN OF THE TODAY'S LECTURE

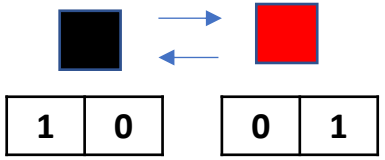
1. Learning the general workflow for the tree inference with molecular data

2. Intro to Bayesian Inference

Felsenstein's pruning algorithm

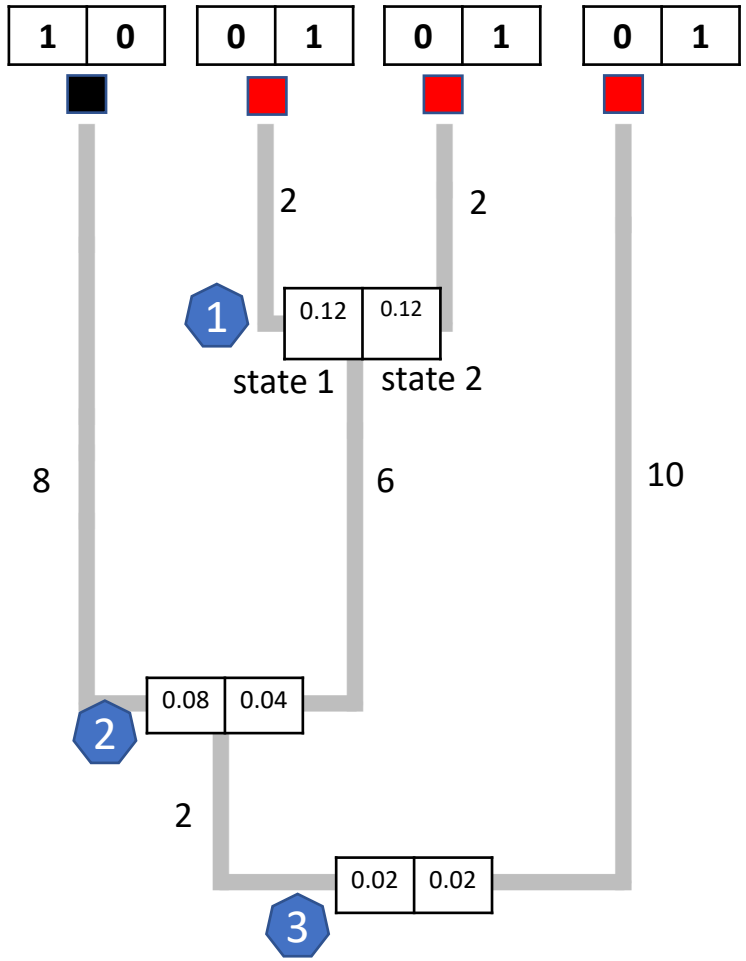
Given values:

state 1 state 2



$$Q=\begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}$$

$\pi = (1/2, 1/2)$



Likelihood (at the root):

L(tree) = Pr(**black**)* π_1+Pr(**red**)* π_2 =
0.02*1/2+ 0.02*1/2 = **0.02**

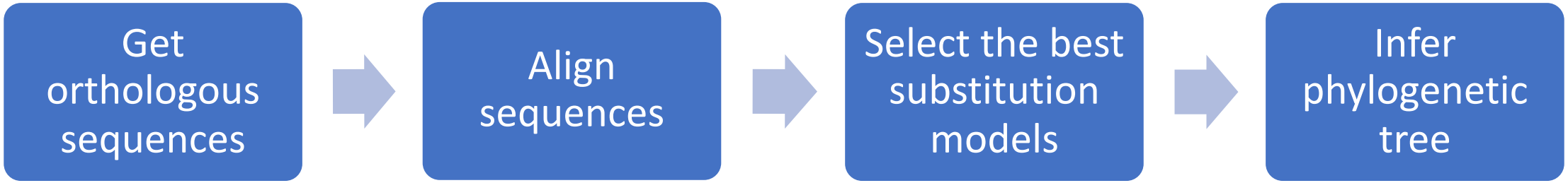
Log Likelihood:
Ln(0.02) = **-3.91**

Quick Demo

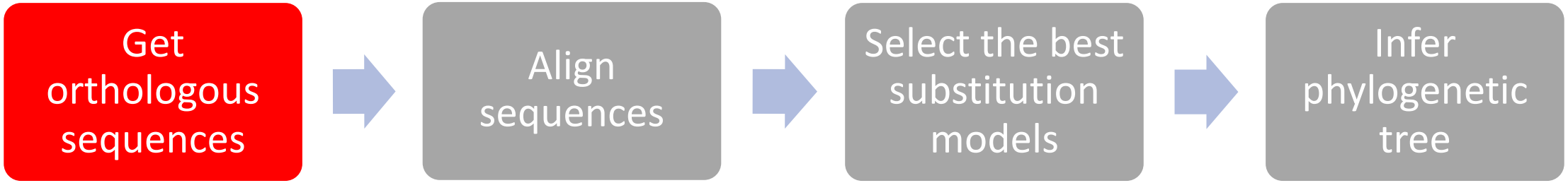
- Phylogeny of dung beetle genus *Helictopleurus* using COI

The general workflow for phylogenetic reconstruction

The workflow for tree reconstruction using molecules



The workflow for tree reconstruction using molecules



Gene Homology

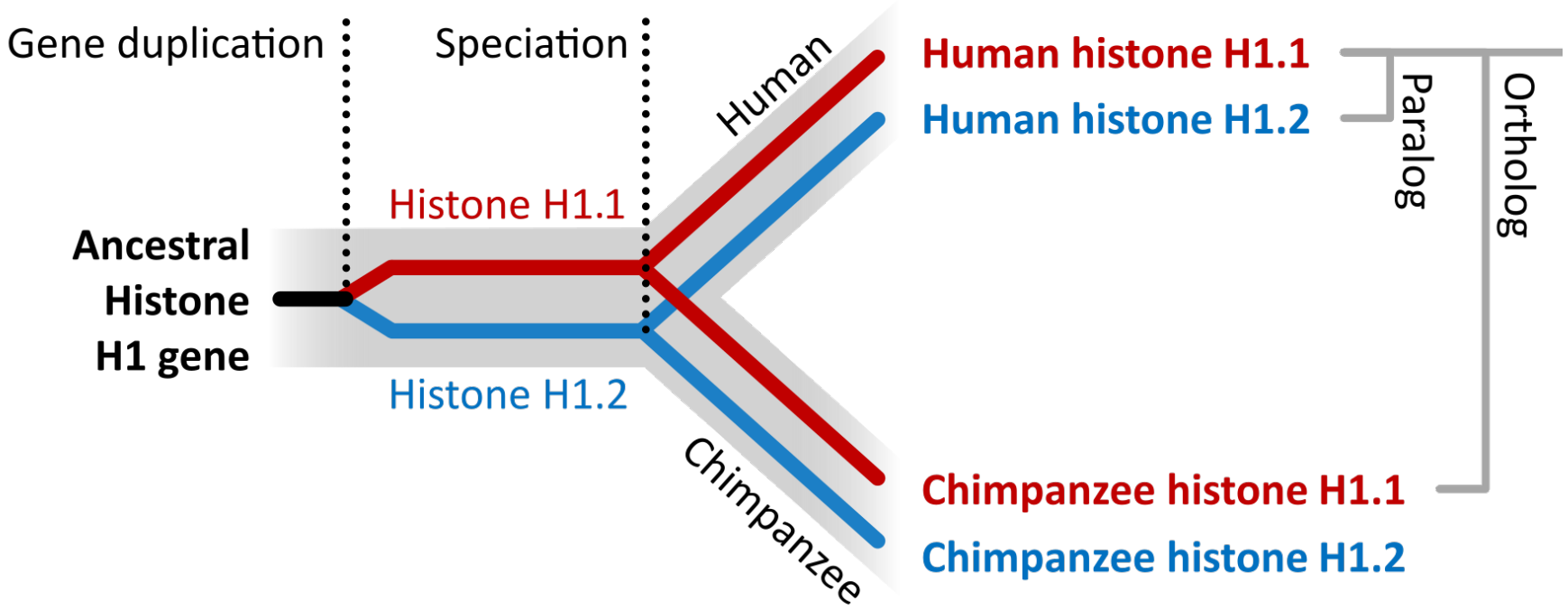
Homology: genes that derive from a common ancestor-gene are called ***homologs***

Orthologous genes are homologous genes in ***different*** organisms

Paralogous genes are homologous genes in *one* organism that derive from **gene duplication**

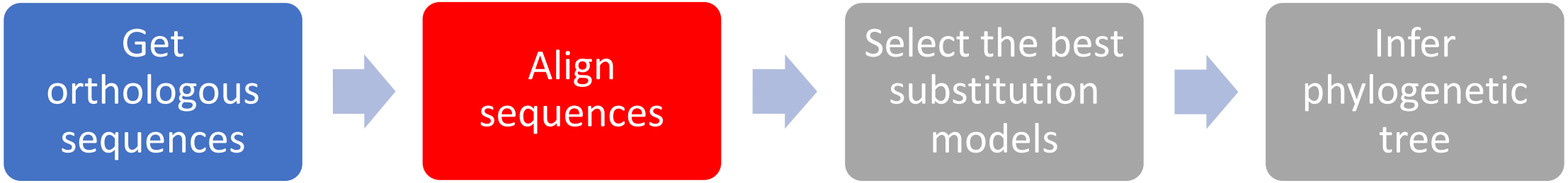
Gene duplication: one gene is duplicated in multiple copies that therefore free to evolve and assume new functions

Gene Homology

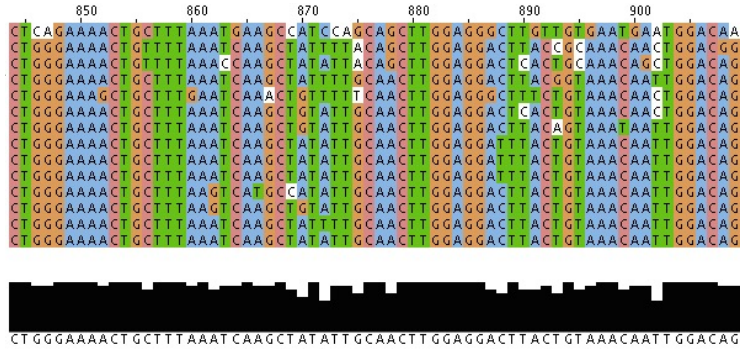


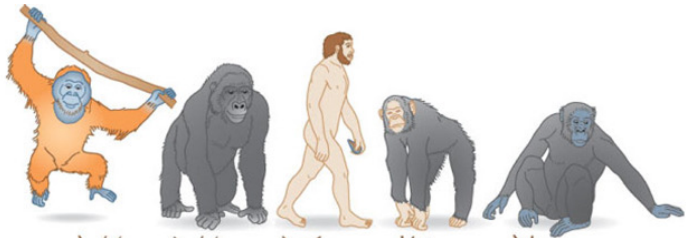
https://en.wikipedia.org/wiki/Sequence_homology

The workflow for tree reconstruction using molecules



Sequence orthology and alignment





Find the similarity between two (or more) DNA-sequences by finding a good **alignment** (match) between them.

Causes for sequence (dis)similarity

mutation: a nucleotide at a certain location is replaced by ***another*** nucleotide (e.g.: A**T**A → A**G**A)

insertion: at a certain location one new nucleotide is inserted inbetween two existing nucleotides (e.g.: AA → A**G**A)

deletion: at a certain location one existing nucleotide is deleted (e.g.: AC**T**G → AC**-**G)

indel: an **in**sertion or a **del**etion

```
tcctctgcctctgccatcat---caaccccaaagt
||||  |||  |||||  |||||     ||||||||||||
tcctgtgcatctgcaatcatgggcaaccccaaagt
```

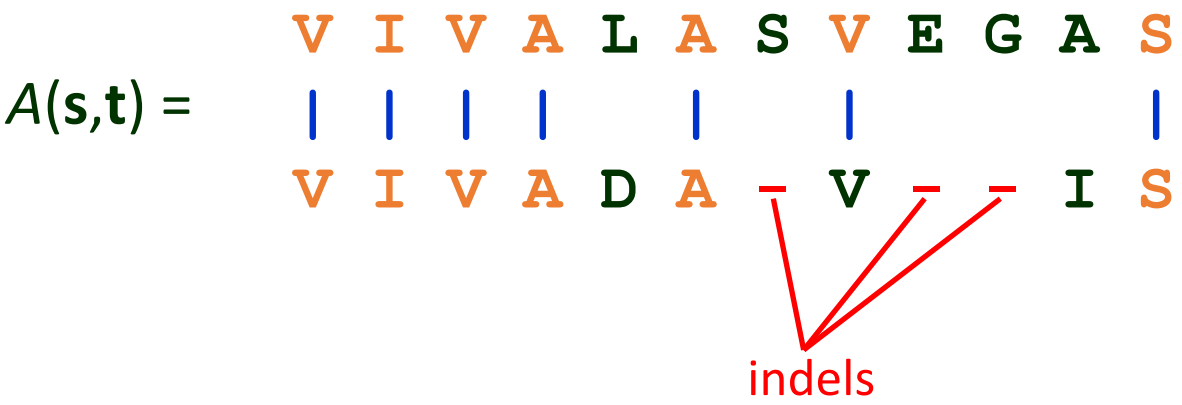
Sequence alignment

- Global alignment
- Local alignment
- Multiple alignment

Global Alignment

Find the **global** best fit between two sequences

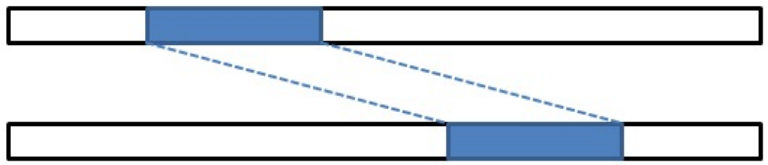
Example: the sequences **s** = VIVALASVEGAS and **t** = VIVADAVIS align like:



Local alignment

Local alignment methods find related regions ***within*** sequences - they can consist of a **subset** of the characters within each sequence.

For example, positions 20-40 of *sequence A* might be aligned with positions 50-70 of *sequence B*.

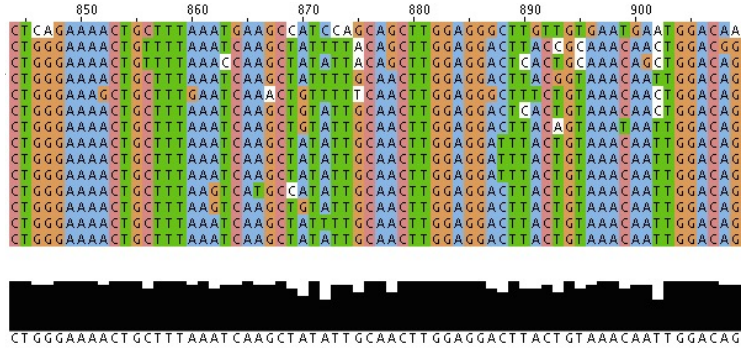


Local Alignment

Multiple alignment

Multiple alignment is an extension of pairwise alignment to incorporate more than two sequences into an alignment.

Multiple alignment methods try to align all of the sequences in a specified set.



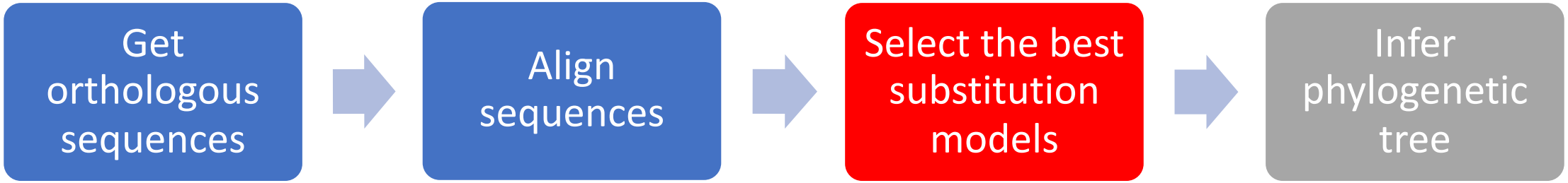
Sequence alignment software

- ***Fast:***
  - ***Blast https://blast.ncbi.nlm.nih.gov/Blast.cgi***
- ***Thorough multiple sequence alignments:***
  - ***Mafft: https://mafft.cbrc.jp/alignment/software/***
  - ***Clustal: https://www.ebi.ac.uk/Tools/msa/clustalo/***
  - ***Geneious: https://www.geneious.com***
- ***Free viewer:***
  - ***Jalview: https://www.jalview.org***

What sequence alignment is important for?

- Prediction of function
- Database searching
- Gene finding
- Sequence assembly

The workflow for tree reconstruction using molecules



DNA substitution models

Jukes & Cantor (1969)

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix} \end{array} \qquad \pi_A = \pi_C = \pi_G = \pi_T = 0.25$$

Kimura (1980)

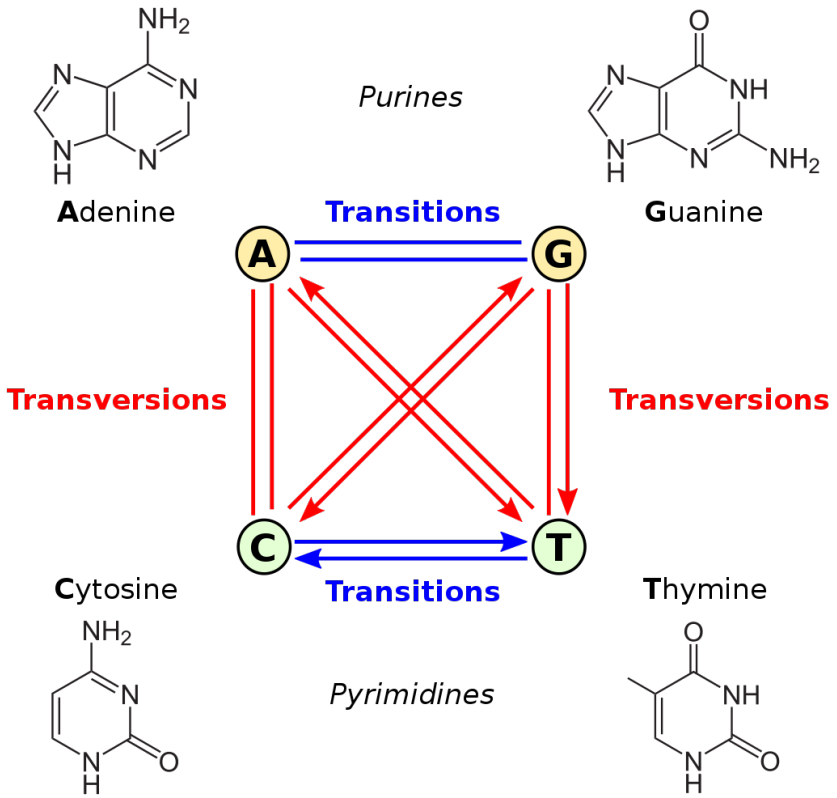
$$Q = \begin{pmatrix} -1 & 1/(\kappa+2) & \kappa/(\kappa+2) & 1/(\kappa+2) \\ 1/(\kappa+2) & -1 & 1/(\kappa+2) & \kappa/(\kappa+2) \\ \kappa/(\kappa+2) & 1/(\kappa+2) & -1 & 1/(\kappa+2) \\ 1/(\kappa+2) & \kappa/(\kappa+2) & 1/(\kappa+2) & -1 \end{pmatrix} \qquad \pi_A = \pi_C = \pi_G = \pi_T = 0.25$$

Hasegawa, Kishino, and Yano (1985)

$$Q = \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{pmatrix} \mu \qquad \text{all } \pi_i \text{ values free}$$

GTR (Tavare, 1986)

$$Q = \begin{pmatrix} - & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & - & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AG}\pi_A & r_{CG}\pi_C & - & \pi_T \\ r_{AT}\pi_A & r_{CT}\pi_C & \pi_G & - \end{pmatrix} \mu \qquad \text{all } \pi_i \text{ values free}$$



The *Q* matrix is normalized:

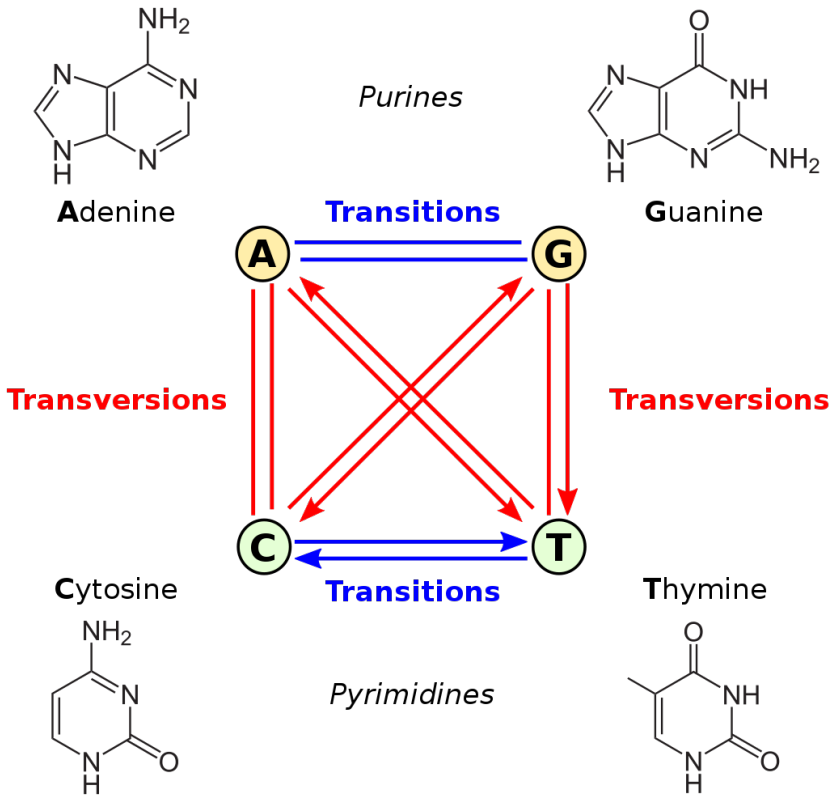
$$-\sum_{k=1}^{4} \pi_i Q_{ii} = 1$$

DNA substitution models

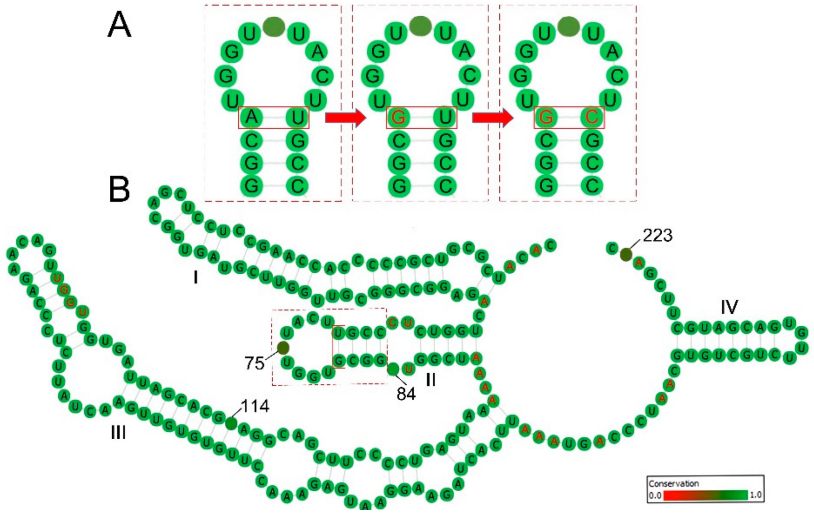
The most general nucleotide model possible is not necessarily time-reversible

$$Q = \begin{pmatrix} - & r_{AC} & r_{AG} & r_{AT} \\ r_{CA} & - & r_{CG} & r_{CT} \\ r_{GA} & r_{GC} & - & 1 \\ r_{TA} & r_{TC} & r_{TG} & - \end{pmatrix} \mu$$

and has 11 parameters



RNA substitution models



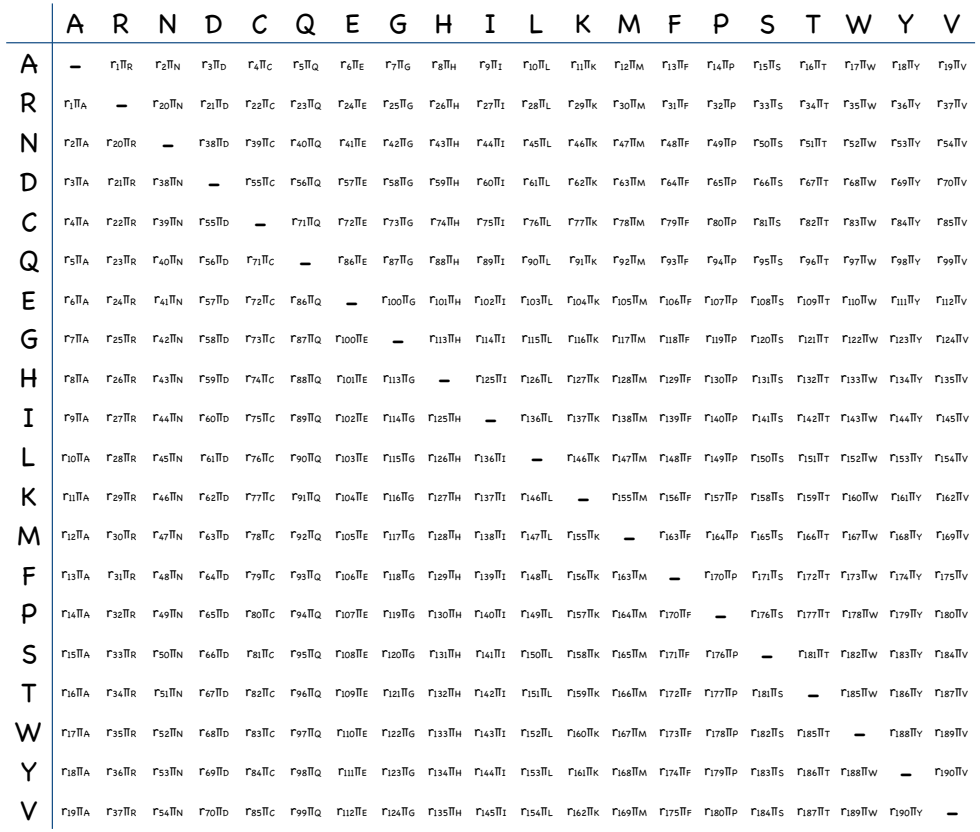
Doublet Model
(Schöniger and von Haeseler, 1994)

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	–	?	?	?	?	0	0	0	?	0	0	0	?	0	0	0
AC	?	–	?	?	0	?	0	0	0	?	0	0	0	?	0	0
AG	?	?	–	?	0	0	?	0	0	0	?	0	0	0	?	0
AU	?	?	?	–	0	0	0	?	0	0	0	?	0	0	0	?
CA	?	0	0	0	–	?	?	?	?	0	0	0	?	0	0	0
CC	0	?	0	0	?	–	?	?	0	?	0	0	0	?	0	0
CG	0	0	?	0	?	?	–	?	0	0	?	0	0	0	?	0
CU	0	0	0	?	?	?	?	–	0	0	0	?	0	0	0	?
GA	?	0	0	0	?	0	0	0	–	?	?	?	?	0	0	0
GC	0	?	0	0	0	?	0	0	?	–	?	?	0	?	0	0
GG	0	0	?	0	0	0	?	0	?	?	–	?	0	0	?	0
GU	0	0	0	?	0	0	0	?	?	?	?	–	0	0	0	?
UA	?	0	0	0	?	0	0	0	?	0	0	0	–	?	?	?
UC	0	?	0	0	0	?	0	0	0	?	0	0	?	–	?	?
UG	0	0	?	0	0	0	?	0	0	0	?	0	?	?	–	?
UU	0	0	0	?	0	0	0	?	0	0	0	?	?	?	?	–

$$q_{ij} = \begin{cases} \kappa\pi_j & : \text{transition} \\ \pi_j & : \text{transversion} \\ 0 & : i \text{ and } j \text{ differ at two positions} \end{cases}$$

(from John Huelsenbeck's presentation)

Amino acid substitution models

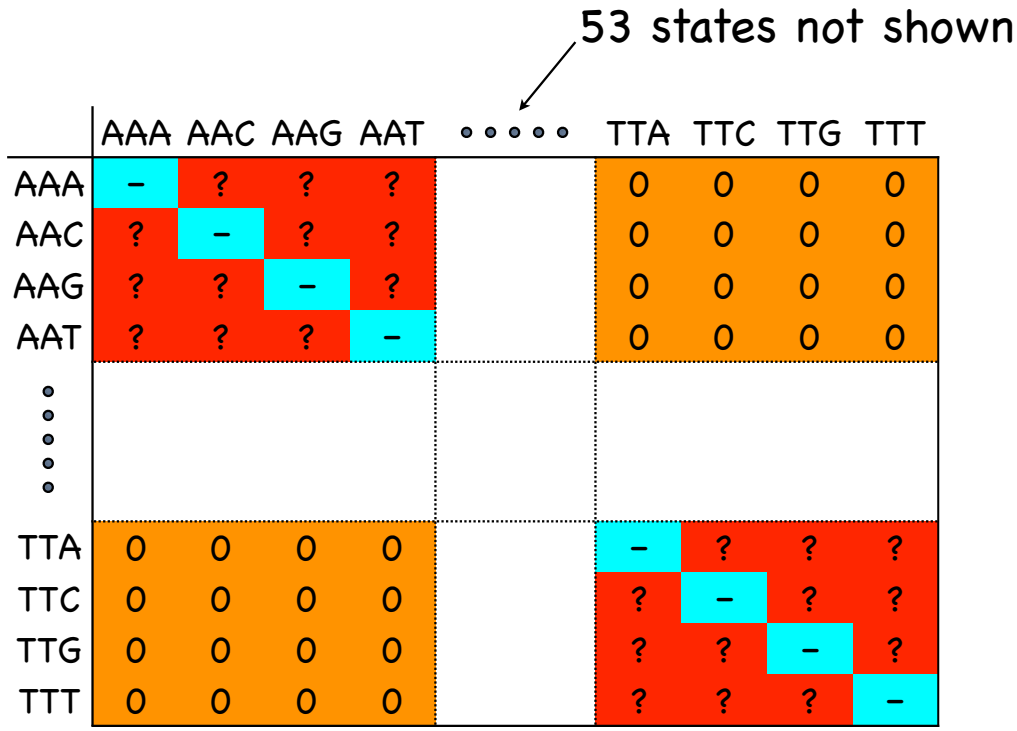


(from John Huelsenbeck's presentation)

Codon substitution models

Codon Model
(Goldman & Yang, 1994; Muse and Gaut, 1994;
Nielsen & Yang, 1998)

$$q_{ij} = \begin{cases} \omega\kappa\pi_j & : \text{nonsynonymous transition} \\ \omega\pi_j & : \text{nonsynonymous transversion} \\ \kappa\pi_j & : \text{synonymous transition} \\ \pi_j & : \text{synonymous transversion} \\ 0 & : i \text{ and } j \text{ differ at 2 or 3 positions} \end{cases}$$



(from John Huelsenbeck's presentation)

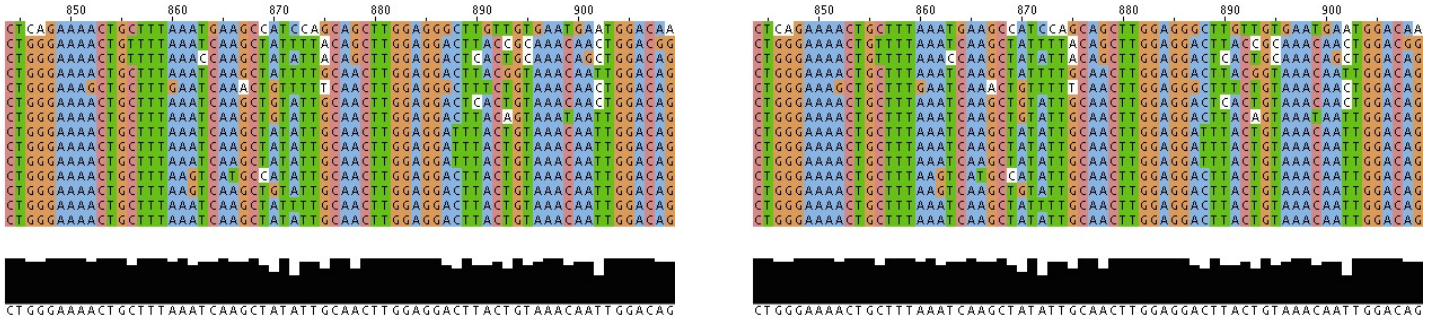
Models for morphological characters

- Mk model (the same as Jukes-Can tor model for DNA)

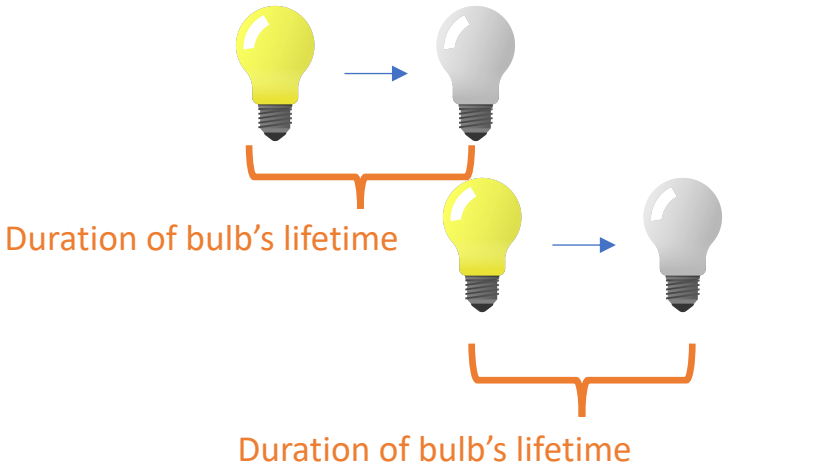
$$\mathbf{Q} = \begin{bmatrix} 1-k & 1 & \dots & 1 \\ 1 & 1-k & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

Rate heterogeneity

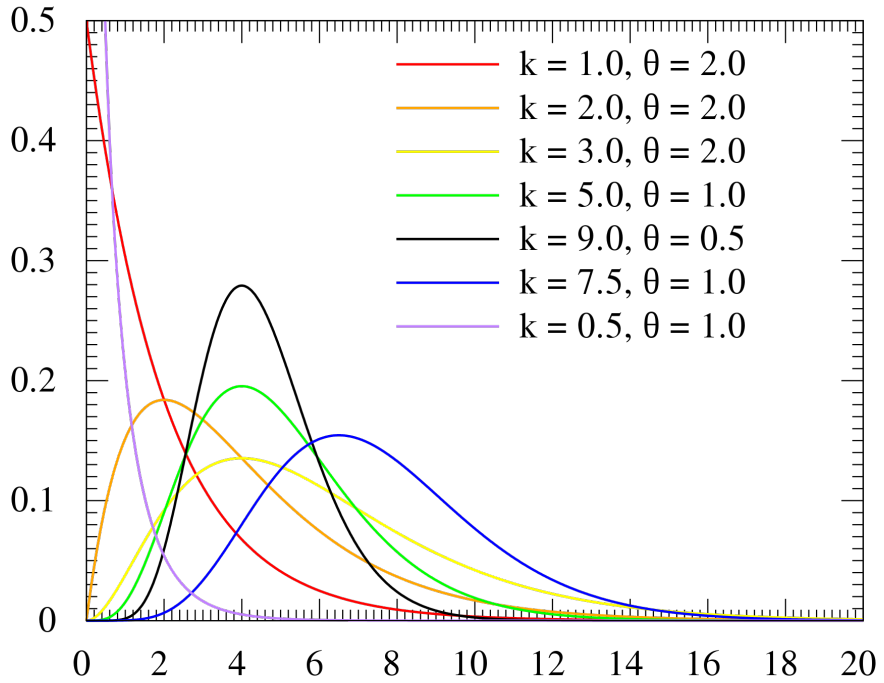
- Sites within gene may evolve at different rates
- By codon position (e.g., 1st and 2nd position vs. 3rd)
- By gene or gene region



Gammadistribution



.....

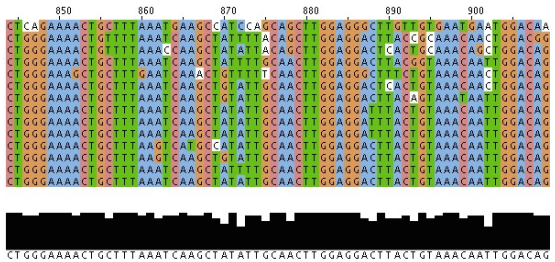


Gamma distribution shows the probability of the waiting-time for the *n*-th event (death) to occur. It is characterized by two parameters.

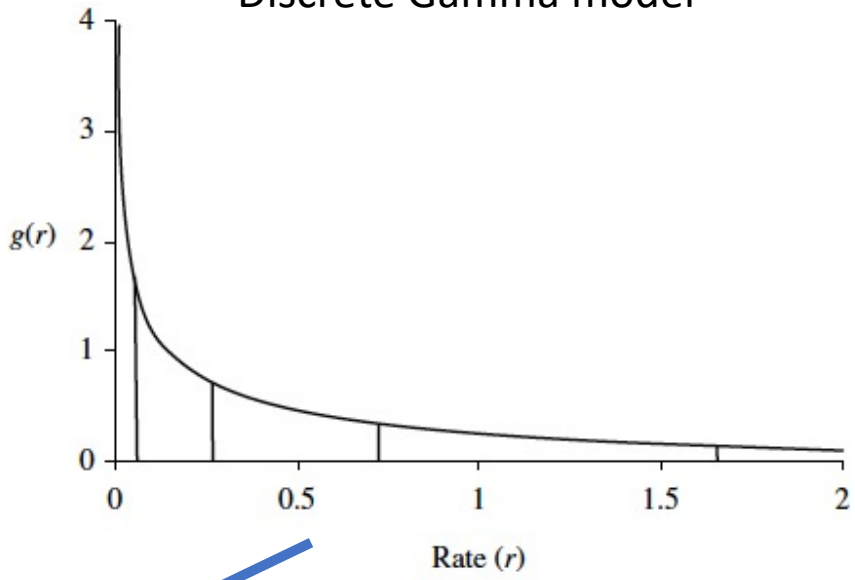
$$f(x) = \frac{1}{\Gamma(k)}\gamma\left(k, \frac{x}{\theta}\right)$$

Using discrete Gamma distribution to model across site rate heterogeneity (e.g., GTR + G)

$$\boldsymbol{Q}=\begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}$$

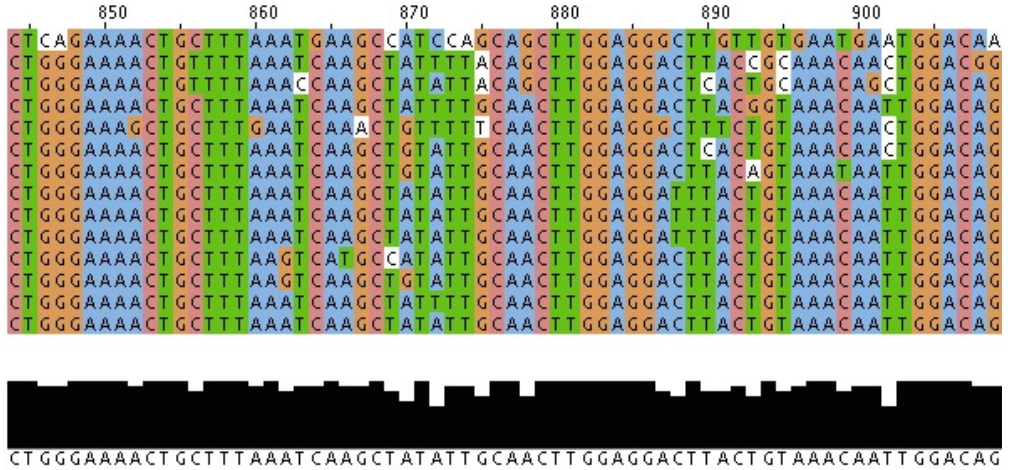


Discrete Gamma model



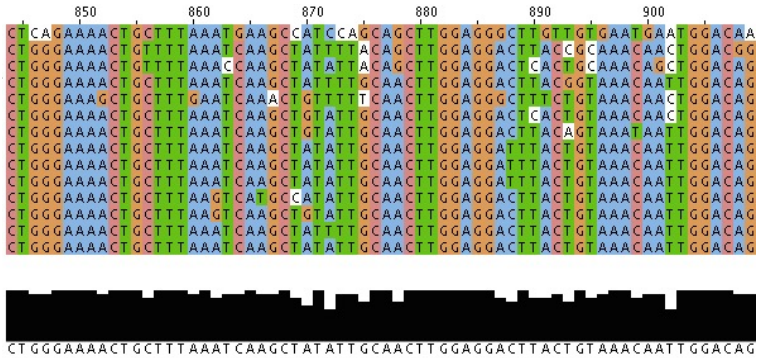
Invariable sites (+I)

- Two categories of sites: variable and invariable
- Invariable do not evolve
- Be careful of the model G+I !

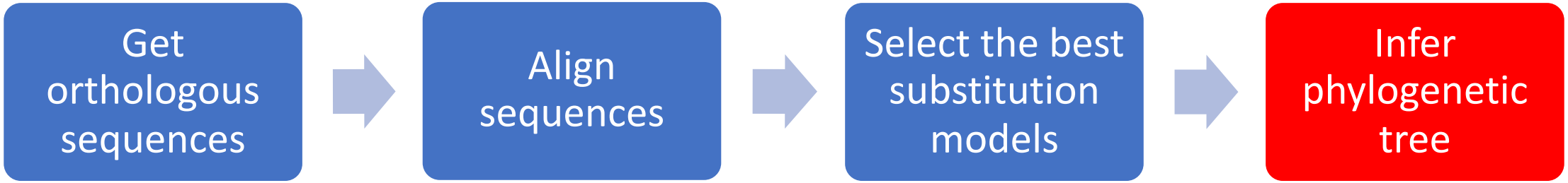


Model Selection

- There are many substitution models
- Create a priori partitions of your dataset by loci and/or codon position for downstream test
- Test possible combinations of partitions against all possible models (IQ tree does it automatically)
- Run tree inference with the best model and partitioning scheme

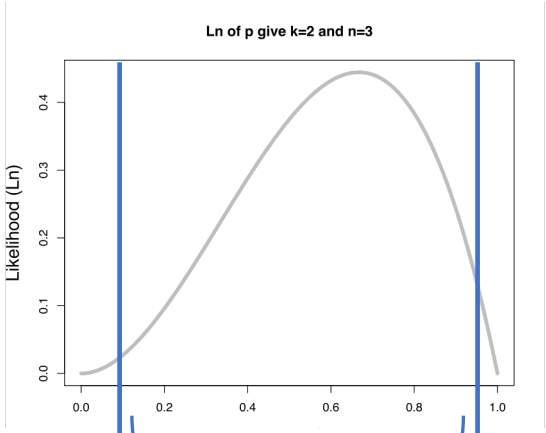


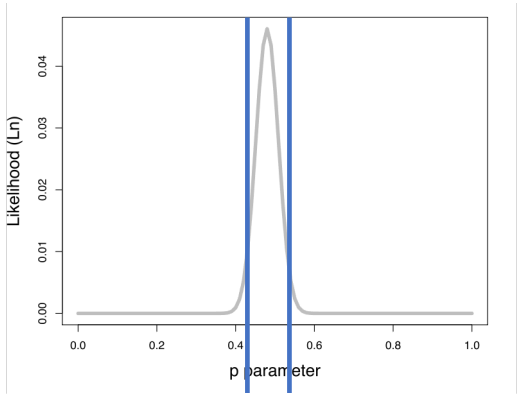
The workflow for tree reconstruction using molecules

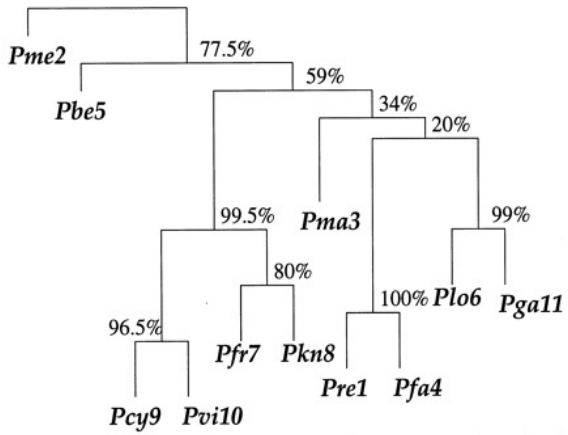


Bootstrap: assessing confidence of a topology

- The confidence interval (CI) interpretation. If we repeat the caculation of CI on numerous samples then CI shows the fraction (e.g., 95%) that includes the true population parameter.
- Bootstrap is a method for estimating confidence intervals of parameters (e.g. tree topology)
- The bootstrap values indicate how many times (say, out of 1000) the same branch was observed when repeating the phylogenetic reconstruction on a re-sampled data.
- Confidence interval is estimated for an associated confidence level
- Confidence level quantifies the level of confidence that the parameter lies in the interval







Felsenstein J(1985) Evolution **39**:783–791.

Infering trees using maximum likelihood

- IQ Tree http://www.iqtree.org
- RAxML https://cme.h-its.org/exelixis/web/software/raxml/index.html

Tree viewer:

- FigTree https://github.com/rambaut/figtree/releases

Summary

- The phylogenetic tree inference consist of four main steps:
  - 1. Getting orthologous sequences
  - 2. Alignment
  - 3. Model selection
  - 4. Tree inference
- There are multiple ways of how each step can be performed that offers flexibility for designing your own analytical pipeline

Bayesian Inference