Phylogenetics

094011
Introductory Course for Ph.D. students

23rd September 2022

Rui Borges
ruiborges23@gmail.com

Phylogenetics

Phylogenetics is the study of evolutionary relationships among biological entities – often species, individuals or genes (which may be referred to as taxa).

- What are the evolutionary relationships or histories among my species/individuals/genes of interest?
- How do sequences evolve (gene, genomes, …)?

Phylogenetics

Phylogenetic pipeline:

1. Observe sequences.
2. Reconstruct evolutionary histories.
3. Learn more about the evolutionary processes.
4. Develop better evolutionary models.

Includes both empirical and theoretical tasks.

Phylogenetics

Phylogenetics enriches our understanding of how genes, genomes, species (and molecular sequences more generally) evolve.

- How do sequences came to be the way they are today.
- Establish general principles that enable us to predict how they will change in the future.

Phylogenetics: applications

- Classification
- Identifying pathogens
- Forensics
- Bioinformatics
- Conservation

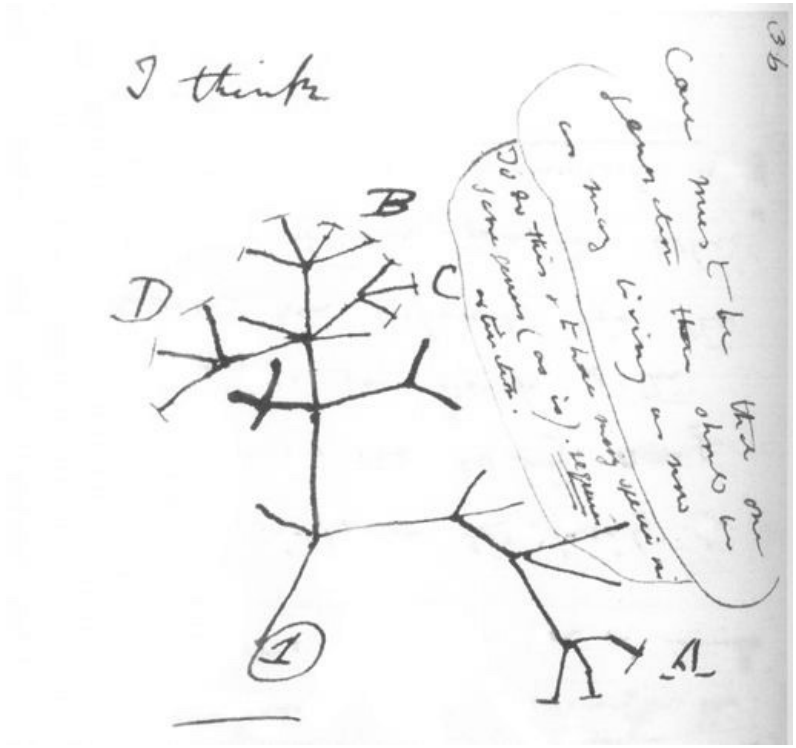
Contents

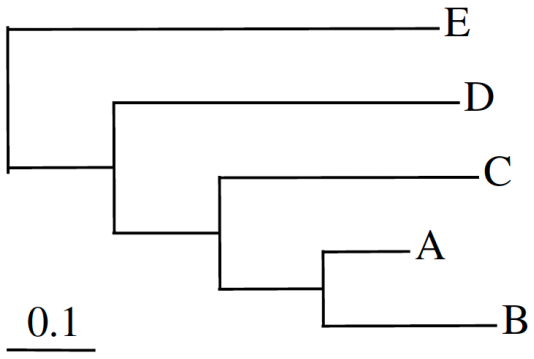
1

Phylogenetic trees: terminology and representation

Phylogenetic trees

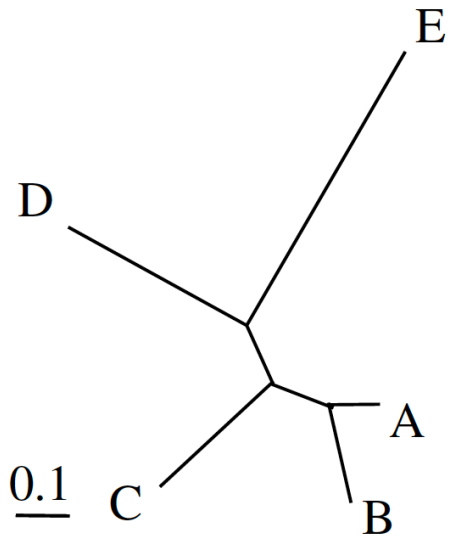
Phylogeny: a representation of the genealogical relationships among species, genes, populations, or even individuals.





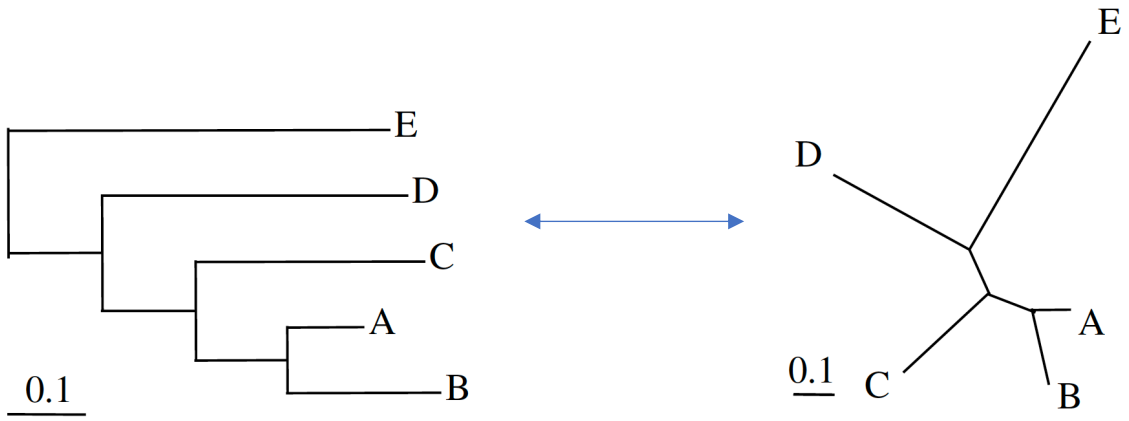
Exercise. Identify the following entities on the phylogenetic tree.

- Node
- Branch
- Branch length
- Distance scale
- Vertex
- Edge
- Root



Rooted and unrooted trees

Unrooted tree: a tree in which the root is unknown or unspecified.



Rooted and unrooted trees

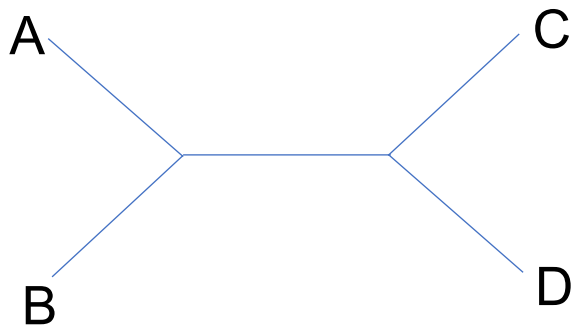
Molecular-clock rooting: if the evolutionary rate is constant over time (i.e., **molecular clock**), we can identify the root and produce rooted trees.

- The clock assumption is most often violated, except for closely related species.

Outgroup rooting: a commonly used strategy is to include distantly related species (i.e., **outgroups**) in tree reconstruction.

- In the reconstructed unrooted tree for all species the root is placed on the branch leading to the outgroups so that the subtree for the ingroups is rooted.

Exercise. Draw all the possible rooted trees for the following unrooted tree:



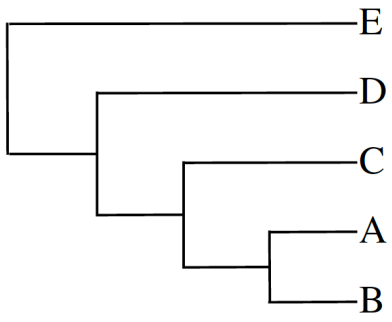
Cladograms and phylograms

The branching pattern of a tree.

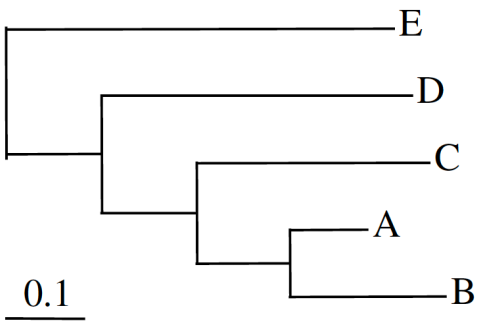
Phylogeny = Topology + Branch lengths

May represent the amount of sequence divergence or the time period covered by the branch.

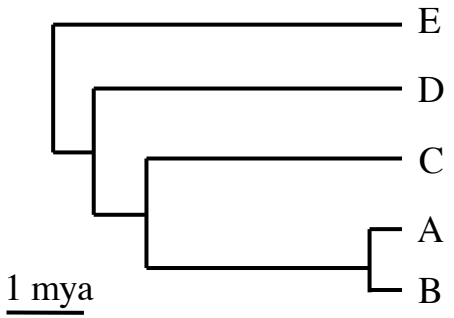
Cladograms and phylograms



Cladogram: tree topology without information about branch lengths



Phylogram: tree showing both the topology and branch lengths



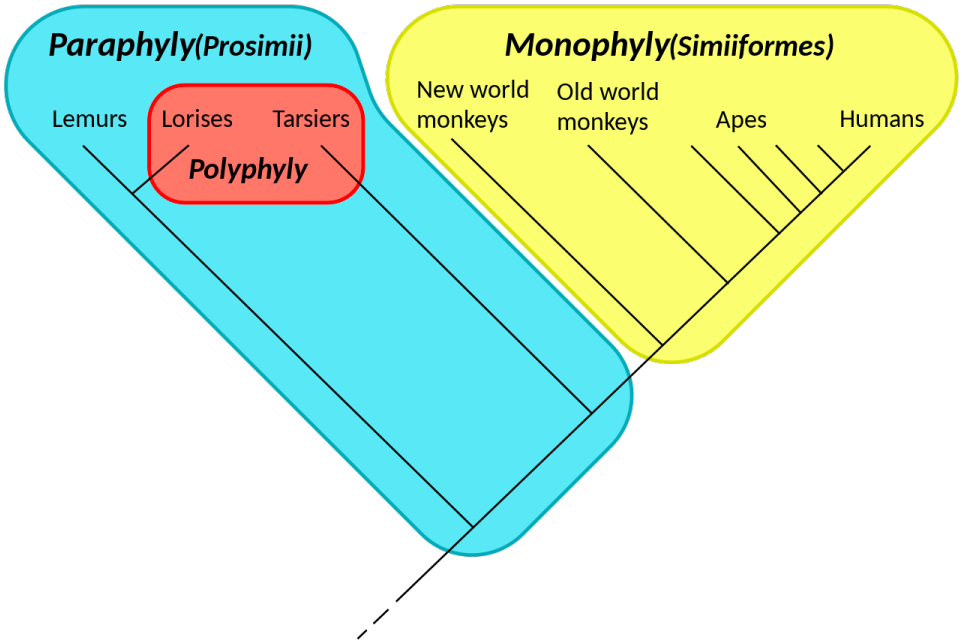
Chronogram (or ultrametric tree): a phylogram that explicitly shows evolutionary time through its branch lengths

Phylogenetic relationships

Monophyletic group: includes the most recent common ancestor of the group and all of its descendants.

Paraphyletic group: includes the most recent common ancestor of the group but not all its descendants.

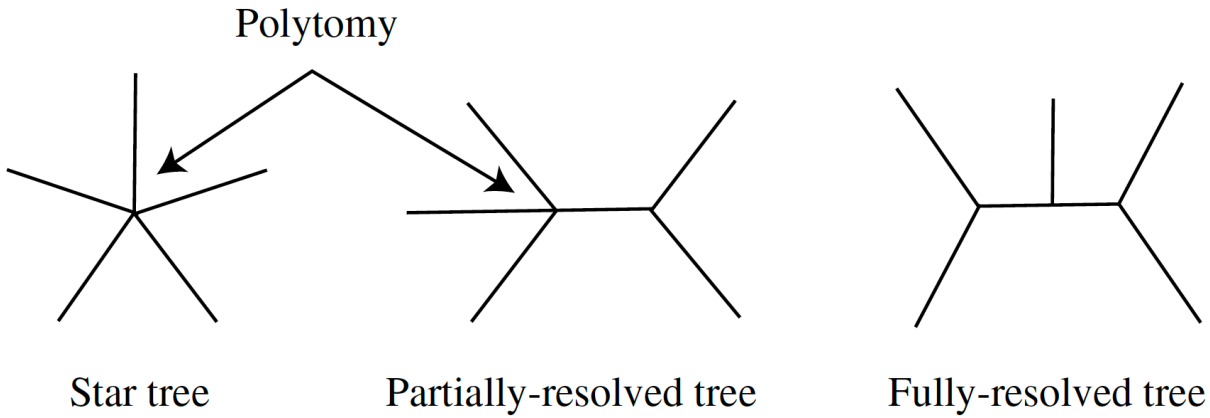
Polyphyletic group: does not include the most recent common ancestor of all members of the group.



Polytomies or multifurcations

Degree of the node: the number of branches connected to a node. Leaves have a degree of 1.

Polytomy (or multifurcation): If the root node has a degree greater than 2 or a non-root node has a degree greater than 3. A tree with no polytomies is called a binary tree, bifurcating tree, or fully resolved tree.



Polytomies or multifurcations

Hard polytomy: a polytomy representing truly simultaneous species divergences.

Soft polytomies: a polytomy representing lack of information in the data to resolve the relationship within a clade (a group of species).

Counting trees

Number of rooted trees

Number of unrooted trees

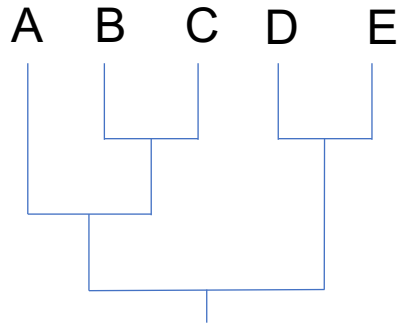
Number of species

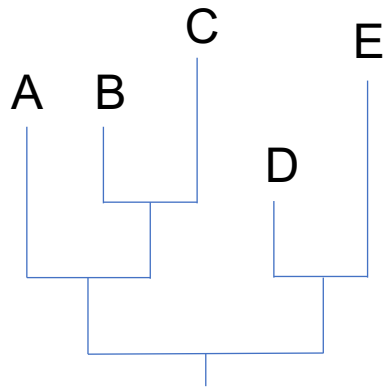
n	T_n	T_{n+1}
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425
20	$\sim 2.22 \times 10^{20}$	$\sim 8.20 \times 10^{21}$
50	$\sim 2.84 \times 10^{74}$	$\sim 2.75 \times 10^{76}$

The newick format

Newick format: for use in computer programs, trees are often represented using the parenthesis notation:

1. uses a pair of parentheses to group sister taxa into one clade.
2. a semicolon marking the end of the tree.
3. branch lengths are prefixed by colons.

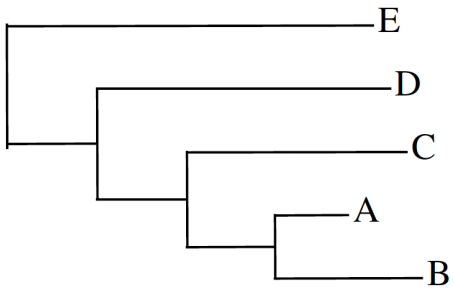


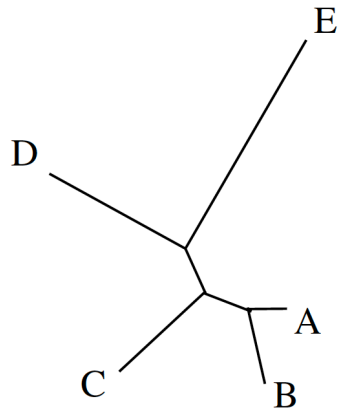



```
((A,(B,C)),(D,E));
```

```
((A:2,(B:1,C:2):1):1,(D:1,E:2.5):1);
```

Exercise. Draw the phylogenetic tree or write in Newick format the following trees.






```
(((A,B),C),(D,E));
```

```
(((A:0.1,B:0.2):0.3,C:0.2):0.1,(D:0.2,E:0.1):0.2);
```

2
Models of sequence evolution

Models of sequences evolution

Distance between two sequences: the expected number of nucleotide substitutions per site.

- ***p*-distance:** a simplistic distance measuring the proportion of different sites.

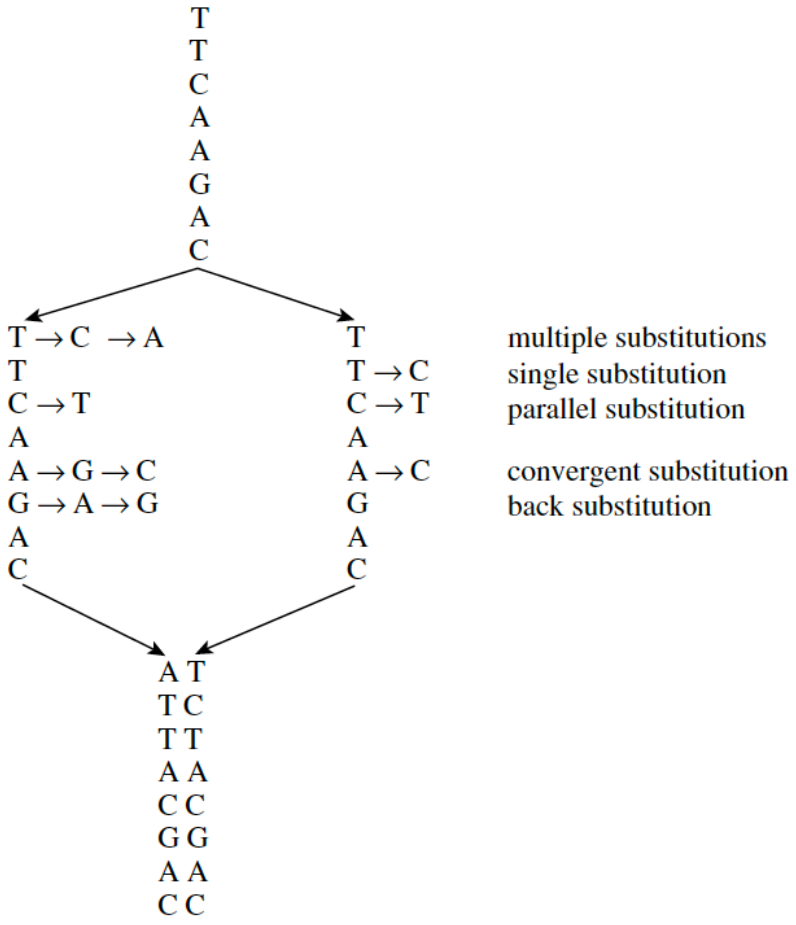
Exercise. Determine the p-distance between these aligned sequences:

```
S1: CGATATGCTGTCAGTCGATC
S2: CGTTATGCTGTCTGTCCATC
```

Models of sequences evolution

P-distance underestimates the number of substitutions that have occurred:

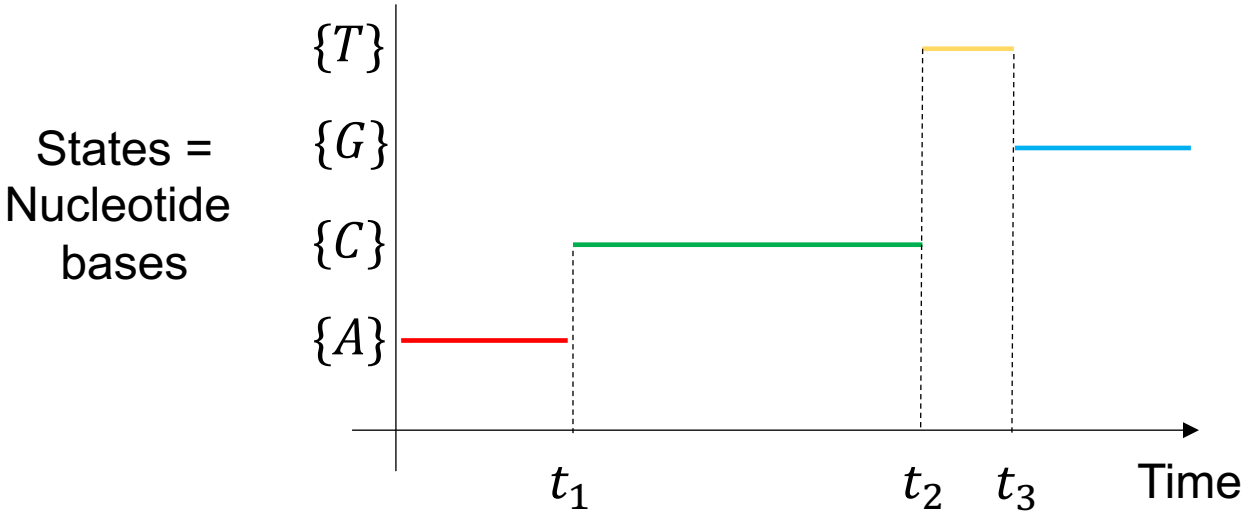
- a variable site may result from more than one substitution.
- a constant site, with the same nucleotide in the two sequences, may harbor **back or parallel substitutions**.
- Multiple substitutions at the same site or **multiple hits** cause some changes to be hidden.



Models of sequences evolution

To estimate the number of substitutions, we need probabilistic models to describe changes between nucleotides:

- **Continuous-time Markov chains:** Substitutions at any particular site are described by a Markov chain, with the four nucleotides to be the states of the chain.



Models of sequences evolution

Substitution rate matrix: list the rates of substitution from nucleotide i to j, with i and $j = A, C, G$ or T.

$$Q = \{q_{ij}\} = \begin{bmatrix} q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{bmatrix}$$

Transition probability matrix: lists the probabilities of a given nucleotide i will become j time t later.

$$P(t) = e^{Qt} = \begin{bmatrix} p_{AA}(t) & p_{AC}(t) & p_{AG}(t) & p_{AT}(t) \\ p_{CA}(t) & p_{CC}(t) & p_{CG}(t) & p_{CT}(t) \\ p_{GA}(t) & p_{GC}(t) & p_{GG}(t) & p_{GT}(t) \\ p_{TA}(t) & p_{TC}(t) & p_{TG}(t) & p_{TT}(t) \end{bmatrix}$$

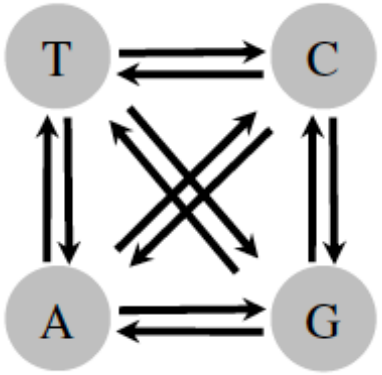
Limiting distribution (or stationary distribution): the probability that the chain is in state j when $t \to \infty$; is often represented by $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$.

Jukes and Cantor (1969) model

The JC69 model assumes that every nucleotide has the same rate:

- λ: rate of changing into any other nucleotide.

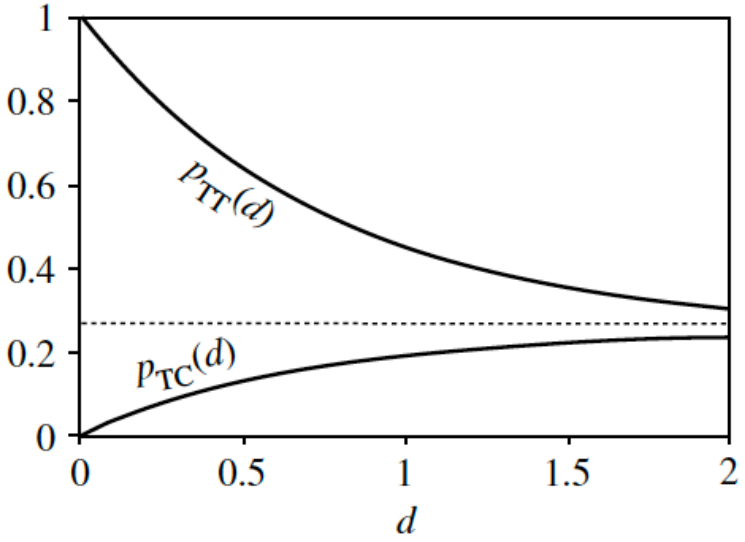
$$Q = \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix}$$



Jukes and Cantor (1969) model

There are two different elements of the transition-probability matrix:

$$P(t) = \begin{cases} p_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} \\ p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \end{cases}$$



Exercise. Calculate $P(t)$ for $t = 0$ and $t \to \infty$. Interpret those results.

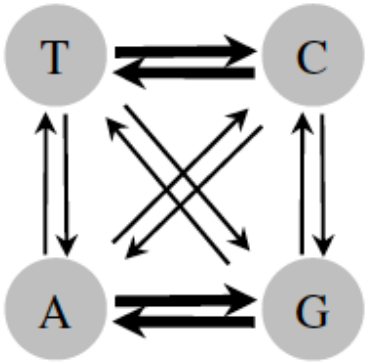
Kimura (1980) model

In real data, transitions often occur at higher rates than transversions:

- **Transitions:** substitutions between the two pyrimidines (T↔C) or between the two purines (A↔G).
- **Transversions:** substitutions between a pyrimidine and a purine (T,C↔A,G).

K80 accounts for different transition and transversion rates.

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

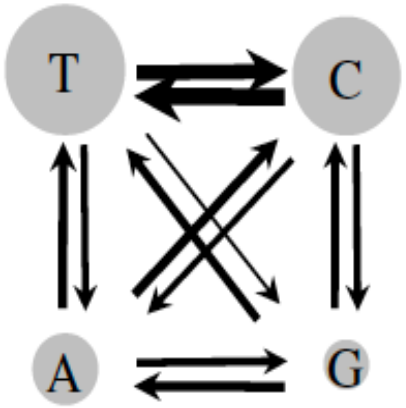


HKY (1985) model

JC69 and K80 have symmetrical substitution rates.

- When the substitution process reaches equilibrium, the sequence will have equal proportions of the four nucleotide: i.e., $\pi = \frac{1}{4}$.
- This assumption is unrealistic for virtually every real data set.

$$Q = \begin{bmatrix} . & \beta\pi_C & \alpha\pi_T & \beta\pi_T \\ \beta\pi_A & . & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & . & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & . \end{bmatrix}$$



General time-reversible model

GTR (or REV) is the most general time-reversible model of evolution:

- It considers a variable base composition π and six different exchangeabilities ρ.

$$Q = \begin{bmatrix} \cdot & \rho_1\pi_C & \rho_2\pi_T & \rho_3\pi_T \\ \rho_1\pi_A & \cdot & \rho_4\pi_G & \rho_5\pi_T \\ \rho_2\pi_A & \rho_4\pi_C & \cdot & \rho_6\pi_T \\ \rho_3\pi_A & \rho_5\pi_C & \rho_6\pi_G & \cdot \end{bmatrix}$$

- A Markov chain is time-reversible if $\rho_{ij} = \rho_{ji}$ for $i \neq j$.
- There is no biological reason to expect the substitution process to be reversible → is a mathematical convenience.

Exercises. Match the models in the first column with their sequence evolution features on the second column.

- JC69
- K80
- HKY
- GTR

- Homogeneous base composition
- Heterogeneous base composition
- A single exchangeability
- Heterogeneous exchangeabilities
- Reversibility

Variable substitution rates across sites

All models discussed so far assume that different sites in the sequence evolve in the same way and at the same rate. This assumption may be unrealistic in real data.

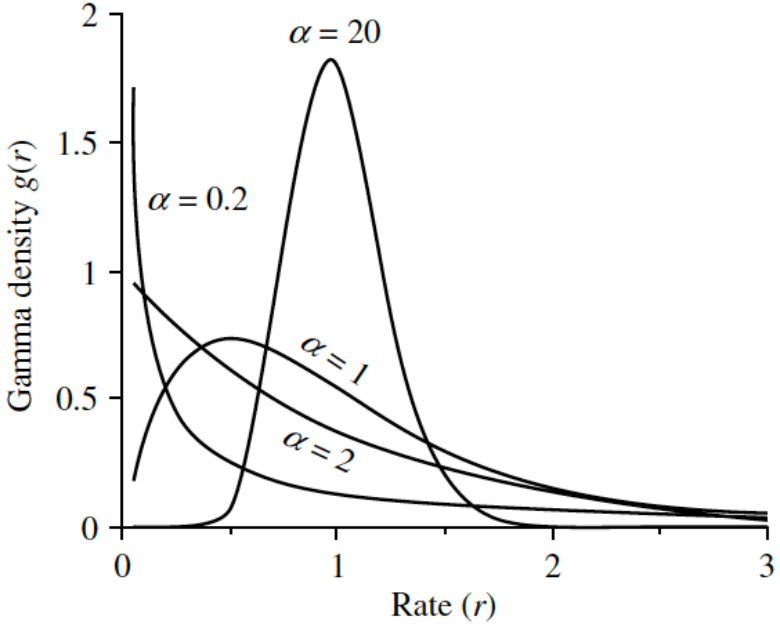
- The mutation rate may vary among sites.
- Mutations at different sites may be fixed at different rates due to their different roles in the structure and function of the gene → different selective pressures acting on them.

Variable substitution rates across sites

One can accommodate the rate variation by assuming a gamma distribution:

$$r \sim G(\alpha) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1}$$

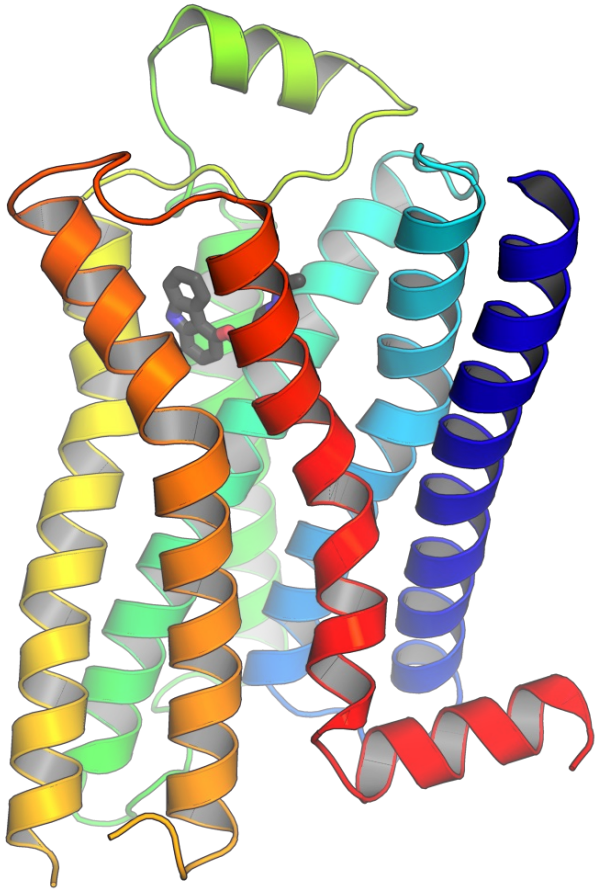
The shape parameter α is then inversely related to the extent of rate variation at sites:



- $\alpha > 1$: distribution is bell-shaped → most sites have intermediate rates around 1.
- $\alpha \leq 1$: distribution has a highly skewed L-shape → most sites have very low rates of substitution, but there are some substitution hotspots.

Exercise. G-protein-coupled receptors are the largest and most diverse group of membrane receptors in eukaryotes. These cell surface receptors act like an inbox for messages in the form of light energy, peptides, lipids, sugars, and proteins. Structurally, these proteins are integral membrane proteins that possess seven membrane-spanning domains or transmembrane helices.

Does it make sense to model across site variation if we are to perform phylogenetic analyses with the GPCRs gene family? If yes, what is the expectation for the value or α?



3
Tree reconstruction methods

Tree reconstruction methods

Distance-based: distances are calculated from pairwise comparison of sequences, forming a **distance matrix**, which is converted into a phylogenetic tree via a clustering algorithm:

- UPGMA (unweighted pair-group method using arithmetic averages) and NJ (neighbor-joining).

Character-based: attempt to fit the characters (nucleotides or amino acids) observed in all species at every site to a tree:

- **maximum parsimony, maximum likelihood** (ML), and **Bayesian**.

Distance methods are often computationally faster than character-based methods and can be easily applied to analyze different kinds of data.

Tree reconstruction methods

Method	Criterion
Maximum parsimony	Minimum number of changes, minimized over ancestral states
Maximum likelihood	Log-likelihood score, optimized over branch lengths and model parameters
Bayesian	Posterior probability, calculated by integrating over branch lengths and substitution parameters

Parsimony

Parsimony principle: is the principle that the simplest explanation that can explain the data is to be preferred.

- In phylogenetics, parsimony means that a hypothesis of relationships that requires the smallest number of character changes is most likely to be correct.

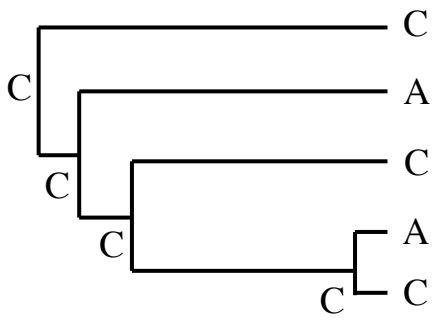
Parsimony

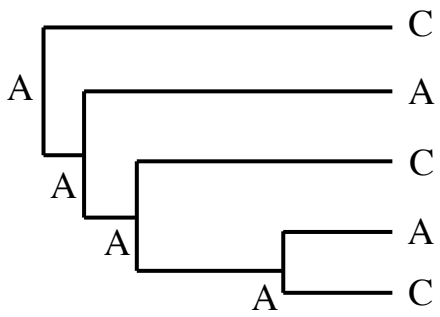
Character length (or site length): the minimum number of character changes at a site.

Tree length (or tree score or parsimony score): the sum of character lengths over all sites in the sequence is the minimum number of required changes for the entire sequence.

Maximum parsimony tree (or the most parsimonious tree): the tree with the smallest tree score is the estimate of the true tree.

Exercise. What is the most parsimonious scenario? Justify by calculating the site length in each case.





Maximum parsimony

Tree length of a tree is defined as:

$$L(T) = \sum_{i=1}^{S} \sum_{j=1}^{2N-3} c(x_j, y_j)$$

- S is the number of sites.
- N is the number of species in the tree.
- x_j and x_j are the states assigned to the nodes at each end of branch j.
- $c(x, y)$ is the cost associated with the change from state x to y.

Maximum parsimony

Some sites do not contribute to the discrimination of trees and are thus non-informative:

- **constant site:** the different species have the same nucleotide → requires no change for any tree.
- **singleton site:** two characters are observed but one is observed only once (e.g. TTTC or AAGA) → requires one change for every tree.

Exercise. Find the most parsimonious tree explaining the evolution of these four species. Consider only the unrooted trees.

```
Taxon        5     10
  S1      TTAGCTACTT
  S2      CTGGCCACTT
  S3      CTAGCTCCCT
  S4      CTGGACCCTT
```

Likelihood

Likelihood: the probability of observing the data when the parameters are given.

$$L(\theta) = p(D \mid T, \theta)$$

To define the likelihood, we have to specify a model by which the data D are generated including the:

- phylogenetic tree T.
- the set of parameter of the model of evolution θ.

Maximum likelihood

Some parameters produce the sequences with higher probability than others:

- we want the tree topology, branch lengths, and model parameters that best explain the observing the sequences.

Maximum likelihood

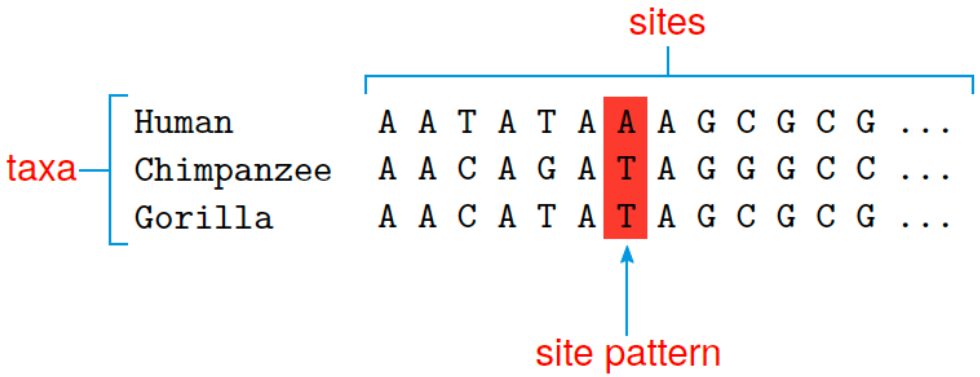
Mathematically speaking, within the maximum likelihood approach we want $\hat{T}$ and $\hat{\theta}$ that maximize the likelihood function:

$$\left\{\hat{T}, \hat{\theta}\right\} = \arg\max_{\theta}\{L(\theta)\}$$

- where $\hat{T}$ and $\hat{\theta}$ are the maximum likelihood estimates (MLEs).

Maximum likelihood: site pattern

A sequence alignment includes information from N taxa and S sites:



- **site pattern:** information from a single site → the bases of the same same share the same evolutionary history.

Maximum likelihood: site pattern

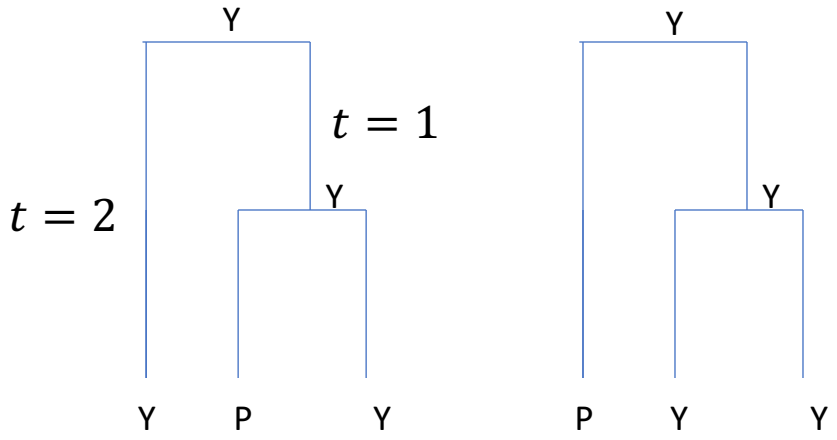
The probability of the whole alignment can be obtained from the probability of each site pattern is

$$p(D|T,\theta) = \prod_{i=1}^{S} p(d_i|T,\theta)$$

- d_i is the i-th site pattern.

Calculating $p(d_i|T,\theta)$ is a small likelihood problem.

Exercise. Calculate the likelihood of each scenario and find the likeliest one.



$$P(t) = \begin{cases} p_{ii}(t) = 0.5 + 0.5e^{-t} \\ p_{ij}(t) = 0.5 - 0.5e^{-t} \end{cases}$$

Maximum likelihood trees

Finding the maximum likelihood tree:

- maximizing the likelihood of an alignments for a given tree and model parameters is feasible but we want the tree that best described the data.
- try out several trees and find the one that maximizes the likelihood function.

Maximum likelihood trees

Problems in finding the maximum likelihood tree:

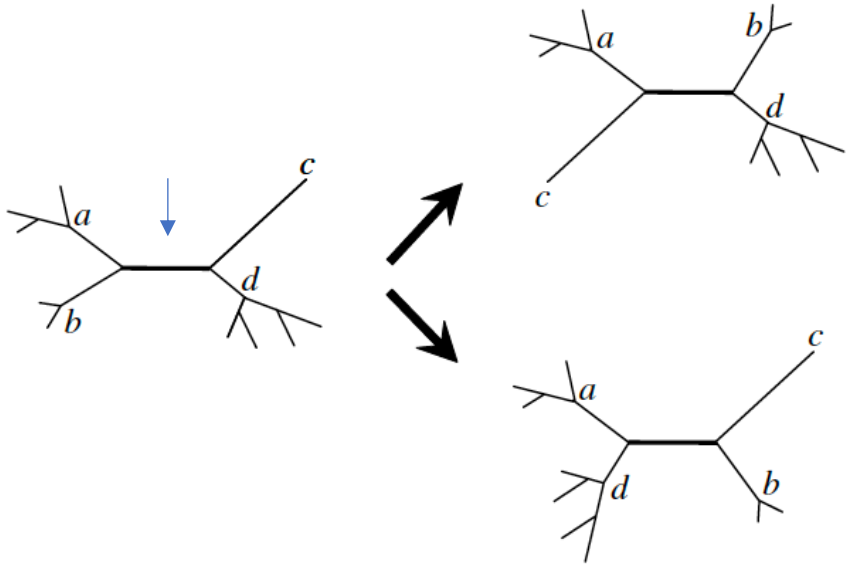
- there is a huge space of possible topologies.
- testing all possible trees is just impossible, even for moderately sized data sets.

n	T_n	T_{n+1}
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425
20	$\sim 2.22 \times 10^{20}$	$\sim 8.20 \times 10^{21}$
50	$\sim 2.84 \times 10^{74}$	$\sim 2.75 \times 10^{76}$

Maximum likelihood trees

Because testing all the possible trees is not computationally feasible, several algorithms are used to suggest reasonable trees.

- **full-tree arrangement operations:** change the structure of a given tree within its neighborhood.



Nearest-neighbor interchange (NNI) algorithm

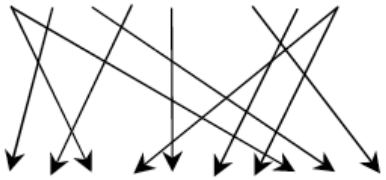
Maximum likelihood trees

Bootstrap: perhaps the most commonly used method for assessing uncertainties in estimated phylogenies:

- **pseudo-alignments** are created by sub setting the alignment.
- **pseudo-trees** are inferred for each pseudo-alignment.

Original alignment

Site	1	2	3	4	5	6	7	8	9	10
human	N	E	N	L	F	A	S	F	I	A
chimpanzee	N	E	N	L	F	A	S	F	A	A
bonobo	N	E	N	L	F	A	S	F	A	A
gorilla	N	E	N	L	F	A	S	F	I	A
orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
gibbon	N	E	N	L	F	T	S	F	A	T

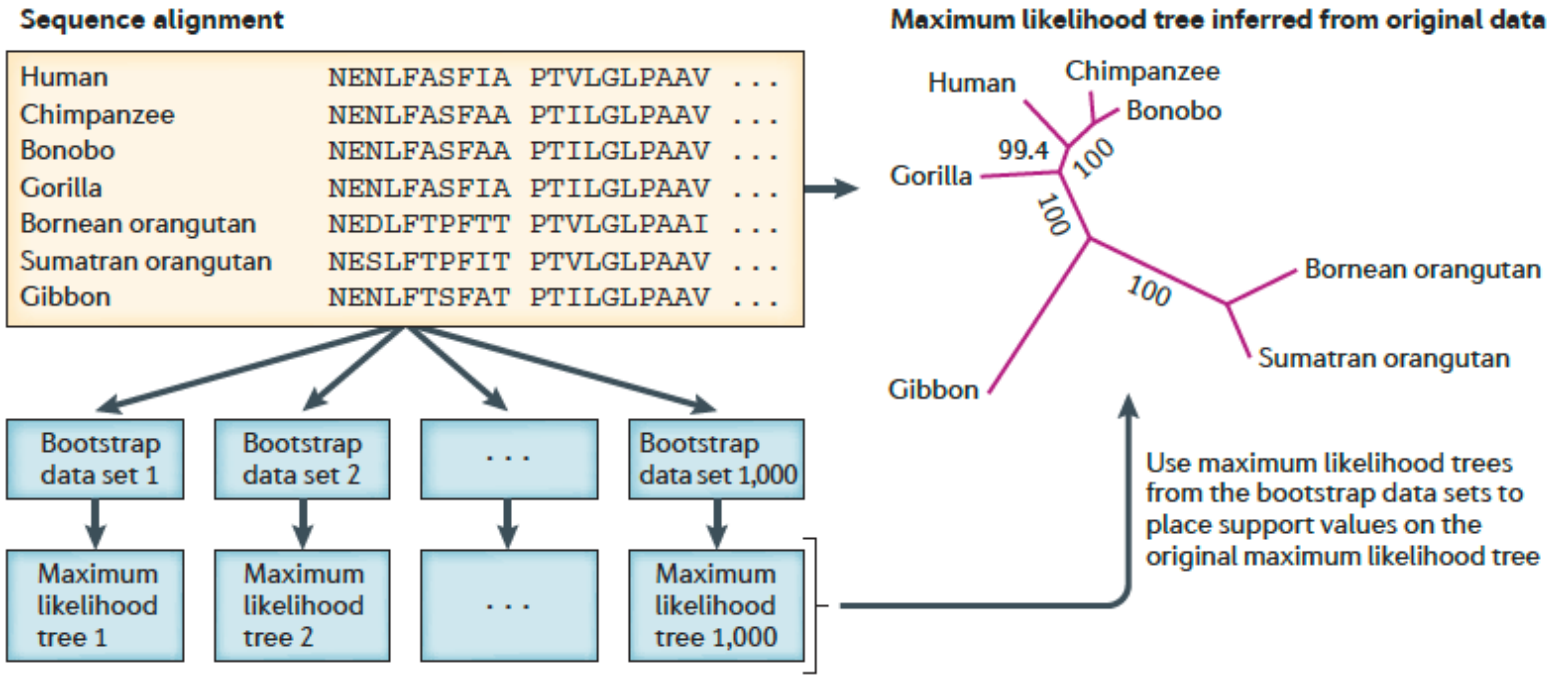


Bootstrap sample

Site	2	4	1	9	5	8	9	1	3	7
human	E	L	N	I	F	F	I	N	N	S
chimpanzee	E	L	N	A	F	F	A	N	N	S
bonobo	E	L	N	A	F	F	A	N	N	S
gorilla	E	L	N	I	F	F	I	N	N	S
orangutan	E	L	N	T	F	F	T	N	D	P
Sumatran	E	L	N	I	F	F	I	N	S	P
gibbon	E	L	N	A	F	F	A	N	N	S

Maximum likelihood trees

Bootstraps represent the number of times a certain clade is present in the pseudo-trees.



Bayesian inference

Bayesian approach to science:

- nothing more than a probability analysis.
- a mathematical formalization of a decision process.
- constitutes a different interpretation of probability.

Thomas Bayes
(1702-1761)

Bayesian inference

Bayesian way of reasoning about probabilities includes:

- prior beliefs.
- information from the data.
- the idea of updated probability.

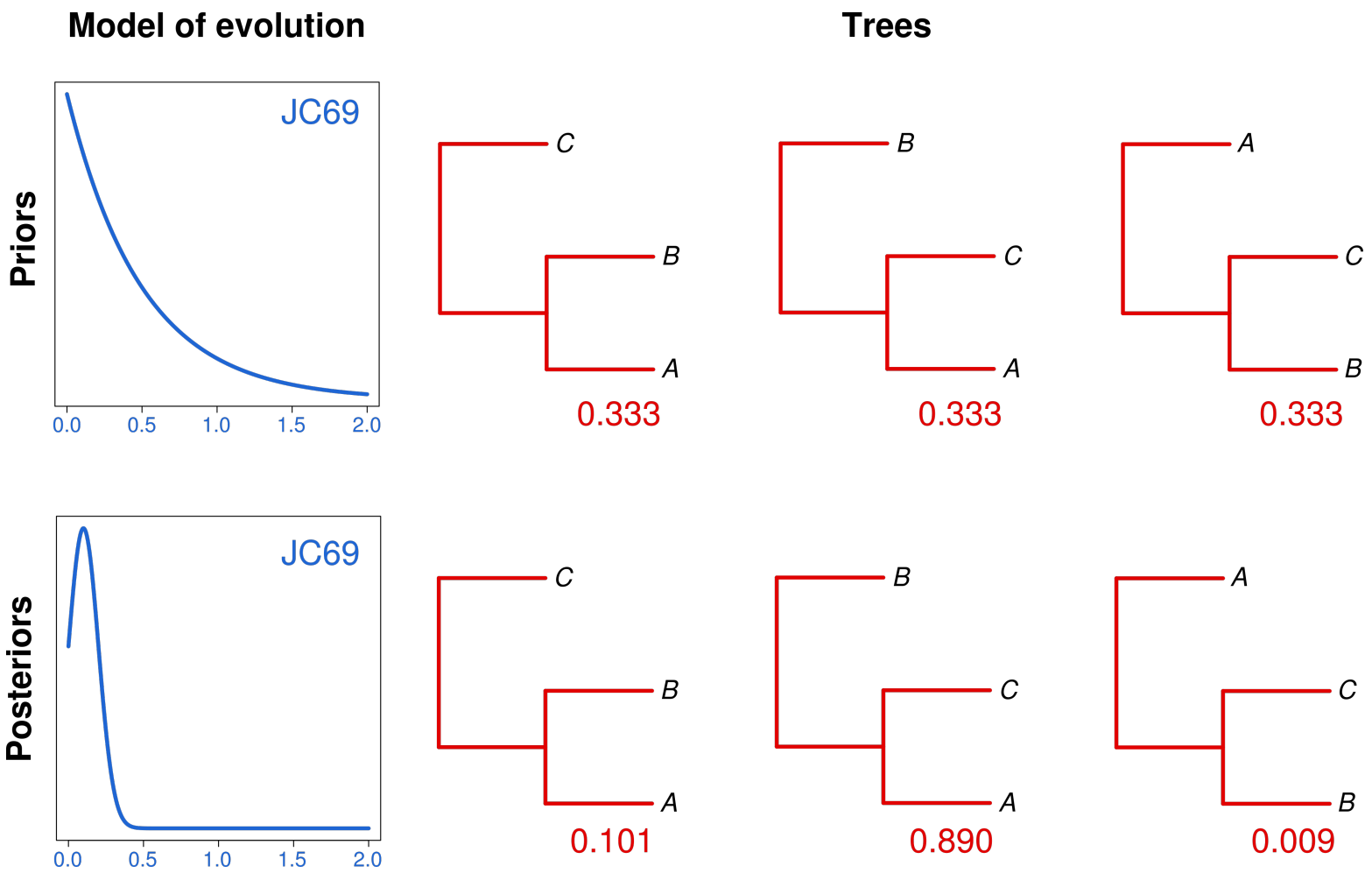
Bayesian inference

Bayes' theorem:

$$p(\theta|D) = p(\theta)p(D|\theta)$$

- $p(\theta)$ is the **prior distribution**.
- $p(D|\theta)$ is the **likelihood**.
- $p(\theta|D)$ is the **posterior distribution** $\rightarrow$ the probability after the prior has been updated with the available data.

Bayesian inference



Estimating the posterior distribution

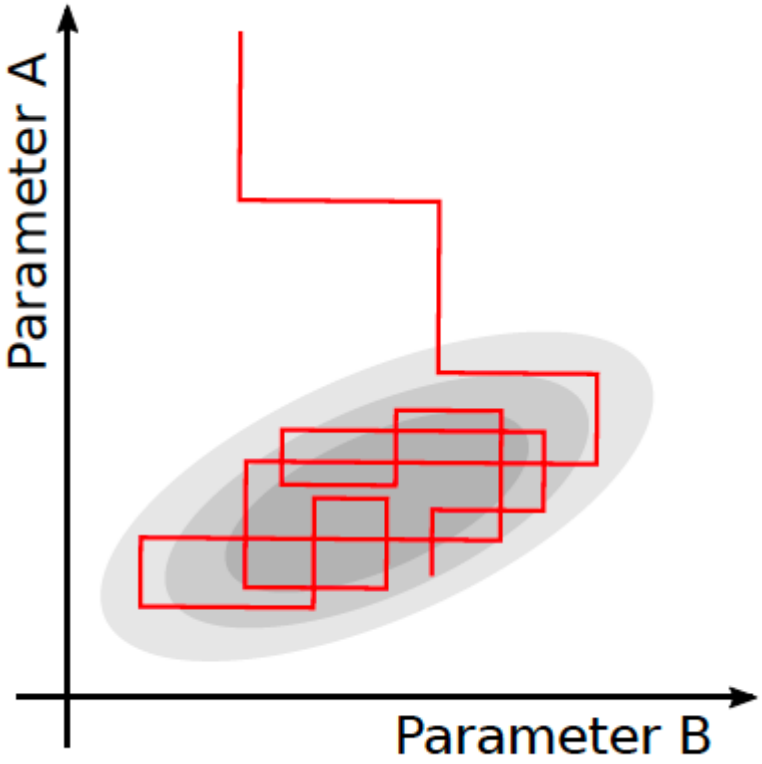
Estimating the posterior distribution in phylogenetic context can be difficult:

- impossible to derive the posterior probability distribution analytically.
- concentrated in a small part of a vast parameter space.

MCMC

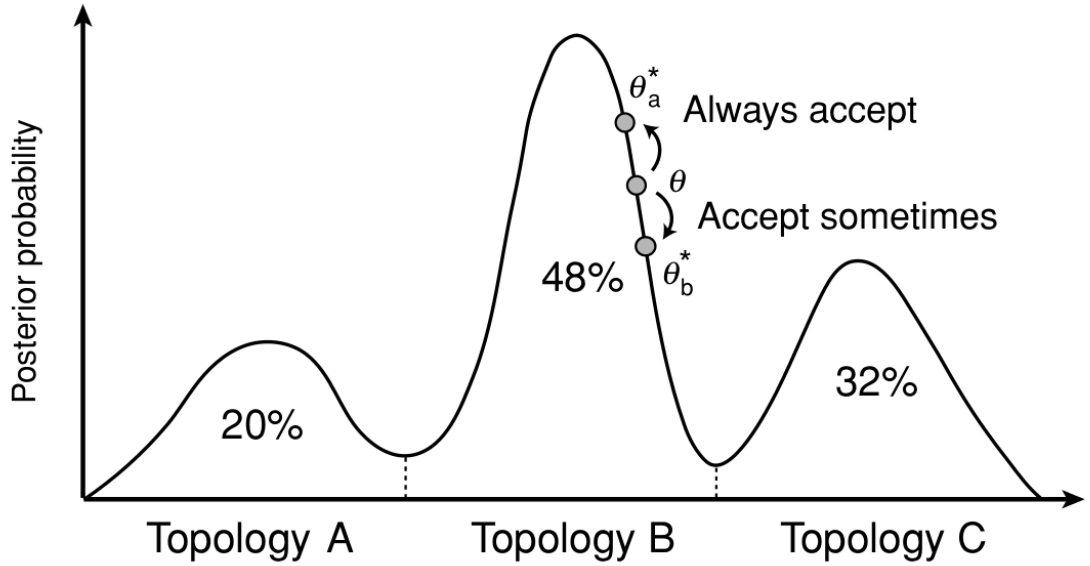
Estimate the posterior distribution using Markov chain Monte Carlo (MCMC) sampling:

- set up a Markov chain that converges onto the posterior probability distribution.
- MCMC represents random samples from the posterior.



MCMC

- Make small random changes on the parameter values.
- Accept or reject those changes according to the appropriate probabilities.



MCMC

MCMC run is a random sample of the posterior distribution:

- the amount of time it spends sampling a particular region is proportional to the posterior probability of that region.
- given that it converged to the target distribution → convergence needs to be monitored.

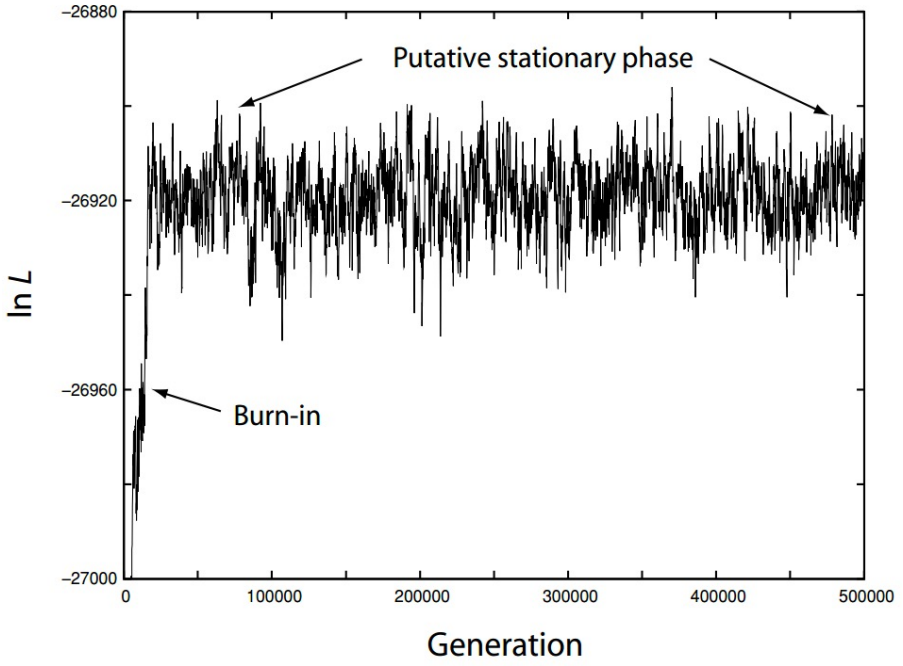
MCMC

Burn-in:

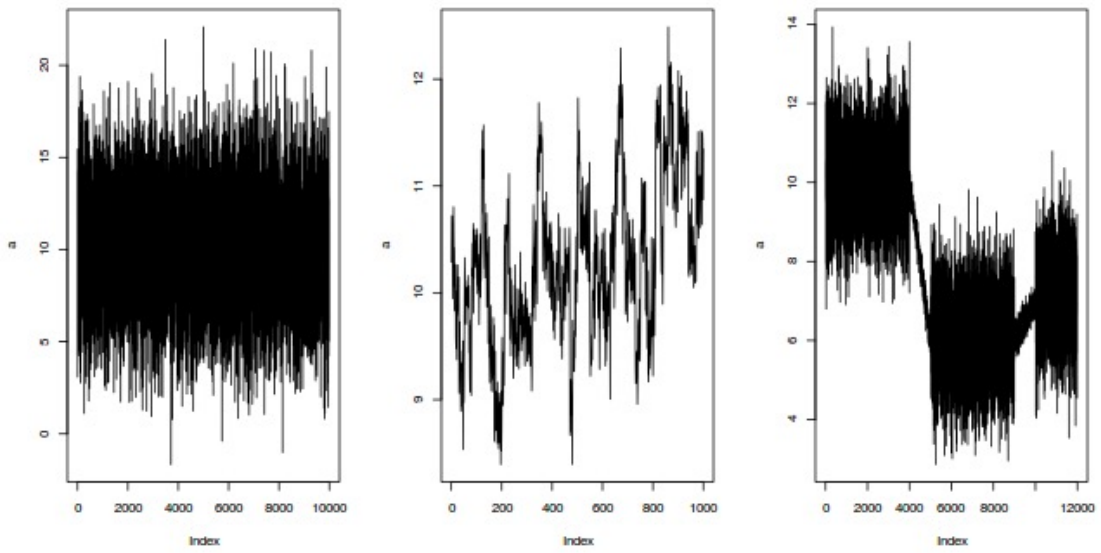
- early phase of the run
- heavily influenced by the starting points
- likelihood increases very rapidly

Stationary phase:

- the chain approaches its stationary distribution
- the likelihood values reach a plateau



Exercise. Select the trace plots that reflect a well converged MCMC. Justify.



Bayesian tree inference

The model parameters and the tree are summarized differently:

- model parameters are usually continuous and can be summarized as any usual statistics: means, median, the credibility interval.
- trees are more difficult to summarize.
- **posterior clade probabilities**: the sum of the posterior probabilities of all trees that contain that clade.

Exercise. Bayesian phylogenetic inferences in a sequence alignment with four species returned the three topologies with the following posterior probabilities (P.p.):

Topology	P.p.
(((Human,Dog),(Chicken,Lizard)),Frog)	0.76
((((Human,Dog),Chicken),Lizard),Frog)	0.17
(((Human,Dog),Chicken),(Lizard,Frog))	0.07

What is the posterior probability of the following clades: (Human,Dog), (Chicken lizard), (Chicken,Frog) and ((Human,Dog),Chicken)?

ML vs. Bayesian tree inference

Maximum-Likelihood trees

- $p(D|\theta)$
- Maximum likelihood tree
- ignores pre-existing information
- bootstrapping
- resamples characters

Bayesian trees

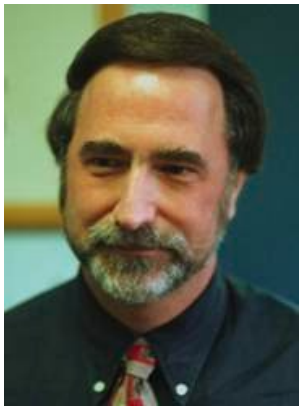
- $p(\theta|D)$
- Maximum *a posteriori* tree
- considers pre-existing information
- MCMC
- resamples parameters

4
Tutorial:
Bayesian Phylogenetic Inference

Forensic application

A gastroenterologist was convicted of attempted second-degree murder by injecting his former girlfriend with blood or blood-products obtained from an HIV type 1 (HIV-1)-infected patient under his care.

Molecular evidence of HIV-1 transmission in a criminal case
Metzker et al. PNAS October 29, 2002 99 (22) 14292-14297



Dr Richard J Schmidt

Janice Trahan

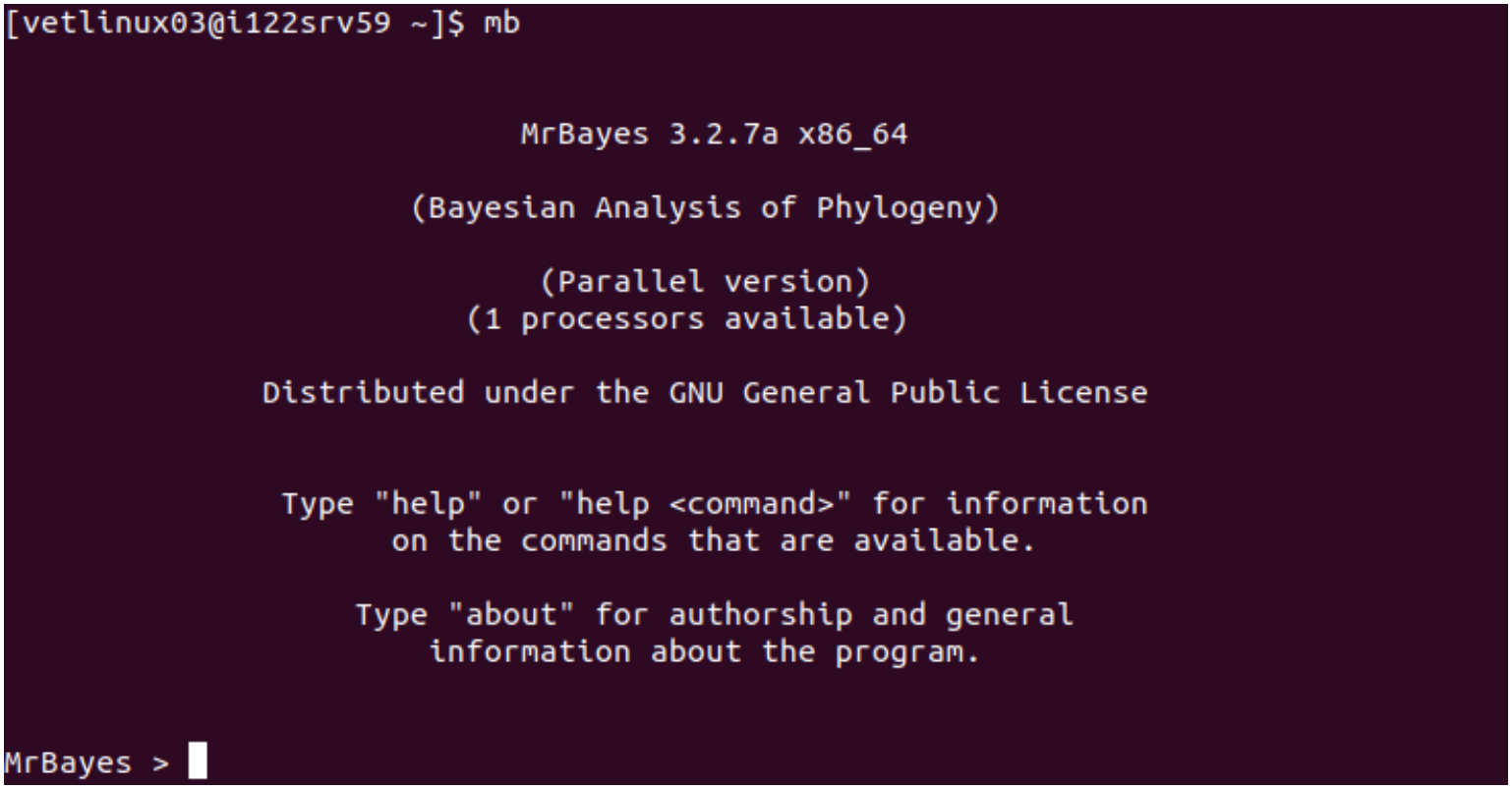
- we want to determine if there is evidence that the victim was directly infected by blood taken from the Gastroenterologist´s patient.

Data

Dataset includes HIV env/gp120 sequences:

- Sequences from the victim all begin with V.
- Sequences from the patient all begin with P.
- Sequences from HIV-infected individuals from the local metropolitan area all begin with LA and were included as controls.
- Two divergent sequences: `M62320_HIVU455` and `K03454_HIVELI`.

MrBayes


[vetlinux03@i122srv59 ~]$ mb
MrBayes 3.2.7a x86_64
(Bayesian Analysis of Phylogeny)
(Parallel version)
(1 processors available)
Distributed under the GNU General Public License
Type "help" or "help <command>" for information
on the commands that are available.
Type "about" for authorship and general
information about the program.
MrBayes >

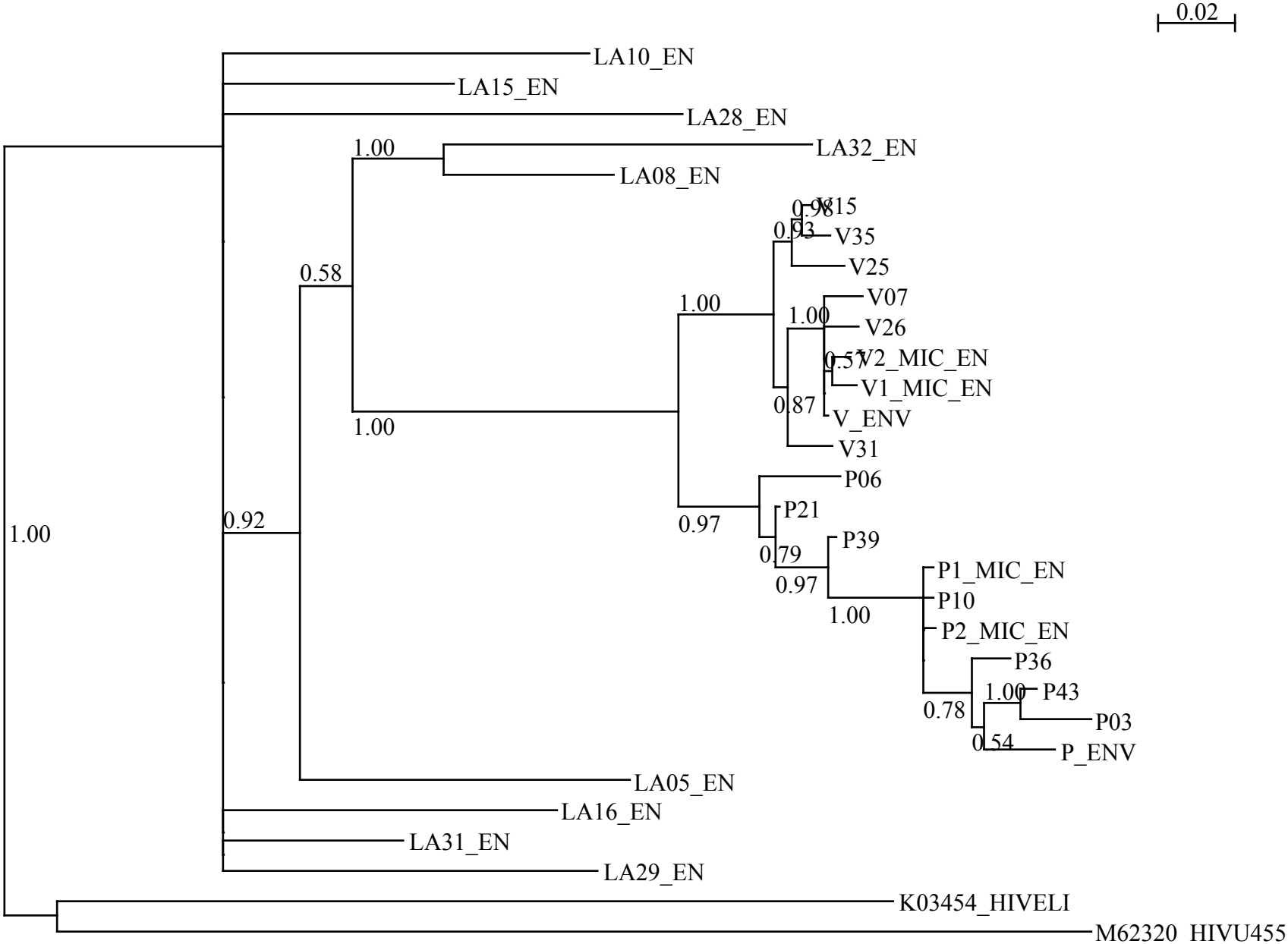
MrBayes script

```
execute HIV_env_gp120.nxs
outgroup M62320_HIVU455

lset nst=6 rates=gamma

mcmcp ngen=100000 samplefreq=100 printfreq=100
mcmc

sump burnin=100
sumt burnin=100
```

Phylogenetic problem

Do we have evidence that the gastroenterologist deliberately infected the victim with HIV-infected blood from one of their patients?

- Support your conclusions using the obtained branch support values.


0.02
LA10_EN
LA15_EN
LA28_EN
LA32_EN
LA08_EN
V15
V35
V25
V07
V26
V2_MIC_EN
V1_MIC_EN
V_ENV
V31
P06
P21
P39
P1_MIC_EN
P10
P2_MIC_EN
P36
P43
P03
P_ENV
LA05_EN
LA16_EN
LA31_EN
LA29_EN
K03454_HIVELI
M62320_HIVU455
1.00
0.58
1.00
0.92
1.00
0.98
0.93
1.00
1.00
0.57
0.87
0.97
0.79
0.97
1.00
0.78
1.00
0.54