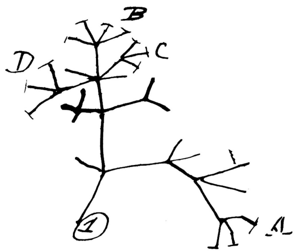


IPS-164 INTRODUCTION TO PHYLOGENETICS 2022

Lecture 8
Reconstructing phylogenetic trees. Part I

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. Calculating likelihood on a tree: Felsentein's pruning algorithm
2. Overview of the main properties of Markov models
3. General workflow for tree inference using DNA

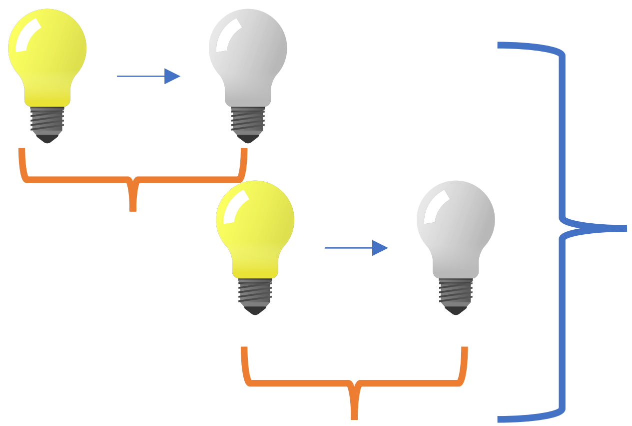
Ingredients to derive continuous-time Markov models

Binomial distribution → Poisson distrubution → Exponential distribution → Continuous-time and discrete-state Markov models

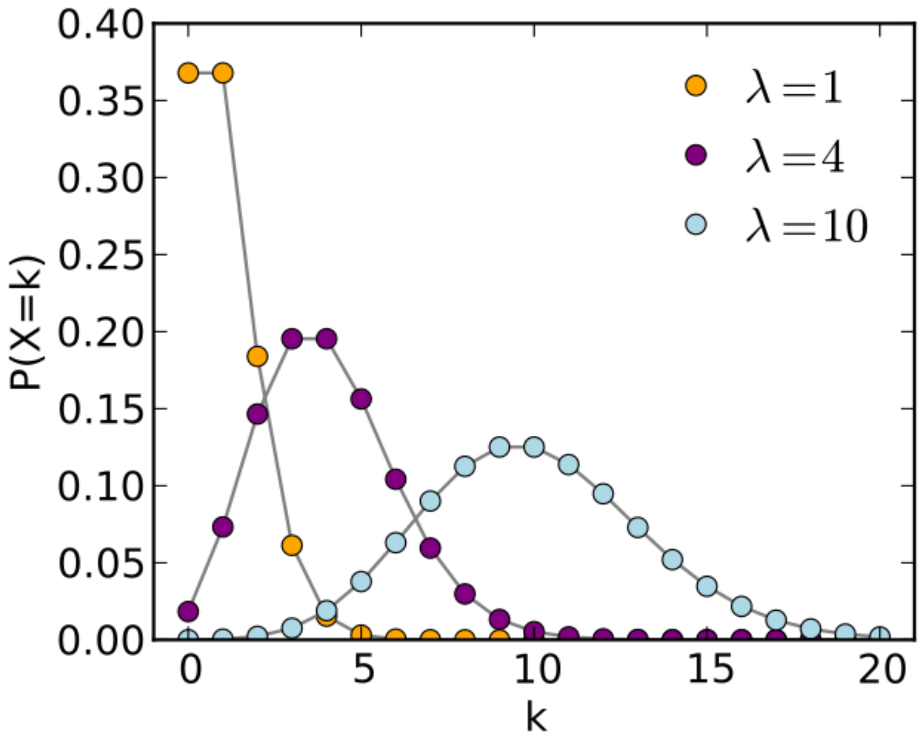
Poisson distribution

$$\text{Poisson}(k\,|\lambda,t) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$

- λ is called the rate parameter
- Poisson distr. shows the number of changes k given λ and time t

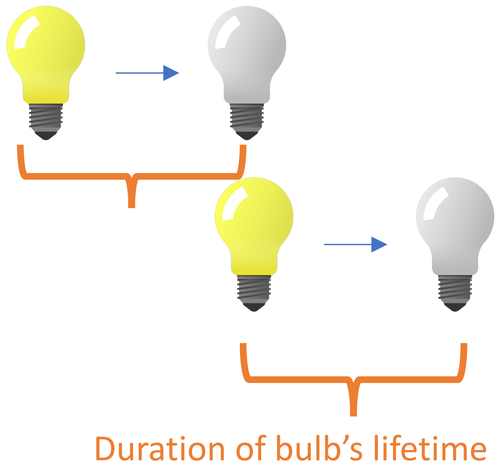


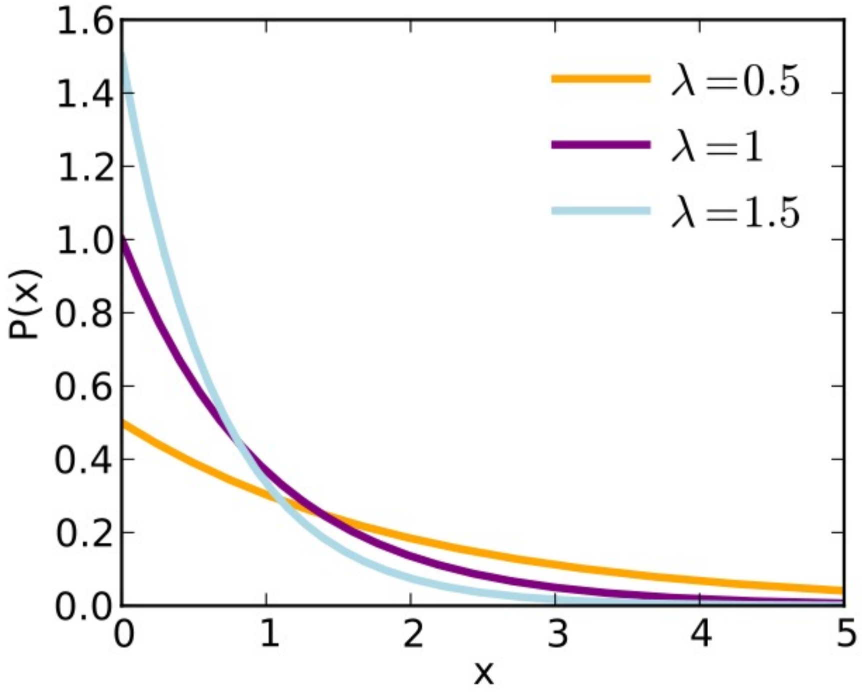
Number of dead bulb's per year



Exponential distribution

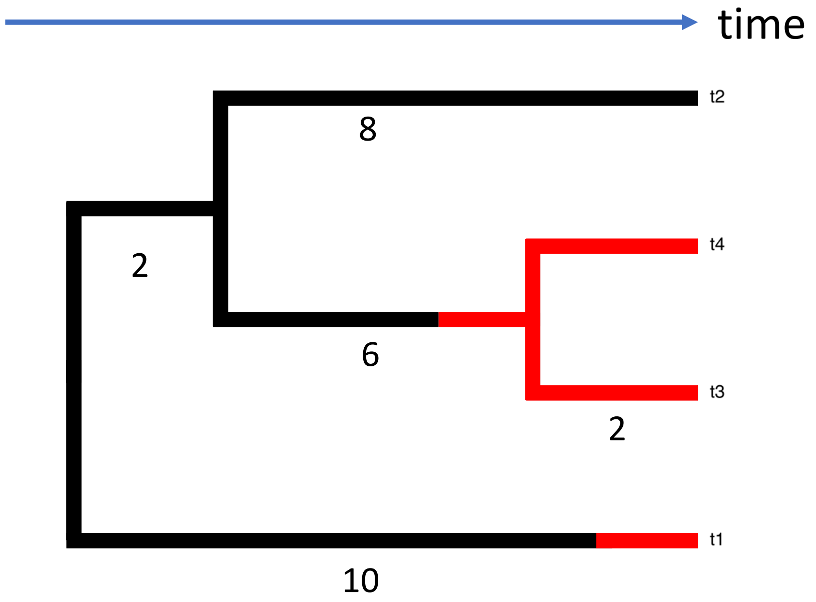
$$\text{Exponential}\,(\,t \mid \lambda) = \lambda e^{-\lambda t}$$





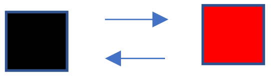
- Exponential and Poisson are the same processes but different aspects
- Same interpretation of the parameter $\boldsymbol{\lambda}$ **(=rate)**
- $\boldsymbol{\lambda}$ is the mean number of changes over time interval in Poisson

Simulating data under Markov models on a tree

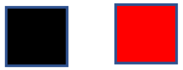


$$\boldsymbol{Q}=\begin{bmatrix}-1 & 1\\ 2 & -2\end{bmatrix}$$

state 1 state 2



Initial vector

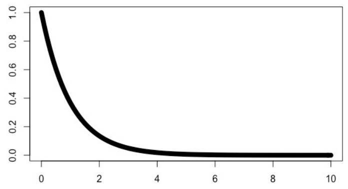


π =(1/2, 1/2)

Random number generator

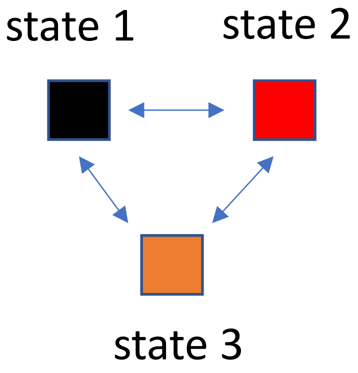
1. Randomly select state at the root from a uniform distribution. RND=0.4 (starting state 1)

2. Draw a random number from Exponential distribution with $\lambda = 1$. RND=2.4



3. Draw a random number from Exp($\lambda = 1$). RND=8
4. Draw a random number from Exp($\lambda = 1$). RND=4.2 (to state 2)
5. Draw a random number from Exp($\lambda = 2$). RND=4.6
6. Draw a random number from Exp($\lambda = 2$). RND=4.9
7. Draw a random number from Exp($\lambda = 1$). RND=9.1 (to state 2)
8. Draw a random number from Exp($\lambda = 2$). RND=3.3

From rates to probabilities



- **Transition rate matrix.** Infinitesimal rates

$$\boldsymbol{Q} = \begin{bmatrix} -0.5 & 0.4 & 0.1 \\ 0.8 & -1 & 0.2 \\ 0.96 & 0.24 & -1.2 \end{bmatrix}$$

- **Probability transition matrix.** Exponentiate rate matrix

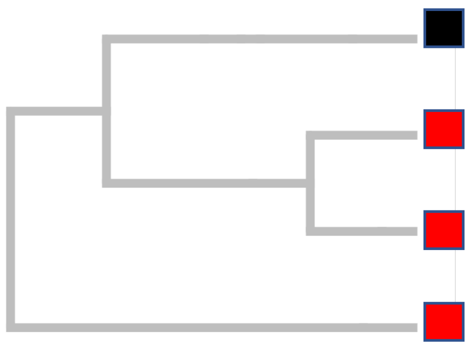
$$P(\boldsymbol{Q}, t) = e^{Qt}$$

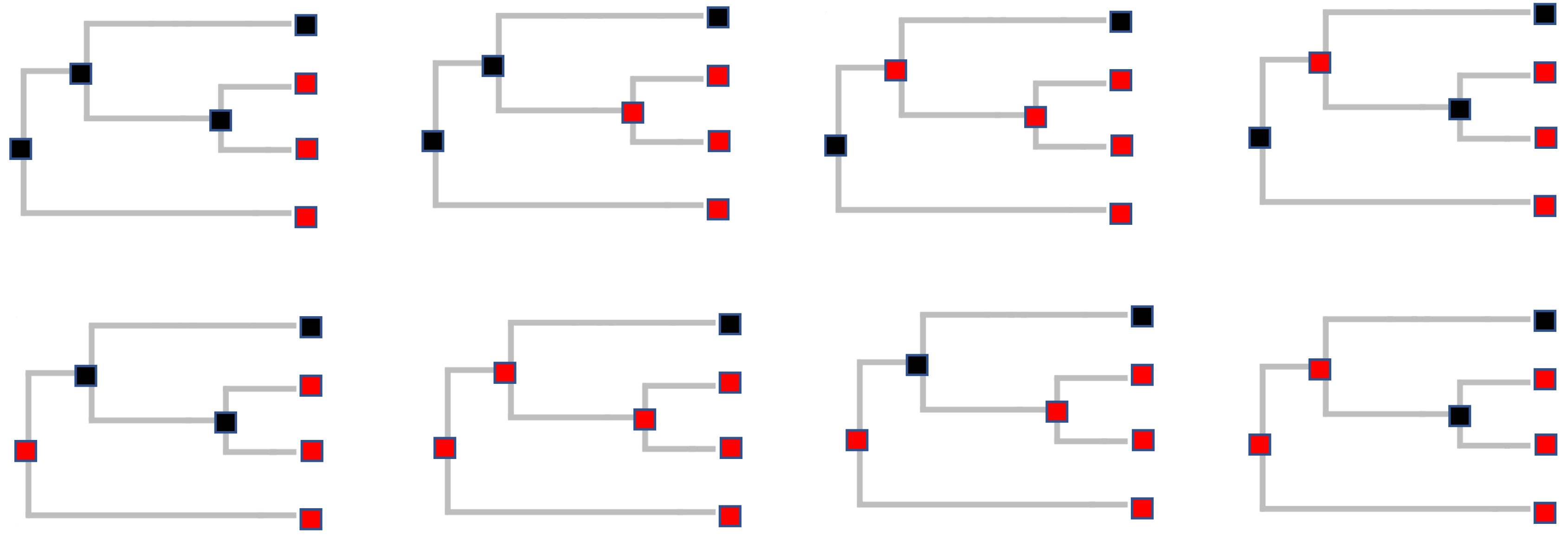
$$e^{Q*1} = \begin{bmatrix} 0.72 & 0.2 & 0.08 \\ 0.46 & 0.46 & 0.08 \\ 0.46 & 0.2 & 0.34 \end{bmatrix}$$

Matrix exponential transforms rates into probabilities:

$$\boxed{e^{Qt} = 1 + \frac{Qt^1}{1!} + \frac{Qt^2}{2!} + \frac{Qt^3}{3!} + \cdots}$$

Inference: estimating tree likelihood





Integrate over all possible combinations of the ancestral states

Let’s calculate likelihoods

- HTH
- H?H

Let’s calculate likelihoods

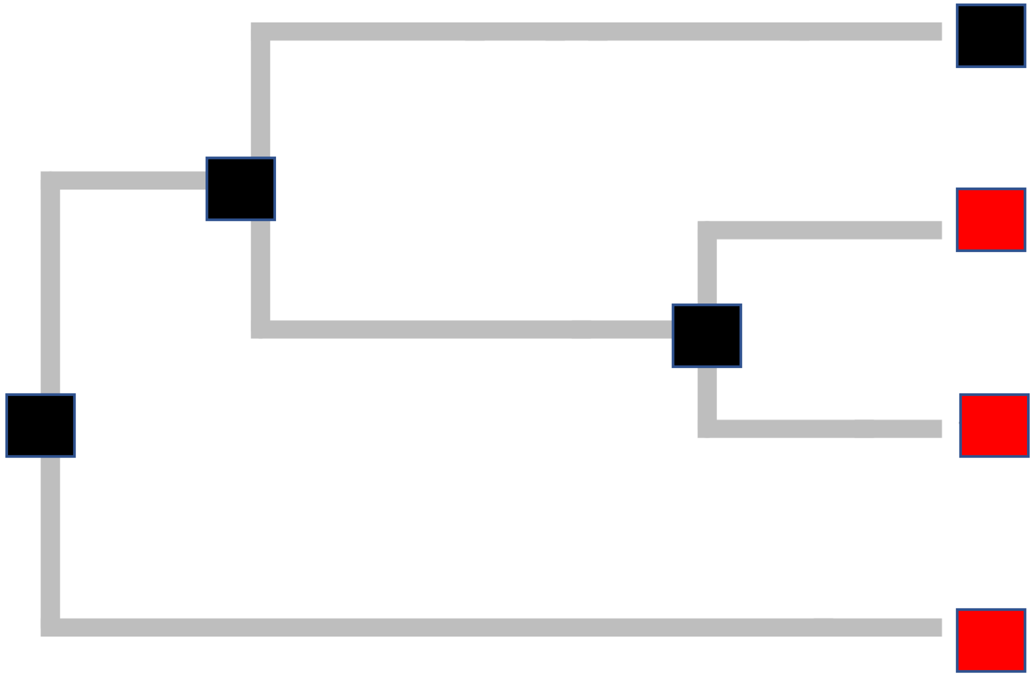
$$\boldsymbol{Q} = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix} \qquad e^{Q*0.1} = \begin{array}{cc} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.91 & 0.09 \\ \underline{0.17} & 0.83 \end{bmatrix} \end{array} \qquad e^{Q*10} = \begin{array}{cc} & \begin{array}{cc} & 2 \end{array} \\ \begin{array}{c} \\ 2 \end{array} & \begin{bmatrix} 0.66 & 0.33 \\ 0.66 & \underline{0.33} \end{bmatrix} \end{array}$$

- S2 -> S1 after t=0.1

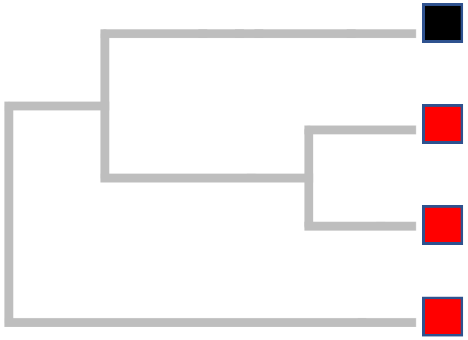
- S2 -> S2 after t=10

Let's calculate likelihoods

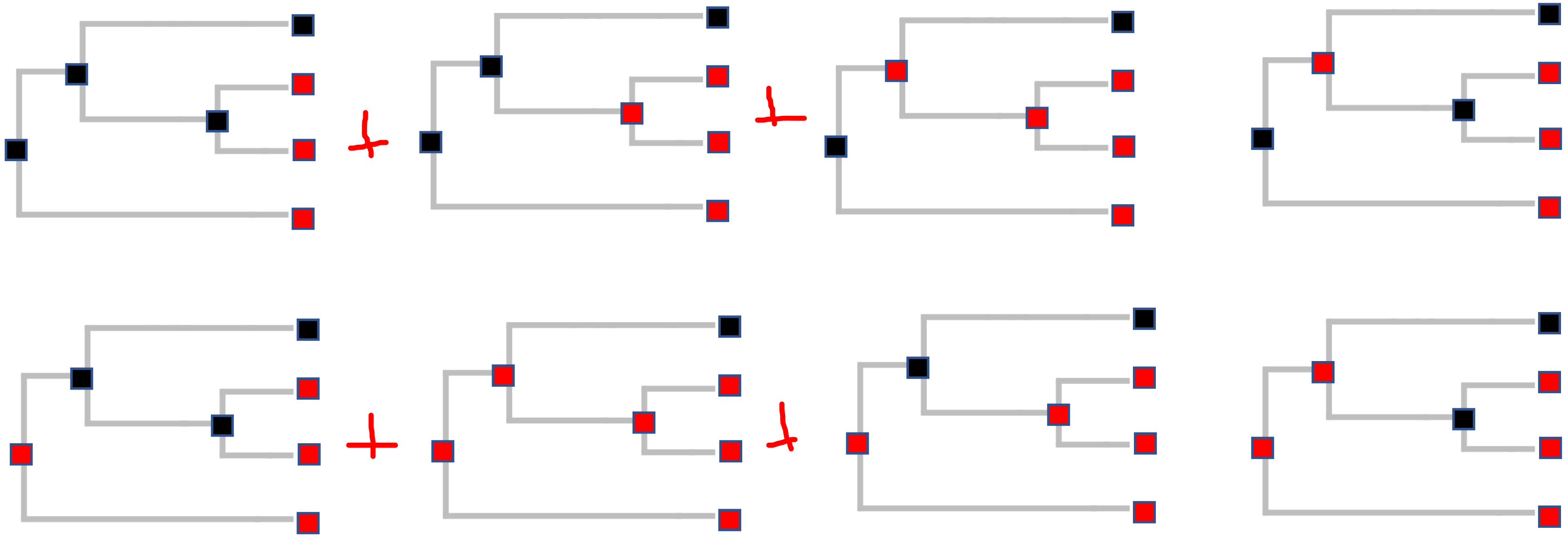
$$\boldsymbol{Q} = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix} \qquad e^{Q*10} = \begin{bmatrix} 0.66 & 0.33 \\ 0.66 & 0.33 \end{bmatrix}$$



Inference: estimating tree likelihood



Our Data

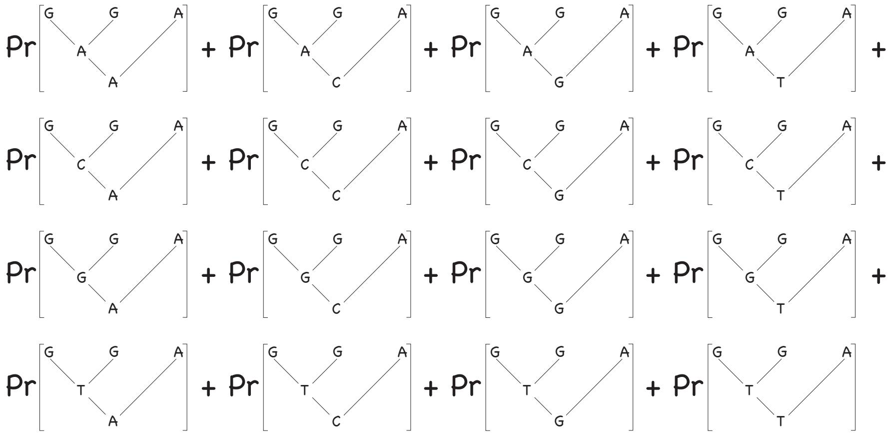


Integrate over all possible combinations of the ancestral states

Likelihood of DNA sequence

```
#NEXUS

begin data;
     dimensions ntax=5 nchar=895;
     format gap=- datatype=dna;
     matrix
     Human      AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTT.....AACCCAAACAACCCAGCTCTCCCTAAGCTT
     Chimpanzee AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTT.....AACCCAAACAACCCAGCTCTCCCTAAGCTT
     Gorilla    AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTT.....AACCCAAACAATTCAACTCTCCCTAAGCTT
     ;
end;
```



1. Calculate likelihood for each site
2. The likelihood of the entire DNA sequence is the product of the likelihoods for each site
3. Or the sum of the log likelihoods for each site

(from John Huelsenbeck’s presentation)

Felsenstein’s coding data at tips

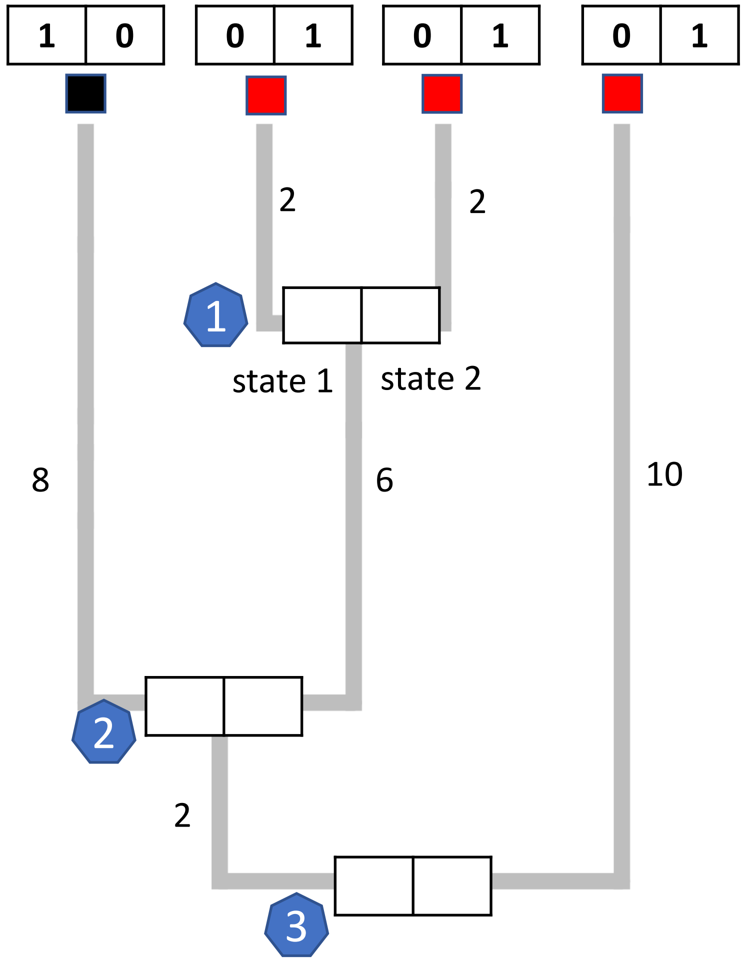
Given values:

state 1 state 2

1	0	0	1

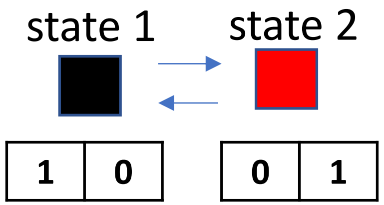
$$\boldsymbol{Q}=\begin{bmatrix}-1 & 1\\ 2 & -2\end{bmatrix}$$

π =(1/2, 1/2)

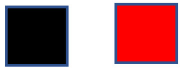


Felsenstein’s pruning algorithm

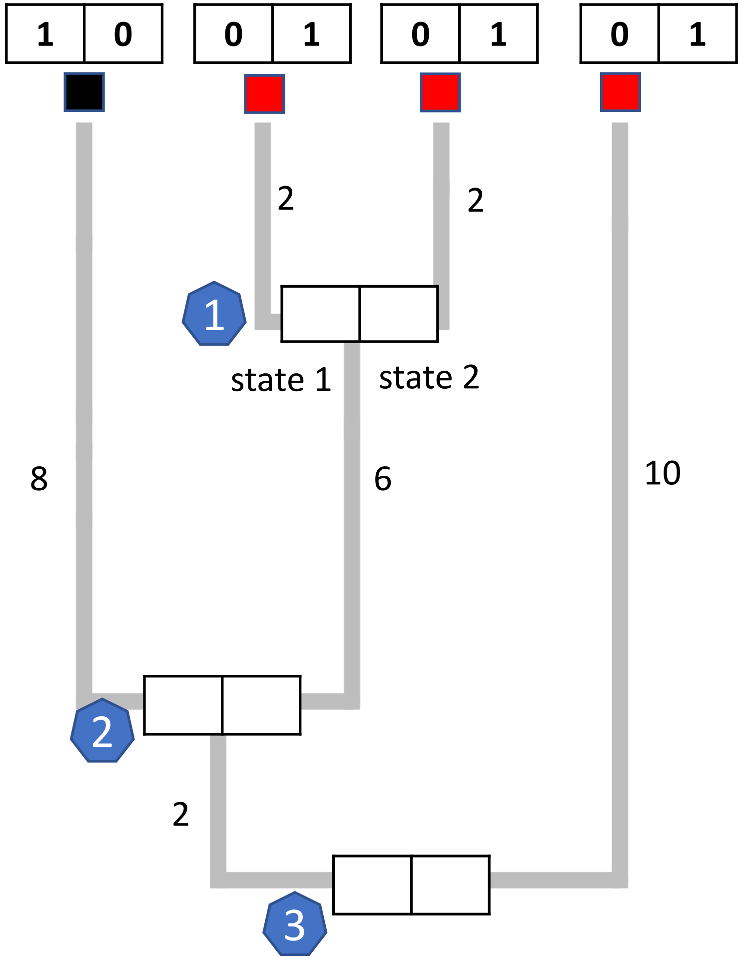
Given values:



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



π =(1/2, 1/2)



Felsenstein's pruning algorithm

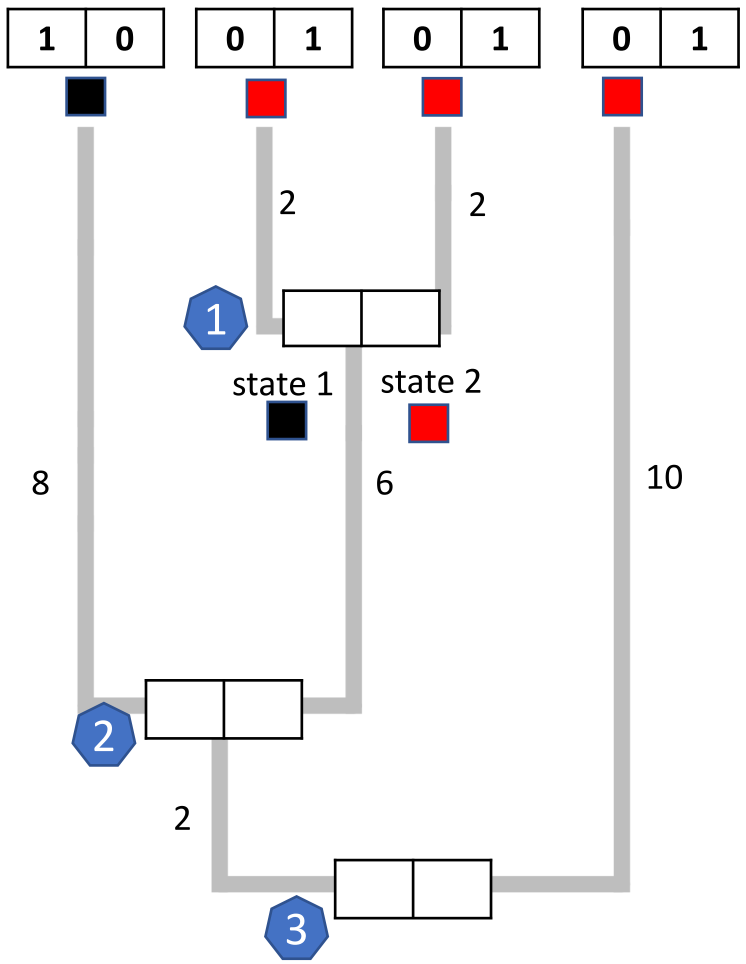
Given values:

state 1 state 2

1	0		0	1

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

π =(1/2, 1/2)



Node 1

At state 1

Left br. $e^{Q*2} = \begin{bmatrix} 0.66 & \mathbf{0.34} \\ 0.66 & 0.34 \end{bmatrix}$

Right br. $e^{Q*2} = \begin{bmatrix} 0.66 & \mathbf{0.34} \\ 0.66 & 0.34 \end{bmatrix}$

Pr(N_1 at **black**)= 0.34*0.34=0.12

At state 2

Left br. $e^{Q*2} = \begin{bmatrix} 0.66 & 0.34 \\ 0.66 & \mathbf{0.34} \end{bmatrix}$

Right br. $e^{Q*2} = \begin{bmatrix} 0.66 & 0.34 \\ 0.66 & \mathbf{0.34} \end{bmatrix}$

Pr(N_1 at **red**)= 0.34*0.34=0.12

Felsenstein’s pruning algorithm

Given values:

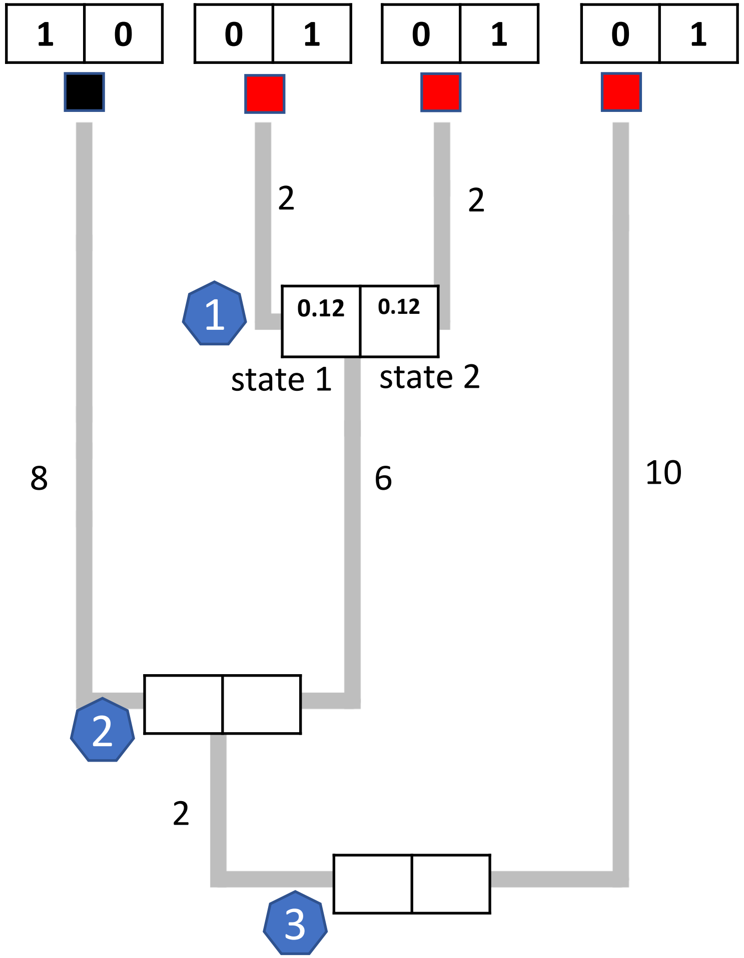
state 1 state 2

1	0

0	1

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

π =(1/2, 1/2)



Felsenstein's pruning algorithm

Given values:

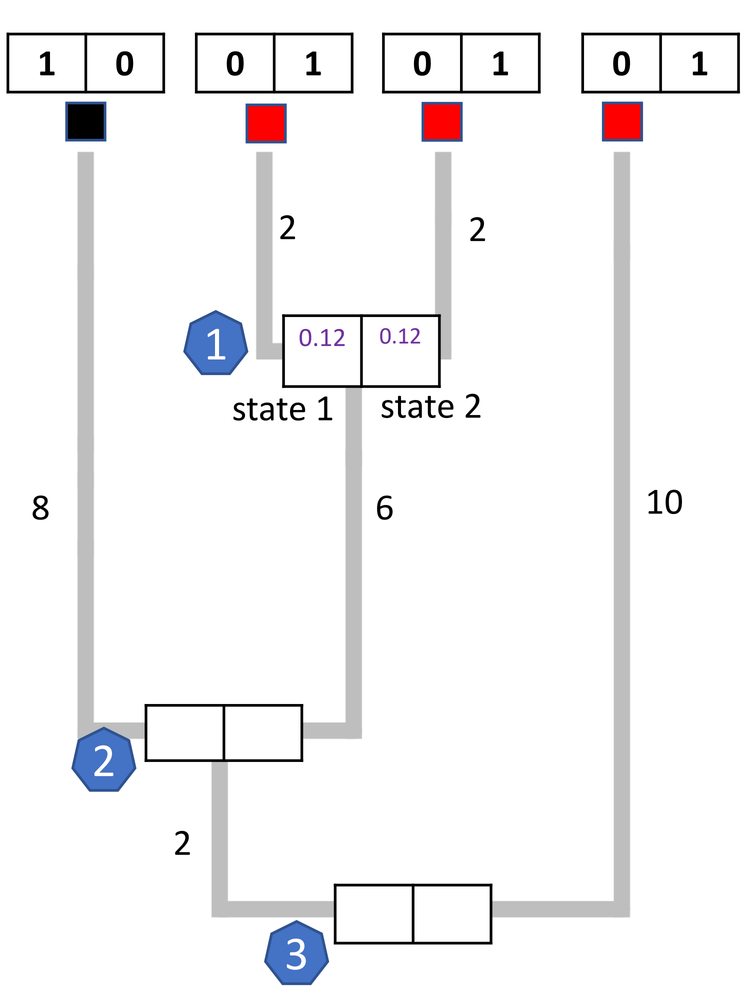
state 1 state 2

1	0

0	1

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

π =(1/2, 1/2)



Node 2

At state 1

Left br. $e^{Q*8}=\begin{bmatrix}\mathbf{0.66} & 0.33\\ 0.66 & 0.33\end{bmatrix}$

Right br. $e^{Q*6}=\begin{bmatrix}\mathbf{0.66} & \mathbf{0.33}\\ 0.66 & 0.33\end{bmatrix}$

Pr(N_2 at **black**)=0.66*(0.66*0.12+0.33*0.12)=0.08

At state 2

Left br. $e^{Q*8}=\begin{bmatrix}0.66 & 0.33\\ 0.66 & 0.33\end{bmatrix}$

Right br. $e^{Q*6}=\begin{bmatrix}0.66 & 0.33\\ \mathbf{0.66} & \mathbf{0.33}\end{bmatrix}$

Pr(N_2 at **red**)=0.66*(0.66*0.12+0.33*0.12)=0.04

Felsenstein's pruning algorithm

Given values:

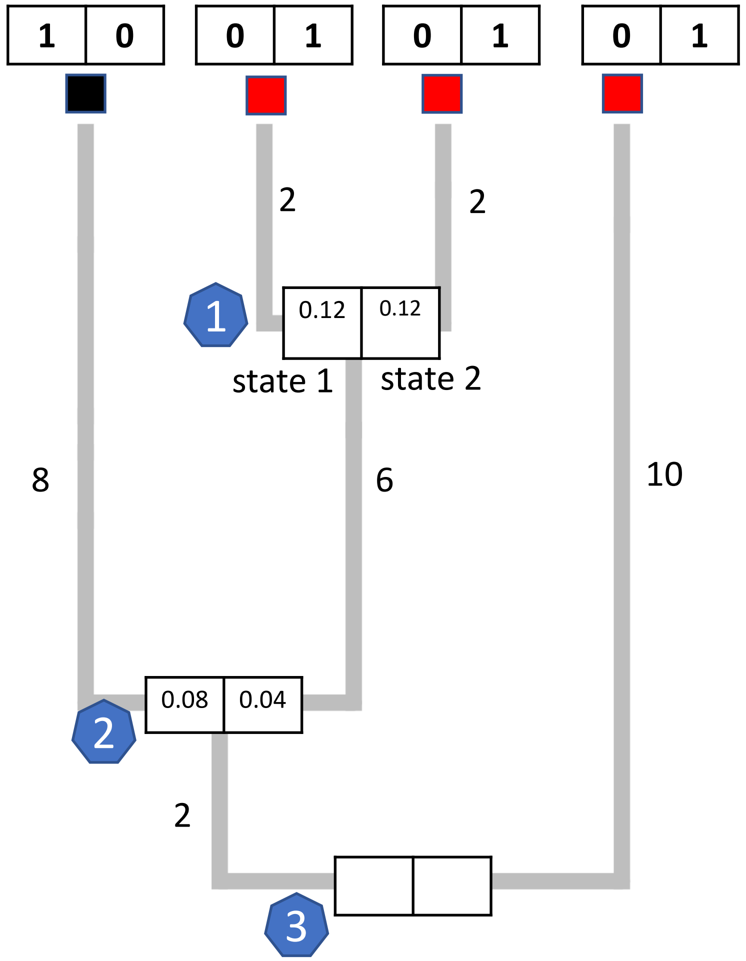
state 1 state 2

1	0

0	1

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

π =(1/2, 1/2)



Felsenstein's pruning algorithm

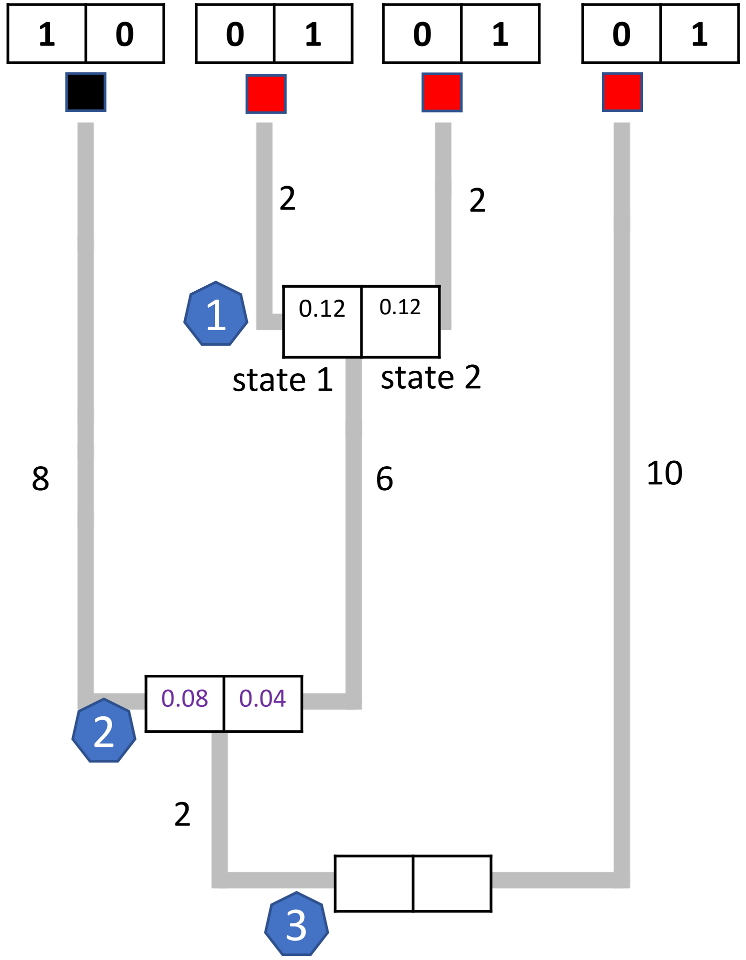
Given values:

state 1 state 2

1	0			0	1

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

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



Node 3

Left br. $e^{Q*2}=\begin{bmatrix}0.66 & 0.34\\ 0.66 & 0.34\end{bmatrix}$

Right br. $e^{Q*10}=\begin{bmatrix}0.66 & 0.33\\ 0.66 & 0.33\end{bmatrix}$

Pr(N_3 at **black**) =
(0.08*0.66+0.04*0.34)*0.33=0.02

Pr(N_3 at **red**)=
(0.08*0.66+0.04*0.34)*0.33=0.02

Felsenstein's pruning algorithm

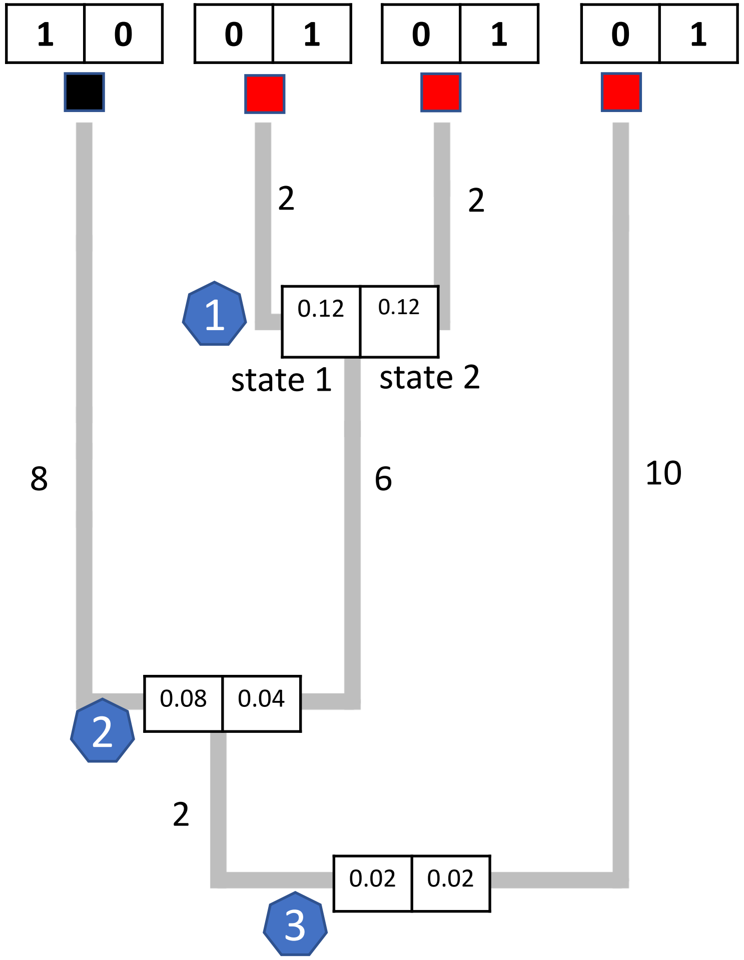
Given values:

state 1   state 2

1	**0**		**0**	**1**

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

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



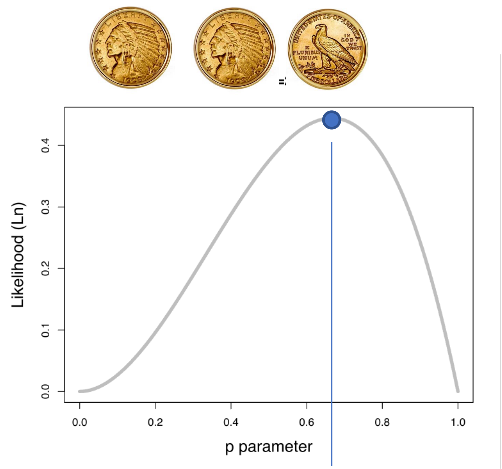
Likelihood (at the root):

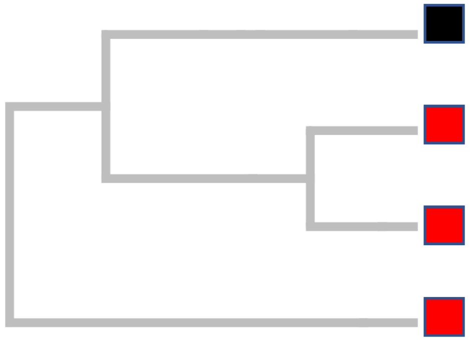
$L(tree)$ = Pr(**black**)* $\boldsymbol{\pi_1}$+Pr(**red**)* π_2 =
0.02*1/2+ 0.02*1/2 = **0.02**

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

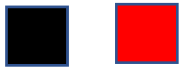
Maximum Likelihood

Find those values of the following parameters that maximize the likelihood function:





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



$$\pi = (\pi_1, \pi_2)$$

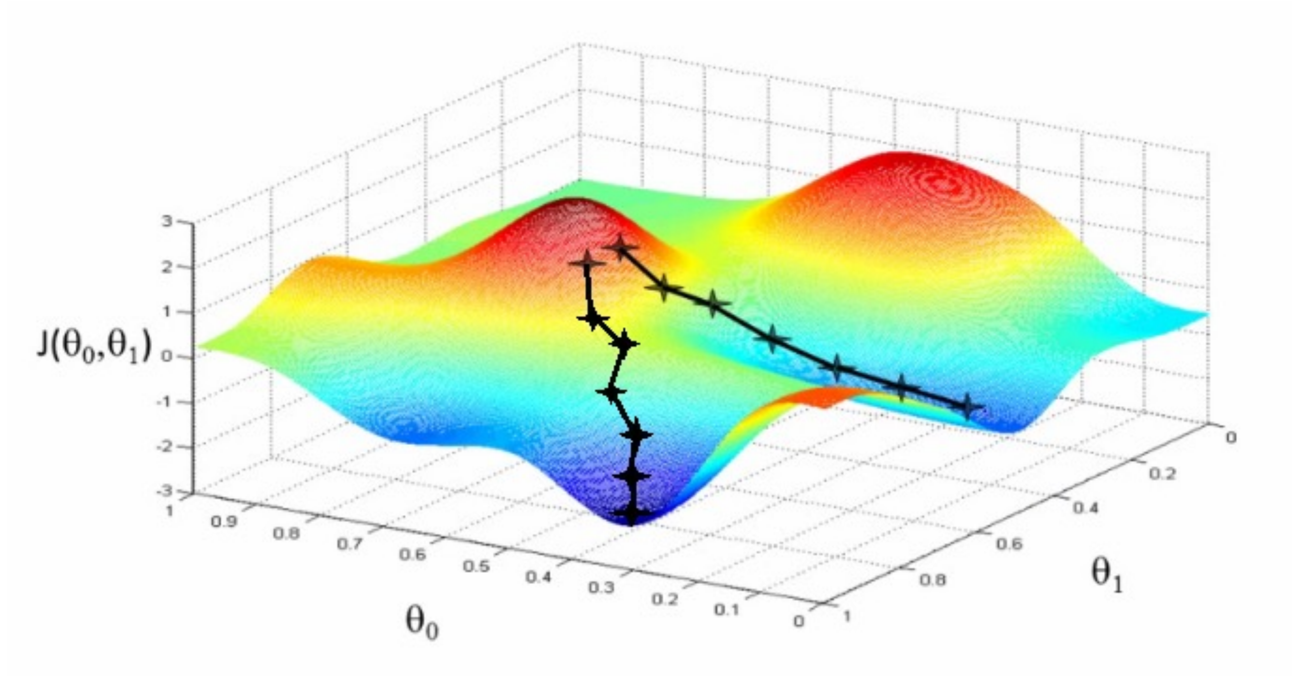
Topology and branch lengths

Rates of the rate matrix

Initial vector at the root of tree

Gradient – ascent algorithm to find a maximum of the likelihood function

1. Start with some initial values
2. Calculate the slope near the neighborhood of the initial values
3. Move along the direction of steepest ascent
4. Maximum is achieved when the slope is zero



Models of DNA evolution

- GTR (Generalised time-reversible model) model (Tavaré 1986)

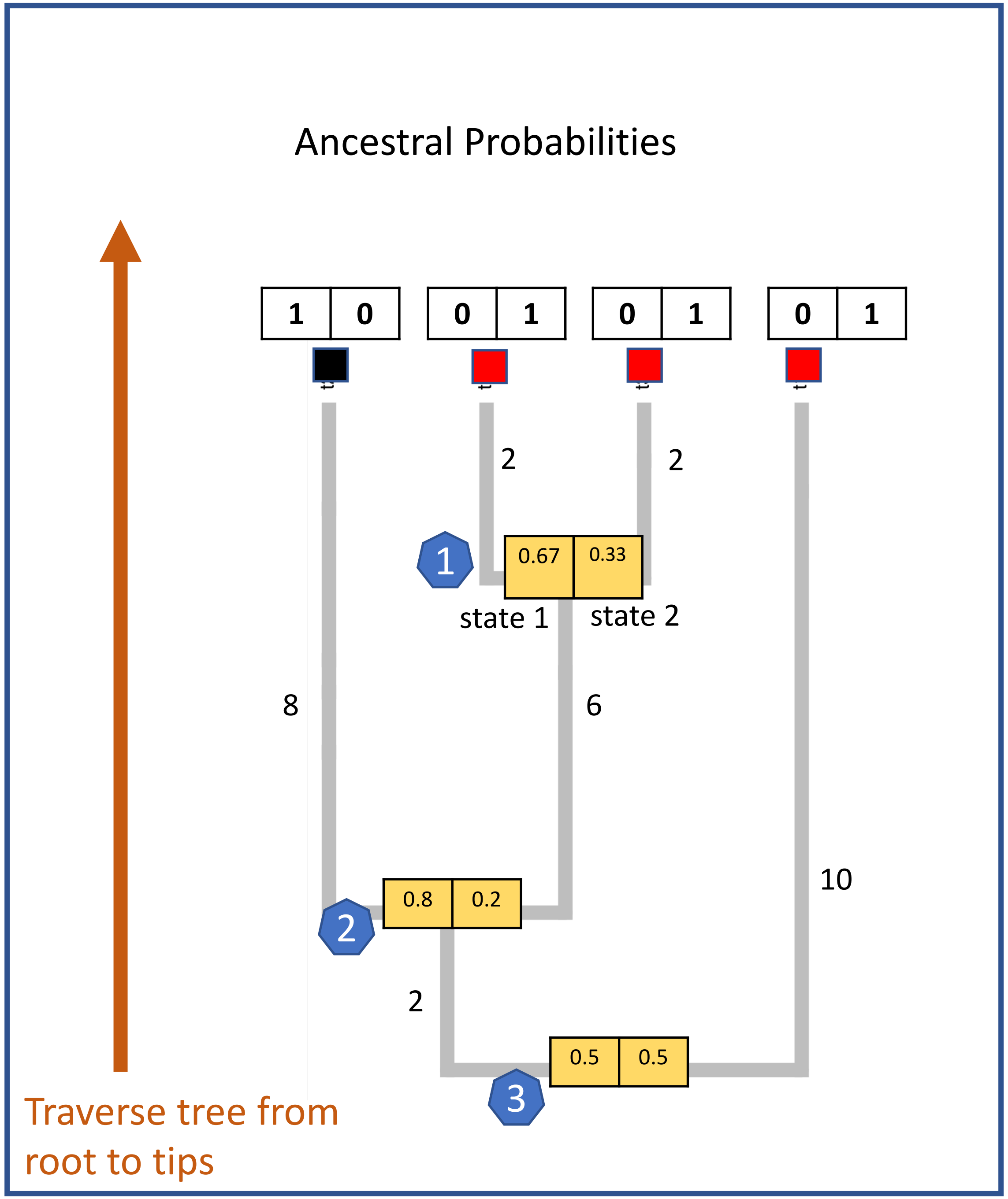
  10 parameters

$$
\begin{array}{cccc}
A & G & C & T
\end{array}
$$

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

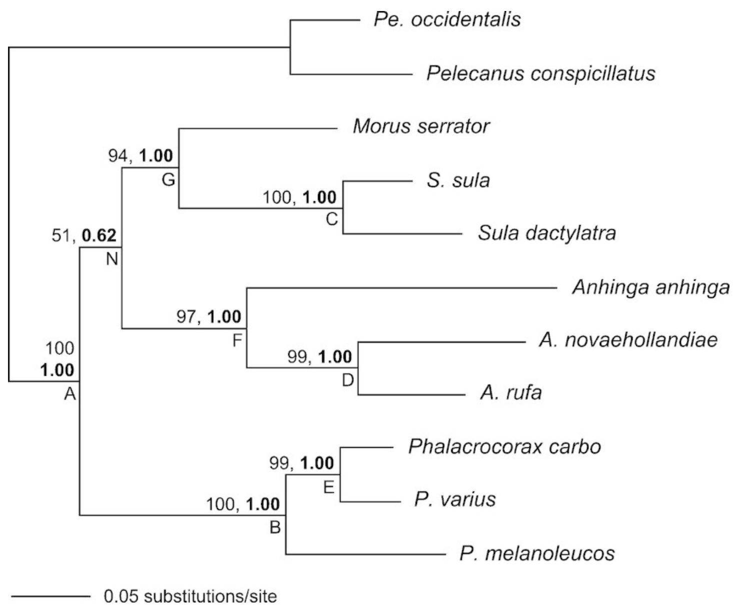
$$\pi_A \neq \pi_G \neq \pi_C \neq \pi_T$$

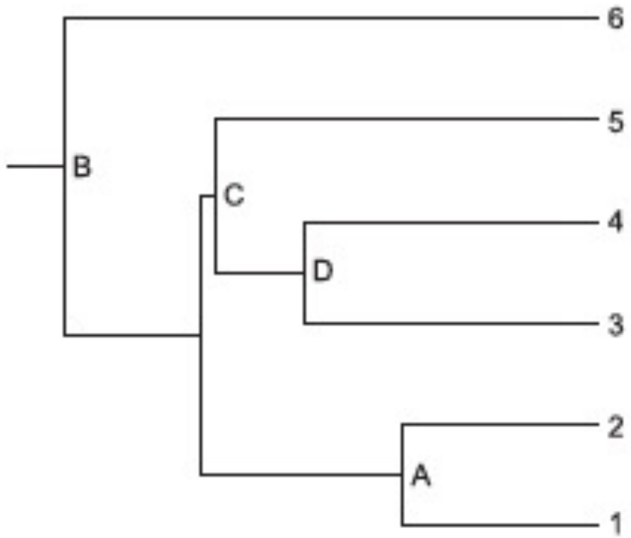
Ancestral state reconstruction



Note the units for the branch length

- ML trees are not ultrametric
- The branch length indicates the the expected number of changes per site/state over time





Units are the expected number of changes

Units are the time

Summary

- We have derived a discrete state Markov model from Binomial distribution
- Discrete state Markov model is the core of almost all phylogenetic approaches that use different type of data (morphology, DNA, proteins, etc.)
- We learnt how infer parameters of Markov model using Felsentein's pruning algorithm

Main terms used in Markov models

- Equilibrium frequencies
- Time-reversibility
- Time-homogeneous vs. time-inhomogeneous

Main terms used in Markov models

- Equilibrium frequencies
- Markov chain is at equilibrium (= stationary distribution, =invariant distribution) when its probabilities remain the same over time

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

Main terms used in Markov models

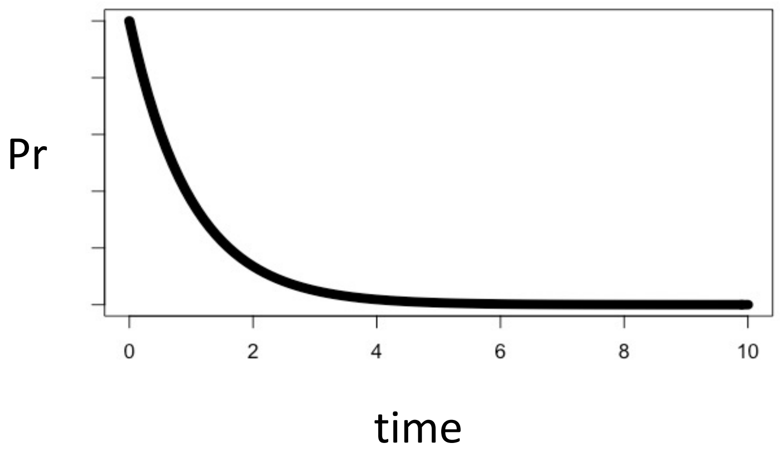
- Equilibrium frequencies
- Markov chain is at equilibrium (= stationary distribution, =invariant distribution) when its probabilities remain the same over time

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

$$e^{Q*0.1} = \begin{bmatrix} 0.91 & 0.09 \\ 0.17 & 0.83 \end{bmatrix}$$

$$e^{Q*10} = \begin{bmatrix} 0.66 & 0.33 \\ 0.66 & 0.33 \end{bmatrix}$$

$$e^{Q*20} = \begin{bmatrix} 0.66 & 0.33 \\ 0.66 & 0.33 \end{bmatrix}$$



Main terms used in Markov models

- Equilibrium frequencies
- Markov chain is at equilibrium (= stationary distribution, =invariant distribution) when its probabilities remain the same over time

Initial vector is not at equilibrium

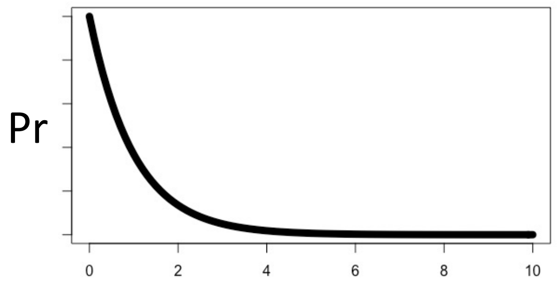
$$\boldsymbol{\pi} = \left(\frac{1}{2}, \frac{1}{2}\right) \quad \boldsymbol{Q} = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}$$

$$e^{Q*0.1} = \begin{bmatrix} 0.91 & 0.09 \\ 0.17 & 0.83 \end{bmatrix}$$

$$\boldsymbol{\pi} e^{Q*0.1} = (0.54, 0.46)$$

Main terms used in Markov models

- Equilibrium frequencies
- Markov chain is at equilibrium (= stationary distribution, =invariant distribution) when its probabilities remain the same over time



Initial vector is at equilibrium

$$e^{Q*10} = \begin{bmatrix} 0.66 & 0.33 \\ 0.66 & 0.33 \end{bmatrix}$$

$$e^{Q*20} = \begin{bmatrix} 0.66 & 0.33 \\ 0.66 & 0.33 \end{bmatrix}$$

$$\overset{\boldsymbol{\pi}}{(0.66, 0.33)} * e^{Q*0.1} = (0.66, 0.33)$$

Time-reversibility

$\boldsymbol{Q}=\begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}$ $\quad \pi = (\pi_1, \pi_2)$

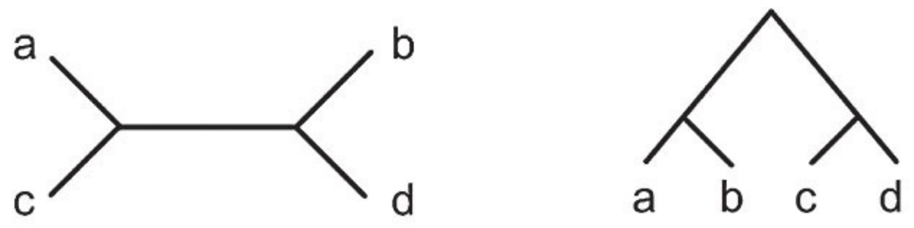
Markov model is time-reversible if $\pi_1 \alpha = \pi_2 \beta$

Why time-reversibility and equilibrium frequencies are important

- All models of the GTR family are time-reversible and have initial vector at equilibrium

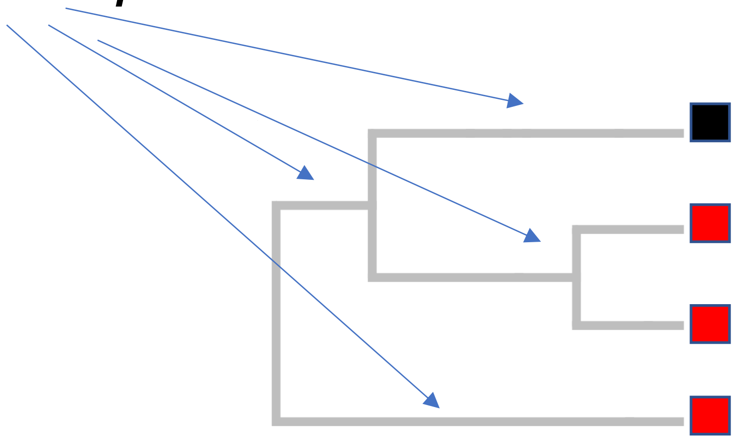
  - They allow to express branch lengths in the expected number of changes per site/state per unit of time

  - They allow working with unrooted trees when calculating likelihood as the likelihood is the same irrespective the placement of the root. Rooted trees require additional parameters.



Time-homogeneous vs. time-inhomogeneous

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



Time homogeneous Markov model means that ***Q*** is constant (=same rates) over time

Non-time-homogeneous and non-stationary models

Syst. Biol. 51(1):32–43, 2002

Inferring the Root of a Phylogenetic Tree

JOHN P. HUELSENBECK, JONATHAN P. BOLLBACK, AND AMY M. LEVINE

Department of Biology, University of Rochester, Rochester, New York 14627, USA;
E-mail: johnh@brahms.biology.rochester.ed u

Abstract.—Phylogenetic trees can be rooted by a number of criteria. Here, we introduce a Bayesian method for inferring the root of a phylogenetic tree by using one of several criteria: the outgroup, molecular clock, and nonreversible model of DNA substitution. We perform simulation analyses to examine the relative ability of these three criteria to correctly identify the root of the tree. The outgroup and molecular clock criteria were best able to identify the root of the tree, whereas the nonreversible model was able to identify the root only when the substitution process was highly nonreversible. We also examined the performance of the criteria for a tree of four species for which the topology and root position are well supported. Results of the analyses of these data are consistent with the simulation results. [Bayesian estimation; hierarchical Bayes; nonreversible models; outgroup; rooting.]

Syst. Biol. 61(6):927–940, 2012

DOI:10.1093/sysbio/sys046
Advance Access publication on April 16, 2012

Fitting Nonstationary General-Time-Reversible Models to Obtain Edge-Lengths and Frequencies for the Barry–Hartigan Model

LIWEN ZOU[1], EDWARD SUSKO[2,*], CHRIS FIELD[2], AND ANDREW J. ROGER[3]

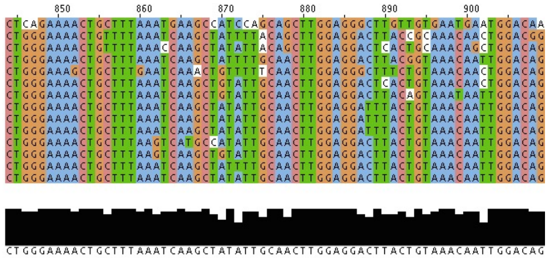
[1]*Bioinformatics Research Center, Department of Genetics, North Carolina State University;* [2]*Centre for Comparative Genomics and Evolutionary Bioinformatics, Department of Mathematics and Statistics; and* [3]*Centre for Comparative Genomics and Evolutionary Bioinformatics, Department of Biochemistry and Molecular Biology, Dalhousie University, Nova Scotia, Canada B3H 3J5;*
**Correspondence to be sent to: Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5;*
E-mail: susko@mathstat.dal.ca

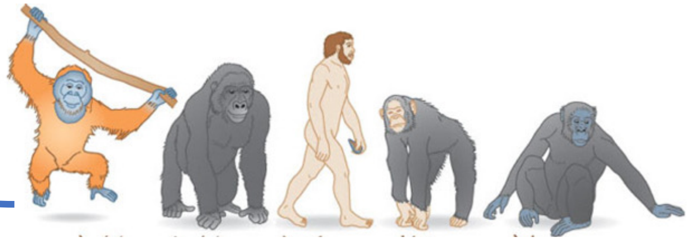
Received 6 November 2011; reviews returned 20 January 2012; accepted 28 March 2012
Associate Editor: Peter Foster

IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era
https://www.biorxiv.org/content/early/2019/11/21/849372.full.pdf

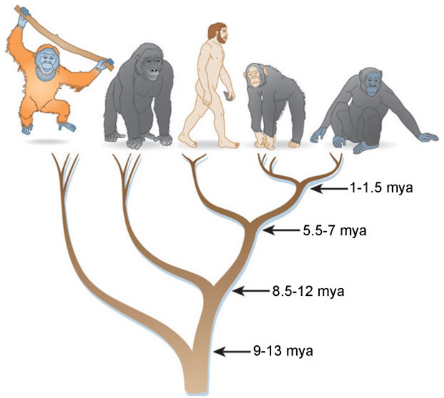
The workflow for phylogenetic reconstruction

Observed data (DNA & Morphology)





inference



The workflow for tree reconstruction using molecules

Get orthologous sequences → Align sequences → Select the best substitution models → Infer phylogenetic tree

Quick Demo

- Phylogeny of dung beetle genus *Helictopleurus* using COI